



Guide pratique pour la production de corpus numérique

Maud Ingarao, Samantha Saïdi

Ce guide peut être téléchargé sur : <http://www.mutec-shs.fr/guides-Mutec>

Guide pratique pour la production de corpus numérique

Tout projet de recherche en SHS visant à produire un corpus¹ numérique, qu'il soit question d'une base de données relationnelle ou XML, d'une iconothèque numérique spécialisée en archéologie, d'une édition critique en ligne ou d'autres choses encore, va réclamer des ressources et des compétences souvent très éloignées du champ disciplinaire d'origine (science politique, histoire...), et donc en général absentes ou partiellement présentes au sein de l'équipe scientifique qui conçoit le projet. Image numérique, ingénierie éditoriale du numérique, modélisation de bases de données, bibliothèques, développement logiciel, archivage numérique pérenne... ces ressources et compétences qui relèvent d'autres métiers sont nombreuses, complexes, et ne peuvent s'improviser.

Ce guide se présente comme une feuille de route qui balaye les grandes étapes d'un projet de production d'une édition critique ou d'un corpus numérique. Pour chacune de ces étapes est proposé un éclairage sur :

- Les différentes méthodes possibles et leurs enjeux
- Les compétences et connaissances métier nécessaires à une réalisation de qualité et pérenne, et des exemples d'acteurs et institutions clés dans ce(s) domaine(s).
- Des exemples puisés dans des projets réalisés ou en cours de réalisation

A qui s'adresse ce guide ?

- Ce guide s'adresse aux chercheur·e·s en SHS qui envisagent de se lancer dans la production d'un corpus numérique, afin de les aider à balayer l'ensemble des chantiers et sous-chantiers impliqués et à détailler les ressources et compétences nécessaires à chacun d'eux. Ce guide voudrait ainsi être une aide pour mieux planifier et budgéter ce type de projet, et pour nouer d'emblée les partenariats et les contacts indispensables à leur succès.
- Il s'adresse dans le même temps aux ingénieur·es et technicien·nes chargés de la mise en œuvre de ces projets au sein des équipes SHS. Dans un paysage professionnel encore relativement peu structuré sur toutes ces questions, MutEC pointe en effet l'isolement des ingénieurs et techniciens au sein des équipes et l'absence de réseaux de métiers structurés comme l'un des principaux obstacles à la production à large échelle de corpus scientifiques numériques de qualité et pérennes. MutEC cherche donc à être un lieu d'analyse, de mise en forme et de capitalisation des méthodes, outils et bonnes pratiques qui émergent dans le champs de la production et de l'édition de corpus numériques scientifiques. La présente feuille de route est l'un des résultats de ce travail.

Remarque : ce guide est modulaire, au sens où tous les projets ne comportent pas nécessairement toutes les étapes présentées (par exemple : un projet d'édition exclusivement textuel ne sera pas concerné par la problématique des images numériques).

¹ Le terme corpus est ici employé dans son acception large.

Ce que ce guide n'est pas

Attention, ce guide n'est pas :

- **Un cours de conduite de projet ni de management par projet.** Si l'on souhaite adopter une démarche projet, on se réfèrera utilement à l'état de l'art en la matière (de nombreux outils théoriques et pratiques de planification et de gestion de projet existent : cahier des charges, diagramme de Gantt, etc.) et l'on pourra décider de doter l'équipe de compétences organisationnelles spécifiques.
- Un état de l'art ou **un guide sur le document numérique et la gestion électronique de document (GED)** en général, qui sont des champs de savoirs théoriques et méthodologiques à part entière pouvant être utilement interrogés ou même intégrés à l'équipe dans le cadre d'un projet de corpus scientifique numérique ;
- **Un cours de génie logiciel ni de gestion de projet informatique**, qui représente également à soi seul un vaste champ de savoirs et de pratiques et des compétences qu'il semble indispensable d'intégrer au sein du projet.
- Un cours ou **un état de l'art sur la numérisation et l'image numérique**, abordées ici très succinctement. Nous conseillons de se reporter aux conseils, outils ainsi qu'au travail de veille du centre national de ressources [CN2SV](#), spécialisée dans l'informatisation de données visuelles (photos, diapos, carnets de terrains, cartes, planches, dessins, croquis, etc.)
- De même, ce guide étant rédigé par des personnes qui travaillent principalement avec des corpus textuels, nous n'abordons pas **la numérisation de vidéo ou de son**. Nous conseillons de se reporter aux conseils, outils ainsi qu'au travail de veille du centre national de ressources [CRDO](#), spécialisé dans la gestion documentaire des ressources orales et la constitution d'un réservoir de données. On peut également se référer au site [Corinthe](#), site dédié à la recherche sur les corpus de langue parlée en interaction proposé par le groupe ICOR de l'UMR 5191.

1. Concevoir un projet : aspects scientifiques, éditoriaux et techniques

Une aventure numérique commence le plus souvent par le repérage, par des chercheur-es spécialistes, d'une ressource (fonds d'archive, ensemble documentaire, collection, etc.) présentant une grande valeur scientifique et/ou patrimoniale, et qu'ils décident alors de "numériser".

Il est alors crucial de franchir une première étape consistant à se demander précisément pourquoi on souhaite numériser cette ressource. Cette réflexion va en effet conditionner de manière fondamentale l'orientation générale et l'ampleur du projet.

Voici ci-après quelques exemples de questions à se poser pour préciser les objectifs visés et ainsi initier la conception du projet.

1.1. Pourquoi (numériser) ce corpus ?

Que vise-t-on avec cette numérisation ?

- La **conservation**, le souci de préservation des documents ? Cette conservation est-elle temporaire ou doit-elle être pérenne ? Toute numérisation est une perte d'information : quelles informations importe-t-il de conserver (image ? couleur ? texte ? son ? etc.)
- La **diffusion** auprès de la communauté scientifique intéressée ? du "grand public" ? de certains publics ? Quels modes d'accès seront intéressants pour ces publics : quelles organisations du corpus, quels modes de navigation, de recherche, quel outillage ? Faut-il prévoir plusieurs organisations concurrentes, conjuguées ? (voir aussi 1.2)
- L'**exploitation**, l'étude scientifique du corpus, c'est-à-dire l'observation, la recherche de certaines propriétés du corpus ? Selon quelles méthodes scientifiques, issues de quelles disciplines (linguistique de corpus, traitement automatique de l'image numérique, ...) ? Y a-t-il alors des formats d'encodage, des standards à respecter pour pouvoir bénéficier des savoir-faire informatiques et des instruments logiciels de ces disciplines ?

A retenir

- La numérisation d'un corpus peut avoir plusieurs objectifs très différents, et chacun de ces objectifs appelle une démarche spécifique, qui n'implique pas les mêmes compétences et ressources scientifiques, éditoriales et techniques que les autres (même si certaines opérations sont communes à tous les projets de numérisation). On peut bien entendu décider qu'on vise tous ces objectifs à la fois, mais il faut alors être certain-e de pouvoir disposer de toute-s les compétences et ressources nécessaires.

Bonne pratique

- Une bonne pratique peut être de se concentrer plutôt sur l'un des objectifs seulement (conservation OU étude scientifique OU autre) tout en veillant à produire des données numériques réutilisables par d'autres équipes qui chercheront à atteindre d'autres objectifs (préférer l'usage de standards, documenter les choix effectués, etc.)

1.2. Diffuser ? Publier ? Mettre en ligne ?

Dès la conception d'un projet, il importe de se demander sous quelle forme on souhaite produire les résultats. Cela ne signifie pas que l'on doit savoir complètement en début de projet tout ce que l'on va découvrir et produire - c'est même le propre de la recherche d'être en situation inverse. Mais il importe d'explicitier le statut scientifique et éditorial que l'on souhaite conférer aux différents types de résultats que l'on va produire. Ce qui est évidemment pour le médium papier (la hiérarchie de statut qui existe entre un article dans une revue à comité de lecture ou une tribune dans un quotidien, une communication dans un séminaire ou dans un colloque, un rapport de recherche ou un ouvrage dans telle collection de telle maison d'édition, etc.) ne l'est pas encore pour les productions numériques.

De même, si le cycle de vie d'un ouvrage papier est aujourd'hui institutionnellement et professionnellement bien balisé, du bon à tirer au catalogage en bibliothèque en passant par l'impression, le dépôt légal, etc., ce n'est pas encore le cas avec les productions numériques.

En somme, c'est toute une culture de l'édition numérique qui est en émergence et n'est pas encore stabilisée, et à laquelle chacun-e peut contribuer en prenant conscience des différentes destinations sociales possibles d'une production numérique, et en s'efforçant de positionner clairement ses propres productions. Voici quelques exemples de questions à se poser pour saisir ces enjeux :

- Veut-on construire un corpus pour répondre à ses besoins de recherche propres - ou ceux de son équipe - sans démarche de diffusion extérieure ? Le but est-il au contraire de mettre des ressources à la disposition de la communauté scientifique ? Du "grand public" ?
- Même si le corpus n'est pas destiné à être diffusé publiquement, quels éléments devront néanmoins être diffusés, voire publiés, pour rendre les résultats scientifiques du projet vérifiables, reproductibles par les pair-es ? Quels éléments devront être visibles par les tutelles et/ou organismes financeurs ?
- Que peut signifier "publier un corpus" ? Quels critères de qualité scientifiques, éditoriaux et techniques faut-il respecter ? Existe-t-il des "labels", des organismes susceptibles de cautionner la qualité d'une production numérique comme le fait une maison d'édition ou une revue pour un article ou un ouvrage papier ?
- Que deviendront les éléments produits une fois le projet terminé ? Qui sera en mesure de les conserver, de les maintenir d'un point de vue informatique, d'en assurer l'accès à long terme ? Comment la communauté scientifique pourra-t-elle se référer à ces productions, les "citer" dans dix ans ? (voir 3. Conserver un corpus numérique)
- Comment protéger la paternité d'un travail diffusé en ligne ? Qu'en est-il des droits d'auteur ?

A retenir

- Le numérique est une technique, un format et un support, qui en soi ne confère aucun statut scientifique ou éditorial à un objet. La destination sociale d'une production numérique doit être pensée et explicitée par son ou ses auteurs.
- Sur internet, le droit d'auteur et la propriété intellectuelle s'appliquent, y compris aux logiciels qui, en Europe, sont protégés par le droit d'auteur. Nul ne peut utiliser une production "trouvée sur internet" sans autorisation expresse des détenteurs des droits.

Bonnes pratiques

- Toujours s'assurer que l'on détient les droits et/ou les autorisations expresses des détenteurs des droits sur les objets utilisés dans le projet. Organiser la gestion de son propre droit d'auteur, expliciter les droits que l'on accorde sur ses productions numériques.
- Se rapprocher du service juridique des institutions qui financent le projet. Dans la plupart des cas ce sont elles qui détiendront les droits patrimoniaux sur le corpus numérique, les logiciels et autres éléments produits dans le cadre du projet.

Par exemple : les licences-types Creative Commons (un jeu de six licences combinant différents types de droits/interdictions : <http://fr.creativecommons.org/contrats.htm>) facilitent cette explicitation. La diffusion sous licence libre (en particulier pour les logiciels) peut être particulièrement adaptée aux productions des laboratoires de recherche. Attention, libre ne signifie pas sans auteur ! Voir par exemple <http://www.projet-plume.org/fr/ressource/declaration-de-berlin>

1.3. Comment s'organiser ? Qu'est-ce que documenter le projet ?

En plus des fichiers numériques eux-même, le projet va produire énormément de documents (CR de réunions, échanges de mails, rapports, comparatifs, tableaux de recette, etc.) La bonne gestion de cette documentation, à la fois pour le suivi du projet et pour sa pérennisation, est cruciale. Elle doit être accessible à l'ensemble des participant-es, structurée, régulièrement synthétisée. Elle doit être pour chacun-e une aide pour se repérer dans le temps et par rapport aux différents chantiers définis. Il peut être stratégique de prévoir une personne ressource dont ce sera le rôle unique. Des outils dédiés à la gestion de projet existent (dotProject, etc.) Leur mise en place est un risque (ces outils ont un coût d'apprentissage, de maintenance, ils exigent une grande rigueur d'utilisation, ils peuvent donc s'avérer sur-dimensionnés et coûter plus qu'ils ne rapportent au projet) mais est indispensable à partir d'un certain niveau de complexité (nombre de personnes impliquées, distance, durée...).

Bonnes pratiques

- Distinguer travail collaboratif et travail coopératif : produire un même objet à plusieurs et organiser la structuration et la circulation de l'information dans la durée au sein d'une équipe sont deux choses différentes, qui demandent des outils différents - ou des usages différents d'un même outil.
- S'assurer de la pérennité de l'hébergement et de la maintenance technique des outils organisationnels choisis : sur quel serveur seront-ils installés, sous la responsabilité institutionnelle de qui ? Qui y aura accès en pratique, qui sait administrer ces logiciels, bases de données, etc. ? Ces compétences seront-elles présentes de manière pérenne dans le projet ?
- S'assurer de la bonne prise en main de ces outils par toute l'équipe : y a-t-il un besoin de formation ? D'assistance ? Faut-il prévoir une homogénéisation des habitudes et des usages de ces outils ?

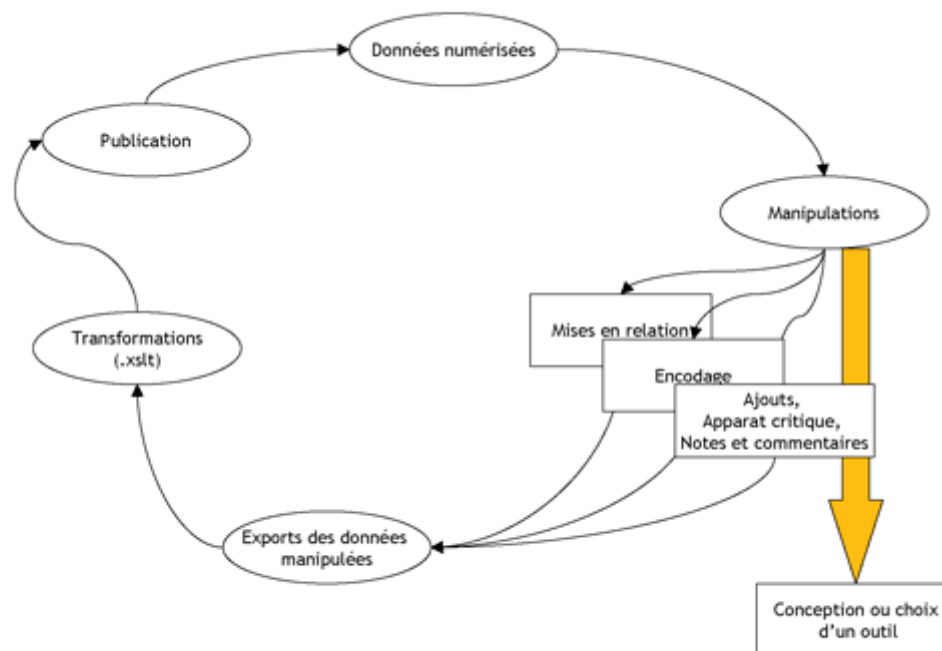
Par exemple : Listes des outils de travail collaboratif et coopératif faisant l'objet de fiches descriptives sur la plate-forme Plume :

[http://www.projet-plume.org/logiciels_valides?tid\[\]=221](http://www.projet-plume.org/logiciels_valides?tid[]=221)

[http://www.projet-plume.org/logiciels_valides?tid\[\]=220](http://www.projet-plume.org/logiciels_valides?tid[]=220)

2. Construire et encoder le corpus

Dans la plupart des cas, le caractère numérique/informatique du projet va induire des tâches préparatoires spécifiques qui vont s'ajouter et se mêler aux impératifs "traditionnels" de la discipline en matière de constitution de corpus.



Avec le numérique, il y a toujours plusieurs manières différentes possibles de réaliser les choses. Commencer par détailler toutes les réalisations (intermédiaires et finales) souhaitées aide fortement à choisir entre une ou une autre voie. Quelques questions clés à se poser pour cela :

1) Dans quelle mesure dois-je numériser les données de mes documents?

- **Mode image : iconographie**
 - Souci de conservation du document sous sa forme originelle? (mode image?)
 - Illustrations
 - La numérisation devra-t-elle permettre une consultation fine de l'image?
 - etc.
- **Mode texte : données textuelles**
 - Quels types d'informations voulez-vous extraire?
 - Quelques types d'opérations voulez-vous effectuer sur les documents sources?
 - En quoi le format numérique facilitera-t-il l'extraction ou des opérations sur les données? Ex : automatisation, rapidité, etc.
 - Besoin d'analyser le texte grâce à un encodage? Fonctionnalités de recherche ?
 - Recherche full-text ?
 - Recherche thématique, structurelle ?
 - Quel mode de consultation ? Unique ou plusieurs parcours de lecture...? Quel mode d'accès ? Ouvert/restreint ?

Description des réalisations intermédiaires et finales :

(Nous n'entrons pas encore dans le détail des modalités de diffusion (public visé, accessibilité, rythme de publication, etc.).

Exemple 1

- Etape a) inventaires dans une base de données : personnes, lieux, événements, etc.
- Etape b) puis requêtes sur cette base de données pour extraction des matériaux à utiliser : listes, graphiques, statistiques, etc.
- Etape c) puis édition : ouvrage papier utilisant ces matériaux | ou | ouvrage électronique utilisant ces matériaux

Exemple 2

- Etape a) numérisation + reconnaissance de caractères sur documents sources OU transcription de documents sources,
- Etape b) encodage/balisateur de ces textes au format XML/TEI,
- Etape c) Et enfin diffusion en ligne de ces textes encodés avec différents points d'entrée possibles dans le texte (listes de personnes, de lieux, moteur de recherche, etc.)

2) Existe-t-il des influences extérieures ?

Devez-vous respecter un standard défini pour la communauté scientifique à laquelle vous appartenez ou à laquelle appartiennent les ingénieurs du projet?

3) Outils, supports informatiques envisagés ?

- Outil de travail (back office) / Outil de publication (front office).
- Documents, sites web relatifs ou inspirateurs au/du projet ?

2.1. Collecter les documents sources et les décrire : les métadonnées

Même si cette étape semble inutile et fastidieuse, il est primordial pour la mémoire et la qualité scientifique du projet, de référencer les documents sources ou secondaires utilisés, que ceux-ci fassent déjà l'objet d'une description détaillée par un organisme spécialisé (bnf, etc.) ou qu'il s'agisse de documents inédits (ex : des archives inédites). A cette étape, il peut être judicieux de faire appel à un spécialiste de la documentation.

Pour les sources déjà décrites, est-il possible de récupérer les notices existantes ?

- si oui, sous quel format ? Base de données, fiches papier...
- si oui, avec quelle norme ? EAD, UNIMARC...

Sinon, voici, quelques questions à se poser pour déterminer les informations à renseigner lors du référencement :

- Observer un standard de métadonnées (par exemple EAD) : quels champs sont pertinents pour la description des sources du projet ? Lesquels ne semblent pas pertinents mais sont pourtant signalés comme obligatoires par le standard ? Y a-t-il des informations qui semblent incontournables dans le projet mais que le standard ne prévoit pas ? Qu'en déduire ?
- Quels éléments clefs de la structure des documents souhaite-t-on garder ?
- Quels types d'informations clefs relève-t-on dans les documents ?

A retenir

- L'absence d'un type d'information souhaité dans un standard ne signifie pas que le standard est inadapté au projet. La plupart des standards prévoient des champs génériques pour l'expression d'informations spécifiques à un corpus donné et non prévues par le standard (exemple en EAD, le champ odd, Other Descriptive Data : <http://www.loc.gov/ead/tglib/elements/odd.html>). Même s'il manque quelques types d'informations souhaités dans le standard, le coût de création d'une structure de notice ad hoc risque fort d'être plus élevé que les bénéfices de l'utilisation du standard aménagé grâce aux champs génériques qu'il prévoit.

Décrire les sources, mode d'emploi

a) Trouver les sources

La recherche de sources primaires et secondaires est au cœur de la formation des chercheurs/euses. Nous n'entrerons donc pas dans le détail ici.

On pourra cependant rappeler quelques questions permettant la délimitation du corpus :

- Quels critères de choix ont présidé à la constitution du corpus?
- Quelle est la volumétrie du corpus (un ou plusieurs documents ?)
- Quelle région géographique délimite ce corpus? (Europe? Royaume Uni? etc.)
- Quelle période délimite ce corpus?
- Exhaustivité?

b) Documenter, référencer

Un problème récurrent dans les projets que nous avons observés est le manque de documentation autour des sources choisies². Même si cette étape semble inutile et fastidieuse, il est primordial pour la mémoire et la qualité scientifique du projet, de référencer les documents sources ou secondaires utilisés, que ceux-ci fassent déjà l'objet d'une description détaillée par un organisme spécialisé (bnf, etc.) ou qu'il s'agisse de documents inédits (ex : des archives inédites). A cette étape, il semble judicieux de faire appel à un spécialiste de la documentation.

Pour les sources déjà décrites, vous pourrez essayer de récupérer les notices existantes :

- catalogage : métadonnées, inventaire : EAD
- si oui, sous quel format ? Base de données, fiches papier...
- si oui, avec quelle norme ? EAD, UNIMARC...³

² Nous nous concentrerons ici sur les sources d'un corpus écrit (textes, iconographique, ...). Pour les corpus Oraux, voir le site [Corinthe](#).

³ Pour aller plus loin, quelques informations utiles sur les normes de description et d'encodage : [IMARK] Pour mieux connaître les formats d'indexation, utiliser le tutoriel IMARK, unité 4 "création et gestion de documents électroniques" : http://www.imarkgroup.org/moduledescription_fr.asp?id=6 * [EAD site] Une documentation très complète sur l'EAD est disponible sur ce site : * [EAD site] Bibliographie EAD : <http://www.archivists.org/saagroups/ead/bibliography.html> * [EAD site] Liste de métadonnées et standards : <http://www.archivists.org/saagroups/ead/metadata.html> * [ArchivisToolKit] Outils pour l'archiviste : <http://archiviststoolkit.org/> * Métadonnées : <http://www.archivists.org/saagroups/ead/metadata.html> * [CNS2SV] Documentation du CN2SV (XML, EAD, EAC) : http://archivesic.ccsd.cnrs.fr/sic_00285219/fr/ * [TEI Initiative] Le "guidelines" de la dernière version P5 <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> * [Poupeau] - Poupeau G. , Réflexions sur l'utilisation de la TEI pour coder les sources diplomatiques à partir de l'exemple du Cartulaire blanc de l'abbaye de Saint-Denis, Le Médiéviste et l'ordinateur, 43, 2004 * [Poupeau] - Poupeau G. , Les possibilités de la TEI P5 pour les sources historiques : l'exemple d'un recueil de chartes, @SIC, 2007. <http://archivesic.ccsd.cnrs.fr/index.php?halsid=2jbnuba7sc22ntvrgcrbpo6n...> * TEI Réseau d'experts TEI (Bertrand Gaiffe) * TEI TELMA : création d'un environnement de production TEI

- est-il possible de récupérer les notices existantes?

Sinon, voici, quelques questions que vous pourrez vous poser pour choisir les informations à renseigner lors du référencement :

- quels champs choisir pour la description?
- quels éléments clefs de cette structure souhaitez-vous garder ?
- quels types d'informations clefs relevez-vous dans vos documents ?

Exemple :

Exemple de référencement bibliographique pour les sources écrites du projet :	
Champs à décrire :	Source décrite
n° de la réf. :	REf01
Titre :	Truc sur les trucs
Auteur :	Prénom Nom
éditeur critique/scientifique :	Prénom Nom
Traducteur :	Prénom Nom
Publication :	monographie
Imprimeurs - maison d'édition :	Nom
Localisation :	Ville
ISBN :	2-58965-563-5
Année de publication :	aaaa
N° de notice :	FRBNF55669944
Langue de la publication	français
Type de support :	
	docs papier : texte en écriture manuscrite, caract. d'imprimerie non standard
	docs papier : texte en caractères d'imprimerie standard, livres, "tapuscrit"
	fichiers numériques images : tif, gif, png, jpg, pdf (texte non sélectionnable)
	fichiers numériques texte : word, odt, pdf avec texte sélectionnable
Type de transcription	
Langue du titre	

Exemple de référencement pour la Transcription ou numérisation de la source pour le projet UNTEL :

Champs à décrire :	Source décrite
Nom du fichier :	définir 1règle de nommage
Encodage du fichier :	ex : iso-8859-1, cp1252
Date :	jj/mm/aaaa
Transcripteur :	Prénom Nom
Support utilisé :	copie BNF, livre publié en 2003
Lieux de conservation actuels sous forme papier	localisation physique
Lieux de conservation actuels sous forme électronique	URL

c) Histoire des documents :

L'histoire d'un document est souvent utile à sa compréhension : contexte historique, géographique, intellectuel, politique, juridique...

- Période(s) : pour mieux comprendre l'évolution de la source
- Thématique(s) : pour mieux la replacer dans un contexte
- Genre(s)
- Inédit(s) ou réédition ?
- A partir d'un manuscrit ou d'un imprimé ?
- Une ou plusieurs versions ?
- Si plusieurs versions / traductions : édition d'une seule version ou plusieurs ?

d) Droits : problèmes juridiques

L'exploitation, la manipulation, la reproduction, la modification ou la diffusion au grand public d'un document demandent le respect du droit (propriété matérielle, intellectuelle, ...). Pour vous guider dans cette étape nous vous renvoyons aux Informations Juridiques proposées par le Ministère de la Culture français : http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_05.htm

Les questions essentielles à se poser pour aborder cette étape

- Quelle est la provenance des documents ? (fonds d'archives, bibliothèques...)
- Au droit de quel pays dois-je me référer ?
- Le document est-il tombé dans le domaine public ?
- Qui est l'ayant droit du document (l'ayant droit n'est pas forcément l'auteur)
- Quel contrat dois-je rédiger pour régler la question avec les ayants droit ? les éditeurs ? Les bibliothèques ?
-

2.2. Extraire les données

Les sources primaires ou secondaires "papier" (textes manuscrits ou imprimés, photographie⁴), ne sont pas utilisables telles quelles, ni pour une analyse pointue employant des traitements automatiques, ni pour

-
- ⁴ Pour aller plus loin : [NumérisationClavaud] document de Florence Clavaud sur les bonnes pratiques : <http://www.cn-telma.fr/bonnes-pratiques.pdf> [MinCulture] document + inventaire type du ministère de la culture : <http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/inventa...> [EchoFabrique | MutEC] cf. GrilleMutECDocNum.pdf pour modèle de workflow de numérisation

une mise en ligne. Une importante étape du projet consistera à extraire les données de ces sources papier pour les rendre numériques, par des procédés aussi simples que la transcription, ou, d'autres, plus complexes, tels la numérisation et le passage au logiciel de reconnaissance de caractères (OCR).

2.2.1 Numérisation et reconnaissance optique de caractères

Il y a de multiples manières de produire les images du corpus, en fonction du matériel utilisé, des formats choisis, etc. La capture est déjà un choix d'encodage de l'information.

Les questions suivantes peuvent aider à préciser comment envisager le scannage.

- A quoi doivent servir les images ? Illustration du produit final ? Recherche automatisée de formes par des logiciels spécialisés ? Si oui, ces logiciels imposent très vraisemblablement des contraintes quand au format et aux propriétés des images pour pouvoir les traiter, lesquelles ?
- Qui a les compétences nécessaires pour orienter les choix de numérisation en mode image (choix à opérer en fonction de l'utilisation qui sera faite des images dans les résultats intermédiaires et finaux) : choix de la résolution, choix du mode chromatique, choix du format d'enregistrement de la copie source, choix du plan de nommage et de classement des fichiers, choix du lieu de stockage des fichiers, choix de l'outil de numérisation (appareil photo numérique, scanner, scanner à livre ouvert), choix des métadonnées de description des fichiers, caractéristiques des documents sources (accessibilité, fragilité, format...) ? Pour valider la qualité de la numérisation, par exemple pour lire un histogramme des couleurs, vérifier qu'il n'y a pas de saturation et donc de perte d'information, etc. ?

A retenir

- Une information non enregistrée à la capture ne pourra plus être créée a posteriori : on peut créer des images en basse résolution à partir d'images en haute résolution, jamais l'inverse. Pour des documents de taille A4 ou approchant, viser une résolution de 600 dpi minimum si possible, et en tout cas ne jamais numériser à moins de 300 dpi ni dans un format compressé.

Bonnes pratiques

- Constituer un jeu de fichiers d'images brutes dans un format non compressé (TIFF par exemple, mais attention, ce format est propriétaire et pose des questions de pérennité), renseigner les métadonnées de ces fichiers (EXIF par exemple) puis travailler sur des jeux secondaires d'images, dont les formats pourront être adaptés aux opérations correspondantes (PNG ou JPEG pour le web, etc.).
- Les besoins en espace disque et en puissance des processeurs seront très vite importants avec les images : évaluer le "poids" global des fichiers images numériques et prévoir au minimum le double (jeu d'images brutes + copie de travail), s'assurer qu'on a les outils adéquats pour manipuler les images, les visualiser, inscrire informatiquement les interventions qu'on effectue sur elles (analyse, classement, métadonnées).

Exploitation et manipulation des images obtenues

- Les images serviront-elles en tant que documents iconographiques? (Illustrations imagées? Objets de commentaire?)
- Les images doivent-elles être converties au format "texte"? Dans ce format le texte sera lisible, caractère par caractère, dans un éditeur de texte et non plus dans un logiciel de retouche d'image

A retenir

- dans le cas d'une conversion au format TEXTE, il faudra choisir un logiciel de reconnaissance de caractère, appelé communément OCR (du terme anglais Optical Recognition Character pour Reconnaissance Optique de Caractères). Le choix du logiciel ne pourra être effectué qu'après localisation et datation des textes à passer en mode texte, le logiciel devant reconnaître les caractères de la langue utilisée et de l'époque où les textes furent produits. En effet, les logiciels de reconnaissance de caractères fonctionnent très bien sur les imprimés de la fin du 20^{ème} siècle, mais moins bien pour la fin du 19^{ème} et le début 20^{ème}

Par exemple : Certains projets européens ont tenté de mutualiser les efforts de plusieurs universités dans le but de développer un logiciel spécifique pour tel ou tel caractère de telle ou telle époque. On peut citer le [projet METAe](#) qui a porté ses efforts sur la reconnaissance des lettres gothiques "Fraktur", beaucoup utilisées en Europe du 18^{ème} jusqu'au milieu du 20^{ème}. Le [projet Debora](#) a quant à lui concentré son travail sur les caractères d'imprimerie du 16^{ème}. Actuellement le [projet Corpus Numériques](#), financé par le cluster 13 de la région Rhône-Alpes, s'intéresse aux caractères du 18^{ème}.

2.2.2 Transcrire

Transcrire ce n'est pas singer le document physique mais de discrétiser et de qualifier l'information. Ce que le lecteur humain distingue parce qu'il s'appuie sur des habitudes culturelles, contextualise ce qu'il lit, la machine ne peut le distinguer. Transcrire, c'est expliciter, prévoir une trace numérique distincte pour chaque type d'information.

A retenir

- Ne pas réinventer sa propre grammaire de transcription. C'est un énorme coût en termes d'essai/erreur pour parvenir à une bonne expressivité, de documentation à maintenir, et cela met en danger la pérennité du corpus. Des modèles de données libres, ouverts et utilisés dans de larges communautés scientifiques existent déjà, qui permettent de bénéficier d'une documentation collectivement maintenue, d'une expertise et d'une mémoire des questions de transcription largement débattues, d'outils mutualisés, etc. Voir par exemple les activités du consortium [TEI](http://www.tei-c.org/index.xml) : <http://www.tei-c.org/index.xml>

Par exemple : l'italique dans une édition papier peut recouvrir deux types d'informations de statut totalement différent : un titre d'ouvrage dans la préface ET une marque d'insistance dans la réplique d'un personnage.

En complément, quelques questions à se poser pour vérifier la qualité scientifique et technique de la transcription : Les règles de transcription ont-elles été explicitées ? (comment encoder les accents, la ponctuation, les variations d'orthographe, la mise en page...) ?

- Par quel dispositif va-t-on contrôler qu'elles sont respectées ?
- Comment ces règles sont-elles inscrites informatiquement ?
- Qui effectue les transcriptions ? Interne/Prestataire externe
- Sous quel format sont-elles été effectuées ? (*.doc, *.txt ...)
- Quel codage de caractère est utilisé ? (utf8, iso ***-***- ...)

2.3 Annoter

L'annotation recouvre deux types d'actes scientifiques de nature différente mais dont la frontière n'est pas toujours évidente : elle peut consister dans la poursuite, l'affinage, de la discrétisation des informations commencée aux étapes précédentes, ou dans l'analyse scientifique, le commentaire sur, la justification d'un choix, etc.

Bonne pratique

- Expliciter où se situe la frontière dans le cas du projet, et veiller à la distinction entre les deux types d'annotation.

Conserver le corpus numérique ?

Les objets numériques ont un cycle de vie propre, leur perpétuation demande un outillage et une organisation humaine spécifiques. Cette question est particulièrement cruciale pour les objets numériques qui ont un caractère scientifique, car se pose le problème de leur citabilité à long terme.

Dans presque tous les cas de projets de corpus numérique, les équipes n'ont pas les moyens en interne de garantir cette pérennité et c'est en soi assez normal : de la même manière chaque équipe de recherche ne possède pas en interne une imprimerie et une bibliothèque pour fabriquer et conserver ses ouvrages papiers... Or pour les objets numériques, l'équivalent professionnel et institutionnel de la "chaîne du livre" n'existe pas (encore).

Voici quelques questions à se poser pour anticiper cet aspect du projet :

- L'institution tutelle est-elle au fait de l'archivage et de la conservation numériques ? Est-elle en mesure de l'assurer avec un minimum de garanties (respect des procédures OAIS, etc.) ?
- Quels sont les acteurs existants qui offrent des garanties suffisantes ?

A retenir

- L'archivage numérique pérenne doit répondre à des normes techniques et organisationnelles précises, comme celles que préconise par exemple le modèle OAIS (http://fr.wikipedia.org/wiki/Open_Archival_Information_System). Cela implique des infrastructures relativement lourdes.

Bonne pratique

- Se rapprocher des tutelles et voir si le projet peut bénéficier de leur politique de préservation de leur patrimoine numérique.