arbido

2014/2 Zur Erschliessung von AV-Medien

Ducry Emmanuel, archiviste aux Archives d'Etat de Genève

Dunant Anouk,

Ciana Xavier,

Blog, métadonnées liées aux images numérisées et archivistique

Traditionnellement, la description archivistique d'un document s'effectue dans une base de données rassemblant les inventaires, catalogues, répertoires, listes et autres éléments de descriptions selon les principes de la norme ISAD-G. Les images numérisées d'un document sont généralement liées à cette description archivistique. Mais qu'en est-il des métadonnées internes aux images numérisées? Le présent article présente le résultat de réflexions sur ces métadonnées bien spécifiques.

Le choix de publier des réflexions sur un blog

Le texte de cet article sur les métadonnées liées aux images numérisées se base en partie sur trois billets de blog consacrés à cette question. Petit rappel des faits: en 2006, les Archives d'Etat de Genève débutent des campagnes de numérisation pour diffuser sur Adhémar, leur base de données en ligne, les séries les plus consultées. Au fil des ans, des nouveaux besoins se font sentir, l'expérience s'accumule et les processus de numérisation et de mise en ligne s'adaptent. En 2013, il nous a semblé important de nous intéresser aux métadonnées directement présentes à l'intérieur des images scannées et mises à disposition du public sur le Web. Après une enquête préliminaire, nous avons constaté qu'il n'existait dans les institutions suisses d'archives aucun standard en vigueur. Nous avons donc mené nos propres recherches et réflexions, puis effectué nos choix.

Souhaitant partager le contenu de nos réflexions au sein de notre communauté professionnelle, nous avons choisi la publication sur le blog «le présent d'hier et de demain», blog personnel professionnel de l'un des auteurs. La publication de billets sur un blog nous semble une manière légère et rapide de communiquer et de mettre en valeur le fruit de ces recherches. Cet outil simple permet de publier un état de réflexion sur un sujet donné qui n'a pas forcément sa place sur un site institutionnel et de susciter des commentaires en retour.

Le blog «le présent d'hier et de demain» a été créé en mai 2012 pour exposer des remises en question sur la pertinence de la théorie des trois âges pour les archives numériques. Des comptes rendus de conférences et de colloques, des réflexions et des retours d'expérience y sont régulièrement publiés. Un blog professionnel permet également de permettre à d'autres auteurs de s'exprimer et de publier le résultat de projets réalisés en commun. C'est ainsi que nous avons rédigé à trois archivistes les billets sur les métadonnées et ce de manière interinstitutionnelle. En deux ans, ce blog, qui contient actuellement 26 billets, a reçu 10 000 visites. Le billet le plus consulté est celui sur «les métadonnées liées aux images numérisées (partie 1)»: on peut en conclure que cette recherche répondait à un besoin réel (3807 consultations).

La question des métadonnées liées aux images numérisées

La problématique abordée ici concerne la numérisation de documents ou registres patrimoniaux à des fins de diffusion. Les deux objectifs principaux qui motivent l'intégration de métadonnées internes aux images sont l'identification de la provenance des documents et l'information sur les conditions d'utilisation des images (comme nous l'indiquons ci-dessus, les métadonnées de description archivistique sont conservées dans le système d'information archivistique).

L'identification ne pose pas de problèmes lorsqu'un document numérisé est examiné dans son contexte, en général sur un site web institutionnel. Lorsqu'un registre d'état civil est par exemple consulté sur le site d'une collectivité publique, l'interface de consultation servant d'outil de recherche fournit les informations nécessaires à l'identification du registre original ainsi que les éléments de contexte nécessaires à sa compréhension (provenance, date, etc.).

Mais que se passe-t-il lorsqu'un document – ou une partie de celui-ci – est extrait de son contexte, puis republié? C'est un axiome du monde numérique: tout document qui peut être lu peut être copié et reproduit. En général, plus un document suscite de l'intérêt, plus il est reproduit et partagé. Et plus les copies sont nombreuses, plus la probabilité est forte que les informations qui accompagnaient le document lors de la publication initiale soient perdues. C'est ici que les métadonnées intégrées jouent un rôle: lorsqu'une personne copie une image comprenant des métadonnées internes, elle reproduit également, même sans le savoir, des informations sur cette image.

Les métadonnées intégrées permettent donc de signaler l'institution qui a numérisé une image ainsi que d'identifier cette image sans ambiguïté (grâce à une cote ou un identifiant unique). Sans informations d'accompagnement ni métadonnées intégrées, il peut être extrêmement difficile de retrouver le fonds ou le dossier d'origine d'une image isolée.

Jeux de métadonnées, encodage et formats d'images

Une bonne compréhension de cette question nécessite de saisir l'imbrication de trois éléments différents *(figure 1)*:

– Les différents jeux de métadonnées internes existants

– Les différentes manières de les encoder au sein des images

– Les différents formats d'images et encodages supportés

1. Les différents standards de métadonnées

Les principaux standards en matière de métadonnées intégrées aux images sont les suivants:

- IPTC: L'International Press Telecom munications Council (IPTC) développe au début des années 1990 l'Information Interchange Model (IIM). Il s'agit d'un jeu de métadonnées applicable à tout type de fichiers (texte, images, multimédia). Il sera essentiellement appliqué dans le domaine de l'image où les métadonnées prévues comportent par exemple: le créateur, le titre, la date, des informations géographiques (pays, région, ville) ou des éléments de description (mots-clés, légende). Au milieu des années 1990, les logiciels tels que Photoshop ont permis d'intégrer ces éléments directement dans les fichiers images. Cette façon de faire a dès lors connu un large

succès.

- EXIF: l'Exif (Exchangeable image file format) est une spécification de formats de fichiers pour les images et sons. La majorité des métadonnées Exif sont techniques, il s'agit d'éléments tels que la taille de l'image, la résolution, la compression ainsi que des données concernant la prise de vue: la date, le temps de pose, la distance focale, l'utilisation d'un flash, ou encore la position GPS de l'appareil. Le grand avantage des métadonnées Exif est l'automatisation: la plupart des appareils photographiques numériques créent des métadonnées Exif dans les images sans aucune intervention de l'utilisateur. De plus, elles sont largement reconnues et peuvent être lues par un grand nombre de logiciels de traitement d'images qui conservent générale ment les métadonnées Exif lors des modifications successives des fichiers.
- Dublin Core: Dublin Core est un schéma de métadonnées génériques bien connu créé en 1995 pour per mettre la description de ressources électroniques. En général utilisé comme métadonnées externes, il peut aussi être utilisé pour ajouter des métadonnées internes aux images.

2. Les différentes manières d'encoder les métadonnées

Les encodages sont les différents moyens techniques qui permettent d'intégrer concrètement les éléments de métadonnées au sein des fichiers images. Certains encodages sont prévus pour un seul jeu de métadonnées, d'autres peuvent en intégrer plusieurs, ou permettre la création de métadon nées adaptées sur mesures.

- TIFF-tags: Largement utilisés du fait de la diffusion du format TIFF, les TIFF-tags ont été définis en 1992 avec la version 6.0 du format TIFF. Le standard comprend 36 métadonnées ou tags «baseline», 60 tags «extension», 74 tags «private» et 58 tags «EXIF». D'autres ensembles de tags ont été développés pour le format DNG, les métadonnées géoréférencées, l'usage médical, etc. Ce système a eu un succès certain, mais outre le fait que seul un nombre limité de tags sont communément affichés par les visionneuses, la prolifération des tags privés a fini par rendre l'extraction des métadonnées de plus en plus complexe.
- XMP: En 2001, Adobe introduit «l'Extensible Metadata Platform» (XMP), un standard basé sur XML et RDF, qui permet d'intégrer des métadonnées dans plusieurs formats de fichiers (TIFF, JPEG, JPEG 2000, PDF, PNG, HTML, PSD, etc.). XMP a été conçu pour être extensible et peut donc accueillir n'importe quel type de métadonnées du moment que cellesci sont exprimées en XML. Dès l'origine, XMP incorpore un certain nombre de standards de métadonnées, tels que Dublin Core, EXIF, VRACore (description d'objets et d'œuvres d'art) ou IPTC-Core successer d'IPTC-IIM décrit ci-dessus *(figure 2)*.

XMP est de plus en plus répandu, les systèmes d'exploitation récents (dès Windows 7) sont notamment capables d'afficher les métadonnées XMP et de les exploiter lors de recherches de fichiers.

3. Les différents formats d'images

Chaque format d'image possède ses spécificités propres et accepte plus ou moins bien certains modes d'encodages. Les formats TIFF et jpeg acceptent ainsi aussi bien l'encodage TIFF-tags que le XMP, alors que jpg2000 n'accepte par exemple que l'encodage XMP.

4. Le choix des Archives d'Etat de Genève (AEG)

La réflexion des AEG a été menée selon un objectif de diffusion des images. Disposant de leur propre atelier de numérisation, il était impératif de ne pas complexifier les processus en cours, ni d'augmenter la charge de travail des opérateurs de scanner tout en réduisant au minimum les interventions à effectuer sur le matériel utilisé. Dans ces conditions, le choix de départ s'est porté sur les deux catégories de métadonnées liées aux images numérisées produites par nos équipements: les métadonnées Exif et IPTC. Le choix du XMP a été abandonné en attendant un remplacement du matériel.

Pour les métadonnées IPTC, les noms du pays, du canton et de l'institution met tant à disposition les images et conservant les originaux ont été considérés comme indispensables pour leur identification. En revanche, ces images étant prévues pour être mises à disposition sur une durée la plus longue possible, les métadonnées susceptibles de changements, comme une adresse web ou l'email de l'institution, n'ont pas été retenues. Une recherche sur le nom d'une institution permet de retrouver facilement ces informations susceptibles de changer régulièrement *(figure 3)*.

On constatera également qu'aucune cote ou identifiant unique ne figure parmi ces champs, cette information apparaissant uniquement dans le nom du fichier. Ce n'est peut-être pas une solution idéale, mais intégrer la cote dans les métadonnées IPTC aurait nécessité un post--traitement que nous souhaitions éviter dans le cadre de ce projet.

Les métadonnées EXIF, essentiellement techniques, relèvent en définitive plus de la conservation à long terme que de la diffusion. Toutefois, tous les appareils d'imagerie numérique produisant ces métadonnées, il aurait été dommage de ne pas les utiliser. Mais quelles métadonnées EXIF sélectionner parmi le vaste panel proposé par ce modèle? Quelques contacts menés auprès de diverses institutions ont démontré des pratiques assez aléatoires. Généralement, on se contente des réglages installés par défaut sur la machine. La question est d'autant plus difficile que l'on entre dans un domaine technique qui devient vite pointu et avec lequel les photographes ont souvent plus d'affinités que les archivistes. Après avoir élaboré un modèle dont la pertinence doit encore être évaluée, la question de la sélection des métadonnées techniques reste toujours ouverte.1

5. Le choix des Archives de la Ville de Genève

Le choix des métadonnées retenues aux Archives de la Ville a été guidé par les trois critères suivants:

- Choisir un standard bien reconnu afin que les métadonnées puissent être lues aisément

- Renseigner un nombre réduit de métadonnées afin de limiter les opérations manuelles potentiellement coûteuses

- Choisir une solution qui permette l'intégration des mêmes éléments de métadonnées dans les différentes versions JPG, TIFF et PDF d'une même image

  Dans ce cadre, notre choix s'est porté sur des métadonnées Dublin Core, intégrées aux fichiers images à l'aide de la norme XMP. Nous n'avons pas retenu l'ensemble des éléments Dublin Core, mais uniquement un nombre limité de métadonnées renseignant les informations qui nous paraissaient essentielles. Ces éléments Dublin Core nous semblaient bien répondre aux objectifs de base: identifier les images et donner le statut juridique *(figure 4)*. Quant à la norme XMP, bien que moins répandue qu'EXIF, elle est maintenant reconnue par un grand nombre de systèmes d'exploitation et de logiciels de visualisation d'images. De plus, elle rend possible l'intégration des métadonnées dans de nombreux formats de fichiers (notamment JPG et PDF).

Conclusion

Au niveau Suisse, en matière de numérisation patrimoniale, l'utilisation de métadonnées intégrées semble peu répandue. Nous l'avons notamment constaté lorsque nous avons sollicité des prestataires pour ajouter des métadonnées lors des numérisations: ils ont dû développer des solutions ad hoc ou tenter de configurer leurs machines sur mesure afin de répondre à nos demandes. Ces demandes – notamment au niveau des métadonnées techniques – n'ont pas toujours pu être complètement satisfaites.

Pour quelle utilité? Contrairement à un avertissement qui serait affiché sur un site web lors de la consultation, la présence de métadonnées intégrées aux images n'est pas forcément évidente pour un utilisateur. Celles-ci ne seront visibles que si le consultant prend la peine d'examiner les propriétés d'une image. On peut dès lors se poser la question de la rentabilité. Il est bien entendu nécessaire de mettre en balance le temps investi par rapport au bénéfice attendu.

Plutôt que de tenter de garder le contrôle du matériel diffusé – une chimère lorsque l'on parle de diffusion numérique – l'intégration des métadonnées correspond à une volonté d'informer sur la provenance d'une image et sur son statut juridique. Ces données permettent ainsi à un utilisateur confronté à une image de revenir à sa source ou de s'assurer des conditions d'utilisation.

Les choix effectués ici ne sont pas définitifs et le débat reste ouvert, de même que la question de la mise à jour des métadonnées que nous n'avons pas abordée ici. Serait-il nécessaire d'établir des recommandations au niveau Suisse, dans le but d'harmoniser les pratiques des différentes institutions?

Emmanuel Ducry

Emmanuel Ducry est historien de formation. Précédemment collaborateur aux Archives de la Ville de Genève ainsi qu’au département des manuscrits de la Bibliothèque de Genève, il travaille depuis 2011 aux Archives d’Etat de Genève (AEG) où il s’occupe notamment d’archivage électronique.

Anouk Dunant

Archives d'Etat de Genève

Xavier Ciana

Archives de la ville de Genève

Abstract

Deutsch
Diese Akronyme beziehen sich auf Metadatensätze und Enkodiersysteme für Informationen, die direkt in Bilddateien abgelegt sind. Auf gesamtschweizerischer Ebene existieren allerdings keine Überlegungen oder eine einheitliche Politik in Bezug auf diesen Bereich. Die Wahl der internen Metadaten, die zu Bilddateien geliefert werden, wird in den allermeisten Fällen den Anbietern der Erfassungsgeräte überlassen. Im Rahmen der Weiterverbreitung der Bilder aus ihren digitalisierten Beständen im Internet haben sich das Staatsarchiv Genf (AEG) und das Archiv der Stadt Genf (AVG) beide gesondert mit der Problematik auseinandergesetzt. Die Überlegungen betrafen die spezifischen internen Metadaten zu den Bildern und nicht die Katalogisierungsmetadaten.

Welche Informationen können derartige Metadaten dem Publikum in Bezug auf die Nutzungsbedingungen, die Identifizierung oder die Herkunft der Bilder bieten? Der Artikel liefert eine Einführung zu den theoretischen Grundlagen, die es erlaubt, die Fragestellungen zu begreifen. Es folgt eine Übersicht über die Wahl, die das AEG und das AVG im Hinblick auf ihre jeweiligen Zielsetzungen und unter Berücksichtigung ihres institutionellen Umfelds getroffen haben. Der Text des Artikels stützt sich teilweise auf drei Einträge im Blog «Le présent d'hier et de demain»*, der sich dieser Thematik widmet. Der Artikel wird eröffnet mit eine paar Gedanken zu den Möglichkeiten, die das Publizieren von Artikeln in einem professionellen Blog bietet.

*****
http://present-hieretdemain.tumblr.com
[abgerufen am 5. Mai 2014] ***(Überstezung: R. Hubler)***