Partenariat Wikimédia France
Bibliothèque nationale de France

Rapport de projet

Version 2.0
« First Breath After Coma »

Sébastien Beyou
Jean-Frédéric Berthelot
Vincent Juhel
Nicolas Vigneron

Octobre 2010
Avril 2012

Table des matières

Introduction

La Bibliothèque nationale de France est la bibliothèque nationale de la République française, héritière des collections royales constituées depuis la fin du Moyen Âge. Première institution chargée de la collecte du dépôt légal, à partir de 1537, elle est la plus importante bibliothèque de France et l'une des plus importantes au monde.

Wikimédia France est une association loi 1901 qui soutient et promeut en France le libre partage de la connaissance, notamment au travers des projets Wikimedia. Elle est reconnue comme *chapter* local par la *Wikimedia Foundation*.

Parmi les projets Wikimedia figurent Wikipédia, encyclopédie, Wikimedia Commons, ressource de fichiers multimédia libres, et Wikisource, une bibliothèque de textes du domaine public en version texte plein, accompagné des numérisation des ouvrages.

Wikimédia France a dialogué durant plusieurs années avec la BnF afin de réflechir à des projets communs. En avril 2010, un partenariat entre ces deux organismes a été rendu public. Il s'agit principalement d'une expérimentation de la correction collaborative par la BnF, qui met à disposition les numérisations et les textes issus de la reconnaissance optique de caractères aux contributeurs de Wikisource. Le rôle de Wikimédia France est de déployer ses meilleurs moyens pour permettre cette tâche et de rendre compte périodiquement des avancées.

L'acheminement des données fournies par la BnF jusqu'à l'étape de correction sur Wikisource a nécessité un important travail technique, qui a essentiellement été pris en charge par une équipe de trois bénévoles de Wikimédia France, pilotée par des membres du conseil d'administration de l'association.

Avant-propos

Ce document a pour but de retracer l'ensemble du cheminement relatif au transfert effectif des livres après la signature du partenariat avec la BnF fin octobre 2009. Il présente la gestion du projet dans son ensemble ainsi que les techniques associées et les raisons de leur choix. Il ne présente que la partie technique, en faisant quasiment abstraction de toute la partie négociations menée dès mars 2008 par Valérie Chansigaud/Valérie75, Alexandre Moatti/Arrakis et Rémi Mathis, sauf lorsque des décisions techniques demandant l'avis de la BnF ont été nécessaires.

Les principaux acteurs dans ce transfert sont (par ordre chronologique et avec le titre officieux de chacun)

- Julien Fayolle/(:Julien:), initiateur et superviseur du projet ;
- Nicolas Vigneron/VIGNERON, consultant en Wikisource et SIB et superviseur local du projet (NCO) ;
- Sébastien Beyou/Seb35, chef de projet et développeur (NCO) ;
- Pascal Martin, développeur ;
- Vincent Juhel/Plyd, développeur principal et vice-chef de projet au mois de mai (NCO) ;
- Jean-Frédéric Berthelot/Jean-Frédéric, consultant en Commons, développeur et vice-chef de projet au moi de mai (NCO).

Ce document a été principalement rédigé par Sébastien Beyou et Jean-Frédéric Berthelot, et retouché et amendé par les autres personnes impliquées dans ce transfert.

En vue d'ensemble, ce projet de transfert devait aller chercher les livres sur un serveur FTP de la BnF, comprenant les images scannées ainsi que les métadonnées et OCR fournis par la BnF, et amener l'ensemble de ces livres sous forme d'un livre DjVu sur Wikimedia Commons pour l'hébergement, la destination principale des livres étant Wikisource en français.

Chapitre 1

Historique du projet

1.1 Débuts et expérimentations

1.1.1 Début du projet

Après la signature du partenariat avec la BnF en octobre 2009, Julien a commencé à regarder les aspects techniques de ce transfert. Fin janvier, le CA se réunit et commence à réfléchir à la mise en œuvre. Le mardi 9 février, Nicolas parle à Sébastien du partenariat avec la BnF, que Julien commence à regarder mais qu'une aide technique serait appréciable. Sébastien commence à se renseigner sur le FTP en masse, la création de DjVu (dont la couche texte qui contiendra l'OCR) et l'import sur Commons. Il est évident qu'il faut un serveur intermédiaire entre la BnF et Commons où l'on pourra créer les livres DjVu.

1.1.2 Création des DjVu

En février, Sébastien étudie le monde DjVu et les programmes à utiliser (librairie DjVuLibre) et écrit un programme (sous forme de Makefile) qui convertit les images png en images ppm, puis crée des DjVu page unique et enfin compile les DjVu page unique en DjVu multipage. Un embryon de programme est disponible fin février pour cette création de DjVu. Il manque encore toute la couche texte dans le DjVu.

1.1.3 Problème de location de serveur

Du 5 au 22 mars, Julien a loué un espace de stockage chez OVH, mais après plusieurs essais (dont lors de l'Assemblée Générale avec Ash Crow) et prises de renseignements auprès de l'assistance utilisateur d'OVH, il est confirmé qu'on n'a pas accès au réseau extérieur depuis le serveur (par mesure de sécurité OVH) : le serveur peut seulement émettre des données, ce qui ne nous convient pas du tout puisque les données doivent transiter en réception et en émission sur ce serveur. Entre temps, il semble que Julien et Christophe Henner essayaient de prendre contact avec Typhon pour obtenir un serveur pour ce projet.

1.1.4 Création des couches texte

En mars, Sébastien se renseigne sur les OCR fournis par la BnF qui sont sous forme de XML ALTO. Ce format est très spécialisé pour le monde des bibliothèques et relativement discret sur Internet (la BnF explique un peu sur son site, le format lui-même est créé et géré par

la Library of Congress, la BnF en utilise une version personnalisée). Il s'agit d'une description de la structure d'une page (colonnes, paragraphes...) avec les coordonnées de chaque zone, jusqu'aux coordonnées de chaque mot. Pour Wikisource, on n'a besoin que du texte sans les coordonnées. Il s'agit donc de parser le XML pour en sortir un texte brut : pour cela, il existe les différentes technologies associées au XML (XSLT, XPath, XQuery, etc). Vu qu'il ne connait pas toutes ces technologies et leur utilisation, Sébastien lance un courriel sur la liste de discussion Wikimédia France le 3 avril, auquel répondent Pascal et Vincent, ce dernier précisant qu'il n'a pas trop le temps de s'en occuper. Lors du Developers Meeting à Berlin du 16 au 18 avril, Sébastien et Pascal font le point de ce qu'il faut faire exactement. Vis-à-vis du problème de serveur, Pascal propose également de fournir un serveur pour le projet, d'autant plus que ses serveurs sont sur le réseau Renater, réseau à haut débit reliant les institutions françaises.

1.1.5 Premiers fichiers DjVu sans OCR et lancement du communiqué de presse

Fin mars, Sébastien crée, à la demande de Rémi, trois fichiers DjVu sans l'OCR. Plusieurs questions se posent : 1) la plupart des livres ont une marge blanche importante (probablement due à la taille de la fenêtre de numérisation) qu'il serait préférable de retirer, ce qui sera fait avec ImageMagick au mois d'avril (option `-trim`) ; 2) L'import sur Commons est limité à 200Mio, ce qui correspond à des livres d'environ 200-300 pages. Le 29 mars, Nicolas intègre Jean-Frédéric dans la boucle, à l'origine pour la création d'un modèle Commons. Prenant connaissance des problématiques d'import, il se renseigne rapidement sur IRC sur l'import côté serveur : quelques rapides réponses en confirment la faisabilité. Le 4 avril, Vincent et Sébastien remarquent qu'il manque des OCR. Le 7 avril, Rémi annonce que la BnF a lancé son communiqué de presse et qu'il faut "rapidement" un livre océrisé. Sébastien extrait à la main l'OCR d'un livre de 28 pages (3 heures de travail).

1.2 Phase opérationnelle

1.2.1 Location du serveur

Le 7 avril, Jean-Frédéric prend contact avec Multichill, gourou des uploads de masse sur Commons. Celui-ci détaille la procédure d'import en masse ne passant pas par l'interface web standard mais par l'insertion directement dans la base de données par un administrateur système (ayant accès aux serveurs de la fondation) : le problème de la taille sera outrepassé de ce fait. Le mardi 27 avril, au cours d'une rencontre de la NCO, Vincent procède à la location d'un serveur chez OVH, ayant cette fois toutes les caractéristiques requises (750 Gio, 4 cœurs, 100 Mbit/s). Sébastien lance le téléchargement des livres le 28 avril depuis le serveur de la BnF et ce téléchargement se finira le dimanche suivant 2 mai.

1.2.2 Relecture des titres de fichiers DjVu

À partir du 28 avril, Sébastien commence à demander un peu partout de l'aide pour vérifier les titres des fichiers DjVu sur Wikisource : NCO, Scriptorium Wikisource (lieu de discussion central du projet), discussions@, wikimediafr-l@...Peu de personnes travailleront effectivement sur cette liste (qui déterminera les titres des fichiers DjVu, et, de par la structure de Wikisource,

le nom des pages de correction de l'OCR, soit, cumulées, 500 000 pages). Sébastien a coupé la liste en 4 afin d'éviter les conflits d'édition. D'ici fin mai, ce seront 4-5 personnes qui reliront 80% des titres (Sébastien a fait une grande partie des pages 1 et 3, Pierre-Yves Mevel une grande partie de la page 2, Nicolas une part importante sur les 3 premières pages, Zaran, Yann et Zyephyrus ont fait une partie de la page 4).

1.2.3 Prise en charge du projet par la NCO

Le mardi 4 mai, au cours de la rencontre de la NCO, Vincent, Jean-Frédéric et Sébastien prennent rendez-vous le lendemain pour avancer sur le projet. Le mercredi 5 mai, comme prévu, la rencontre a lieu chez Sébastien de 19h30 à 00h00. Après avoir testé un peu la technologie XSLT depuis le début de la semaine, et après avoir rencontré des problèmes dû au manque de souplesse du XSLT, il est décidé d'utiliser un "vrai" langage de programmation : Perl, Java ou Python. Jean-Frédéric a écrit un petit script en Perl, Vincent connait bien Python ; c'est Python qui sera choisi, en particulier à cause de ses possibilités de manipuler facilement du XML, et, dans un avenir un peu plus lointain, parce que Pywikipedia, la célèbre librairie de bots MediaWiki, est écrite en Python. La page de requête d'import en masse sur Commons a été ouverte le 2 mai par Sébastien, et Jean-Frédéric avance alors ce soir sur la forme des pages Commons et Wikisource, il s'occupera ensuite de la plupart des discussions liées à l'import sur Commons. Il est posé la question des licences à utiliser (domaine public partout ?).

Vincent crée le programme de base pour transformer l'ALTO en couche texte DjVu, basiquement en extrayant tous les mots et en les mettant les uns à la suite de autres, et crée deux formats de texte proches : un format à inclure dans le DjVu, et un format wikitexte qui pourra être utilisé pour initialiser les pages de Wikisource (prise en charge des en-têtes (inclus dans des sections `noinclude`) et des fins de ligne (à supprimer dans le format wikitexte)). Il prend en charge également l'exception que constitue les césures de mots. On dispose de très peu d'informations sur la structure que doit avoir la couche texte DjVu, on n'a même pas trouvé les spécifications DjVu, et on se base sur des "reverse-engineering" de DjVu existants. Vincent ajoute également la fonction chargée de créer les pages d'index de Wikisource et de description de Commons, à partir des métadonnées. Sébastien inclut son précédent Makefile de création des (images) DjVu dans le programme Python, et se charge de lire le fichier de métadonnées inclu dans chaque livre (XML RefNum propre à la BnF). Un débat fait rage lors de cette soirée sur l'introduction de watermarking (filigrane) sur les documents, Rémi ayant demandé dans l'après-midi s'il était possible d'en inclure un. Vincent est contre et Jean-Frédéric et Sébastien considérent qu'il n'est pas du ressort des programmeurs de prendre une telle décision. Pascal n'ayant pas eu le temps de se pencher sur le problème de cette couche texte, le programme utilisera alors la transformation Python de Vincent.

1.2.4 Amélioration du programme et refonte des couches texte

Le mercredi 12 mai, Vincent, Jean-Frédéric et Sébastien reprennent rendez-vous pour continuer et finir (!), la soirée a duré de 19h00 à 3h00 chez Jean-Frédéric. Jean-Frédéric a trouvé les spécifications DjVu et demandé des renseignements (sur IRC et sur le Scriptorium en anglais) sur la couche texte à inclure dans les DjVu, peu de personnes savent finalement ce qu'une couche texte DjVu est censée contenir, puisqu'on s'était posé la question de savoir si on mettait du wikitexte (entêtes `noinclude`) ou strictement le texte brut. Il s'avère alors qu'il serait préférable d'utiliser toute l'information contenue dans les ALTO pour créer des DjVu

de très bonne qualité, sans wikitexte pour ne pas "polluer" en cas de reprise extérieure des DjVu. Jean-Frédéric et Sébastien se pencheront sur la compréhension des spécifications DjVu (très très mal écrites) et par reverse-engineering d'une couche texte existante (pour compenser la mauvaise écriture des spécifications DjVu), afin notamment de comprendre le système de coordonnées des mots et les unités de mesure utilisé par l'ALTO (ambiguïté entre le 1/10mm et le pixel) puisque les nombres écrits dans l'ALTO ne correspondent pas aux tailles des images (réduction ?). Ils regarderont aussi ce qui concerne le trimming (coupure automatique des marges blanches par ImageMagick (option `-trim`) sans que la quantité coupée ne soit connue) pour le transformer en cropping intelligent (idée de Jean-Frédéric), en cherchant dans l'ALTO les positions minimales et maximales des cadres puis en coupant (option `-crop` de ImageMagick), ce qui permet ensuite d'en déduire la position de chaque mot dans la nouvelle image coupée (ce qui n'était pas possible auparavant). Vincent avance sur la robustesse du processus par une journalisation et une distribution multi-processeur des tâches (pour utiliser les 4 cœurs à notre disposition), et étudie également Pywikipedia qui initialisera les pages d'index de Wikisource et les pages de wikitexte des livres. Il crée un programme Python chargé de cette tâche (alors non-testé).

1.2.5 Création des `pagelist` et liste des auteurs

Lors des RBLL, Sébastien et Nicolas travaillent sur la création des balises `pagelist`, afin d'extraire depuis les métadonnées RefNum de chaque livre les paginations (romaine, arabe, non-paginé). Ce fichier servira ensuite d'argument à l'initialisation des pages d'index sur Wikisource. Sébastien extrait une liste des auteurs que Nicolas complète pour Commons et Wikisource.

1.2.6 Derniers détails

Le mardi 18 mai, lors de la soirée « Mardi, c'est sushi » chez Nicolas, après avoir appris que les "vraies" images n'étaient pas dans les dossiers C et D mais dans des `.tif` multipage, Vincent a modifié le programme pour utiliser ces images de résolution bien meilleure (2500×3500 ou 3500×4500). Jean-Frédéric travaille sur le peaufinage du watermarking qu'il a étudié la veille. Sébastien crée le bot BnFBoT qui initialisera les pages d'index et, peut-être, les pages de wikitexte. Nicolas contacte Zephyrus sur IRC qui donne le statut de bot à BnFBoT. Sébastien a pris une journée de congés le mercredi afin de finir (!) et de lancer la production de livres. Finalement il passera une grosse partie de la journée à continuer de revoir les titres de DjVu sur Wikisource et à parlementer avec la communauté Wikisource pour expliquer ce que fera BnFBoT, chose que feront aussi Jean-Frédéric et Nicolas à partir de mercredi.

1.2.7 Tests réels sur Wikisource

Les samedi 22 mai et dimanche 23 mai, Sébastien lance quelques livres et continue de traiter les exceptions, notamment l'échappement propre des caractères, les problèmes d'encodage UTF-8/ASCII, ainsi que diverses exceptions qui apparaissent sur les tests d'autres livres. Le dimanche, quatre livres sont créés par Sébastien selon le process standard, versés sur Commons, et dont les pages d'index sont initialisées par BnFBoT (la communauté Wikisource n'ayant pas formellement accepté l'initialisation des pages wikitexte). De plus, Julien annonce que Multichill a généré des pages "Creator" sur Commons à insérer dans la description des livres

sur Commons. Le samedi, Jean-Frédéric finalise le filigrane, et s'entretient longuement avec Multichill sur IRC pour peaufiner l'import Commons.

1.2.8 Prolongations et tirs au but

Le 24 mai, Sébastien annonce sur le Scriptorium que le programme final tourne, et donne quatre fichiers de tests à qualifier par les wikisourciers [1]. Les critiques émises ne sont pas celles attendues : des souciers commentent sur des erreurs dans les titrages, titrages extraits de la liste censément vérifiée au cours de ce mois. Selon eux, un gros travail sur cette liste est nécessaire.

Le 27 mai, Adrienne met sur le Wiki Membres le texte de la convention entre Wikimédia France et la BnF.

Le 6 juin, lancement de la création des livres pour déceler des erreurs de production. De nombreuses erreurs (place sur le disque, etc.) font planter le processus, en grande partie des erreurs de mémoire insuffisante lors des conversions TIFF → ppm, ainsi que quelques erreurs mineures sur les process.

Le 17 juin, Sébastien et Jean-Frédéric se retrouvent chez ce dernier pour une petite GLAM-party, pour corriger deux derniers bugs. Le 18 juin, au matin, la production est lancée... pour une semaine de traitement. Le 19 juin, désillusion : 30% des bouquins partent en erreur, toujours à cause d'erreurs de mémoire. Le WE du 26, Jean-Frédéric se renseigne sur IRC pour trouver un interlocuteur pour le versement des bouquins sur Commons, et est orienté vers Tim Starling. Le 27 juin, devant la taille gigantesque des bouquins et après s'être replongé dans les tréfonds de DjVu, Sébastien suggère de passer en bitonal plutôt qu'en multicolore, ça fait gagner un facteur 10 en taille des DjVu (les plus gros livres de 1 Gio passent à une taille de l'ordre de 100 Mio) et beaucoup de temps de création des DjVu (le temps total pour la génération passe de 1 semaine à 1 jour) grâce à la transformation directe des TIFF en DjVu, sans passer par les gros fichiers ppm qui faisaient utiliser le disque dur du serveur). Le 28 au soir, GLAM-party chez Vincent. Le 1er juillet, Jean-Frédéric contacte Tim Starling pour lui parler de l'opération. Le 3 juillet, les bouquins sont générés et prêts. Le 7 juillet, Tim s'enquiert de l'apostrophe typographique dans les noms de fichiers.

Le 10 juillet, alors que Sébastien et Jean-Frédéric sont à Gdańsk pour Wikimania, ils apprennent par Multichill que les bouquins sont arrivés sur Commons.

Sébastien consacre une bonne partie des deux jours qui suivent à préparer l'initialisation des pages sur Wikisource.

1.3 Travail complémentaire

1.3.1 Interlude

La relecture des livres reste alors à faire par les Wikisourciers. L'équipe, qui a encore d'autres idées en réserve, hésite à arrêter là ou bien poursuivre le travail. Le problème de la création des statistiques sera de toute façon à traiter. Lors de la réunion hebdomadaire de la NCO du 20 juillet, Sébastien et Jean-Frédéric y réfléchissent avec l'aide de Nicolas. Sébastien a également l'idée d'utiliser les propriétés des OCR de la BnF afin d'extraire en masse les illustrations des bouquins, idée qui séduit Jean-Frédéric.

1. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Mai_2010#Fichiers_de_test_BnF

Extrait du compte-rendu de la NCO du 20 juillet :

Wikimédia France s'est engagée à fournir des rapports trimestriels incluant des statistiques sur l'avancement du projet. Il nous faut donc réfléchir à des métriques pertinentes à présenter à la BnF. En mode brainstorming, Sébastien et Jean-Frédéric lancent de nombreuses idées :

- Nombre de page par niveau d'achèvement (*Quality Level*)
- Nombre de livres par niveau d'achèvement
- Estimation de la quantité de travail fournie par les wikisourciers :
  - temps passé : si cela n'est pas possible en soi (la fin d'édition est horodatée, mais pas le début ; sans compter le travail off-line, etc.), une approche détournée partielle est possible. En émettant l'hypothèse qu'un contributeur wikisource corrigera souvent plus d'une page à la suite (session), en analysant le profil de contributions des contributeurs BnF, on peut établir estimer le temps passé sur la correction d'une page donnée comme le temps écoulé depuis la modification précédent cette correction. A mettre en balance avec les deux points suivants.
- difficulté de la correction : estimation en calculant la distance entre le texte d'origine (couche OCR) et le texte final.
- Analyse des différents bouquins (nombre de pages, indice de confiance OCR, date de publication, etc.) et détermination de classes.
- Analyse des profils de contributeur BnF : anonymes, enregistrés, noobs, power-users — nombre de contribs, type de contributions (taille de la modif, utilisation de modèles [déterminer des classes de modèles 'accessibles' ou 'pour power-users']). Déterminer le nombre de nouveaux contributeurs au corpus BnF, en tracer l'évolution, calculer le taux de rétention.
- Projection de toutes les métriques sur les deux axes précédents (quels bouquins attirent quels types de contributeurs, etc.)
- Analyse de l'évolution du corpus BnF : scans supplantés, supprimés, ajout de couche texte.

Bien entendu, reste à voir la faisabilité et l'implémentation effective de ces idées. Jean-Frédéric et Sébastien vont envoyer un e-mail à la t0p s3cr3t l1st spécial GLAM. Les professionnels qui y rôdent auront sûrement des idées pertinentes, et les développeurs présents pourront les mettre en application.

Par ailleurs, il est grand temps de boucler le rapport technique commencé par Sébastien il y a bientôt deux mois. Ce document servira de base pour d'autres documents : retour d'expérience pour le wiki Outreach, best practice pour Meta, diffusion sur les ML spécialisées... De plus, il est aussi prévu de fournir à la BnF un tel rapport.

L'équipe technique et le consultant WS se fixent rendez-vous lundi soir prochain pour une soirée rédaction. D'ici là, les apports des autres acteurs du projet seraient les bienvenus...

Le 21 juillet, Jean-Frédéric teste le versement d'un ouvrage au format TIFF. Le même jour, Adrienne s'enquiert de la situation du serveur et de la pertinence de le garder un mois de plus. Sébastien fait le point sur le travail restant envisagé, et souligne que tous ces points tiennent du bonus, dès lors que les scans sont sur Commons et les pages sur Wikisource.

Le 26 juillet, Jean-Frédéric, Sébastien, Nicolas et Vincent (alias les BnFboyZ) se réunissent

pour revoir le rapport, décider du renouvellement ou non du serveur et rendre public le programme.

Extrait du compte-rendu de la NCO du 27 juillet :

Rapport

Le rapport a été principalement rédigé par Sébastien, puis complété par Jean-Frédéric et Vincent, au mois de mai. Le midi, Jean-Frédéric a ajouté les diverses péripéties arrivées depuis jusqu'à l'heureuse conclusion. Nicolas a relu le brouillon.

Il est prévu de faire deux rapports : l'un pour l'association et la BnF, et un autre public, placé sur Meta. Au passage, Vincent s'inquiète de la dispersion des ressources entre divers wikis (Commons, Wikisource, etc.). Dans l'idée de pallier ce problème, une page sur Meta est créée. L'idée est qu'elle serve de hub pour toutes ces pages dispersées.

Code

Vincent s'emploie à créer un projet propre pour le code développé, et réécrit notamment les commentaires en anglais. L'idée est de mettre le code à disposition de tous (GitHub, Google Code ou le FishEye du toolserver). Il est également placé sous licence GPL. Étant toutefois très spécifique à la livraison BnF, il est peu probable qu'il soit utilisé comme tel ; en revanche il peut sans nul doute servir de modèle pour un projet similaire.

Renouvellement

Une nouvelle fois se pose la question du renouvellement du serveur dont la location arrive à terme. Après mûre réflexion, l'équipe technique décide de prolonger d'un mois. En effet, d'autres choses sont encore possibles : exploiter les tables de matières en HTML qui accompagnent certains livres, récupérer en masse toutes les illustrations des livres grâce aux positions indiquées dans les OCR de la BnF, et conserver les fichier TIFF haute qualité (19 Gio) pour un éventuel versement sur Commons.

Les jours qui suivent, les BnFboyZ harcèlent Rémi Mathis pour savoir s'il est bien permis de verser les fichiers TIFF, ce à quoi Rémi répond par l'affirmative le 28 juillet.

1.3.2 Match retour

Le 5 août, sur les routes de Suisse, Sébastien, Jean-Frédéric et Nicolas réfléchissent au projet BnF. Sébastien et Jean-Frédéric écrivent notamment les spécificiations de l'*ImageExtractor*, programme d'extraction massive de toutes les illustrations des livres grâce aux informations contenues dans les OCR. Jean-Frédéric développe et teste le programme dans les semaines qui suivent, et le finalise le 26 août. La production est lancée le 27 août, et fournit 22 799 images (environ 245 Mio compressé).

le 19 août, Sébastien demande à Tim Starling le versement en masse des TIFF, sous les mêmes modalités que le précédent.

Le 25 août, Rémi fait le point sur le travail restant à effectuer. Grandement fatiguée de ces mois de développement, l'équipe technique réagit avec peu d'enthousiasme.

En l'absence de réponse de la part de Tim Starling, Sébastien procède le 27 août au

versement des fichiers TIFF originaux sur Wikimedia Commons, via robot Pywikipedia.

Le 28 août, Sébastien déclare le projet de développement terminé. Les objectifs de livrables à la communauté fixés par l'équipe technique ont alors tous été atteints. Vincent confirme que le serveur est rendu à la date du 30 août, soit après quatre mois de location.

1.3.3 And in the end...

Rémi relance l'équipe technique le 22 septembre sur deux sujets. D'une part, il invite à une réflexion sur « la méthode à utiliser afin de lier les articles de Wikipédia sur les personnes, les oeuvres, les noms géographiques, etc. aux fiches autorité de la BnF ». D'autre part, il informe que la BnF demande des statistiques sur le travail déjà effectué, et s'enquiert de la disponibilité et de la faisabilité de telles données.

Le 4 octobre, Sébastien annonce qu'il a réalisé une large part des statistiques envisagées par l'équipe technique deux mois plus tôt. Le 12, il informe la liste de diffusion de l'association de ses premières avancées. Le 17, il annonce la publication d'une version admissible (*Release candidate*), demandant les retours des membres de l'association et des contributeurs de Wikisource. Le 24 octobre, Jean-Frédéric et Sébastien se réunissent pour avancer sur les deux rapports. Sébastien publie la version définitive du rapport d'avancement le soir même.

1.4 Résumé et principaux jalons

Suite à des discussions de plusieurs années, le partenariat avec la BnF s'est concrétisé en octobre 2009. L'équipe technique a d'abord été constituée de un puis trois membres. Après trois mois de recherche et développement, la phase de développement a duré quatre mois (dont deux mois et demi pour la livraison principale). Après une pause (d'environ un mois), un mois a été consacré à la seconde livraison (rapport d'avancement) et à la complétion de la documentation.

Octobre 2009
: Signature du partenariat

Février 2010
: Arrivée de Sébastien

Fin mars/début avril
: Arrivée de Vincent et Jean-Frédéric

7 avril
: Annonce du partenariat [2]

27 avril
: Location du serveur

Mai-juin
: Essentiel du travail par l'équipe technique

1er juillet
: Demande de versement

8 juillet
: Versement des DjVu sur Wikimedia Commons

10 – 12 juillet
: Initialisation des pages sur Wikisource

27 août
: Versement des TIFF sur Wikimedia Commons et extraction massive des images

28 août
: Fin du projet de développement

30 août
: Fin de location du serveur

Fin septembre – octobre
: Réalisation du rapport d'avancement et complétion du rapport de projet

24 octobre
: Publication du premier rapport d'avancement

31 octobre
: Publication d'une pré-version du rapport de projet

2. `http://www.wikimedia.fr/wikim%C3%A9dia-france-signe-un-partenariat-avec-la-bnf`

Chapitre 2

Aspects techniques

2.1 Généralités

OCR (*Optical Character Recognition* — Reconnaissance optique de caractères)
Procédé informatique de transformation d'une image de texte imprimé en fichier texte.

DjVu
Format de fichier ouvert pour l'archivage de documents numériques, largement utilisé sur Wikisource.

TIFF (*Tagged Image File Format*)
Format de fichier pour l'archivage d'images numériques.

ALTO (*Analyzed Layout and Text Object*)
Format basé sur XML, décrivant le texte issu d'une opération de reconnaissance optique de caractère.

Pywikipedia
Bibliothèque de programmes écrits en langage Python permettant d'agir sur les wikis et tout particulièrement les sites Wikimedia.

ImageMagick
Suite logicielle de manipulation d'images matricielles, essentiellement en ligne de commande.

DjVuLibre
Implémentation libre de DjVu, incluant notamment des utilitaires de manipulation.

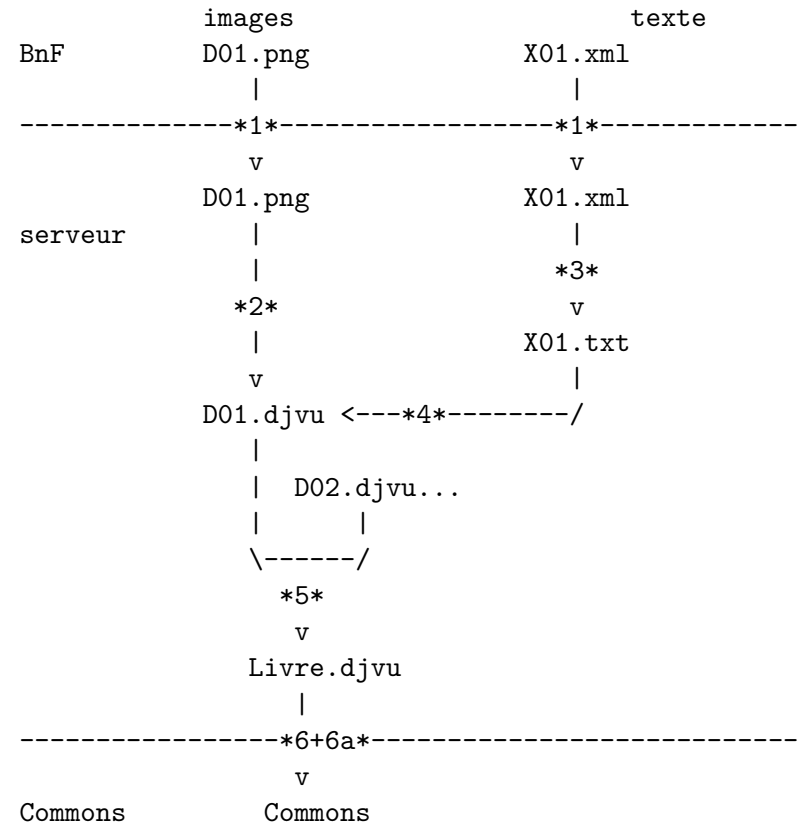
Versement côté serveur (*Server-side upload*)
Téléversement sur un wiki effectué par un administrateur système, sans passer par l'interface web, et outrepassant de fait toute limite imposée sur cette interface.

2.2 Chaîne de traitement

Une page dédiée sur le wiki des membres a servi de page de travail au cours du projet [1]. Beaucoup de points ont été découverts, éclaircis et résolus (ou contournés) qu cours du projet. Cela est dû à de nombreux facteurs énoncés dans la partie "Difficultés" ci-dessus, principalement les technologies mises en œuvre, leur utilisation et la volumétrie importante de données.

2.2.1 Vue d'ensemble

Voici le schéma (simpliste) fait au début (16 avril dans l'historique de la page projet, mais formalisé bien avant par Sébastien, sauf que non partagé puisque non écrit :-)



1. téléchargement de la BnF vers un serveur à nous
2. création de djvu pages uniques avec uniquement une couche image
3. extraction du texte depuis le XML ALTO
4. intégration de la couche texte dans les djvu pages uniques
5. assemblage des djvu pages uniques en un djvu livre (toutes les pages)
6. upload sur Commons
   (a) insertion éventuelle de métadonnées relatives au fichier
7. création des pages de description sur Wikisource
8. initialisation des pages sur Wikisource (en-têtes mis en `noinclude` par exemple)

1. `http://membres.wikimedia.fr/index.php/Groupes_de_travail/Musées_et_Bibliothèques/Coopération_BnF/Technique`

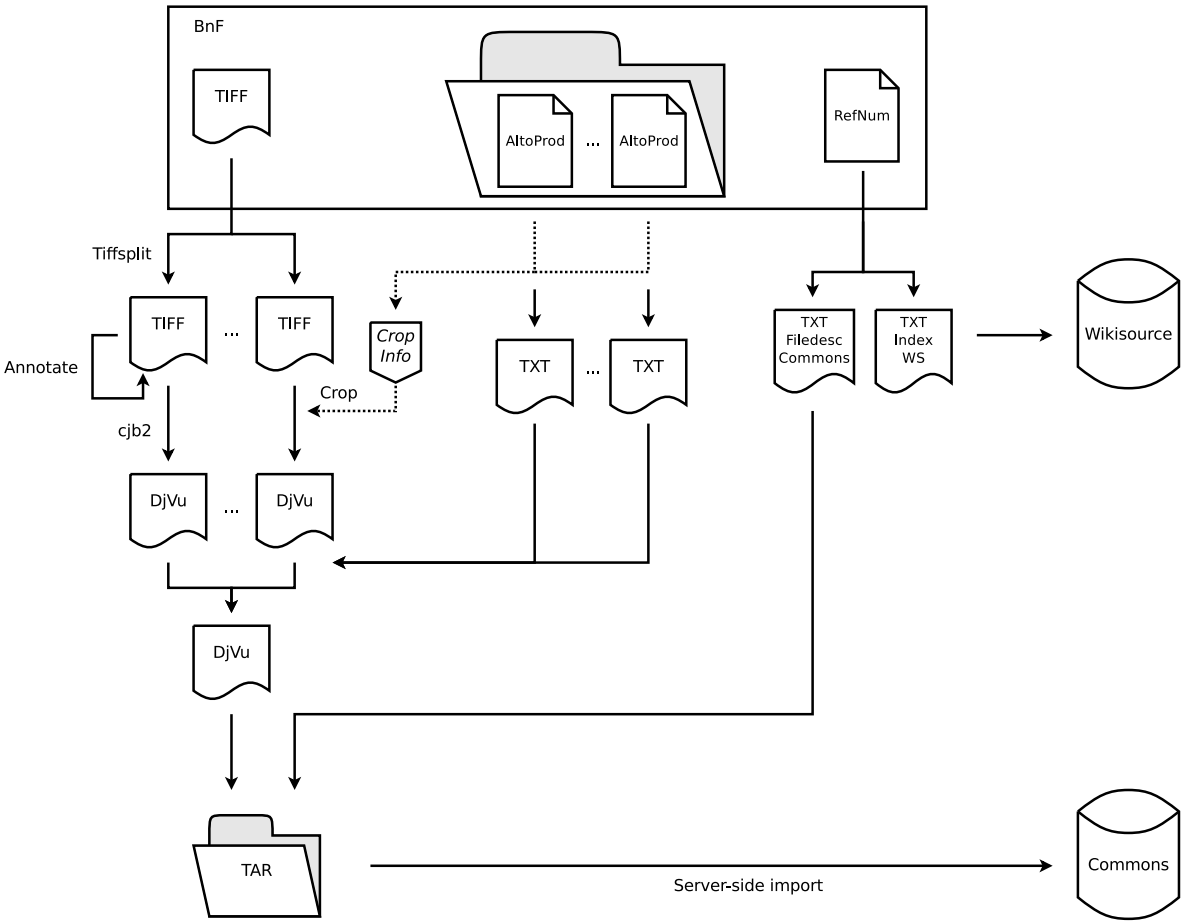


FIGURE 2.1 – Vue d'ensemble de la chaîne de traitement.

2.2.2 Entrées

La BnF a fourni les données relatives à 1416 livres. À chaque livre était associé un dossier, contenant les données suivantes :

- Le livre numérisé en images, au format TIFF multipage ;
- Un fichier XML "RefNum"[2][3], contenant les métadonnées globales sur le livre, ainsi que les métadonnées relatives à chaque page du livre (associées au fichier TIFF) ;

Dans certains cas, les dossiers de livre contenaient également :

- Un sous-dossier "X", contenant n fichiers XML "ALTO"[4][5], contenant l'OCR pour chaque page de l'ouvrage ;
- La table des matières de l'ouvrage, au format HTML ;
- Un sous-dossier "C", contenant n images de résolution moyenne au format PNG.

359 livres ne contenaient pas de sous-dossier "X", ie ne disposaient pas d'OCR.

2.2.3 Sorties

La chaîne de traitement produit, pour chaque livre :

- Une archive TAR contenant :
  - Un fichier DjVu, contenant les images et l'OCR (voir annexe A.1) ;

– Un fichier texte (syntaxe wiki) contenant les informations de description du fichier pour Wikimedia Commons (voir annexe A.2) ;
– Un fichier texte (syntaxe wiki) contenant les informations de description du fichier pour Wikisource (voir annexe A.3) ;

L'archive au format TAR constituait le livrable pour l'administrateur système de la Wikimedia Foundation, qui se chargeait du versement sur Wikimedia Commons.

Le fichier texte de description Wikisource est utilisé en marge de la chaîne de traitement, par robot PyWikipedia.

2.2.4 Processus principal

Algorithme 1: Programme principal

Entrées : Liste des livres à traiter
Lecture des métadonnées depuis fichier
Création des répertoires
si *un seul livre est à traiter* **alors**
    Traiter le livre
sinon
    Lecture de la liste de livres à traiter
    pour chaque *livre* **faire** en parallèle sur quatre processus
        Traiter le livre

Algorithme 2: Traitement unitaire

Entrées : Référence du livre (à la fois identifiant et nom de répertoire)
si *l'identifiant n'est pas dans le fichier de métadonnées* **alors quitter**
Récupérer les métadonnées du livre donné
try
    `bookProcess()`
catch
    `cleanUp()`
    `bookProcess()`

Algorithme 3: bookProcess()

```
Entrées : Dossier, Dossier initial, Dossier d'écriture, Métadonnées
Récupération des métadonnées pertinentes
Découpage du TIFF multi-page en n pages dans un répertoire temporaire
si l'OCR existe alors
    pour chaque page faire
        Détermination de l'étendue de la page
    Détermination du masque de découpe optimal
    pour chaque page faire
        Génération de la couche texte
si l'OCR existe alors
    pour chaque page faire
        si c'est la première page alors Annoter la page
        Découpage suivant le masque optimal
sinon
    pour chaque page faire
        si c'est la première page alors Annoter la page
        Trimming
    Détermination de la taille maximale
    pour chaque page faire
        Redimensionnement
pour chaque page faire
    Conversion au format DjVu unipage (encodage cjb2)
Agrégation des différentes pages en un DjVu multipage
si l'OCR existe alors
    Ajouter la couche texte
Renommer le fichier DjVu
Générer les sorties Commons et Wikisource
Préparer l'archive TAR
```

2.2.5 Détail

Découpage des fichiers TIFF

Le découpage du fichier TIFF multipage est effectué avec une commande qui génère dans un dossier donné autant de fichiers TIFF que de pages. Celles-ci sont immédiatement renommées.

Commande 2.1 – `tiffsplit`

```
tiffsplit image.tif
```

Annotation en première page

Une mention est ajoutée en première page de chaque ouvrage, « Source : Gallica - Bibliothèque nationale de France ». Cet élément est ajouté tant par demande de la BnF que par bonne pratique bibliothécaire d'indication de la source.

Commande 2.2 – `-annotate` de ImageMagick

```
convert <image_TIFF> -pointsize 20 -gravity South -annotate
+0+30 "Source : Gallica - Bibliothèque nationale de France"
<image_TIFF>
```

Conversion en fichier DjVu

La conversion en fichier DjVu est fait grâce à la commande `cjb2` de la suite DjVuLibre.

Commande 2.3 – `cjb2` de DjVuLibre

```
cjb2 -dpi <resolution> <image_TIFF> <Fichier_DjVu>
```

Ajout de la couche texte

L'ajout de la couche texte au DjVu multipage est fait grâce à la commande `djvused` de la suite DjVuLibre, à laquelle on passe un script indiquant les transformations à effectuer.

Commande 2.4 – `djvused` de DjVuLibre

```
djvused -s -f <script> <Fichier_DjVu>
```

Changement de coordonnées

En ALTO comme en DjVu, une page est découpée en différentes zones rectangulaires, dont la position est indiquée.

Dans la référence ALTO, le point d'origine est le point le plus en haut et à gauche de la page[1]. La position des différents éléments, assimilables à des rectangles, est définie par :

- `HPOS` l'abscisse du coin supérieur gauche ;
- `VPOS` l'ordonnée du coin supérieur gauche ;
- `WIDTH` la largeur ;
- `HEIGHT` la hauteur.

À l'inverse, dans la référence DjVu, le point d'origine est le point le plus en bas et à gauche de la page. La position des différents éléments est définie par :

- `Xmin` l'abscisse des coins de gauche ;
- `Xmax` l'abscisse des coins de droite ;
- `Ymin` l'ordonnée des coins inférieurs ;
- `Ymax` l'ordonnée des coins supérieurs ;

L'objectif de cette étape est de convertir les coordonnées de chacun des mots depuis la référence ALTO vers la référence DjVu.

Les mesures suivantes sont disponibles :

- `HPOS`, `VPOS`, `WIDTH` et `HEIGHT` sont les coordonnées du mot indiquées dans l'ALTO, dans le référentiel ALTO ;
- `XminPage`, `YminPage` et `tailleY` sont les coordonnées du masque de découpage, dans le référentiel ALTO.

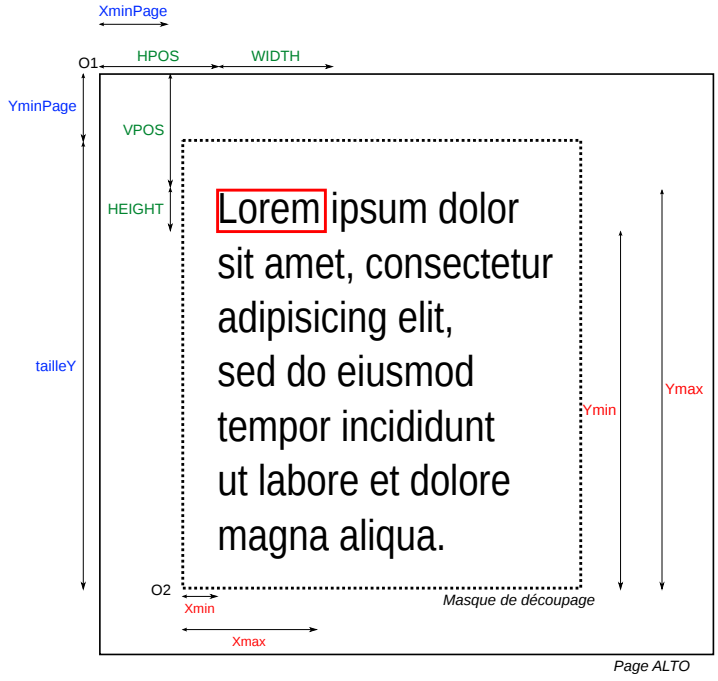


FIGURE 2.2 – Changement de coordonées.

Dès lors on a :

$$
\begin{aligned}
Xmin &= HPOS - XminPage \\
Xmax &= Xmin + WIDTH \\
Ymax &= tailleY + YminPage - VPOS \\
Ymine &= Ymax - HEIGHT
\end{aligned}
$$

2.3 Programmes annexes

2.3.1 *ImageExtractor*

L'*ImageExtractor* est un programme procédant à l'extraction automatique des illustrations des ouvrages, en se basant sur les données ALTO.

Fondement

Comme vu en section 2.1, ALTO est un format basé sur XML de description du texte issu d'une reconnaissance optique de caractères. En plus de la position de chaque mot du texte reconnu, il décrit également les positions des différents éléments graphiques et illustratifs (image ou dessin)[1], suivant la syntaxe suivante :

```
<GraphicalElement ID="PAG_1_GE000002" HPOS="984" VPOS="70" HEIGHT="268" WIDTH="340"
TYPE="manuscript"/>
<Illustration ID="PAG_3_IL000001" HPOS="758" VPOS="694" HEIGHT="370" WIDTH="400">
```

Processus

Algorithme 4: Processus de l'*ImageExtractor*

Entrées : Métadonnées
pour chaque *livre* **faire**
  Découpage du TIFF multi-page en n pages dans un répertoire temporaire
  pour chaque *page* **faire**
    Lecture du XML compressé
    si *le livre contient des illustrations* **alors**
      Création d'un répertoire de sortie
      si *plus d'une illustration* **alors** Changement de la convention de nommage
      pour chaque *illustration trouvée* **faire**
        Récupération des informations de positionnement
        Extraction de l'image

Résultats

Le programme a extrait 22 799 images.

Chapitre 3

Livrables

À l'issue du projet, l'équipe technique a livré les éléments suivants :

- Les fichiers DjVu sur Wikimedia Commons [1] ;
- Les fichiers TIFF sur Wikimedia Commons [2] ;
- Les images extraites sur le Toolserver ;
- Une page de centralisation des ressources, sur Meta-Wiki [3] ;
- Une page de présentation du projet sur Wikimedia Commons [4] ;
- Une page d'aide aux nouveaux contributeurs sur Wikisource [5] ;
- Les sources logicielles du projet, sous licence libre GPL [6] ;
- Un rapport d'avancement mis à jour périodiquement [7] ;
- Le présent rapport de projet.

1. http://commons.wikimedia.org/wiki/Category:Books_provided_by_the_BNF
2. http://commons.wikimedia.org/wiki/Category:Books_provided_by_the_BNF(TIFF)
3. http://meta.wikimedia.org/wiki/BnF
4. http://commons.wikimedia.org/wiki/Commons:Bibliothèque_nationale_de_France
5. http://fr.wikisource.org/wiki/Wikisource:Aider_pour_la_livraison_BnF
6. https://svn.toolserver.org/svnroot/seb35/BnF_import/
7. http://toolserver.org/~seb35/bnf-stats/Stats_Wikisource_BnF-WMFr.pdf

Annexe A

Format des sorties

A.1 Caractéristiques du fichier DjVu

La chaîne de traitement génère des fichiers DjVu encodés en DjVuBitonal (ou JB2), algorithme de compression en bitonal (noir et blanc). Pour les ouvrages disposant d'un OCR, une couche texte est présente, indiquant la position mot par mot.

A.2 Format de la sortie Wikimedia Commons

A.2.1 Texte

```
{{Book
|Author       = <Creators>
|Translator   =
|Editor       = <Editeur>
|Title        = {{lang|fr|<Titre>}}
|Subtitle     =
|Series title =
|Edition      =
|Publisher    =
|Printer      =
|Date         = <Date>
|City         =
|Language     = {{language|fr}}
|Description  = {{fr|1=[[:s:fr:Livre:<DjVu-nom>]]}}
|Source       = {{ARK-BNF|<Ark>}}
|License      = PD-scan
|Image        =
}}
{{BNF cooperation project}}

[[Category:Scanned French books in DjVu]]
[[Category:<Year> books]]
<AutresCategories>
```

Creators
: Les auteurs du livre, selon la syntaxe des modèles dits "*Creator*" de Wikimedia Commons : `{{Creator:Prenom Nom}}`

Editeur
: L'éditeur du livre

Titre
: Titre du livre

Date
: Date de publication du livre

DjVu-nom
: Nom du fichier DjVu

Ark
: L'identifiant ARK *Archival Resource Key*

Year
: Année de publication du livre

AutresCategories
: Catégories additionnelles d'auteurs

A.2.2 Modèles employés

Cette sortie exploite les modèles de Wikimedia Commons suivants :

`{{Book}}`
: Boîte descriptive pour un livre

`{{Creator}}`
: Boîte descriptive pour un auteur

`{{lang}}`
: Indication de langue (notamment pour les synthétiseurs vocaux et l'indexation correcte des inclusions de mots en langues différentes par les moteurs de recherche)

`{{language}}`
: Modèle d'internationalisation, pour afficher une langue en se basant sur son code ISO-639

`{{fr}}`
: Indication de langue, précédée de la mention de la langue (« Français »)

`{{PD-scan}}`
: Bandeau de statut de la ressource multimédia au regard du droit d'auteur, indiquant qu'il s'agit de la numérisation d'un ouvrage du domaine public

`{{ARK-BNF}}`
: Création d'un lien externe vers le site Gallica, en se basant sur l'identifiant pérenne *Archival Resource Key* [6][7]

`{{BNF cooperation project}}`
: Bandeau indiquant le partenariat entre Wikimédia France et la Bibliothèque nationale de France, et donnant des liens vers des pages plus détaillées.

A.3 Format de la sortie Wikisource

```
{{:MediaWiki:Proofreadpage_index_template
|Type=<Typepubli>
|Series=
|Titre=<Titre>
|Volume=
|Auteur=<Auteurs>
|Traducteur=
|Éditeur scientifique=
|Éditeur=<Editeur>
|School=
|Lieu=
|Année=<Annee>
|Clé=
|Source=djvu
|Image=
|Avancement=C
|Pages=<Pagelist>
|Tomes=
|Sommaire=
}}
```

Typepubli
: Type de la publication

Titre
: Titre du livre

Auteurs
: Les auteurs du livre

Editeur
: L'éditeur du livre

Annee
: Année de publication du livre

Pagelist
: Balise `<pagelist/>`, indiquant la numérotation des pages

Annexe B

Index des programmes

La chaîne de traitement est appelée avec `mainBnF.py`, et est constituée des programmes suivants :

mainBnF.py
: Programme d'appel, qui traite un livre ou bien boucle sur tous les livres en parallèle.

bookprocess.py
: Apppelle les scripts et commandes traitant un unique livre.

altoparser.py
: Lit un fichier ALTO et génère un texte en syntaxe DjVu et un texte en syntaxe Wikisoure.

coverparser.py
: Analyseur syntaxique du fichier RefNum et des métadonnées pour générer l'index des pages du livre.

Des programmes indépendants sont appelés en marge de la chaîne de traitement :

createPagelist.py
: Génère les métadonnées `<pagelist />`.

readResol.py
: *generates the metadata resolution (dpi) to resolution.txt, that will be injected in metadatabooks.txt, to finally be included in DjVu.*

treat_list.php
: Fusionne les différentes sources de métadonnées en un seul fichier `metadatabooks.txt`, utilisé par la chaîne de traitement principal.

packOCR.sh
: Gère les fichiers requis par le robot pour initialiser les pages sur Wikisource.

Annexe C

Dialogue avec les communautés

Durant ce projet, l'équipe technique et ses pilotes ont communiqué avec les communautés.

Sur Wikimedia Commons, l'essentiel de la discussion s'est déroulée sur une page dédiée à la préparation du versement de masse des fichiers DjVu. Les discussions se sont étalées du 5 mai au 16 août [1].

Sur Wikisource, les discussions se sont généralement tenues sur le Scriptorium, lieu de dialogue central, ainsi que sur la page de discussion de Sébastien, principal opérateur du robot [2].

- Demande d'information sur un point technique, 5 avril [3] ;
- Annonce de la publication du communiqué de presse, 7 avril [4] ;
- Deamdne d'aide pour la vérification des titres, 11 avril [5] ;
- Demande d'information sur un point technique, 5 mai [6] ;
- Discussions sur l'opportunité d'intialiser l'espace Page, 19 mai – 27 mai [7] ;
- Demande de qualification de fichiers de test, 23 mai – 28 mai [8]
- Information sur l'avancée, 19 juin [9]
- Information sur l'avancée, 2 juillet [10]
- Réflexions sur la qualité de la livraison BnF, 18 juillet – 20 juillet [11]
- Demande d'avis sur les publications et la documentation, 19 juillet [12]
- Information sur l'avancée et demande d'aide sur le rapport d'avancement, 17 octobre [13]

1. http://commons.wikimedia.org/wiki/Commons:Batch_uploading/Bibliothèque_Nationale_de_France
2. http://fr.wikisource.org/wiki/Discussion_utilisateur:Seb35
3. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Avril_2010#Partenariat_avec_la_BNF
4. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Avril_2010#BNF_:_Publication_du_communiqu.C3.A9_de_presse
5. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Avril_2010#BnF_:_titre_des_fichiers
6. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Mai_2010#Documentation_de_pagelist
7. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Mai_2010#Utilisateur:BnFBoT
8. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Mai_2010#Fichiers_de_test_BnF
9. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Juin_2010#Nouvelles_de_la_BnF
10. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Juillet_2010#BnF_:_d.C3.A9nouement
11. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Juillet_2010#Qualit.C3.A9_des_documents_de_la_BNF
12. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Juillet_2010#Accueil_de_nouveaux_contributeurs
13. http://fr.wikisource.org/wiki/Wikisource:Scriptorium/Octobre_2010#Statistiques_pour_la_BnF

Bibliographie

[1] Bibliothèque nationale de France, *Charte de conversion OCR des collections de la BnF*, février 2009, http://www.bnf.fr/documents/charte_ocr.pdf

[2] Bibliothèque nationale de France, *Numérisation et métadonnées*, 5 août 2010, http://www.bnf.fr/fr/professionnels/num_metadonnees/s.num_metadonnees_documents.html

[3] Bibliothèque nationale de France, *Schéma RefNum*, 27 janvier 2010, http://bibnum.bnf.fr/ns/refNum.xsd

[4] Bibliothèque nationale de France, *Conversion OCR et format ALTO*, 12 janvier 2010, http://www.bnf.fr/fr/professionnels/num_conversion_texte/s.num_conversion_texte_ocr.html

[5] Bibliothèque nationale de France, *Schéma ALTO*, 4 décembre 2007, http://bibnum.bnf.fr/pfp/ns/alto_prod.xsd

[6] Bibliothèque nationale de France, *Les identifiants pérennes à la BnF*, juillet 2006, http://bibnum.bnf.fr/identifiants/

[7] Emmanuelle Bermès, Bibliothèque nationale de France, *Des identifiants pérennes pour les ressources numériques : l'expérience de la BnF*, 5 mai 2006, http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf

[8] LizardTech, *Lizardtech DjVu Reference*, novembre 2005, http://djvu.org/docs/DjVu3Spec.djvu