Le Guide des Bonnes Pratiques Numériques

Version 1, Décembre 2009

Ce guide (et ses versions ultérieures) peut être téléchargé sur
http://www.tge-adonis.fr/wiki/index.php/guides

Contributeurs au Guide : Laurent Dousset (INSHS & CREDO), Jean-Luc Minel (TGE Adonis & MoDyCo), Stéphane Pouyllau (TGE Adonis & CN2SV), Richard Walter (TGE Adonis & IRHT)

Remerciements : Les auteurs du guide remercient Shadia Kilouchi, Michel Jacobson et Gautier Poupeau, auteurs de différents documents sur lesquels ils se sont appuyés.

Table des Matières

Introduction

Le passage au numérique est devenu une priorité et souvent même une nécessité dans le paysage actuel de la recherche et de sa patrimonialisation. Numériser afin de traiter et d'analyser autrement et plus en profondeur les données, afin d'archiver l'existant et de créer un patrimoine scientifique pour les générations à venir, afin d'être en phase avec les pratiques qui se généralisent ailleurs, afin de valoriser la production scientifique, afin de pouvoir travailler à distance et en collaboration sur des corpus ; voici, parmi bien d'autres, les objectifs qui motivent les laboratoires, chercheurs et enseignants-chercheurs d'effectuer ce passage.

Malgré les opportunités incontestables qu'offre le numérique, il est aussi synonyme de confusion et de propagation de pratiques, formats et standards les plus divers et peu transparents. Nombreux sont ceux qui se retrouvent, après des efforts humains et financiers parfois considérables, avec un corpus numérique inexploitable quelques années plus tard seulement car les formats ont changé ou n'existent plus. Souvent aussi des corpus numérisés sont incompatibles avec les plateformes et logiciels les plus courants dans le monde de la recherche et de l'archivage numérique. Le passage aux pratiques numériques n'est ainsi pas automatiquement synonyme d'archivage pérenne et de potentiel d'exploitation ; encore faut-il que ces pratiques numériques soient en phase avec celles qui se généralisent ou qui sont adoptées par les acteurs dominant le paysage.

Ce *Guide des Bonnes Pratiques* est une première version dans la définition des formats et standards conseillés. Il est une première étape dans l'accompagnement indispensable des unités et chercheurs qui souhaitent entamer le passage au numérique. Il doit aussi répondre à ceux qui souhaitent harmoniser leurs corpus numériques avec ceux d'autres initiatives.

Rédigé par le TGE Adonis sous l'impulsion de l'Institut des Sciences Humaines et Sociales, du Bureau des Très Grands Équipements du CNRS et du Ministère de l'Enseignement Supérieur et de la Recherche, ce guide est toutefois amené à évoluer. Dans sa version actuelle il nous fait un inventaire des formats et pratiques qui se révèlent les plus stables et les plus interopérables, c'est-à-dire des formats qui ouvrent le potentiel de l'échange d'information et la compatibilité mutuelle.

Chaque communauté scientifique connait toutefois des besoins particuliers qui évoluent au fur et à mesure de la progression scientifique et numérique. Si ce *Guide* est une première étape indispensable au passage coordonné dans le numérique, il est aussi amené à évoluer et à spécifier avec un particularisme croissant l'inventaire des formats et pratiques qui garantissent cette interopérabilité et la pérennisation des données.

Les données numériques : un pense-bête

Quelques questions avant de commencer

Avant de commencer un projet, il est nécessaire d'en déterminer les objectifs :

- **Conserver** : quelle est la durée de conservation des données (courte, moyenne ou longue) ?
- **Diffuser** : quel est le public qui doit ou peut avoir accès à ces données ?
- **Exploiter** : quelles en sont les raisons scientifiques (ou autres) et en relation avec quels autres acteurs et données ?

Les types de réponses apportées à ces questions devront vous guider dans la mise en place de l'espace numérique, des relations que vous établirez avec d'autres acteurs, en particulier dans les domaines de l'exploitation et de la conservation, et des choix technologiques que vous allez opérer.

Recommandations générales pour la numérisation

Les étapes principales de chaque projet numérique sont les suivantes :

- Sélectionner les documents à traiter (corpus de fait ou créé, cohérence du regroupement, respect du contenu et des droits).
- Définir des modes opératoires (recopie brute, corrections).
- Choisir des formats d'enregistrement adaptés (non propriétaires, avec standards officiels ou de fait, indépendance vis-à-vis des logiciels et des plates-formes).
- Définir un plan de nomenclature des fichiers (déterminer les noms des fichiers).
- Numériser avec un cahier des charges adapté aux spécificités de l'objet.
- Indexer et décrire les métadonnées. Souvent cette étape demande un investissement humain considérable. Il importe
  a) de se conformer à une initiative existante et là aussi adaptée aux spécificités de l'objet (voir plus loin dans ce document),
  b) de ne pas perdre le lien entre les métadonnées et le fichier numérique produit.
- Annoter et commenter éventuellement les ressources produites.
- Archiver de manière pérenne : l'OAIS (Open Archival Information System) est un modèle pour la gestion et l'archivage à long terme de documents numériques. Norme ISO 14721 :2002.
- Construire des entrepôts de données : l'OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), couplé aux descriptions utilisant le *Dublin Core* simple (DC Element Set), est une solution simple, minimale, qui permet de faire de l'interopérabilité.

Le protocole OAI-PMH

Vous pouvez stocker vos données dans un entrepôt et les manipuler au travers d'une base de donnée, par exemple compatible avec le langage de requêtes SQL. Mais dans tous les cas il est souhaitable de dupliquer ces données dans un entrepôt spécifique qui permet l'interopérabilité (l'échange avec d'autres fournisseurs de données).

Cet entrepôt doit être interrogeable par des requêtes conformes au protocole d'échange OAI-PMH (*Open Archives Initiative – Protocol for metadata harvesting*) qui demande de publier des métadonnées structurées en XML et conformes au Dublin Core (voir plus bas).

À l'origine, le protocole OAI-PMH a été mis au point par l'Open Archives Initiative pour faciliter l'échange et la visibilité des données stockées dans les archives ouvertes, entrepôts d'articles scientifiques mis à disposition par les chercheurs eux-mêmes. Il s'est peu à peu diffusé dans d'autres domaines d'applications de par sa simplicité et la disponibilité de nombreux outils.

Le protocole OAI-PMH implique deux acteurs :

- Le **fournisseur de données** (*data provider*) qui expose, grâce à une interface Web spécifique, les métadonnées des différents enregistrements contenus dans son entrepôt. Il s'agit là des données produites par les chercheurs, laboratoires, etc.
- Le **fournisseur de services** (*service provider*) qui moissonne un ou plusieurs entrepôts, en utilisant les interfaces exposées par le fournisseur de données, afin d'offrir aux utilisateurs des interfaces de recherche ou de navigation. Le moteur Isidore initié par Adonis, comme d'autres moteurs de recherches (Crevilles.org, OAIster, Driver-Community, etc.) pourront ainsi moissonner les données conformes au standard OAI-PMH.

Dans un entrepôt OAI, chaque ressource stockée correspond à un « enregistrement » (ou « record »). Chaque enregistrement est obligatoirement décrit en Dublin Core simple.

En plus de cette description en Dublin Core simple, chaque enregistrement peut être décrit suivant un ou plusieurs formats de métadonnées dont le choix est laissé à l'appréciation de l'administrateur de l'entrepôt. Les différents formats de métadonnées utilisés par l'entrepôt peuvent être connus par le moissonneur grâce à une requête spécifique.

Un entrepôt peut être organisé en différents ensembles d'enregistrements (« set »). Un enregistrement peut appartenir à plusieurs ensembles. Les différents ensembles peuvent être organisés hiérarchiquement. Par exemple, vous pouvez imaginer avoir des objets particuliers (des descriptions de photographies, par exemple), qui sont regroupés dans un ensemble/set (toutes les photographies d'un photographe particulier).

Recommandations particulières pour la description : Unicode

Les métadonnées descriptives doivent être encodées en Unicode UTF-8. Unicode est une norme développée par le Consortium Unicode, qui vise à donner à tout caractère de n'importe quel système d'écriture un nom et un identifiant numérique, et ce de manière unifiée, quelle que soit la plateforme informatique ou le logiciel. Le choix d'UTF-8 garanti au mieux que vos données seront lisibles sur n'importe quel système d'exploitation ou plateforme, si on dispose d'une police de caractères adéquate

Cette norme concerne l'encodage des caractères et non leur visualisation qui a besoin d'une police adaptée. Le choix d'UTF-8 n'a ainsi pas de répercussion sur la police que vous allez utiliser pour visualiser vos données sur l'écran. D'autres encodages existent (ISO 8859-1 et ASCII par exemple) mais ils sont beaucoup moins complets, particulièrement pour les langues anciennes ou rares.

Nomenclature des fichiers numériques

Une identification claire et en prévision de la réalisation d'inventaires doit être respectée. L'utilisation d'identifiants uniques est très importante : il s'agit, dès le nom du fichier, d'avoir une nomenclature unique. Ceci permet d'éviter la confusion entre fichiers. Nous faisons référence au document *Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques* édité par la BNF :

> « *Attribution d'un identifiant unique : dans un environnement informatique où chaque fichier doit* pouvoir être "adressé" de façon univoque, un nom (ou numéro) unique devra être attribué à chaque document à numériser (par exemple : XX_000001). Volumaison: les différentes parties (volumes, bobines, cassettes…) d'un tout (le document, la cote, la référence…) font également l'objet d'une identification grâce à la subdivision d'un identifiant unique (par exemple : XX_000001_V1_1 ou XX_000001_V1_n, si n parties). D'autres choix pourront être faits, mais il est impératif de reporter automatiquement l'identifiant unique attribué sur le boîtier (s'il y en a un) et sur le support lui-même. En cas de volumes importants et pour la gestion ultérieure des supports, l'usage de codes à barres est souhaitable.
>
> Il est possible, selon les supports et les renseignements dont on dispose, d'affiner encore à ce stade l'identification en précisant les notions de faces (cassette audio) ou de pistes (CD). Si ces informations ne sont pas disponibles ou suffisamment fiables, elles seront renseignées ultérieurement lors du transfert. »

Donc, quelle que soit la hiérarchie de dossiers et sous-dossiers dans lesquels vous allez placer vos fichiers numérisés ou les fichiers de métadonnées (voir ci-dessous) qui les accompagnent, veillez à ce que chaque fichier porte un nom unique. Veillez aussi à ne jamais utiliser des caractères spéciaux dans les noms des fichiers et évitez également les espaces. N'utilisez donc, dans vos noms de fichiers, que les lettres et chiffres a…z et 0…9. Le signe _ (underscore) est autorisé et recommandé pour distinguer des entités au sein du nom du fichier

mais en cas d'utilisation sur le web l'underscore peut être confondu avec le soulignement propre au lien hypertexte.

Les métadonnées

Ce chapitre est issu du travail du CRDO Paris

Les métadonnées sont des données qui décrivent d'autres données. On les appelle aussi des descripteurs. Dans les bibliothèques classiques, les documents sont décrits à l'aide de notices bibliographiques où l'on identifie les auteurs, les éditeurs, les titres, les dates de parution, etc. Ces notices sont utiles tant aux bibliothécaires pour la gestion de leur fond, qu'aux usagers pour retrouver un ouvrage.

Pour un document numérique, et plus particulièrement dans le cadre d'une diffusion par Internet, ces notices portent le nom de « métadonnées », alors que les documents eux-mêmes sont nommés « ressources ».

Les documents électroniques prennent de plus en plus d'importance dans notre vie quotidienne et leur nombre ne fait qu'augmenter. Rechercher une « ressource » spécifique est devenu une tâche à la fois complexe et indispensable d'autant plus que cette recherche s'effectue maintenant dans des architectures distribuées (les « ressources » ne se trouvent pas toutes au même endroit physique, sur le même serveur). C'est dans ce contexte que les préoccupations de standardisation et de normalisation des pratiques de codage et d'échange de métadonnées trouvent leurs origines.

Des métadonnées génériques : le Dublin-Core

En 1995, à Dublin (Ohio), des représentants de communautés diverses, issus du monde des bibliothèques, de l'informatique et du web, se réunissent pour définir un noyau commun de métadonnées : le Dublin Core Metadata Initiative (DCMI), abrégé souvent comme « Dublin-Core » ou DC.

Le Dublin-Core est un ensemble de 15 descripteurs de portée très large et de sens très générique. Certains ont trait au contenu, d'autres à la propriété intellectuelle, d'autres enfin à l'instanciation. Cet ensemble de descripteurs a été normalisé au sein de l'ISO en 2003 sous le nom d'ISO Standard 15836-2003. Les 15 descripteurs sont les suivants :

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights

- Source
- Subject
- Title
- Type

Des informations supplémentaires sur ces descripteurs peuvent être trouvées sur la page suivante : http://dublincore.org/documents/dces/

Ces éléments de base peuvent dans certains cas être jugés insuffisamment précis, il est alors possible d'utiliser un autre ensemble de « qualifiers » qui en précisent l'acception. Dublin-Core définit deux classes de qualifieurs :

- Les « refinements » qui rendent plus spécifique le sens d'un élément. Par exemple, à la place de l'élément « date » il est possible d'utiliser un de ces refinements : created, valid, available, issued, modified, dateAccepted, dateCopyrighted, dateSubmitted.
- Les schémas d'encodage, et les vocabulaires contrôlés comme par exemple le schéma « Point » qui permet de définir les propriétés d'un point géographique (coordonnées: longitude, latitude, altitude, référentiel, nom).

Le DC peut servir de base au Dublin Core dit qualifié dans lequel il est possible de typer les métadonnées, en utilisant les types de données proposés par le DCMI ou ses propres types de données définis dans un schéma XML (cf. ci-dessous).

Un fichier XML est un fichier texte mais dans lequel des balises, suite de caractères délimités par des chevrons, comme par exemple <Exemple_balise> encadrent et structurent les zones de texte qui contiennent l'information. Par exemple, si nous voulions délimiter le titre d'un ouvrage, en utilisant les balises du Dublin Core, nous écririons :

```
<dc:title>
        La Géographie locale du notaire languedocien
</dc:title>
```

Un schéma XML est un ensemble de rubriques (balises) qui sont prédéfinies et propres à ce schéma. La définition d'un schéma XML est assez similaire à la définition des champs dans une table de base de données.

Des schémas génériques

METS

C'est un schéma de structuration pour rassembler des métadonnées de description et de gestion pour un ensemble de documents (et non pas un seul document).

Mis au point et maintenu par la Library of Congress, METS (Metadata Encoding and Transmission Standard) est un schéma XML dit d'empaquetage des métadonnées. Il vise à décrire des objets numériques complexes, rassemblant au sein d'un fichier XML unique les métadonnées descriptives, les métadonnées administratives et les métadonnées de structure.

Un fichier suivant le schéma XML METS est composé de sept parties :

- METS header (metsHdr) permet d'indiquer les références du fichier METS (les métadonnées du fichier de métadonnées), en particulier le producteur du fichier ;
- Description Metadata Section (dmdsec) permet de renseigner les métadonnées descriptives de l'objet principal décrit par le fichier METS et éventuellement des objets le composant. Exemple : un fichier METS décrit un fond d'estampes, on peut à la fois décrire le fond dans une section de métadonnées descriptives et avoir autant de sections qu'il y a d'estampes ;
- Administrative Metadata Section (amdSec) permet de renseigner l'ensemble des métadonnées administratives de l'objet principal et éventuellement des objets le composant, c'est-à-dire les métadonnées techniques, les métadonnées juridiques, les métadonnées sur la source des fichiers, les métadonnées décrivant le processus de numérisation et les migrations (au sens large : passage des données de l'analogique au numérique ou du numérique au numérique) ;
- File Section (fileSec) permet de décrire l'emplacement physique de chaque fichier, rassemblé par groupe de même nature et il est aussi possible d'inclure à cet endroit le contenu du fichier ;
- Structural Map (structMap) permet d'organiser selon une structure hiérarchique les objets composant l'objet principal décrit dans les parties dmdSec, amdSec et/ou fileSec. Il est possible de décrire plusieurs cartes de structure ;
- Structural Map Linking (structLink) permet de décrire les liens éventuels entre des divisions appartenant à des cartes de structure différentes ;
- Behaviour section (behaviourSec) permet d'indiquer des comportements entre différents objets décrits dans le fichier METS.

Des informations supplémentaires sur ces spécifications peuvent être retrouvées sur la page suivante :
http://www.loc.gov/standards/mets/mets-schemadocs.html

Le format METS sépare les différents types de métadonnées, ce qui permet d'organiser et de relier les objets décrits dans les différentes sections qu'il est possible de répéter selon les besoins et ce de façon indépendante d'une quelconque structure hiérarchique. Les différentes sections correspondant à un même objet sont reliées par un système d'identifiants et de références aux identifiants.

Par ailleurs, il offre un système d'enveloppes (mdWrap) qui permettent de renseigner les métadonnées descriptives ou administratives dans le format XML qui paraît le plus adapté. Ainsi il est possible de décrire l'objet principal ou les objets le composant dans les différents formats de métadonnées existants. Ce système introduit une grande souplesse pour utiliser le format de métadonnées correspondant aux besoins.

Le modèle RDF

Mis au point au W3C dans le cadre des activités du Web sémantique, RDF

(Resource Description Framework) n'est pas à proprement parlé un schéma de métadonnées. Il constitue un modèle de description des données structurées, inspiré de la logique des prédicats de premier ordre et de la théorie des graphes.

Sa capacité à être généralisé et sa souplesse offrent des perspectives intéressantes dans la description de ressources. En particulier, il n'impose pas aux différents producteurs de se mettre d'accord strictement sur une structure de métadonnées ou de se limiter à un plus petit dénominateur commun pour assurer l'interopérabilité.

Ainsi, il permet aux producteurs de compléter les métadonnées en Dublin Core par d'autres vocabulaires très simplement.

Les ressources

Open Archives Initiative : http://www.openarchives.org
OAI-PMH : http://www.openarchives.org/OAI/openarchivesprotocol.html
Dublin Core : http://dublincore.org
Unicode : http://www.unicode.org
METS : http://www.loc.gov/standards/mets
RDF : http://www.w3.org/RDF

À la suite de ces quelques commentaires généraux et références globales, nous allons aborder, type par type, les différents documents qui peuvent composer votre projet de numérisation.

Les données textuelles

Les types de données

Les données textuelles peuvent refléter des contenus très variables :

- Textes linéaires
- Textes structurés
- Textes typés (poème en vers, théâtre, etc.)
- Liste de mots
- Dictionnaires
- Etc.

Il s'agit donc aussi bien de textes « bruts » intéressants par leur contenu que de ressources linguistiques qui sont déjà organisées par une logique scientifique ou documentaire.

La numérisation

Les sources textuelles peuvent être de format très différent. On ne numérise pas de la même façon un atlas linguistique, un manuscrit médiéval ou une collection d'ouvrages reliés, au format identique ou non.

Une chaine de traitement spécifique doit être mise en place selon chaque format. Elle est toujours basée sur la captation par une image du contenu (pour le traitement de cette image, voir la partie « Images fixes »). Pour le dire rapidement, on numérise en format image la page d'un ouvrage, la feuille de manuscrit etc.

Selon les besoins et la qualité de l'original numérisé, une OCR (*Optical character recognition*) pourra être effectuée sur le contenu pour transformer le contenu de l'image en du texte éditable. Cette « océrisation » n'est pas pertinente pour du texte avec une langue non reconnue par le logiciel ORC, ou avec une écriture manuscrite difficilement déchiffrable par ce même logiciel. Des programmes de reconnaissance par effet d'entrainement existent. Il est alors possible de tester la capacité du logiciel sur une petite partie du corpus sur laquelle un apprentissage des formes rencontrées est effectué. Puis, selon la régularité du graphisme, il est possible d'appliquer cet apprentissage sur l'ensemble du corpus numérisé. Cette méthode est efficace pour de gros volumes, mais une vérification manuelle reste souhaitable, sinon nécessaire. L'OCR simple ou par effet d'entrainement ne remplacera jamais une relecture attentive et donc « humaine ».

Les métadonnées

Pour permettre une exploitation ultérieure, il est nécessaire :

a) De choisir un format structuré et construit,

b) D'accoler à ce format structuré un modèle de données ou une documentation sur les différentes catégories créées.

En d'autres termes, une fois que vous avez numérisé et éventuellement « océrisé » vos documents textes, il vous faut en plus décrire le contenu (créer des descriptifs) de vos textes qui permettent de les organiser, de les classer, des les moissonner et de les exploiter : créer des métadonnées. Ces métadonnées sont le plus souvent, comme ce document l'a déjà expliqué, exprimées dans le format XML et suivant un schéma/encodages prédéfinis (Dublin Core par exemple). Un schéma ou encodage définit les champs nécessaires pour décrire vos documents. Différentes initiatives d'encodage existent, dont la principale est aussi celle que nous recommandons pour les textes : la TEI.

La TEI

La TEI (Text encoding initiative) est un modèle XML pour l'édition structurée et l'échange de tout type de texte. La TEI a été lancée en 1987 et elle est supportée par un consortium TEI. Un conseil TEI est chargé de l'amélioration du modèle et des aspects techniques de cette initiative qui en est à sa 5^{e} version (P5).

Elle est plus souple qu'un schéma XML classique car elle propose un ensemble de recommandations (« Guidelines ») et d'éléments particuliers rassemblés dans des modules distincts (« Tag sets ») qui s'adaptent à des besoins particuliers. Elle est largement utilisée en sciences humaines et sociales et sert aussi pour indiquer la structure sémantique d'un contenu.

Les recommandations

- R01) Pour le texte, il est indispensable de coder en UTF-8 (voir plus haut) pour garantir le bon stockage de tous les caractères du contenu.

- R02) Pour un balisage en TEI, suivre les Guidelines for Electronic Text Encoding and Interchange, les adapter à son corpus et documenter ses choix.

Les ressources

TEI : http://www.tei-c.org

Le centre de ressources numériques TELMA (Traitement électronique des manuscrits et des archives) : http://www.cn-telma.fr

Le centre de ressources numériques CNTRL (Centre national de ressources textuelles et lexicales) : http://www.cnrtl.fr

Les données iconographiques - images fixes

Les types de données

Les données iconographiques fixes recouvrent les :

- photographies (diapositives, négatif, tirages positifs)
- documents visuels fixes :
    - documents 2D numérisés
    - illustrations
    - plans, croquis, dessins etc.
    - cartes anciennes ou plus récentes, excluant les cartes construites automatiquement à partir de coordonnées et données géographiques

Numérisation et stockage

Les images numériques fixes entrent dans deux catégories principales : les images matricielles (ou « pixelisées ») et les images vectorielles (« orientées objet »). Les images matricielles prennent la forme d'une grille ou matrice, où chaque « élément d'image » (pixel) a un emplacement unique dans la matrice et une valeur de couleur indépendante pouvant être modifiée séparément. Les fichiers vectoriels fournissent un ensemble d'instructions mathématiques utilisées par un programme de dessin pour construire une image. En général, le processus de numérisation génère une image matricielle, les images vectorielles étant plus souvent le produit d'un logiciel de dessin. Par exemple, Photoshop ou Gimp créent et lisent en règle générale des images matricielles, alors que Illustrator crée et lit des images vectorielles. Les images vectorielles peuvent être converties en images matricielles. L'inverse n'est que difficilement possible.

Lors de la création et du stockage d'images matricielles, deux facteurs doivent être pris en considération : le format de fichier et les paramètres de qualité. Les images matricielles devraient en principe être stockées sous une forme non comprimée générée par le processus de numérisation, sans aucun retraitement. Les images matricielles doivent être créées et enregistrées sous l'un des formats suivants : Tagged Image File Format (TIFF), Portable Network Graphics (PNG), Graphical Interchange Format (GIF) ou JPEG Still Picture Interchange File Format (JPEG/SPIFF).

Deux paramètres fondamentaux doivent être pris en compte :

- La résolution spatiale : la fréquence à laquelle des échantillons de l'original sont capturés par le dispositif de numérisation, exprimée sous la forme d'un nombre d'échantillons par pouce (spi) ou plus communément sous la forme de pixels par pouce (ppp dans l'image numérique qui en résulte). Il s'agit là de la densité d'information (le nombre de points) enregistrée par unité de surface. Plus cette densité est haute et plus l'image numérisée est de bonne qualité. La densité pour les pages web est normalement de 72ppp. L'impression se sert normalement de densités oscillant entre 300 et 600ppp.

Il est important de noter que plus l'original est petit, et plus la densité (les ppp) devra être élevée.

- La résolution des couleurs (profondeur de bits) : le nombre de couleurs (ou de niveaux de luminosité/gris) disponibles pour représenter différentes couleurs (ou tons de gris) dans l'original, exprimé en nombre de bits. Par exemple, une résolution de couleurs de 8 bits signifie que 256 couleurs différentes sont disponibles.

La sélection des paramètres de qualité nécessaires à la numérisation d'une ressource est déterminée par la taille de l'original, la quantité de détails présents dans l'original et les utilisations prévues de l'image numérique. Numériser une diapositive de 35mm exige une résolution plus élevée que dans le cas d'une lithographie de 6x4 car la diapositive est plus petite et plus détaillée. Si l'une des utilisations de l'image d'une aquarelle requiert de pouvoir analyser d'infimes détails de coups de pinceaux, la résolution nécessaire est plus élevée que pour le seul affichage de l'image à l'écran. Plus la qualité de l'image numérisée est haute, et plus le fichier sera lourd, mais plus, également, vous pourrez agrandir l'image sans perdre de la qualité visuelle.

Les images devraient être créées à la résolution adaptée et à la profondeur de bits la plus élevée possible, à un cout acceptable et en demeurant pratiques et maniables au vu des utilisations envisagées. Chaque équipe projet doit identifier le niveau minimal de qualité et de densité d'informations dont elle a besoin. A titre d'exemple, une résolution de 600 points par pouce (ppp) et une profondeur de bits de 24 bits couleur ou de 8 bits à échelle de niveaux de gris devraient être envisagées pour les impressions photographiques. Une résolution de 2400 ppp devrait être appliquée pour des diapositives de 35 mm afin de capturer la plus grande densité d'informations. (Source : EMII DCF)

Dans certains cas, par exemple lors de l'utilisation d'appareils photo numériques de moindre qualité, il peut être indiqué de stocker les images sous un format JPEG/SPIFF, comme alternative au format TIFF. Les images seront alors plus petites et de plus basses qualité. De telles images peuvent être utiles pour la présentation de photographies d'évènements pour un site Internet, par exemple. Mais l'utilisation de tels appareils photos n'est pas recommandée pour la numérisation de contenu à grande échelle.

Les métadonnées

La photographie elle-même n'est pas encore une information exploitable dans le domaine du numérique. Pour ce faire il faut ajouter des métadonnées. Plusieurs méthodes existent qui peuvent être complémentaires. Soit les métadonnées sont incrustées dans le fichier de l'image, soit elles accompagnent le fichier d'image dans un fichier parallèle.

Les métadonnées EXIF

L'Exchangeable Image File (Exif) est un format crée en octobre 1995 par le Japan Electronic Industry Development Association (JEIDA). La version 2.1 des spécifications date du 12 juin 1998 et la version 2.2 a été publiée en avril 2002.

Le format Exif, bien que n'étant pas établi par une organisation internationale de standardisation, reste un format incontournable puisque la majorité des constructeurs d'appareils photographiques numériques l'utilisent. Il peut être également exprimé selon le standard MIX en XML.

Ce format définit un schéma de métadonnées permettant le stockage des informations techniques concernant les paramètres de prise de vue et les réglages des appareils photographiques numériques lors de la capture numérique.

Ces données sont fournies automatiquement par l'appareil photographique numérique et sont contenues dans le fichier image lui-même. Voici la liste des principaux champs Exif :

- Tag name : Description
- MakerNote : Données constructeur
- File Size : Taille du fichier
- Mime Type : Type MIME du fichier (ex : image/jpeg)
- ExposureTime : Temps d'exposition en s
- FocalLength : Distance focale en mm
- ExifImageWidth : Dimensions de l'image
- ExifImageLength
- X-Resolution : Résolution de l'image
- Y-Resolution
- Date and Time (Original) : Date et heure de l'original
- DateTimeDigitized : Date et heure de numérisation
- Tags Relating to GPS : Toutes les données relatives aux coordonnées GPS.

Quelques logiciels open source ou gratuits permettant d'afficher, éditer et extraire les métadonnées Exif :

- Exifer
- Exif Reader
- ExifTool
- ExifPro Image Viewer
- Exiv2
- IrfanView
- Photo Studio
- XnView

Métadonnées IPTC

L'International Press Telecommunications Council (IPTC) est une organisation internationale créée par les agences de presse en 1965, dont la mission est d'établir un standard normalisé de stockage des métadonnées relatives aux images de presse pour en faciliter l'échange.

IPTC/IIM

Les travaux de l'IPTC ont abouti à la mise en place d'un schéma normalisé des métadonnées des images de presse : l'IPTC/IIM.

XMP

Adobe a créé en 2001, un schéma qui utilise une expression en RDF simplifié de champs totalement paramétrables et donc extensible à des besoins particuliers. Mais ce schéma XMP est propriété d'Adobe.

IPCT-Core 1.1

IPTC Core redéfinit en XMP les métadonnées IPTC/IIM, c'est-à-dire les champs IPTC habituels plus quelques champs nouveaux. IPTC-Core n'est pas une norme ouverte, mais un standard de fait.
(Cf. http://www.iptc.org/cms/site/index.html?channel=CH0089)

Quelques logiciels open source ou gratuits permettant d'afficher, éditer et extraire les métadonnées IPTC, XMP et IPCT-Core :

- Exifer
- ExifTool
- Exiv2
- IrfanView
- PhotoThumb IPTCExt
- Rodeo Info (Mac OSX)
- XnView

Les recommandations

- R01) Formats et jeux de données : Il est préférable et conseillé de faire deux jeux de données :
  1. un au format TIFF non compressé pour la conservation,
  2. un au format JPG en qualité maximale pour une exploitation sur le web.

  De manière générale, les images photographiques ou documents numérisés doivent être créés au format TIFF.
  Vous pouvez numériser l'ensemble de vos images en haute résolution et format TIFF, dupliquer l'ensemble puis utiliser un logiciel comme *mogrify* pour créer votre second jeu d'images (http://www.imagemagick.org).

- R02) Taille des données : Une numérisation à *300 dpi est le minimum pour le format TIFF*. Pour le jeu destiné au web, au regard des possibilités de stockage actuel et de débit (2009), la taille d'exploitation web peut être équivalente à celle des TIFF. (La résolution de l'image standard pour le web étant de 72 ppp, vous pourrez agrandir votre image en ligne d'un facteur 4 environ sans perdre de qualité).

- R03) Les métadonnées descriptives des images peuvent être exprimées selon les standards :
  1. EXIF (métadonnées techniques),
  2. IPTC-Core (métadonnées descriptives).

  Nous encourageons cependant la structuration des métadonnées selon les vocabulaires DC Element Set (15 champs, voir plus haut) ou du DC:Terms dans des fichiers indépendants aux fichiers images eux-mêmes :

- dans un fichier XML séparé et nommé selon le nom de fichier de l'image dont seule l'extension (les derniers trois caractères) change : (NomDuFichierDelImage.xml).

Les ressources

- Centre de ressources numériques CN2SV (Centre National pour la numérisation de sources visuelles) : http://www.cn2sv.cnrs.fr/
- Didacticiel d'imagerie numérique de Cornell University : http://www.library.cornell.edu/preservation/tutorial-french/contents.html
- RC MINERVA : http://www.minervaeurope.org/interoperability/digitisationguidelines.htm
- IPTC : http://www.iptc.org/cms/site/index.html;jsessionid=a6fFGl6cnmYe?channel=CH0089
- Exif : http://www.exif.org
- IPTC : http://www.iptc.org
- DC : Terms : http://dublincore.org/documents/dcmi-terms

Les données iconographiques - images animées et films

Les types de données

Tous les supports vidéos :

- Bande 2 Pouce Quadruplex,
- Bande ½ Pouce,
- Vidéo Cassette ¾ Pouce,
- Vidéo Cassette ½ Pouce « substandard »,
- Vidéo Cassette ½ Pouce professionnelle,
- Vidéo Cassette 8 mm,
- Vidéo Cassette ¼ Pouce,
- Vidéodisque.

Tous les films argentiques :

- 8 mm, Super 8 mm, 9,5 mm,
- 16 mm,
- 35 mm.

Dangerosité de certains matériaux

En matière de sécurité en vue de la numérisation, il est important, pour les films argentiques d'en connaitre la composition. *En effet les films nitrate de cellulose, qui peuvent s'auto-enflammer, sont à identifier avec la plus grande précaution.* Nous renvoyons le lecteur au document *Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques* de la BNF, paragraphe 3.2 :

> « Une analyse visuelle et olfactive (syndrome du vinaigre) de la boîte et de son contenu permettent de se faire une idée de l'état de conservation du document : poussière, moisissures, état des étiquettes et de leur colle, ratures synonymes d'une réutilisation d'un support enregistrable sont autant d'indices de problèmes éventuels et de la nécessité d'un dépoussiérage ou d'un nettoyage, voire plus, avant lecture. On prendra garde également aux dangers inhérents à certains supports, comme les films nitrate de cellulose, susceptibles de s'auto-enflammer. En cas de doutes, et afin d'éviter tout risque de contamination croisée, il devra être fait appel à un spécialiste pour un diagnostic précis. »

Les métadonnées

MPEG-7

La norme MPEG-7 décrit les caractéristiques de contenu audio et vidéo de telle sorte que les utilisateurs puissent rechercher, parcourir et extraire ce contenu de manière effective et efficace. Elle combine :

- des métadonnées sur le contenu (titre, créateur, droits, renseignements sur les personnes, les objets et les évènements représentés dans le fichier multimédia, etc.) ;
- des métadonnées techniques sur le fichier.

MPEG-7 est une norme ISO élaborée par le MPEG (Moving Picture Experts Group - Groupe d'experts sur les images animées). Au regard des difficultés d'implémentation de la norme MPEG-7, nous formulons une RC plus bas.

La numérisation

Préparation du corpus

Il convient dans un premier temps de déterminer :

- La datation des données
- La nature des données :
  - Extraits de films, émissions,
  - Films complets,
  - Rushes.

Il s'agit de points importants qui vont entraîner des structurations de corpus différents.

La numérisation elle-même

En matière de numérisation, il convient de produire :

- Une version de conservation :
  - Numérisation avec compression sans pertes :
    - MJPEG
    - MJPEG 2000
    - MPEG2 4 :2 :2
  - Choix de codecs, c'est-à-dire un type d'encodage : MPEG2 MP@ML ou MPEG4 (H264). Le choix du codec a des influences sur la qualité de l'encodage/compression et la taille du fichier final. Certains codecs, propriétaires, sont à éviter. Pour plus d'informations vous pouvez consulter http://en.wikipedia.org/wiki/Codec et les pages associées.
- Une version pour diffusion web :
  Issue de la version de conservation, la version pour le web pourra être en :
  - REAL (realplayer)
  - QuickTime
  - FLV (flash)

Les recommandations

Nous faisons toujours référence au document *Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques* édité par la BNF.

- RC01) Formats et jeux de données pour la conservation :
  MJPEG, MJPEG 2000 (débit > à 25 mgb/s).

- RC02) Formats et jeux de données pour le web :
  FLV, Quicktime et REAL.

- RC03) Métadonnées :
  Tout comme pour les images fixes, nous encourageons la structuration des métadonnées selon les vocabulaires DC Element Set (15 champs) ou du DC:Terms, enregistrés dans un fichier XML séparé et nommé selon le nom de fichier de l'image dont seule l'extension (les derniers trois caractères) change : (NomDuFichierDelImage.xml).

Les ressources

- Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques :
  http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/cahier_charges_numerisation.pdf

- Institut national de l'audiovisuel (INA) : http://www.ina.fr

- Logiciel d'annotation de films, Ligne de Temps de l'IRI :
  http://www.iri.centrepompidou.fr/fr/atelier.html

Données sonores

Les types de données

Nous entendons par données sonores, l'ensemble des données audios : enregistrements de parole, de conversations ou de musiques.

Les métadonnées

(Cf. aussi la section générale sur les métadonnées plus haut)

La sonorité concerne différents types de données :

- les données qui ne sont pas ou pas seulement à proprement parler linguistiques (musique, bruits divers etc.)
- les données qui peuvent faire l'œuvre d'un traitement linguistique.

Parfois, un même document contient les deux types de données. Lorsqu'il ne s'agit pas de faire un traitement linguistique des données, les mêmes recommandations que pour les « images animées » sont également de vigueur (cf. ci-dessus mais voir aussi la recommandation RC01 ci-dessous). Pour les corpus linguistiques, on suivra également les recommandations suivantes.

Open Language Archive Community (OLAC)

OLAC est une organisation internationale regroupant un certain nombre d'institutions et d'individus préoccupés par le partage et la diffusion de ressources de nature linguistique. Le but d'OLAC est d'organiser cette communauté afin qu'elle puisse échanger facilement des documents. Pour cela OLAC a fait deux choix stratégiques dès son démarrage en 2000: celui du Dublin-Core qualifié auquel il a ajouté 5 attributs liés à des vocabulaires contrôlés pour en préciser le sens et l'adapter à la pratique de la communauté, et celui de l'OAI pour la diffusion de ces métadonnées. Les ajouts au Dublin-Core sont les suivants:

- Un attribut « language » peut être ajouté aux éléments « subject » et « language ». Sa valeur doit être prise dans le catalogue *Ethnologue*.
- Un attribut « linguistic-field » peut être ajouté à l'élément « subject ». Il doit prendre sa valeur dans une liste fermée (phonetics, phonology, pragmatics, psycholinguistics...).
- Un attribut « discours-type » peut être ajouté aux éléments type et subject (liste fermée).
- Un attribut « linguistic-type » peut être ajouté à l'élément type (liste fermée).
- Un attribut « role » peut être ajouté aux éléments « contributor » et « creator ». Il doit prendre sa valeur dans une liste fermée (recorder, researcher, signer, singer, speaker, transcriber, translator...)

Pour organiser la communauté, OLAC tient aussi les rôles d'agrégateur (OAI) et de fournisseur de service, puisqu'il maintient une liste de fournisseurs de ressources qu'il vérifie puis qu'il moissonne régulièrement et sur laquelle il offre un moteur de recherche.

Les recommandations

Nous faisons là aussi référence au document *Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques* édité par la BNF. Les préconisations sont les suivantes pour les données audio :

- RC01) Version pour conservation :
  - Numérisation sans compression,
  - Format de fichier "normalisé" : WAV ou BWF,
  - Quantification : 16, 24 bits ou plus,
  - Fréquence d'échantillonnage : 44.1, 48, 96, 192 kHz ou plus,
  - Copie dite « droite » : absence de traitement,
  - Importance du choix du convertisseur analogique / numérique (tests préalables).

- RC02) Version pour diffusion web :
  - Format de fichier : MP3, OGG.
  - Débit à ajuster en fonction du mode de diffusion envisagé.

Les ressources

- Centre de ressources numériques CRDO (centre de ressources pour la description de l'oral – Aix) : http://crdo.fr

- Centre de ressources numériques CRDO (centre de ressources pour la description de l'oral – Paris : http://crdo.risc.cnrs.fr/exist/crdo

- Projet TELEMETA : http://www.imageson.org/document1045.html

- Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques : http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/cahier_charges_numerisation.pdf

- OLAC : http://www.language-archives.org/

La bibliographie

- Corpus oraux, Guide des bonnes pratiques, O. Baude (dir.), PUO, 2006. http://www.cnrseditions.fr/Linguistique/5584-Corpus-oraux-Olivier-Baude.html.