Bibliothèque Numérique
Franco-Berbère

Bibliothèque Numérique Franco - Berbère

Guide des bonnes pratiques de numérisation

Ce Guide est fait à partir des ressources produites par le TGE Adonis (http://www.tge-adonis.fr/ressources/guides)
Il est amendé et spécifié pour le projet Bibliothèque Numérique berbère par Alain Vaucelle.

TELECOM SudParis

Table des matières

1. INTRODUCTION 4

2. LE PROJET NUMERIQUE DE LA BIBLIOTHEQUE NUMERIQUE BERBERE 4

2.1 DECIDER 4

2.2 ORGANISER 5

2.3 NUMERISER 5

2.3.1 Le plan de nommage des fichiers numériques 5

2.4 STRUCTURER 6

2.5 EXPLOITER 6

2.6 DIFFUSER 7

2.7 PERENNISER 7

2.8 DROITS JURIDIQUES 7

2.9 LISTE DES RECOMMANDATIONS 8

3. ORGANISATION, CLASSEMENTS, CATEGORIES & MOTS CLES 9

3.1 LES CATEGORIES 9

3.1.1 Les sous-catégories 9

3.2 LOCALISATION GEOGRAPHIQUE 10

3.3 LES MOTS CLES 10

3.4 LES CODES LANGUES 12

4. METADONNEES, PROTOCOLE OAI-PMH 13

4.1 METADONNEES : GENERALITES 13

4.2 METADONNEES ET SCHEMAS GENERIQUES 14

4.2.1 Le Dublin Core 14

4.2.2 Liste des types MIME 16

4.3 LE PROTOCOLE OAI-PMH 18

4.4 LES RESSOURCES 18

5. LES BASES DE DONNEES 19

5.1 GENERALITES 19

5.2 LES METADONNEES 19

5.3 L'ENTREPOT : LA PLATEFORME OMEKA 19

5.4 LISTE DES RECOMMANDATIONS 20

5.5 LES RESSOURCES 20

6. LES DONNEES TEXTUELLES 20

6.1 LES DONNEES 20

6.2 LA NUMERISATION 20

6.3 LES METADONNEES 21

6.4 LA TEI (TEXT ENCODING INITIATIVE) 21

6.5 LISTE DES RECOMMANDATIONS 21

6.6 LES RESSOURCES 22

6.7 FORMATS ET METADONNEES TEXTES DU PROJET BNFB 22

6.7.1 Nommage du fichier 22
6.7.1 Nom de l'extension 22
6.7.2 Format et résolution des fichiers 22
6.7.3 Métadonnées 22

7. LES DONNEES ICONOGRAPHIQUES – IMAGES FIXES 22

7.1 LES DONNEES 22

7.2 LA NUMERISATION 23
7.2.1 Le format de fichier 23
7.2.2 Paramètres de qualité 23

7.3 LES METADONNEES 24
7.3.1 Métadonnées techniques 24
7.3.2 Les métadonnées techniques : les EXIF 24
7.3.3 Métadonnées descriptives 24
7.3.4 Les métadonnées descriptives : les XMP et les IPTC 25

7.4 LES RESSOURCES 25

7.5 FORMATS ET METADONNEES DES IMAGES FIXES DU PROJET BNFB 26
7.5.1 Nommage des fichiers 26
7.5.1 Nom de l'extension 26
7.5.2 Format et résolution des fichiers 26
7.5.3 Métadonnées 26

8. LES DONNEES ICONOGRAPHIQUES – IMAGES ANIMEES ET FILMS 26

8.1 LES DONNEES 26

8.2 LES CONTENEURS ET LES CODECS 26

8.3 LA NUMERISATION 27

8.4 LES METADONNEES 27

8.5 LISTE DES RECOMMANDATIONS 27

8.6 LES RESSOURCES 27

8.7 FORMATS ET METADONNEES DES IMAGES ANIMEES DU PROJET BNFB 28
8.7.1 Nommage du fichier 28
8.7.2 Nom de l'extension 28
8.7.3 Le conteneur vidéo 28
8.7.4 Le codec vidéo 28
8.7.5 Résolution des vidéos 28
8.7.6 Métadonnées 28

9. LES DONNEES SONORES 28

9.1 LES DONNEES 28

9.2 LA NUMERISATION 29

9.3 LES METADONNEES 29

9.4 LISTE DES RECOMMANDATIONS 30

9.5 LES RESSOURCES 30

9.6 FORMATS ET METADONNEES DES FICHIERS SONORES DU PROJET BNFB 30
9.6.1 Nommage du fichier 30
9.6.2 Nom de l'extension 30
9.6.3 Conteneur audio 30
9.6.4 Codec audio 30
9.6.5 Métadonnées 30

10. GLOSSAIRE 31

11. TABLEAU RECAPITULATIF 34

1. 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 comme procédé de communication et d'exploitation de l'information, numériser comme procédé de conservation de l'information. Numériser afin de traiter et d'analyser autrement et plus en profondeur les données, 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 à effectuer ce passage. S'ajoute aux opérations de numérisation le fait qu'aujourd'hui de plus en plus de données sont numériques dès leur création.
Malgré les opportunités incontestables qu'offre le numérique, il est aussi parfois synonyme de confusion et de propagation de pratiques, formats et supports les plus divers et les moins 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 les outils de lecture et d'exploitation n'existent plus. Souvent aussi des corpus numérisés se révèlent inexploitables sur des plateformes ou avec des logiciels couramment utilisé dans le monde de la recherche et de l'archivage numérique.
Le passage aux pratiques numériques n'est pas ainsi automatiquement synonyme de potentiel d'exploitation, ni d'archivage à long terme. Encore faut-il que les pratiques numériques soient en phase avec celles qui sont adoptées par les acteurs du paysage.
D'où l'importance de choisir des formats interopérables, c'est-à-dire des formats qui permettent l'échange des données et des informations: « l'objet » devient partageable et diffusable, car sa description sur le plan informatique respecte des formats standards ou bien des normes, soit encore des références internationalement reconnues.

Ce Guide des Bonnes Pratiques constitue une première étape dans l'accompagnement des chercheurs qui se lancent dans le numérique. En s'intéressant de manière détaillée aux formats et en faisant le point sur les pratiques les plus récentes, ce guide peut aussi répondre aux besoins de ceux qui souhaitent harmoniser ou mettre à niveau leurs corpus numériques existants.

2. LE PROJET NUMERIQUE DE LA BIBLIOTHEQUE NUMERIQUE BERBERE

Un projet numérique est toujours un projet collectif, car il requiert des compétences varies (problématique de recherche, informatique, documentation, communication, etc.).
Tout projet numérique demande d'adopter une méthodologie de projet.
Les différentes étapes d'un projet numérique peuvent être structurées à travers les verbes suivants:

Décider
Organiser
Numériser
Structurer
Exploiter
Diffuser

Il faut y ajouter le verbe « Pérenniser », qui rassemble un ensemble de questions qu'il importe de se poser dès le début du projet, et qui implique des décisions et des actions, pratiquement à chaque étape du projet.

2.1 DECIDER

Il s'agit ici de collecter les idées, de définir l'objectif et de préciser tous les aspects et tous les enjeux du projet. On recommande l'application de la méthode QQOQCCP :

« Qui fait Quoi, Où, Quand, Comment, Combien, Pourquoi », qui permet de poser l'ensemble des questions nécessaires. Il importe de répondre à chacune de ces questions avec le plus de précisions possibles.
À cette étape du projet, trois questions importantes sont à traiter: **la question des droits juridiques** (qui touchent les données du projet, notamment en matière de diffusion), la question du stockage et la question de l'archivage.

2.2 ORGANISER

On prépare ici le pilotage du projet. Celui-ci passe par la préparation d'une note de cadrage, qui établit l'organisation précise du projet. Elle en permet la validation. Par la suite, il faut construire le cahier des charges fonctionnel (voir cahier des charges du projet BNB).

2.3 NUMERISER

Numériser des données revient à :
Établir les spécifications techniques
Numériser ou faire numériser
Contrôler la qualité des résultats

Plus précisément, il s'agit de:
Sélectionner les documents à traiter (corpus de fait ou créé). Points importants : cohérence du regroupement, respect du contenu et des droits.
Établir un cahier des charges adapté aux spécificités de l'objet : il permet de préciser la demande à la personne en charge de la numérisation mais aussi de servir de document « tiers » lors du contrôle de la qualité.
Définir les modes opératoires (recopie brute, corrections ; utilisation de logiciels avec entraînement ou non, etc.).
Choisir des formats d'enregistrement non propriétaires, respectant des normes internationales ou bien étant des standards de fait.
Définir un plan de nommage des fichiers ; voir ci-après.
Toujours contrôler la qualité des résultats obtenus (relecture, observation attentive, etc.).

Le cahier des charges devra, au besoin, préciser les conditions de livraison des originaux ainsi que des copies numérisés, les conditions d'assurance sur les originaux et les conditions de réutilisation des données numérisées par le prestataire. Il pourra aussi parfois être utile de préciser le degré de confidentialité à respecter.

2.3.1 Le plan de nommage des fichiers numériques

Une identification claire des fichiers numériques doit être préparée, à la fois en prévision du traitement et pour en réaliser l'inventaire. Il y a toujours nécessité de fournir un inventaire des fichiers.
Il faut utiliser pour chaque fichier un nom unique. Cela permet d'éviter une confusion :
entre les fichiers et leurs différentes versions ;
entre les fichiers numérisés et les fichiers de métadonnées.
Un nom unique devra être attribué à chaque document à numériser. Par exemple : XX_000001
Les différentes parties (volumes, bobines, cassettes, numéro de la face (cassette), nom de la piste (CD), etc.) doivent être identifiées en utilisant une subdivision du nom unique. Par exemple : XX_000001_V1_1 ou XX_000001_V1_n, si n parties.
Le nom unique doit toujours être reporté sur le boîtier (s'il y en a un) et sur le support lui-même.
Quelle que soit la hiérarchie choisie de dossiers et sous-dossiers (pour les fichiers numérisés ou pour les fichiers de métadonnées), chaque fichier doit avoir un nom unique. Pour écrire le nom unique, tous les caractères ne sont pas admis :
Il faut utiliser uniquement les lettres latines a…z et les chiffres 0…9, et ne jamais utiliser l'espace ni de caractères accentués.
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).

Pour le projet BNB, le plan de nommage des fichiers est le suivant tout en minuscule :
Nom du lieu de la numérisation_numero d'ordre_n (n si plusieurs parties)
Exemple : oujda_01
niamey_10_1 (si plusieurs parties)
niamey_10_V1_1 (si volume et plusieurs parties)
Note : On associera un fichier de métadonnées DC et autres métadonnées à chaque archives avec le même nom en .xml.

2.4 STRUCTURER

Structurer les données signifie :
Analyser et modéliser
Choisir des formats
Les enrichir

Ce qui est déterminant dans les choix de structuration des données, c'est l'exploitation qui est visée. En d'autres termes, ce qui doit être explicité, circonscrit, discrétisé est déterminé en partie par les objectifs du projet. Eventuellement, il y aura un compromis à établir entre ce que l'on souhaite et le temps/les moyens/les outils dont on dispose.
Beaucoup de projets numériques sont fondés sur l'utilisation de métadonnées. Pour celles-ci, il faut choisir des formats interopérables, ouverts, fondés sur des standards internationaux. Il importe donc de connaître les initiatives et de choisir celle(s) qui est(sont) adoptées aux objets numériques traités.
Comme toute information textuelle, les métadonnées doivent être encodées en Unicode UTF-8[1]. Ce format garantit l'affichage de tous les caractères, quels qu'ils soient. L'encodage en Unicode/UTF-8 ne couvre que l'aspect encodage des caractères, c'est-à-dire principalement du contenu textuel des métadonnées.
Sauf cas particuliers, les métadonnées seront inscrites dans un fichier différent de celui qui contient la donnée numérisée. On conservera le lien entre le fichier de la donnée numérisée et le fichier des métadonnées qui lui ont été associées, en gardant le même nom de fichier à l'extension près (seuls les trois derniers caractères seront différents). Enfin, on documentera les catégories utilisées et les ressources produites.
Il faut distinguer le codage des catégories de métadonnées (standards : dublin-core, PREMIS, OLAC, MARC, EAD, etc.), le codage de leur organisation (standard METS), du format utilisé pour les écrire (fichiers XML).

2.5 EXPLOITER

L'exploitation des données s'inscrit toujours dans la problématique de recherche ou la problématique documentaire. Elle s'appuie uniquement sur ce qui aura été structuré.
Exploiter ses données revient à les outiller, puis à les utiliser et les interpréter. On indique ci-dessous différentes pistes d'exploitations possibles. Elles ne sont pas exclusives les unes des autres :

a) Ouverture des données ou mise à la disposition de la communauté.

Exemples :
Mise en place d'entrepôts respectant le protocole OAI-PMH
Mise en place de descriptions archivistiques de fonds (« instruments de recherche ») respectant le standard EAD.
Documentation sémantique du système de balisage utilisé dans les fonds textuels, en format ODD pour la TEI.
Représentation des données sous la forme d'un graphe RDF et leur exposition suivant les principes du « web de données »

b) Traitement en ingénierie linguistique :

[1] Unicode est un standard développé par le Consortium Unicode et synchronisé depuis 1993 sur la norme ISO/IEC 10646. Unicode 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. Cette norme concerne l'encodage des caractères et non leur visualisation, qui elle a besoin d'une police adaptée. Le choix d'UTF-8 n'a pas ainsi de répercussion sur la police que l'on utilise pour visualiser les données à l'écran. Il ne faut pas confondre les encodages de l'Unicode (UTF-8, 16, 32, etc.) avec les encodages d'autres codes comme ceux de l'ASCII et l'ISO-88591.

Mise à disposition des résultats d'analyses linguistiques (analyse des formes, statistiques, etc.)

c) Construction de réseaux sociaux

d) Géolocalisation

Etc.

2.6 DIFFUSER

Il importe d'envisager à travers une stratégie éditoriale la forme ou bien les différentes formes que vont revêtir la publication et la valorisation des données. Site web spécifique, espace partagé pour une communauté particulière, ouverture à l'ensemble de la communauté scientifique, autant de possibilités, parfois non exclusives, qu'il faut préciser.

Barrière mobile

Les éditeurs de données numériques ont mis au point la notion de barrière mobile : l'accès aux données peut être différé dans le temps (une ou plusieurs années). Quel que soit le moment auquel on donnera accès aux données, le respect des standards et des normes reconnus internationalement demeure un impératif.

2.7 PERENNISER

Il importe tout d'abord de déterminer quelle est la durée d'utilité des données pour l'organisme qui les a produites (courte, moyenne ou longue). Au delà de cette durée où l'organisme qui a produit la donnée l'utilise (exploitation scientifique par exemple) ou la garde par devers lui pour des raisons juridiques, fiscales, etc. il convient de déterminer si les données ont un intérêt historique qui justifie leur conservation définitive par une institution dont c'est la mission.

Pendant la période d'utilisation courante de la donnée, il faut penser à son stockage.

Une sauvegarde des données « hors de ses murs », pour laquelle on veillera à ce que les données soient enregistrées de manière régulière (x fois par y heures / jours / semaines), de même qu'à une sauvegarde sur plusieurs supports stockés sur des sites différents représentent de bonnes pratiques permettant de lutter contre les pertes accidentelles. « Moins » de contraintes pèsent sur le choix des formats et sur le codage des données si l'on s'en tient strictement à du stockage sur du court terme.

Par contre, si l'on souhaite garantir l'accès à ses données sur le « long » terme, le simple stockage ne suffit pas. Il faut alors passer à l'archivage numérique à long terme[2], ce qui nécessite de la gestion, de la surveillance et le renouvellement des supports d'enregistrement, mais aussi l'absence de formats propriétaires et un bon codage initial des données.

Pour archiver ses données à long terme, il faut là encore se conformer aux standards internationaux. On recommande le modèle OAIS (Open Archival Information System) pour la gestion et l'archivage à long terme.

2.8 DROITS JURIDIQUES

Des principes généraux, régissent l'utilisation et la création des données numériques
(https://www.sup.adc.education.fr/bib/acti/num/numdef.htm,
http://www.culture.gouv.fr/culture/mrt/numerisation/fr/dll/juridi.html)

Toute création qui fait œuvre d'originalité (œuvre écrite, musicale, photographique, typographique...) est protégée par des **droits d'auteur :**

1°) les droits patrimoniaux

2°) le droit moral de l'auteur

1°) Durée des **droits patrimoniaux**

- un seul auteur : jusqu'à 70 ans après son décès
- une œuvre de collaboration : 70 ans après la mort du dernier auteur
- une œuvre collective : 70 ans après la publication

[2] L'archivage à long terme doit garantir que les données, bien évidemment à la condition que celles-ci respectent un certain nombre de formats, seront encore accessibles dans 10, 20, 30 ans et plus.

2°) Durée du **droit moral de l'auteur**

En France le droit moral de l'auteur est **perpétuel** il concerne :

- le droit à la paternité (mention du nom de l'auteur obligatoire)
- le droit au respect de l'œuvre (ne pas porter atteinte à l'intégrité de l'œuvre)

Deux cas se présentent :

a) Les œuvres tombées dans le domaine public ;

b) Les œuvres protégées par le droit d'auteur.

Œuvres tombées dans le domaine public

Les œuvres peuvent être librement exploitées lorsque la durée des droits patrimoniaux a été respectée, il faut cependant veiller d'une part à ne pas porter atteinte au droit moral de l'auteur, et d'autre part préférer une édition ancienne. Car s'il s'agit d'une réédition récente d'une œuvre tombée dans le domaine public, il faut prendre soin de ne pas reproduire les éléments originaux de l'édition moderne photographies, illustrations nouvelles ...).

Nous privilégierons les œuvres tombés dans le domaines public pour lesquelles aucun droits d'auteur n'est requis.

Concernant l'exploitation de ces documents, ils seront impérativement sous :

Licences Creative Commons, leur code et leur symbole[1,2]

Désignation complète du contrat	Terme abrégé désignant la licence	Symboles désignant la licence	Type de licence
Paternité Pas d'utilisation commerciale	CC-**BY-NC**		Licence de libre diffusion

Paternité [BY] (*Attribution*) : l'œuvre peut être librement utilisée, à la condition de l'attribuer à l'auteur en citant son nom.
Pas d'utilisation commerciale [NC] (*Noncommercial*) : le titulaire de droits peut autoriser tous les types d'utilisation ou au contraire restreindre aux utilisations non commerciales (les utilisations commerciales restant soumises à son autorisation).

http://fr.wikipedia.org/wiki/Licence_Creative_Commons

2.9 LISTE DES RECOMMANDATIONS

RECOMMANDATIONS

R1 : Appliquer en début de projet la méthode QQOQCCP pour définir, de la manière la plus détaillée qui soit, tous les aspects du projet numérique.

R2 : Organiser son projet demande de construire un cahier des charges fonctionnel et d'établir un planning prévisionnel.

R3 : Pour numériser des données, établir des spécifications techniques, choisir éventuellement un prestataire, contrôler la qualité des résultats.

R4 : Préparer un plan de nommage. Attribuer un nom unique à chaque document à numériser. Utiliser uniquement les lettres a…z, les chiffres 0…9, éventuellement le signe _ (underscore).

Ne jamais introduire d'espace dans le « nom unique » ni de caractères accentués.

R5 : Pour l'écriture des métadonnées :

a) se conformer à une initiative existante, adaptée aux spécificités de l'objet.

b) écrire les métadonnées en utilisant l'Unicode UTF-8.

c) Sauf cas particuliers, inscrire les métadonnées dans un fichier différent de celui qui contient la donnée numérisée.

d) Toujours conserver le lien entre le fichier de la donnée numérisée et le fichier des métadonnées qui lui ont été associées (même nom à l'extension près).

e) Toujours documenter les catégories utilisées et les ressources produites.

R6 : Exploiter ses données en respectant les protocoles d'échanges interopérables, reconnus comme standards sur le plan international.
R7 : Construire une stratégie éditoriale qui offre au moins deux points de vue sur les données.
R8 : Même si les données vont être soumises à une barrière mobile, les traiter en respectant les standards et les normes reconnus internationalement.
R9 : Pour archiver de manière pérenne des données, se conformer au modèle OAIS.

3. ORGANISATION, CLASSEMENTS, CATEGORIES & MOTS CLES

3.1 LES CATEGORIES

Les différentes catégories sont (liste non exhaustive) :

1. Sciences sociales
2. Langue
3. Littérature
4. Histoire
5. Art - Archéologie
6. Géographie

3.1.1 Les sous-catégories

Sous-catégories 1

Anthropologie
Ethnologie-ethnographie
Sociologie
…

Sous-catégories 2

Grammaire
Morphologie
Phonétique-phonologie
Syntaxe
Sémantique
Vocabulaire
Lexique
Sociolinguistique
Contact de langues
Libyco-berbère

Sous-catégories 3

Littérature de tradition orale
Littérature écrite
Conte
Poésie
Chanson
Devinette
Proverbe
Roman
Théâtre
Cinéma

Sous-catégories 4

Antiquité
Moyen Age
Temps modernes
Histoire contemporaine
Ethnohistoire
Histoire locale

Sous-catégories 5

Fouilles
Epigraphie
Déchiffrement

Sous-catégories 6

Territoires
Topographie
Géographie historique
Population
Ressources
Economie

3.2 LOCALISATION GEOGRAPHIQUE

Afrique du Nord
Sahara
Algérie
Maroc
Tunisie
Libye
Mali
Niger
Mauritanie
Etc.

3.3 LES MOTS CLES

1950-1959	1960-1969	1970-1979
20e siècle	Acculturation	ADGHAGH
ADHAGH	ADRAR	AFRICAIN
AFRIQUE	Afrique du Nord	AGAWAWA
AHAGGAR	AHNET	AIR
AIT	AIT-FRAH	AL-A`LA
AL-ALA	AL-ANDALUS	ALEKSOD
ALGERIE	ALGERIEN	Aliénation (Psychologie sociale)
Almohade	Alphabet	AL-TAGR
Alternance de code	ANALYSE	Analyse sémiotique
ANGLAIS	Anglais (Langue)	Animal
ANTHOLOGIE	ANTHROPO-BIOLOGIQUE	ANTIQUITE
ARABE	Arabe (Langue)	ARABES
ARAGON	ARCHITECTURE	Architecture marocaine
Architecture vernaculaire	ART	ARTICLE
Arts décoratifs	ASSIMILATION	ATLAS
AURES	AUTOBIOGRAPHIE	AVENTURE
AYACHE	AYR	AYT
AZAOUAK	AZAWAK	BARBARESQUE
BARBARESQUES	BENI-AISSA	BERBER
BERBERE	Berbère (Langue)	BERBERIE
BERBERITE	BEURBRIEUNE	BIBLIOGRAPHIE
BIOGRAPHIE	BORORO	BOUCLIER
CENTRAL	CENTRE	Chacal
CHAMITIQUE	Changement social	Chanson folklorique
CHANT	Chaouïa (Peuple berbère)	CHLEUHS
CHRETIEN	CHRISTIANISME	CIVILISATION
Civilisation islamique	COGNITIF	COGNITION
COLLECTION	Colonisation	COMPARATIVE

COMPLEXE	CONDITION	Condition féminine
Condition sociale	Conflit culturel	CONNAISSANCE
CONTE	Conte folklorique	Costume
COUTUME	Coutume sociale	CULTUREL
Danse	DECORATIF	DECORATION
DERNIER	DIALECT	DIALECTE
DIALECTOLOGIE	Diaspora juive	DICTIONNAIRE
Dictionnaire (Descripteur de forme)	Dominance	Droit coutumier
ECOLE	EGLISE	Empire islamique
Emprunt (Langage)	ENCYCLOPEDIE	ENFANT
ENSEIGNEMENT	ESPAGNE	ETAT
ETATS	Ethnicité	ETHNIE
ETHNIQUE	ETHNOGRAPHIE	ETHNOLOGIE
ETRANGER	ETUDE	ETYMOLOGIE
Étymologie	EVOLUTION	EXPEDITION
Fable	FAMINE	Femme
Femme arabe	FETE	Folklore berbère
Formation des mots	FORTIFICATION	FRAH
FRANCAIS	Français (Langue)	France
FRANCISTE	Frontière	GENEALOGIE
GENETIQUE	GEOCHIMIE	GEOCHIMIQUE
GEOCHRONOLOGIE	GEOCHRONOLOGIQUE	Géographie physique
GEOGRAPHIQUE	GEOLOGIE	GEOMORPHOLOGIE
GNEISS	GRAMMAIRE	GRAND
Graphie	GROUPE	HABITAT
HABITATION	HAMITIQUE	HAUT
Haut Atlas (Maroc)	HAUT-ATLAS	HEMATOLOGIE
HIATUS	HISTOIRE	Histoire politique
Historicité	HOGGAR	IDENTITE
Identité culturelle	ILLABAKAN	Immigrante
INSTITUT	INSTITUTION	Intellectuel
INTERFERENCE	IRJEN	ISLAM
ISLAMIQUE	JENIBA	Juif
Justice (Droit)	KABYLE	Kabyle (Langue)
KABYLIE	Kabylie (Algérie)	KEL
KEL-DENNEG	KHAMITIQUE	LANGAGE
LANGUE	Langues en contact	LEXICOGRAPHIE
LEXICON	LIBYCO-BERBERE	Libye
Lien social	LINGUISTIQUE	LITHOLOGIE
LITHOSTRATIGRAPHIE	LITTERATURE	Littérature marocaine
Littérature orale	LYBICO	MAHGREB
MAISON	MARCHE	MAROC
MAROCAIN	Massif de l'Aurès (Algérie)	MAURE
Mémoire collective	Merveilleux	MESLAILLE
METAMORPHIQUE	METAMORPHISME	Minorité
Minorité linguistique	MOEURS	MONDE
MONTAGNAIS	MONTAGNE	Morphologie (Linguistique)
Morphosyntaxe	MOT	Motif décoratif
MOUVEMENT	Moyen Âge	Moyen-Orient
Multilinguisme	MUR	MURAL
MURALE	MUSIQUE	Musique de oud
Musique marocaine	MUSULMAN	NATIONAL
NATIONALISME	NDHIR	Néologisme

NIGER	Nom (Linguistique)	NOMADE
NORD	Ntifa (Peuple berbère)	OASIS
OCCIDENTAL	OCCIDENTALE	OEUVRE
Ogre	ORIENT	ORIENTAL
ORIGINE	Orthographe	OU
OUARGLA	OUELLEMINDEN	OUEST
OUI	OUM	Oussaid, Brick
PASTORAL	PATOIS	PAYS
PAYSAN	PEINTURE	Perception spatiale
PETROGRAPHIE	PETROLOGIE	PEUPLE
Peuple	PEUPLEMENT	PHILOLOGIE
PHONETIQUE	PHONETIQUE BERBERE	PHONOLOGIE
PHRASE	PIERRE	Pluriel
POEME	POESIE	POETE
POLITIQUE	Politique étrangère	Politique linguistique
POLYCYCLIQUE	POPULATION	Population nomade
Pouvoir	PRECAMBRIEN	Préfixe
PRIMITIF	PSYCHOLINGUISTIQUE	PSYCHOLOGIE
QUESTION	Question nationale	RACE
RACIAL	RECHERCHE	RECUEIL
REFERENCE	REGION	REGIONAL
RELATION	Relations interethniques	REPERTOIRE
Représentation sociale	REPRESSION	REPUBLIQUE
Résistance politique	ROCHE	ROMAN
ROYAUME	SAHARA	Sahara (Désert)
SAHEL	SAHELIEN	SARQ
SECHERESSE	SEGHROUCHEN	SEGUEMTE
SEKSAWA	Sémiotique	SGOUGOU
SHILHA	SOCIAL	SOCIETE
Sociolinguistique	SOUS	STRATIGRAPHIE
STRATIGRAPHIQUE	STRUCTURAL	STRUCTURE
Structure sociale	STUDY	SUPERIEUR
SUPPRESSION	SYNTAXE	TAMACHEK
Tamacheq (Langue)	Tamazight (Langue)	TAMAZIGT
TARGUI	TASSILI	Tassili des Ajjer
Tchad	TECTONIQUE	TEXTE
Thème littéraire	TIMIMOUN	TISSINE
Tissu	TIZEURA	TOMBOUCTOU
TOUAREG	TOUAREG BERBERE	Tradition
TRADITION	TRADITIONNEL	TRIBU
TROGLODYTE	TROGLODYTIQUE	TUNISIE
VERBE	VERSION	VIE
Vie quotidienne	Vie sociale	Violence
Vocabulaire	ZAIAN	ZOUAOUA
ZWAWA		

3.4 LES CODES LANGUES

Language Family Trees (http://www.ethnologue.com)
Afro-Asiatic, Berber
Afro-Asiatic (374)
Berber (25)
Eastern (3)
Awjila-Sokna (2)
Awjilah [auj] (Libya)

Sawknah [swn] (Libya)
Siwa (1)
Siwi [siz] (Egypt)
Northern (17)
Atlas (3)
Judeo-Berber [jbe] (Israel)
Tachelhit [shi] (Morocco)
Tamazight, Central Atlas [tzm] (Morocco)
Kabyle (1)
Kabyle [kab] (Algeria)
Zenati (12)
East (3)
Ghadamès [gha] (Libya)
Sened [sds] (Tunisia)
Nafusi [jbn] (Libya)
Ghomara (1)
Ghomara [gho] (Morocco)
Mzab-Wargla (4)
Tagargrent [oua] (Algeria)
Tamazight, Temacine [tjo] (Algeria)
Taznatit [grr] (Algeria)
Tumzabt [mzb] (Algeria)
Riff (2)
Senhaja de Srair [sjs] (Morocco)
Tarifit [rif] (Morocco)
Shawiya (1)
Tachawit [shy] (Algeria)
Tidikelt (1)
Tamazight, Tidikelt [tia] (Algeria)
Chenoua [cnu] (Algeria)
Tamasheq (4)
Northern (1)
Tamahaq, Tahaggart [thv] (Algeria)
Southern (3)
Tamajaq, Tawallammat [ttq] (Niger)
Tamajeq, Tayart [thz] (Niger)
Tamasheq [taq] (Mali)
Zenaga (1)
Zenaga [zen] (Mauritania)

4. METADONNEES, PROTOCOLE OAI-PMH

4.1 METADONNEES : GENERALITES

Les métadonnées sont des données qui décrivent d'autres données. La notion de métadonnée renvoie dans les faits à des éléments et à des notions de nature différente.

Avant l'ère du numérique, les documents des bibliothèques étaient décrits à l'aide de notices bibliographiques dans lesquelles on identifiait les auteurs, les éditeurs, les titres, les dates de parution, etc. Ces notices étaient utiles tant aux bibliothécaires pour la gestion de leur fonds, qu'aux usagers pour retrouver un ouvrage. À ces notices étaient accolés des « descripteurs », soit des mots-clés chargés de spécifier le contenu des documents.

Avec l'ère du numérique, les notices se sont informatisées et normalisées. Dans le domaine de l'informatique documentaire, les métadonnées correspondent maintenant aussi bien aux éléments des notices bibliographiques (auteur, titre, éditeur, etc.) qu'aux descripteurs (mots-clés). Les documents identifiés par les notices sont désormais appelés « ressources ».

En parallèle, depuis plusieurs années s'est développé, sous l'impulsion du Web, un langage de structuration de l'information utilisant des balises : le XML[3]. Ce langage permet de décrire un document, de spécifier, d'ajouter, de catégoriser des informations à celui-ci. On utilise alors une suite de caractères délimitée par des chevrons, par exemple <Exemple_balise>, qui encadre une « information ».
Avec le RDF (Resource description framework), soutenu par le W3C (organisme international gérant les évolutions du web), le mode de représentation change. Il ne s'agit plus ici d'annoter, d'ajouter des informations, puis de les interpréter, mais plutôt de structurer cette description – qui est toujours une annotation interprétative – dans un langage qui ne connaît qu'une structure simple: le triplet « sujet-objet-prédicat », cette structure pouvant être représenté par un graphe. La représentation contient en elle-même son propre système d'interprétation (qui est ici fondé sur une adaptation de la logique des prédicats du 1er ordre).

Par la suite, on distinguera :
les métadonnées techniques qui sont des métadonnées liées à la technicité associée à la numérisation de la donnée (par exemple, le temps d'exposition d'une photo) ;
les métadonnées descriptives ou documentaires qui sont des métadonnées qui renvoient à la problématique de recherche ou documentaire, à l'utilisation des données, à leur administration ;
les métadonnées génériques qui sont des métadonnées qui s'appliquent à n'importe quel type d'objet numérisé ;
le schéma qui est un ensemble de balises (rubriques) prédéfinies ;
Dans le cas de XML (Extensible Markup Language), la définition d'un schéma est assez similaire à la définition des champs dans une table de base de données. XML peut être considéré comme l'une des façons de mettre en forme la structuration de l'information dans les champs.

4.2 METADONNEES ET SCHEMAS GENERIQUES

4.2.1 Le Dublin Core

En 1995, à Dublin (Ohio), des représentants issus du monde des bibliothèques, de l'informatique et du web, se sont réunis pour définir un noyau commun de métadonnées : le Dublin Core Metadata Initiative (DCMI), abrégé souvent en Dublin Core ou encore 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 (diffuseur de la ressource)
Coverage (couverture géographique ou temporelle de la ressource)
Creator (auteur de la ressource)
Date (événement dans la vie de la ressource)
Description
Format (format mime ; dimension physique ; durée de la ressource) ;
Identifier (identifiant unique de la données : URL, DOI, n° ID, etc.)
Language (normalisé selon l'ISO 639. Pour préciser un « parler » non codifié on ajoutera un qualifier (voir ci-dessous) au code ISO déjà défini. Pour coder une langue non encore codifiée par l'ISO 639 (par exemple une langue sacrée subsaharienne dont on aura vérifié l'absence de code), on utilisera « ber » suivi d'une extension).
Publisher (éditeur)
Relation (lien vers d'autres ressources)
Rights
Source
Subject
Title

[3] « eXtensible Markup Language » (XML) est une recommandation du W3C qui prend sa source dans la norme SGML (ISO 8879:1986).

Type (nature descriptive de la données : événement, corpus, fonds de chercheurs, film, image, photo, etc., cf. http://dublincore.org/documents/dcmi-typevocabulary/)
Ces éléments de base peuvent dans certains cas être insuffisamment précis. Il est alors possible d'en utiliser d'autres. Ce sont les qualifieurs (qualifiers), qui précisent l'acception. Deux ensembles de qualifieurs ont été proposés :

1. Les raffineurs (refinements) qui précisent le sens d'un élément. Par exemple, à la place de l'élément « Date », il est possible d'utiliser un de ces raffineurs : Created,
Valid, Available, Issued, Modified, DateAccepted, DateCopyrighted, DateSubmitted.
2. Les schémas d'encodage et les vocabulaires contrôlés. Par exemple le schéma
« Point » qui permet de définir les propriétés d'un point géographique (coordonnées : longitude, latitude, altitude, référentiel, nom).

Le Dublin Core 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.

Élément	Élément (anglais)	Commentaire
1. Titre (métadonnée)	Title	Titre principal du document
2. Créateur (métadonnée)	Creator	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document
3. Sujet (métadonnée) ou mots clés	Subject	Mots-clefs, phrases de résumé, ou codes de classement
4. Description (métadonnée)	Description	Résumé, table des matières, ou texte libre. Raffinements : table des matières, résumé
5. Éditeur	Publisher	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document
6. Contributeur	Contributor	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document. Chaque contributeur fait l'objet d'un élément Contributor séparé
7. Date (métadonnée)	Date	Date d'un évènement dans le cycle de vie du document
8. Type de ressource	Type	Genre du contenu

9. Format	Format	Type MIME, ou format physique du document
10. Identifiant de la ressource	Identifier	Identificateur non ambigu : il est recommandé d'utiliser un système de référencement précis, afin que l'identifiant soit unique au sein du site, par exemple les URI ou les numéros ISBN. Raffinement : Is Available At
11. Source	Source	Ressource dont dérive le document : le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI
12. Langue (métadonnée)	Language	La langue du contenu intellectuel de la ressource. Il est recommandé d'utiliser une des valeurs définies dans la RFC 3066 qui, avec la norme ISO 639, définit des codes de langues primaires à deux et à trois caractères, ainsi que des sous-codes facultatifs. Exemples : « en » ou « eng » pour l'anglais, « fr » ou « fre » pour le français, « akk » pour l'akkadien et « en-UK » pour l'anglais utilisé au Royaume-Uni. L'élément Language s'applique aussi bien aux ressources textuelles ainsi qu'aux ressources audio et vidéo qui comportent des passages parlés. Pour une ressource textuelle, il s'agit de la langue dans laquelle elle a été écrite. Pour les œuvres multilingues, il y a autant d'éléments Language que de langues utilisées.
13. Relation (métadonnée)	Relation	Lien avec d'autres ressources. De nombreux raffinements permettent d'établir des liens précis, par exemple de version, de chapitres, de standard, etc.
14. Couverture (métadonnée)	Coverage	Couverture spatiale (point géographique, pays, régions, noms de lieux) ou temporelle
15. Droits (métadonnée)	Rights	Droits de propriété intellectuelle, Copyright, droits de propriété divers

4.2.2 Liste des types MIME

Le type MIME (Multipurpose Internet Mail Extensions) est un standard qui a été proposé par les laboratoires *Bell Communications* en 1991 afin d'étendre les possibilités du courrier électronique (mail), c'est-à-dire de permettre d'insérer des documents (images, sons, texte, ...) dans un courrier.
Depuis, le type MIME est utilisé d'une part pour typer les documents attachés à un courrier mais aussi pour typer les documents transférés par le protocole HTTP. Ainsi lors d'une transaction entre un serveur web et un navigateur internet, le serveur web envoie en premier lieu le type MIME du fichier envoyé au navigateur, afin que ce dernier puisse savoir de quelle manière afficher le document.
Un type MIME est constitué de la manière suivante :
Content-type: type_mime_principal/sous_type_mime
Une image GIF a par exemple le type MIME suivant : **Content-type: image/gif**
En voici une liste relativement complète.

Type MIME	Type de fichier	Extension associée
application/acad	Fichiers AutoCAD	dwg
application/clariscad	Fichiers ClarisCAD	ccad
application/drafting	Fichiers MATRA Prelude drafting	drw
application/dxf	Fichiers AutoCAD	dxf
application/i-deas	Fichiers SDRC I-deas	unv

application/iges	Format d'échange CAO IGES	igs,iges
application/octet-stream	Fichiers binaires non interprétés	bin
application/oda	Fichiers ODA	oda
application/pdf	Fichiers Adobe Acrobat	pdf
application/postscript	Fichiers PostScript	ai,eps,ps
application/pro_eng	Fichiers ProEngineer	prt
application/rtf	Format de texte enrichi	rtf
application/set	Fichiers CAO SET	set
application/sla	Fichiers stéréolithographie	stl
application/solids	Fichiers MATRA Solids	dwg
application/step	Fichiers de données STEP	step
application/vda	Fichiers de surface	vda
application/x-mif	Fichiers Framemaker	mif
application/x-csh	Script C-Shell (UNIX)	dwg
application/x-dvi	Fichiers texte dvi	dvi
application/hdf	Fichiers de données	hdf
application/x-latex	Fichiers LaTEX	latex
application/x-netcdf	Fichiers netCDF	nc,cdf
application/x-sh	Script Bourne Shell	dwg
application/x-tcl	Script Tcl	tcl
application/x-tex	fichiers Tex	tex
application/x-texinfo	Fichiers eMacs	texinfo,texi
application/x-troff	Fichiers Troff	t,tr,troff
application/x-troff-man	Fichiers Troff/macro man	man
application/x-troff-me	Fichiers Troff/macro ME	me
application/x-troff-ms	Fichiers Troff/macro MS	ms
application/x-wais-source	Source Wais	src
application/x-bcpio	CPIO binaire	bcpio
application/x-cpio	CPIO Posix	cpio
application/x-gtar	Tar GNU	gtar
application/x-shar	Archives Shell	shar
application/x-sv4cpio	CPIO SVR4n	sv4cpio
application/x-sv4crc	CPIO SVR4 avec CRC	sc4crc
application/x-tar	Fichiers compressés tar	tar
application/x-ustar	Fichiers compressés tar Posix	man
application/zip	Fichiers compressés ZIP	man
audio/basic	Fichiers audio basiques	au,snd
audio/x-aiff	Fichiers audio AIFF	aif,aiff,aifc
audio/x-wav	Fichiers audio Wave	wav
image/gif	Images gif	man
image/ief	Images exchange format	ief
image/jpeg	Images Jpeg	jpg,jpeg,jpe
image/tiff	Images Tiff	tiff,tif
image/x-cmu-raster	Raster cmu	cmu
image/x-portable-anymap	Fichiers Anymap PBM	pnm
image/x-portable-bitmap	Fichiers Bitmap PBM	pbm
image/x-portable-graymap	Fichiers Graymap PBM	pgm
image/x-portable-pixmap	Fichiers Pixmap PBM	ppm
image/x-rgb	Image RGB	rgb
image/x-xbitmap	Images Bitmap X	xbm
image/x-xpixmap	Images Pixmap X	xpm
image/x-xwindowdump	Images dump X Window	man
multipart/x-zip	Fichiers archive zip	zip

multipart/x-gzip	Fichiers archive GNU zip	gz,gzip
text/html	Fichiers HTML	htm,html
text/plain	Fichiers texte sans mise en forme	txt,g,h,c,cc,hh,m,f90
text/richtext	Fichiers texte enrichis	rtx
text/tab-separated-value	Fichiers texte avec séparation des valeurs	tsv
text/x-setext	Fichiers texte Struct	etx
video/mpeg	Vidéos MPEG	mpeg,mpg,mpe
video/quicktime	Vidéos QuickTime	qt,mov
video/msvideo	Vidéos Microsoft Windows	avi
video/x-sgi-movie	Vidéos MoviePlayer	movie

4.3 LE PROTOCOLE OAI-PMH

L'OAI-PMH (Open Archive Initiative - Protocol for Metadata Harvesting) est un protocole standard interopérable qui permet d'avoir accès aux métadonnées d'un entrepôt numérique. Il a été mis au point en 1999 par l'Open Archives Initiative pour faciliter l'échange et la visibilité des données stockées dans les archives ouvertes (des entrepôts d'articles scientifiques mis à disposition par les chercheurs eux-mêmes). Il s'est peu à peu diffusé dans d'autres domaines du fait de sa simplicité et de la disponibilité de nombreux outils.
Ce protocole est particulièrement utile dans le cas des bases de données. Si d'un côté, les données contenues dans ces bases peuvent être affichées dynamiquement dans des pages web, faire l'objet de requêtes particulières, etc., il est toujours souhaitable de les dupliquer et de les rassembler dans un entrepôt. Quand cet entrepôt est muni de métadonnées respectant le protocole OAI-PMH, on assure l'interopérabilité et par suite le moissonnage par des moteurs de recherche.
Le protocole OAI-PMH implique ainsi 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.
- 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.

La plateforme ISIDORE (http://rechercheisidore.fr), initiée par le TGE Adonis, comme d'autres moteurs de recherche pourront ainsi moissonner les métadonnées exposées selon le standard OAI-PMH.
Dans un entrepôt OAI-PMH, chaque ressource stockée correspond à une notice ou encore « enregistrement » (ou « record »). Chaque notice ou enregistrement est obligatoirement décrit à l'aide de métadonnées respectant a minima le Dublin Core[4] simple. Il est possible de décrire les métadonnées, en plus du Dublin Core simple, à l'aide de vocabulaires plus riches : DC Terms, MODS, OLAC, etc.
Ces enregistrements peuvent être rassemblés en différents ensembles (« set ») et un enregistrement peut appartenir à plusieurs ensembles. Les différents ensembles peuvent être organisés hiérarchiquement.
Par exemple, on peut avoir des objets particuliers (des descriptions de photographies), qui sont regroupés dans un ensemble/set (toutes les photographies d'un photographe particulier).
Les différents formats de métadonnées utilisés par l'entrepôt, celles publiées en Dublin Core simple comme les autres, sont accessibles au moissonneur grâce à une requête spécifique.

4.4 LES RESSOURCES

Dublin Core : http://dublincore.org/
http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_dublin_core.html
Dublin Core - Informations sur les 15 descripteurs :
http://dublincore.org/documents/dces/
METS : http://www.loc.gov/standards/mets
METS - Schémas : http ://www.loc.gov/standards/mets/mets-schemadocs.html
METS - Liste d'outils dédiés : http ://www.loc.gov/standards/mets/mets-tools.html
MODS : http://www.loc.gov/standards/mods

[4] OAIster (http://www.oclc.org/oaister/), Driver-Community (http://www.drivercommunity.eu/), etc.

OAI-PMH : http://www.openarchives.org/OAI/openarchivesprotocol.html
Open Archives Initiative : http://www.openarchives.org
RDF : http://www.w3.org/RDF
RDF - Liste d'outils dédiés et classés par catégories :
http://www.w3.org/2001/sw/wiki/Category:Tool
Unicode : http://www.unicode.org

5. LES BASES DE DONNEES

5.1 GENERALITES

Une base de données, du point de vue des humanités numériques, doit être considérée comme un réservoir organisé de données, dont on affiche une ou plusieurs « vues » à un instant T.
Les « vues » peuvent être constituées par les données « brutes » et/ou par les métadonnées qui y ont été associées. Les vues sont affichées par différents moyens :
via une édition électronique statique : par exemple un fichier en format pdf ;
via une édition en flux d'information ;
via une interface web de recherche.
Dans les deux derniers cas, on parle d'édition dynamique des données.
Avec le numérique et l'obsolescence des formats, la durée plus courte des projets de recherche, une réflexion approfondie doit être conduite, dès lors qu'on souhaite mettre à disposition dans un avenir plus ou moins proche ses données à l'ensemble de la communauté.

La réflexion doit donc inclure AU TOUT DEBUT DU PROJET un travail permettant d'avoir une idée claire sur la pérennisation des données que l'on traite. Dans un tel cadre, l'édition d'une base de données ne peut se limiter à l'édition d'un formulaire de recherche et à l'élaboration d'une maquette graphique.
L'application d'une méthodologie de projets liée aux objets numériques est ici aussi nécessaire, et il faudra veiller particulièrement :
à l'utilisation de standards internationaux pour le codage des données ;
à l'utilisation de formats « ouverts ».
Il existe bien évidemment d'autres bases de données qui n'ont pas vocation à être éditées : des bases de données « personnelles », pour organiser la recherche en train de se faire. Cependant, il importe de noter que bon nombre de ces projets évoluent très souvent vers des projets d'édition et de mise à disposition des données à l'ensemble de la communauté.
Dans ce cas, quand on a choisi au départ un format propriétaire (qui se justifiait donc dans le cadre du projet de départ), très souvent, le chercheur doit refaire dans un autre format ce qui a déjà été fait. Cet investissement, considéré sous un angle strictement technique, est souvent jugé trop onéreux et inintéressant. La conséquence en est que des données seront effectivement mises à disposition de la communauté mais… pour un temps a priori très limité. Et la pérennisation des données ne sera en aucun cas assurée.
Aussi on recommande de choisir des formats ouverts (non propriétaires), respectant des standards internationaux, y compris pour les bases de données « personnelles ».

5.2 LES METADONNEES

Une base de données ayant vocation à stocker et à organiser des données de différentes natures (taille, format, etc.), on se reportera aux chapitres consacrés aux différents types de données.

5.3 L'ENTREPOT : LA PLATEFORME OMEKA

Des données numérisées peuvent être manipulées à travers une base de données, compatible par exemple avec le langage de requêtes SQL. Cependant, pour que cette base de données puisse être moissonnable par des moteurs de recherche, on recommande de transformer ces bases de données en entrepôt de données.
Cela entraîne de nouvelles tâches : définir les objets que l'on va considérer comme des « ressources » de l'entrepôt, établir une correspondance entre des champs de la base et des rubriques du Dublin Core et éventuellement d'autres schémas de métadonnées, implémentation des fonctions utiles au protocole OAI-PMH, etc.).

L'entrepôt de données doit être construit en respectant le protocole OAI-PMH. À chaque élément de la base de données, il faudra associer des métadonnées, qui seront exprimées en Dublin Core simple, soit un jeu de 15 métadonnées supplémentaires.
L'entrepôt OMEKA pour le projet BNBF :
http://bnfb.cartago-alliance.org/

Pour administrer les contenus :
http://bnfb.cartago-alliance.org/admin/users/login

Le blog du projet : http://wp-bnfb.cartago-alliance.org/

5.4 LISTE DES RECOMMANDATIONS

RECOMMANDATIONS
R10 : Pour les bases de données, y compris les bases de données « personnelles », choisir des formats ouverts reconnus internationalement.
R11 : Transformer sa base de données en entrepôt de données.
R12 : Pour un entrepôt de données, utiliser le protocole OAI-PMH et entrer les métadonnées en utilisant le Dublin Core simple.

5.5 LES RESSOURCES

OAI-PMH : http://www.openarchives.org/OAI/openarchivesprotocol.html
Open Archives Initiative : http://www.openarchives.org

6. LES DONNEES TEXTUELLES

6.1 LES DONNEES

Les données textuelles recouvrent aussi bien des textes « bruts » que des textes structurés, et ce, quelle que soit leur structure formelle, narrative, etc. : romans, poèmes en vers, pièces de théâtre, interviews, transcriptions de conversations, lettres, listes de mots, dictionnaires, etc. Ces textes peuvent être intéressants aussi bien du point de vue de leur contenu que du point de vue des aspects linguistiques qu'ils comportent.

6.2 LA NUMERISATION

Mise en place d'une chaîne de traitement
On met en place une chaîne de traitement qui doit être adaptée à chaque format. En effet, on ne numérise pas de la même manière un manuscrit médiéval, une collection d'ouvrages reliés, etc.
Captation
On effectue la captation du texte via une « image fixe ». En d'autres termes, on numérise en format image la page d'un ouvrage, la feuille d'un manuscrit, etc. Pour plus d'informations, on se reportera à la section 5.
Application d'un logiciel de reconnaissance optique de caractères : logiciel OCR / océrisation
Selon les besoins, on peut utiliser un logiciel de reconnaissance automatique de caractères. Ce programme permet de transformer le contenu de l'image en un texte éditable.

L' « océrisation » est pertinente :
quand on a de gros volumes de données à traiter ;
si la qualité de l'original le permet ;
si la langue des données textuelles est reconnue par le logiciel OCR ;
dans le cas d'une écriture manuscrite, si celle-ci est « facilement » déchiffrable.
Il existe des programmes OCR comportant des modules d'entraînement. On peut alors tester et entraîner le logiciel sur une petite partie du corpus, puis décider, en fonction des résultats, de l'utiliser sur l'ensemble du corpus.

Attention !
Dans tous les cas, c'est-à-dire, quels que soient le volume traité, le logiciel OCR retenu (avec ou sans entraînement), relire attentivement le résultat de l' « océrisation ».
Le format « texte seul » avec l'encodage UTF-8
Pour du texte sans mise en page, il est recommandé d'utiliser le format « texte-seul » avec l'encodage UTF-8.

6.3 LES METADONNEES

Il existe de nombreuses manières d'exploiter et d'afficher des données textuelles : simple exposition dans un environnement de site web ; qualification sophistiquée des différents éléments des « textes »; analyse automatique du corpus, extraction de données, etc. De plus, les mêmes données textuelles peuvent être exploitées de différentes façons. Tout dépend de la problématique de recherche, des résultats auxquels on souhaite arriver, des résultats que l'on souhaite exposer, etc.

Si l'exploitation des données recourt à des métadonnées, il faut impérativement :
Choisir pour les métadonnées un format structuré et construit.
Accoler à ce format un modèle de données et/ou une documentation sur les différentes catégories créées.
Le modèle de données et/ou la documentation des différentes catégories sont essentiels, car ils garantissent le fait que les données pourront encore être exploitées, éventuellement par d'autres, dans les années qui viennent.
Le modèle de données le plus couramment utilisé est le format XML, qui est un standard international. Les métadonnées exprimées dans ce format suivent un schéma prédéfini, qui définit les règles concernant l'usage des balises.
Parmi les modèles de données respectant le format XML, on trouve :
Metadata Encoding and Transmission Standard (METS)
Hypertext Markup Language (HTML)
Text Encoding Initiative (TEI)
On présente ci-dessous la TEI, modèle que l'on recommande quand on souhaite structurer (et par la suite, analyser, exploiter, exposer, etc.) des données textuelles.

6.4 LA TEI (TEXT ENCODING INITIATIVE)

La TEI a été lancée sur le plan international comme projet de recherche en 1987.
Depuis 2000, sa gestion et son évolution sont supportées par un consortium à but non lucratif. Un Conseil Technique TEI est chargé de l'amélioration du modèle et de ses aspects techniques.
La TEI fournit un modèle très riche de composants pour tous les types de textes. On peut ainsi choisir ceux qui seront pertinents dans le cadre du projet scientifique. Ces composants devront être articulés dans un schéma.
Plus souples qu'un schéma XML classique, les « TEI Guidelines », qui en sont à leur cinquième version (P5), proposent un ensemble de recommandations particulières, rassemblées dans des modules distincts. Elles comprennent essentiellement des indications sur la structure sémantique du contenu textuel, et une documentation formelle des balisages employés. Ces recommandations peuvent être adaptées en fonction de besoins particuliers.
Le système est largement utilisé en sciences humaines et sociales, où une communauté très large et active se charge de son évolution et de son exploitation pour l'édition des sources primaires (manuscrits, documents historiques, etc.), les ressources lexicales et linguistiques (dictionnaires, corpus, etc.), les textes littéraires et politiques, etc.

6.5 LISTE DES RECOMMANDATIONS

RECOMMANDATIONS
R13 : Pour garantir le bon codage de tous les caractères, choisir le codage UTF-8.
R14 : Pour la structuration des données textuelles, utiliser la TEI. Suivre les « TEI Guidelines », les adapter à son corpus et documenter ses choix.

6.6 LES RESSOURCES

Centre de ressources numériques CNTRL (Centre national de ressources textuelles et lexicales) : http://www.cnrtl.fr
Centre de ressources MEET (Mutualisation d'expériences d'encodages textuels)
http://meet-adonis.fr
Centre de ressources numériques TELMA (Traitement électronique des manuscrits et des archives) : http://www.cn-telma.fr
METS : http://www.loc.gov/standards/mets/
TEI : http://www.tei-c.org

6.7 FORMATS ET METADONNEES TEXTES DU PROJET BNFB

6.7.1 Nommage du fichier

Nom du lieu de la numérisation_numero d'ordre_n (n si plusieurs parties)
Exemple : oujda_01
niamey_10_1 (si plusieurs parties)
niamey_10_V1_1 (si volume et plusieurs parties)

6.7.1 Nom de l'extension

*.pdf

6.7.2 Format et résolution des fichiers

2 jeux en pdf : 300 dpi et 72 dpi
Eventuellement un OCR

6.7.3 Métadonnées

Dublin core
TEI si nécessaire

7. LES DONNEES ICONOGRAPHIQUES – IMAGES FIXES

7.1 LES DONNEES

Les données iconographiques - images fixes recouvrent aussi bien :
des photographies : diapositives, négatifs, tirages positifs, photos numériques ;
des documents visuels fixes : documents 2D numérisés, illustrations, plans, croquis, dessins, cartes anciennes ou plus récentes (à l'exclusion des cartes construites automatiquement à partir de coordonnées et données géographiques).
Elles rassemblent donc des données nativement numériques et des données non numériques au départ que l'on aura numérisées.
Images matricielles et images vectorielles
Sur un plan technique, on distingue :
les images matricielles ou « bitmap » ou « raster »
les images vectorielles « images 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 ; chacune de ces valeurs peut être modifiée indépendamment.
Les images vectorielles fournissent un ensemble d'instructions mathématiques utilisées par un programme de dessin pour construire une image.
Des logiciels tel que Photoshop ou Gimp créent et lisent en règle générale des images matricielles, alors qu'Illustrator crée et lit des images dites vectorielles. Les images vectorielles peuvent être converties en images matricielles. L'inverse n'est que difficilement possible et implique le plus souvent la reprise du dessin, à la main, dans un éditeur.

7.2 LA NUMERISATION

En général, le processus de numérisation génère une image matricielle, les images vectorielles étant le plus souvent le produit d'un logiciel de dessin.
Les différents choix qui doivent s'opérer lors de la numérisation sont fonction du document, de son format, du rendu tout comme de l'usage ultérieur que l'on souhaite faire ensuite du fichier.
Pour les images matricielles, deux éléments doivent être pris en considération : le format de fichier et les paramètres de qualité. De manière générale, les images doivent être créées à la meilleure résolution possible et à la profondeur de bits la plus élevée possible, mais en veillant à ce que les fichiers obtenus soient pratiques et maniables en fonction des utilisations envisagées.

7.2.1 Le format de fichier

Les images matricielles peuvent 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).
Il importe par ailleurs de créer toujours deux jeux de données, l'un sera utilisé pour la conservation et l'archivage ; l'autre sera exploité dans le cadre du web. Dans le premier cas, la numérisation devra être effectuée sans compression, donc sans perte.
On recommande alors le format TIFF. Dans le second cas, il pourra y avoir compression, à la condition cependant que la qualité reste maximale. On recommande alors le format JPEG.
Il est possible de créer le premier jeu au format TIFF puis d'utiliser un logiciel comme mogrify (http://www.imagemagick.org) pour créer un second jeu d'images.

7.2.2 Paramètres de qualité

La sélection des paramètres de qualité lors de 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. Doivent être prises en compte :

a) La résolution spatiale

La résolution spatiale établit la fréquence avec laquelle des échantillons de l'original sont capturés par le dispositif de numérisation. Elle est 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 enregistrée par unité de surface. Plus cette densité est haute, plus l'image numérisée sera de bonne qualité. La densité pour les pages web est normalement de 72 ppp. L'impression se sert normalement de densité oscillant entre 300 et 600 ppp.

b) La résolution des couleurs ou 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, est exprimé en nombre de bits. Par exemple, une résolution de couleurs de 8 bits signifie que 256 couleurs différentes sont disponibles.
Le choix de la profondeur de bits dépend du support de départ : numériser une diapositive de 35mm exige une résolution plus élevée que celle 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 doit être plus élevée que pour le seul affichage de l'image à l'écran.
Attention !
Plus la qualité de l'image numérisée est grande, plus le fichier sera lourd à manipuler.
Cependant, plus la qualité de l'image numérisée est grande, plus l'image pourra être agrandie sans que l'on perde en qualité visuelle.
À 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.
Dans certains cas, par exemple lors de l'utilisation d'appareils photo numériques de moindre qualité, les images peuvent être stockées sous un format JPEG/SPIFF, comme alternative au format TIFF. Les images seront alors plus petites et de plus basse qualité. De telles images peuvent être utiles pour la présentation de photographies d'événements pour un site internet. Mais l'utilisation de tels appareils photos n'est pas recommandée pour une numérisation à grande échelle.

7.3 LES METADONNEES

Sans métadonnées une image numérique est vierge de toute information. Il est impossible de déterminer quel en a été le contexte de production (auteur, lieu, date, etc.) et de ce fait l'image reste inexploitable. Il est donc nécessaire d'intégrer dans le document un certain nombre d'informations – des métadonnées - qui pourront ensuite être exploitées dans la chaîne de production en aval (récupérées lors de la création de bases de données, moissonnées par des moteurs de recherche, etc.).

7.3.1 Métadonnées techniques

Pour les photos numériques, un certain nombre de métadonnées sont automatiquement générées par les appareils photographiques eux-mêmes et contenues dans le fichier image lui-même. Ce sont ce qu'on appelle des métadonnées techniques. Elles concernent les paramètres de prise de vue, et les réglages des appareils photographiques numériques. Ces métadonnées peuvent être affichées, éditées ou extraites grâce à des logiciels libres.
Ces métadonnées n'ont pas vocation à être transformées, certains éléments pouvant même être endommagés en cas de modification. C'est pourquoi on recommande d'établir un autre jeu de métadonnées. Ce seront des métadonnées descriptives.

7.3.2 Les métadonnées techniques : les EXIF

EXIF est une spécification de format de fichier pour les images utilisées par les appareils photographiques numériques. Elle a été établie par le Japan Electronic Industry Development Association (JEIDA). La dernière version 2.3. a été publiée en avril 2010.
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'utilise. Il peut être également exprimé selon le standard MIX en XML.
Il permet le stockage d'un large éventail d'informations techniques concernant les paramètres de prise de vue et les réglages des appareils photographiques numériques lors de la capture numérique:
date et heure de la prise de vue ;
réglage de l'appareil (marque, modèle de l'appareil mais aussi d'autres informations comme la vitesse d'obturation, la longue de focale, la sensibilité, etc.) ;
informations géographiques provenant d'un éventuel système GPC connecté à l'appareil.
Ces données sont fournies automatiquement par l'appareil photographique numérique et sont contenues dans le fichier image lui-même.

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 secondes
FocalLength : distance focale en millimètres
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.
Plusieurs logiciels permettent d'afficher, d'éditer et d'extraire les métadonnées EXIF :
Exifer, Exif Reader, ExifTool, ExifPro Image Viewer, Exiv2, IrfanView, Photo Studio, XnView, etc.

7.3.3 Métadonnées descriptives

Dans tous les cas (images numérisées, images nativement numériques), il importe d'indiquer des métadonnées descriptives. Sous le terme de métadonnées descriptives, on rassemble des métadonnées de type technique (cf. ci-dessus), mais aussi des métadonnées de type documentaire, ciblant le contenu du document.

De manière générale, les métadonnées descriptives sont indiquées dans un autre fichier que celui du fichier de l'image, qui a le même nom que celui du fichier de l'image, à l'extension près.

7.3.4 Les métadonnées descriptives : les XMP et les IPTC

XMP (Extensible Metadata Platform)
Le format de métadonnées XMP a été lancé par la société Adobe en 2001. Adobe possède donc cette marque et en contrôle les spécifications.
Ce format structure de l'information et permet de l'enregistrer sous la forme d'un fichier XML, qui peut être inclus dans l'image (ce n'est pas ce que nous recommandons) ou bien stocké à part. Il utilise une expression en RDF simplifié de champs totalement paramétrables et adaptables à des besoins particuliers. Le format XMP est compatible avec le format IPTC via le format d'IPTC-Core (voir ci-dessous).
IPTC/IIM
Élaboré par le monde de la presse et des agences photographiques dans les années 1990, via l'International Press Telecommunications Council (IPTC)[5] pour leurs besoins spécifiques d'échange d'images et d'informations, le format IPTC/IIM (IIM pour Information Interchange Model) permet d'encapsuler les informations à propos du document. Les métadonnées souvent utilisées sont le nom de l'auteur ou du photographe, des informations sur le copyright et les descriptions. S'il est principalement utilisé pour les images de presse, il peut aussi être appliqué à d'autres types de documents, tels que le texte ou d'autres médias.
IPTC-Core
Dans sa dernière version, l'IPTC a adopté le format XMP. Le format de métadonnées IPTC-Core redéfinit ainsi en XMP les métadonnées IPTC/IIM et offre la possibilité d'ajouter aux champs IPTC standards de nouveaux champs. IPTC-Core n'est pas une norme ouverte, mais un standard de fait.
Il existe plusieurs logiciels qui permettent d'afficher, d'éditer et d'extraire les métadonnées XMP, IPTC/IIM et IPCT-Core : Exifer, ExifTool, Exiv2, IrfanView, PhotoThumb IPTCExt, Rodeo Info (Mac OSX), XnView, etc.
En parallèle, on recommande de compléter la description du document en utilisant le Dublin Core. Les métadonnées seront stockées dans un fichier séparé de format XML, dont le nom sera identique à celui du fichier de l'image à l'extension près : NomDuFichierDelImage.xml

RECOMMANDATIONS

R15 : Réaliser au moins deux jeux de données :

a) pour la conservation et l'archivage, le jeu de données sera numérisé sous une forme non comprimée, sans traitement supplémentaire : en haute résolution au format TIFF ;

b) pour une exploitation sur le Web, le jeu de données sera numérisé au format JPEG en qualité maximale.

R16 : Toujours réaliser plusieurs jeux de métadonnées. Pour les métadonnées techniques, utiliser EXIF ; pour les métadonnées descriptives, utiliser IPTC-Core.
R17 : Utiliser le Dublin Core pour compléter la description du document iconographique.
R18 : Stocker les métadonnées descriptives dans un fichier portant le même nom que le fichier de l'image à l'extension près.

7.4 LES RESSOURCES

Dublin Core : http://dublincore.org/ ; http://dublincore.org/documents/dcmi-terms
Didacticiel d'imagerie numérique de Cornell University :
http://www.library.cornell.edu/preservation/tutorialfrench/contents.html
EXIF : http://www.exif.org
IPTC : http://www.iptc.org
IPTC - Métadonnées :
http://www.iptc.org/cms/site/index.html;jsessionid=a6fFGl6cnmYe?channel=CH0089
Recommandations MINERVA :

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

http://www.minervaeurope.org/interoperability/digitisationguidelines.htm

7.5 FORMATS ET METADONNEES DES IMAGES FIXES DU PROJET BNFB

7.5.1 Nommage des fichiers

Nom du lieu de la numérisation_numero d'ordre_n (n si plusieurs parties)
Exemple : oujda_01
niamey_10_1 (si plusieurs parties)
niamey_10_V1_1 (si volume et plusieurs parties)

7.5.1 Nom de l'extension

*.tiff
*.jpeg

7.5.2 Format et résolution des fichiers

2 jeux d'images fixes
1 tiff – 300dpi
1 jpeg - 72 dpi

7.5.3 Métadonnées

IPTC
Logiciel pour les métadatas ITPTC (lecture et écriture)
http://www.embedmydata.com/?omeka
Logiciel pour les metadata IPTC : Irfant view http://www.irfanview.com/
Utiliser les plugins Irfanview pour lire et écrire les métadatas

Dublin Core

8. LES DONNEES ICONOGRAPHIQUES – IMAGES ANIMEES ET FILMS

8.1 LES DONNEES

Les données considérées ici sont des supports vidéo ou bien des films argentiques.
Pour les supports vidéos, il s'agit de :
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
Pour les films argentiques, il s'agit de :
8 mm, Super 8 mm, 9,5 mm
16 mm
35 mm

8.2 LES CONTENEURS ET LES CODECS

Les données multimédias (audiovisuelles/sonores) numériques font appel à deux notions techniques :
Codec (codeur/décodeur) : programme permettant d'encoder à la capture puis de décoder à la lecture les données. Il permet également de compresser et/ou décompresser ce signal. Il ne doit pas cependant être confondu avec le procédé de compression/décompression, même s'il permet de réaliser à la fois

l'encodage et la compression. Il est le plus souvent utilisé comme algorithme de compression pour réduire la taille d'un flux (vidéo/audio).
Conteneur : fichier « enveloppe » contenant des données destinées à être archivées et les informations relatives à l'interprétation de ces données. Il peut contenir divers types de formats, ce qui permet de gérer les différents flux audio et/ou vidéo codés à l'aide de codecs, ainsi que d'autres types de données : chapitrage, soustitres, métadonnées, description des flux contenus, etc.

8.3 LA NUMERISATION

Quand on prépare le corpus, il importe 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 différentes du corpus.
Faire attention à la dangerosité de certains matériaux.
Il faut connaître la composition des films argentiques. En effet, les films nitrate de cellulose[6], qui peuvent s'auto-enflammer, sont à identifier avec la plus grande précaution. Voici ce que préconise la BNF, dans son document Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques :
« 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. »
Lors de la numérisation, il convient de produire une version de conservation ET une version de diffusion. Pour plus de détails, on se reportera aux guides de la BNF et du TGE Adonis cités plus bas.

8.4 LES METADONNEES

Pour les métadonnées, on peut utiliser la norme MPEG-7. C'est une norme ISO élaborée par le MPEG (Moving Picture Experts Group - Groupe d'experts sur les images animées). Elle permet de décrire 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 techniques sur le fichier ;
des métadonnées documentaires (titre, créateur, droits, renseignements sur les personnes, les objets et les évènements représentés dans le fichier multimédia, etc.).
Cette norme reste cependant difficile à implémenter. C'est pourquoi, tout comme pour les images fixes, on recommande la structuration des métadonnées à l'aide du Dublin Core. Les métadonnées seront enregistrées dans un fichier XML séparé, portant le même nom que le fichier du document audiovisuel, à l'extension près.

8.5 LISTE DES RECOMMANDATIONS

RECOMMANDATIONS
R18 : Réaliser au moins deux versions numérisées, l'une pour la conservation l'autre pour la diffusion.
R20 : Pour les métadonnées, utiliser le Dublin Core. Les métadonnées sont enregistrées dans un fichier XML séparé, portant le même nom que le fichier du document audiovisuel à l'extension près.

8.6 LES RESSOURCES

Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques, BNF :

[6] Seul, les films antérieurs aux années 1960 (films d'archives anciens) sont susceptibles d'être des films nitrates car après cette date le nitrate a été remplacé par l'acétate (savety film) non auto-inflammable. »
Nota : certaines pellicules photographiques anciennes sont aussi en nitrate de cellulose.

http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/cahier_charges_numerisation.pdf
Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles, TGE ADONIS :
http://www.tge-adonis.fr/ressources/guides
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

8.7 FORMATS ET METADONNEES DES IMAGES ANIMEES DU PROJET BNFB

8.7.1 Nommage du fichier

Nom du lieu de la numérisation_numero d'ordre_n (n si plusieurs parties)
Exemple : oujda_01
niamey_10_1 (si plusieurs parties)
niamey_10_V1_1 (si volume et plusieurs parties)

8.7.2 Nom de l'extension

*.mp4

8.7.3 Le conteneur vidéo

Pour notre projet on retiendra :
MP4 (ISO/IEC 14496-14)
Conteneur multimédia issu de la norme MPEG-4 : MPEG-4 ASP, MPEG-4 AVC (vidéo) et AAC (audio). Il supporte tous les types de contenus multimédia (plusieurs pistes son, vidéo, sous-titres, etc.) et des contenus avancés (« Rich Media » ou BIFS (BInary Format for Scenes) : menus de type DVD, graphisme 2D/3D, etc.). Il est également distribuable en streaming.

8.7.4 Le codec vidéo

H.264 (MPEG-4 AVC : Advanced Video Coding)
Codec développé par le groupe MPEG et édité par l'UIT-T (Union Internationale des Télécommunications). Il offre de multiples techniques permettant d'améliorer le taux de compression par rapport au MPEG-2 et une meilleure qualité d'affichage. Il est également adapté à une très grande variété de réseaux et de systèmes : TNT, VOD, Blu-ray, téléphonie mobile (iPhone, iPad, etc.) et streaming.

8.7.5 Résolution des vidéos

HD : 1920x1080 en progressif
SD : 720x576 en progressif

8.7.6 Métadonnées

Dublin Core

9. LES DONNEES SONORES

9.1 LES DONNEES

Par données sonores, on entend l'ensemble des données audio: enregistrement de parole, de conversations ou de musique. Elles peuvent être considérées du point de vue aussi bien de leur contenu que du traitement linguistique dont elles peuvent faire l'objet.
Parmi les supports les plus couramment rencontrés dans les fonds d'archives, les bibliothèques, les musées ou les autres services culturels, on trouve[7] :
le cylindre
le disque « 78 tours »

[7] La liste est issue du guide de numérisation édité par la BNF.

le disque à gravure directe
le disque microsillon
la bande magnétique
la cassette
la micro cassette
le DAT (Digital Audio Tape)
le miniDisc
le CD audio (Compact Disc)

La nature du contenu et la typologie des supports sont en étroite relation avec le contexte de leur production. L'usage de tel ou tel support ne présage cependant en rien du caractère unique et de l'importance du contenu.

Attention !

L'évaluation de la qualité des supports (de leur boîte et de leur contenu) est primordiale avant d'entreprendre leur lecture et de s'engager dans un projet de numérisation.

9.2 LA NUMERISATION

Il convient de distinguer les données sonores analogiques et les données sonores nativement numériques.
Pour un document analogique, afin de permettre une exploitation la plus riche possible en termes d'information, il sera important de rechercher les niveaux les plus élevés en matière de quantification et de fréquence d'échantillonnage.
Pour un document nativement numérique, il n'est pas nécessaire de le copier dans un format de qualité supérieure à celui d'origine.
Quelle que soit l'origine du document (analogique ou nativement numérique), qu'il s'agisse de conservation, diffusion, ou de restauration, il importe de suivre les recommandations techniques associées à chaque type d'opération.

Il faut distinguer la restauration du support (bandes collantes, cassures, etc.) de la restauration du contenu.
Pour la conservation et la diffusion, il importe de faire plusieurs jeux de numérisation.

9.3 LES METADONNEES

Les métadonnées associées aux documents sonores peuvent être récupérées de manière automatique (lors de la numérisation) ; elles sont sinon indiquées manuellement.
Comme pour les autres types de données traitées dans ce guide, il est essentiel que les métadonnées soient renseignées selon des normes et des standards reconnus.
C'est la problématique de recherche et les objectifs du projet qui conduisent à se déterminer par rapport aux différents formats possibles.
Plusieurs schémas sont disponibles pour exposer les métadonnées. On peut utiliser des schémas génériques.
Pour une exploitation linguistique, on suivra les recommandations de l'Open Language
Archive Community (OLAC), qui est une organisation internationale dont l'objectif est le partage et la diffusion de ressources de nature linguistique. Celle-ci recommande l'utilisation du Dublin Core qualifié, auquel ont été ajoutés 5 attributs dont les valeurs sont liées à des vocabulaires contrôlés. Il s'agit de :
1. l'attribut « language » ajouté aux éléments DC « subject » et « language » et dont la valeur doit être prise dans le catalogue Ethnologue (http://www.ethnologue.com/) devenu depuis la nouvelle norme ISO-639 d'abréviation des langues sur 3 caractères ;
2. l'attribut « linguistic-field » ajouté à l'élément DC « subject ». Il doit prendre sa valeur dans une liste fermée (phonetics, phonology, pragmatics, psycholinguistics…) ;
3. l'attribut « discourse-type » ajouté aux éléments DC « type » et « subject », à choisir dans une liste fermée (drama, formulaic_discourse, interactive_discourse, language_play, oratory, narrative, procedural_discourse, report, singing, unintelligible_speech) ;
4. l'attribut « linguistic-data-type » ajouté à l'élément DC « type », à choisir dans une liste fermée (lexicon, primary_text, language_description) ;
5. l'attribut « role » peut être ajouté aux éléments « contributor » et « creator ». Il doit prendre sa valeur dans une liste fermée (recorder, researcher, singer, speaker, transcriber, translator…).

Le fichier rassemblant les métadonnées doit être stocké de manière séparée. C'est un fichier de format XML, dont le nom sera identique à celui du fichier sonore à l'extension près.

9.4 LISTE DES RECOMMANDATIONS

RECOMMANDATIONS

R19 : Numériser un document analogique au niveau le plus élevé en matière de quantification et de fréquence d'échantillonnage.
R20 : Formater un document nativement numérique en choisissant un format de qualité égal à celui d'origine.
R21 : Pour la conservation et la diffusion, prévoir plusieurs jeux de numérisation.
R22 : Pour une exploitation linguistique, renseigner les métadonnées dans le format OLAC.
R23 : Les métadonnées sont enregistrées dans un fichier XML séparé, portant le même nom que le fichier sonore, à l'extension près.

9.5 LES RESSOURCES

Centre de ressources numériques pour la description de l'oral :
Aix : http://crdo.fr
Paris : http://crdo.risc.cnrs.fr/exist/crdo/
Corpus Oraux, Guide des bonnes pratiques, O. Baude (dir.), PUO, 2006
http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Corpus_Oraux_GBP%202006_version_imprimee.pdf?CV=5584&type1=Ouvrage
Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques, BNF, août 2009 :
http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/cahier_charges_numerisation.pdf
OLAC : http://www.language-archives.org/
Projet TELEMETA : http://telemeta.org

9.6 FORMATS ET METADONNEES DES FICHIERS SONORES DU PROJET BNFB

9.6.1 Nommage du fichier

Nom du lieu de la numérisation_numero d'ordre_n (n si plusieurs parties)
Exemple : oujda_01
niamey_10_1 (si plusieurs parties)
niamey_10_V1_1 (si volume et plusieurs parties)

9.6.2 Nom de l'extension

*.m4a

9.6.3 Conteneur audio

1 fichier de conservation : **192 kbit/s @ 44,1 kHz pour conservation**
1 fichier de consultation : **128 kbit/s @ 44,1 kHz pour diffusion**

9.6.4 Codec audio

MPEG-4 AAC - technologie Advanced Audio Coding (AAC)

9.6.5 Métadonnées

Dublin Core

10. GLOSSAIRE

Attribut
En langage XML, un attribut apporte des informations sur l'élément qui le contient. Le nom des attributs, les éléments qui les contiennent et le type de valeurs que les attributs peuvent contenir, peuvent être précisés dans des schémas (schémas XML, DTD, relaxng, etc.).
Ex. <element type= "exemple">. Ici l'attribut « type » est renseigné par la valeur « exemple », qui précise le nom de la balise <element>.

Balise
Une balise est un élément fondamental des langages d'encodage. Elle fonctionne comme un marqueur syntaxique. Une balise permet d'identifier et de qualifier des contenus. Par exemple une balise <title> désigne un titre. Le nom d'une balise est formellement écrit entre deux chevrons. Les balises fonctionnent par paire : une balise ouvrante et une balise fermante.
Ex :<title>La géographie du notaire languedocien</title>

Compression
Compresser une information numérique permet d'en réduire la taille d'où :
un gain d'espace ;
une réduction de l'infrastructure inhérente (volume d'une baie de stockage, d'archivage, ou de consommation électrique, etc.) ;
une réduction du temps de transfert lors de téléchargements.
En ce qui concerne les sources orales ou visuelles, la compression pourra cependant entraîner l'augmentation du temps de traitement pour la lecture, qui est en général compensée par de plus faibles volumes à traiter.

Les compressions peuvent être
a) sans perte : l'opération est dite « réversible ». Le document produit après compression puis décompression est identique au document original ;
b) avec pertes : dans ce cas la fidélité de l'écoute et de la visualisation d'une séquence audiovisuelle peut être plus ou moins altérée selon le « codec » choisi.
La compression sans perte est idéale pour l'archivage à long terme : au gain de place s'ajoute la fidélité de la restitution.
Pour les images numériques, chaque pixel est défini par une série d'informations digitales : plus il y a d'informations, plus l'image est de bonne qualité mais plus le fichier est lourd. Le « codec » généralement utilisé pour compresser les images et en réduire le poids est le JPEG.

Dublin Core
C'est une norme internationale très souvent employée pour la description et l'échange de métadonnées. Son langage simplifié (15 balises ou « descripteurs ») permet de faciliter l'échange de métadonnées génériques et l'interopérabilité entre différents projets.

EXIF (Exchangeable Image File)
Format de métadonnées images. Les métadonnées EXIF sont générées automatiquement par les appareils de prise de vue numériques. Selon les fabricants, on retrouve à peu près les mêmes types d'informations concernant la prise de vue (date, heure, diaphragme, vitesse, focal, avec ou sans flash).

Format informatique
Un format informatique est une convention pour représenter une donnée sous forme numérique. Il peut être spécifié, ouvert, normalisé, standardisé ou propriétaire.

Format normalisé ou standardisé
Un format est normalisé ou standardisé quand sa description est adoptée par un organisme de normalisation ou de standardisation. Parmi ces organismes dans le domaine des technologies de l'information, on citera :
AFNOR – Association française de normalisation – http:// www.afnor.org/
ISO – Organisation Internationale de normalisation – http://www.iso.org/
OASIS – Organization for the Advancement of Structured Information Standards – http://www.oasis-open.org/
W3C – World Wide Web Consortium – http://wwww.w3.org/

Format ouvert
L'article 4 de la loi française n°2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique, définit un format ouvert : « on entend par standard ouvert, tout protocole de communication,

d'interconnexion ou d'échange et tout format de données interopérable et dont les spécifications techniques sont publiques et sans restriction d'accès ni de mise en oeuvre ».
Un format ouvert est légalement exempté de droits d'utilisation et sa description est publique. Il est alors compréhensible et interopérable. Il est compréhensible car sa description ou spécification est publique : tout le monde peut alors prendre connaissance de la manière dont les informations sont organisées au niveau de ce format. Il est alors possible à partir de cette connaissance de créer une variété de programmes ou d'équipements qui l'exploitent. On dit d'un tel format qu'il est interopérable. Les notions de format ouvert et de format libre sont très proches. Un format sera qualité de libre uniquement si aucune restriction juridique ne lui est applicable. Un format qui n'est pas « ouvert » est naturellement dit « fermé ».

Format propriétaire

Un format est dit propriétaire si son cadre d'utilisation est contrôlable par une personne ou une entité juridique. Ce droit peut s'établir par exemple via le droit d'auteur, le brevet ou le copyright. Cependant, même si l'utilisation du format est contrôlable, cela ne signifie pas qu'elle soit obligatoirement contrôlée. Ainsi le format PDF est ouvert, car ses spécifications sont libres d'accès et que son propriétaire Adobe Systems, société de droit privé, autorise des programmes tiers à réutiliser son format.
Ce format est donc ouvert même s'il est propriétaire. Ces deux notions ne sont pas antinomiques.
Le terme propriétaire est souvent et abusivement employé pour désigner un format dont l'utilisation est fortement restreinte par les droits que possède son propriétaire.
Si tel est le cas – et si la spécification n'est même pas consultable – on parle de format fermé. Un format qui n'est pas « propriétaire » est un format dit « libre ».

Format spécifié

Un format est dit spécifié lorsqu'il est suffisamment décrit pour en développer une implémentation complète. La spécification est souvent trouvée sous la forme d'un fichier au format PDF ou TEXT, en une ou plusieurs langues. Elle contient des informations qui nécessitent le plus souvent une bonne connaissance en informatique.
Il n'y a pas d'adresse particulière regroupant toutes les spécifications. Elles se trouvent, le plus souvent, sur le site internet du propriétaire du format ou sur celui de l'organisme qui a édité une norme à son sujet.

HTML (Hyper Text Mark-Up Language)

Langage de balisage qui permet d'écrire de l'hypertexte et de structure sémantiquement le contenu de pages web. Il contient un nombre fixe de balises.

Interopérabilité

Des ressources numériques sont dites interopérables quand elles sont décrites et exposées à l'aide de formats standardisés ou normalisés. Elles peuvent alors être recherchées, identifiées, exposées, partagées, réutilisées, etc.

IPTC (International Press Telecommunications Council)

Consortium réunissant les principales agences de presses et photographiques dont la principale activité consiste à développer et à maintenir des standards permettant l'échanges des données. Il a notamment mis au point le format IPTC/IIM (pour Information Interchange Model). Ce format de métadonnées documentaires est spécifique aux images.

JPEG (Joint Photographic Experts Groups)

Né de la fusion en 1986 de plusieurs groupes de professionnels de l'industrie de l'image, ce groupe à donné son nom à une norme ouverte de compression d'images numériques, le JPEG. C'est un format de compression à perte (il élimine donc des informations) mais son taux de compression est réglable, ce qui permet un bon compromis entre le taux de compression et la qualité de l'image compressée.

Métadonnée

Une métadonnée est une donnée servant à définir ou à décrire une autre donnée, qui représente le document numérique résultant de la transformation de la source première. Les métadonnées sont à la base des techniques du web sémantique. Elles doivent donc être rédigées en tenant compte des standards, car ce sont elles qui permettent l'accès aux données et qui garantissent l'interopérabilité. Une métadonnée isolée est difficilement concevable. Elles sont le plus souvent associées en « jeux de métadonnées » metadata set qui sont structurés et modélisés (exemple le DC, METS etc…

METS (Metadata Encoding and Transmission Standard)

Schéma XML permettant la description d'objets numériques. Il est particulièrement destiné aux échanges entre institutions patrimoniales et est conforme aux recommandations de l'OAIS (Open Archival Information System). Ce schéma est maintenu actuellement par la Bibliothèque du Congrès.

Numérisation
Au sens le plus répandu, la numérisation est la conversion d'un signal (vidéo, image, audio, caractère d'imprimerie, impulsion, etc.) en une suite de nombres permettant de représenter cet objet en informatique ou en électronique numérique.
OAIS (Open Archival Information System)
Modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme des documents numériques. C'est une norme ISO (référence 14721). Il permet de décrire les fonctions, les responsabilités et l'organisation d'un système qui souhaite préserver de l'information.
OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting)
Protocole standard interopérable qui permet d'avoir accès aux métadonnées d'un entrepôt de données numériques.
RDF (Resource Description Framework)
C'est un modèle de représentation de données, qui relie des triplets (sujet-prédicatobjet) dans un graphe. Il permet de décrire de façon formelle des ressources web et leurs métadonnées et des traiter informatiquement. Développé par le W3C, RDF est le langage de base du web sémantique.
TEI (Text Encoding Initiative)
Il s'agit d'un langage XML permettant de structurer des données et des métadonnées textuelles. C'est un système de modélisation textuelle, permettant la construction des schémas XML pour la structuration des données et des métadonnées textuelles, très répandu dans le domaine scientifique.
Unicode
Norme visant à donner de manière unifiée à tout caractère de n'importe quel système d'écriture, un nom et un identifiant numérique.
W3C (World Wide Web Consortium)
Organisme de standardisation fondé en octobre 1994, par Tim Berners-Lee (principal inventeur du web). Le consortium est chargé de promouvoir la compatibilité des technologiques du web. Il a été fondé au MIT/LCS (Massachusetts Institute of Technology/Laboratory for Computer Science) avec le soutien de l'organisme de défense américain DARPA et la Commission européenne.
XML (Extensible Markup Language)
Langage de balisage extensible. C'est un langage qui permet de structurer sous forme d'arborescence les données. À la différence du langage HTML dont le nombre de balises est fixe, le langage XML peut accueillir autant de nouvelles balises que nécessaire.
XMP (eXtensible Metadata Platform)
Format de métadonnées documentaires particulièrement utilisé pour les images.

11. TABLEAU RECAPITULATIF

	Textes	**Image fixe**		**vidéo**	**Fichier sonore**
Résolution archivage	300 dpi	300 dpi		Si HD : 1920x1080 Progressif Si SD : 720x576 Progressif	192 kbit/s @ 44,1 kHz pour conservation
OCR	Si possible				
Résolution diffusion OMEKA	72 dpi	72 dpi		Non décidé	128 kbit/s @ 44,1 kHz pour diffusion
Codec				H264 (pas de compression)	MPEG-4 AAC - technologie Advanced Audio Coding (AAC)
Extension de fichiers	pdf	tiff pour archivage jpeg pour OMEKA		mp4	m4a
Metadonnées	Dublin Core	Dublin Core XMP (éditer avec xnview)		Dublin Core	Dublin Core avec attributs
Extension Métadonnées	xml	xml		xml	xml

Télécom SudParis
Département ARTEMIS – Alain Vaucelle
9 rue Charles Fourier
91011 Évry Cedex - France
+33 6 79 05 13 77
Alain.vaucelle@telecom-suparis.eu
www.telecom-sudparis.eu