Groupe de travail Accessibilité de la presse ancienne
Réunion 1

arald
livre et lecture
en Rhône-Alpes

Compte rendu

Date : mardi 10 novembre 2015
Horaires : 10h-12h
Lieu : Bibliothèque municipale de Lyon, la Part-Dieu, salle de la Découverte

Participants :

Grégory Charbonnier, Public et nouvelles technologies, Archives municipales de Saint-Etienne
Benoît Encelle, Professeur associé en science informatique, LIRIS, Université Lyon 1-INSA
Guillaume Hatt, Responsable du service informatique, Bibliothèque municipale de Grenoble
Anne-Laurence Hostin, Bibliothécaire, Archives départementales de l'Ardèche
Luc Maumet, Responsable de la médiathèque, Association Valentin Haüy
Anne Meyer, Responsable documentation régionale, dépôt légal, Bibliothèque municipale de Lyon
Denis Reynaud, Responsable de l'équipe LIRE, Gazettes européennes du 18e siècle

Alizé Buisse, Assistante archivage et accessibilité numérique, Arald
Antoine Fauchié, Chargé de mission numérique, Arald
Delphine Guigues, Chargée de mission bibliothèques et patrimoine écrit, Arald

Participants excusés :

Dominique Barbet-Massin, Responsable des collections médiévales, chargée de la politique de numérisation, Bibliothèque municipale de Grenoble
Rachel Brault, Chargée de numérisation, Bibliothèque du patrimoine de Clermont Communauté
Pascale Ferrand, Chargée de ressources documentaires, LIRE, Gazettes européennes du 18e siècle
Taos-Hélène Hani, Responsable des secteurs patrimoine et image & son, Médiathèque publique et universitaire de Valence
Cyril Longin, Directeur, Archives municipales de Saint-Étienne
Jean-Philippe Moreux, Expert du service numérisation, Département de la conservation, Bibliothèque nationale de France

Experts associés : *(ne participant pas aux réunions mais apportant leur regard extérieur)*

Vanessa van Atten, Chargée de mission Publics empêchés, Service du livre et de la lecture, Ministère de la Culture et de la Communication
Bruno Bachimont, Directeur à la Recherche, Sorbonne Universités, Université de Technologie de Compiègne. Directeur scientifique, Institut national de l'Audiovisuel
Carlo Blum, Responsable du département des nouvelles technologies de l'information, Bibliothèque nationale du Luxembourg
Carole Duguy, Espace numérique, Bibliothèque municipale de Lyon
Yves Maurer, Responsable des projets de numérisation, Bibliothèque nationale du Luxembourg
Bruno Mayorgas, Chargé des collections patrimoniales, Médiathèque municipale de Vienne
Eric Nunes, Responsable de la numérisation et valorisation du patrimoine, Bibliothèque-Médiathèque de Nancy
Laurent Tissier, formateur et rééducateur en informatique basse-vision, Formation et insertion des déficients visuels (FIDEV)

Introduction :

Présentation du fonctionnement de l'expérimentation :
- un groupe de travail qui se réunit trois fois (novembre, décembre 2015 et janvier 2016) et qui construit l'expérimentation
- un comité d'experts associés qui réagit à distance aux avancées du groupe de travail via la lecture des comptes rendus des réunions

Les objectifs de la première réunion du groupe de travail sont les suivants :
- définition des bases et des orientations de l'expérimentation ;
- définition du ou des publics ciblés ;
- définition du mode de fonctionnement de l'expérimentation

Présentation des membres du groupe de travail et de leurs intérêts concernant l'expérimentation :

Guillaume Hatt : responsable informatique au sein de la Bibliothèque municipale de Grenoble. Il explique que des titres de la bibliothèque municipale font partie de la campagne de numérisation pilotée par l'Arald entre 1996 et 2012. Actuellement les titres grenoblois sont numérisés mais pas encore accessibles. Guillaume Hatt travaille également sur l'accessibilité numérique du portail de la Bibliothèque de Grenoble, projet qui sera concrétisé en 2016.

Anne-Laurence Hostin : bibliothécaire aux Archives départementales d'Ardèche ; elle est partie prenante et a suivi la campagne de sauvegarde de la presse ancienne régionale depuis dix ans. Les archives départementales de l'Ardèche sont actuellement en train de numériser en XML ALTO.

Grégory Charbonnier : responsable du public et des nouvelles technologies au sein des archives municipales de Saint-Etienne. Les archives ont numérisé en jpeg et PDF les titres stéphanois du 19e siècle. Actuellement la numérisation se fait en XML ALTO (pour les journaux de la période 1914-1918) car il s'agit des recommandations de la Bibliothèque nationale de France. Il est intéressé par les nouvelles fonctionnalités de recherche offertes par le format XML ALTO.

Denis Reynaud : responsable de l'équipe de l'UMR LIRE qui dispose d'un gros volume de presse numérisée, principalement 18e siècle ; mais aussi d'autres ressources (bibliothèque de sites ; liens, etc.). La numérisation a débuté il y a quinze ans, en jpeg uniquement.

Benoît Encelle : chercheur en informatique au sein du laboratoire LIRIS (Laboratoire d'informatique en image et systèmes d'information) et plus spécifiquement du groupe SICAL qui mène différents projets sur les approches et les outils pour améliorer les capacités à interagir dans un contexte collaboratif.

Luc Maumet : responsable de la médiathèque Valentin Haüy spécialisée pour les personnes empêchées de lire en raison d'un handicap. Actuellement la médiathèque a accès aux fichiers sources des éditeurs, en XML ou PDF.

Anne Meyer : responsable des fonds locaux incluant les périodiques régionaux à la Bibliothèque municipale de Lyon. Au départ il n'existait pas de numérisation en format alto (*Progrès Illustré, Revue du lyonnais*). Par la suite une centaine de titres ont été numérisés en alto et mis en ligne sur le portail *Numelyo*. La valorisation de cette presse passe par des dossiers thématiques (dossiers de

presse sur des contenus choisis tels que les journaux anarchistes, la mode, le football, etc.). Anne Meyer indique que le journal se prête au numérique, notamment grâce au dépouillement et à la singularisation de l'article. Les formats utilisés à la bibliothèque municipale de Lyon sont : alto/TEI , MODS, et du METS pour encapsuler les journaux (métadonnées).
Actuellement les serveurs sont saturés du fait de l'import massif de documents issus des numérisations effectuées par Google. Une migration vers d'autres serveurs a due être mise en place ; elle est actuellement en cours. *Numelyo* n'est plus alimenté depuis trois ans ; malgré des contenus disponibles pour une mise en ligne. Pour pallier ce problème de stockage, d'autres fonds (moins volumineux) ont été mis en ligne comme le fonds photographique rhônalpin.
Anne Meyer indique que le passage à l'alto a été motivé par la possibilité de faire de la correction de l'OCR.

Jean-Philippe Moreux : expert OCR et formats éditoriaux du service Numérisation de la BnF. Faisant suite au projet européen Europeana Newspapers, la BnF a fait évoluer ses exigences pour la numérisation de la presse : formats METS/ALTO avec structuration en mode article et typage des contenus. Par ailleurs, la BnF produit des contenus numériques accessibles depuis 2013, via des programmes de rétroconversion OCR vers EPUB 3 et XML DTBook. Améliorer la valorisation des contenus numériques patrimoniaux, ainsi que leur accès par tous les publics, sont des enjeux majeurs de la bibliothèque numérique Gallica.

1. Point d'étape sur le portail *Mémoire et Actualité en Rhône-Alpes* et son évolution vers *Lectura+*

> Bref rappel du projet

Actuellement *Mémoire et Actualité en Rhône-Alpes* donne accès à trente-cinq titres de presse ancienne régionale (400 000 pages), interrogeables en texte intégral (PDF océrisés). En 2016 *Mémoire et Actualité* sera fusionné avec le site *Lectura*, le portail des huit bibliothèques des villes-centres de Rhône-Alpes, pour former *Lectura+*. L'espace *« presse ancienne »* sera maintenu et développé grâce à l'expérimentation actuellement en cours. La dimension de l'accessibilité numérique du portail *Lectura+* et des contenus a été retenue comme perspective majeure de développement en 2016.

> Fonctionnalités existantes et manquantes

Actuellement l'accès à la presse ancienne sur *Mémoire et Actualité* est possible via un moteur de recherche simple ou avancée. Les résultats peuvent être affinés géographiquement, par date, par titre, etc..
La recherche plein texte est possible mais sectionne systématiquement les résultats par page (un résultat correspond toujours à une page d'un journal).

Les manques et les limites du portail *Mémoire et Actualité* sont les suivants :
- problème avec le lecteur Adobe reader qui n'est pas toujours compatible avec les nouvelles versions des navigateurs (notamment Mozilla Firefox ou Chrome/Chromium). Dépendance au lecteur de PDF qui pose problème en terme d'accessibilité. Actuellement lors de la recherche en texte intégral les termes recherchés n'apparaissent pas de manière surlignée dans le lot de résultats
- pas de téléchargement du numéro entier d'un titre. Indexation à la page qui pose des problèmes

quant à l'accès et à la consultation des titres
- Anne Meyer indique qu'il manque également une mise en contexte lors des recherches. Les termes recherchés ne sont pas expliqués et l'utilisateur est obligé de vérifier par lui-même si le moteur de recherche a identifié correctement le mot qu'il cherchait (selon la bonne occurrence). Par exemple s'il s'agit d'un nom propre, d'un nom commun, etc.

Question posée : comment rendre la presse ancienne régionale, patrimoine écrit particulièrement riche, plus accessible ?

Denis Reynaud met en avant ses doutes concernant l'océrisation et se demande s'il est réellement possible de faire mieux que Google. Anne Meyer répond qu'actuellement *Arkhênum* (le prestataire numérisation de la Bibliothèque municipale de Lyon), fait mieux que Google. Elle ajoute que les bibliothécaires ont un réel besoin de presse, qui est un support essentiel pour leur travaux. L'OCR n'est pas parfait, mais il permet néanmoins d'améliorer les recherches par mots-clés en mode intégral.

Luc Maumet insiste sur l'importance de l'OCR en terme d'accessibilité. Avoir à disposition un texte propre est une condition nécessaire pour rendre un document accessible. Il explique que pour les personnes voyantes, l'association des yeux et du cerveau permet de reconstruire une information correcte à partir d'un texte comportant des anomalies. Or, pour les personnes déficientes visuelles, ces anomalies sont beaucoup plus dures à tolérer de manière auditive.
Penser rendre accessible toute la presse ancienne n'est pas réalisable immédiatement selon lui. La prise en compte des besoins des personnes déficientes visuelles constituerait déjà une avancée énorme. La mise en accès d'une sélection réduite de journaux du 19e siècle aurait déjà un grand intérêt et permettrait par exemple de satisfaire la curiosité des personnes déficientes visuelles.
Luc Maumet tient aussi à différencier, au niveau du public déficient visuel, le public de chercheurs spécialisés qui accède déjà aux contenus par des moyens détournés (appel à une personne tierce voyante qui retranscrit les informations nécessaires par exemple).

Recommandations :

- faire montre d'une considération pour le handicap
- travailler sur un corpus réduit mettant en avant la diversité de la presse ancienne. Chercher à rendre accessible la presse ancienne au grand public via quelques exemplaires choisis et non à un public de chercheurs
- travailler sur une meilleure qualité de l'OCR

Question posée : faut-il développer une version tablette et/ou mobile pour l'accès de la presse ancienne ?

Les versions tablettes/mobiles des sites sont souvent beaucoup plus synthétiques et épurées. Elles peuvent donc devenir des entrées en terme d'accessibilité. Les versions tablettes ont par exemple des conséquences sur le graphisme, les contenus. Il s'agit de proposer une réponse reconfigurable et modulable de la version en-ligne initiale.

Il est également possible de réfléchir à des solutions dédiées, adaptées précisément au type de contenu. Par exemple pour les feuilletons littéraires, mettre en place des productions de voix humaines.

Au sein des Archives municipales de Saint-Étienne, la numérisation a été calquée sur le format et le système d'indexation papier déjà en place. Il s'agit avant tout, pour le document numérique, de respecter le document physique et de suivre le système de cotation existant. Pour la période 1914-1918 notamment, les numéros ayant été indexés et rangés au volume, un numéro peut correspondre à six mois de parution.

2. Présentation du format XML/ALTO : de nouvelles perspectives

Présentation de trois projets utilisant le format XML ALTO, du moins avancé au plus avancé :

- Le *Kiosque Lorrain (http://www.kiosque-lorrain.fr)*

Plateforme de la médiathèque de Nancy pour la valorisation de la presse quotidienne régionale de Meurthe-et-Moselle (nouvelle version lancée en 2015). Actuellement cinq titres sont en ligne (60 000 pages). Indexation à la page et au fascicule.
Visionneuse : Omeka (licence libre).
Prestataire technique : Moobee (Nancy)
Formats utilisés : METS/ALTO
Contact : Eric Nunes (Responsable de la numérisation et de la valorisation du patrimoine, Bibliothèque-Médiathèque de Nancy)

Principaux problèmes en terme d'accessibilité :
- OCR de mauvaise qualité ;
- structuration minimale de la page, pas de reconnaissance des articles
- fond noir

- *eLuxemburgensia (http://www.eluxemburgensia.lu)*

Portail de la Bibliothèque nationale du Luxembourg pour la mise en ligne du patrimoine imprimé luxembourgeois lancé en 2009 et progressivement enrichi. Actuellement vingt-huit titres de presse sont accessibles en mode texte et dix titres sont accessibles en mode image. Une version tablette est également disponible.
Visionneuse : open source
Prestataire technique : CCS et Jouve
Contact : Yves Maurer (Responsable du projet numérisation) et Carlo Blum (Responsable du département des nouvelles technologies de l'information)

Avancées en terme d'accessibilité :
- structuration qui reconnaît l'article et le sectionne
- correction des titres (manuelle)
- navigation dans la page via un index hiérarchisé incluant tous les articles
- identification de typologies (article, publicité, météo, avis mortuaire)

- *Trove (http://trove.nla.gov.au/newspaper/)*

Bibliothèque en ligne de la Bibliothèque nationale d'Australie lancée en 2009. Actuellement, au sein

de la section « Digited Newspapers » on accède à plus de 700 journaux australiens (13,5 millions de pages).

Avancées en terme d'accessibilité :
- correction de l'OCR par les utilisateurs via un compte utilisateur
- reconnaissance des articles sur plusieurs pages successives
- fonctionnalités pour conserver les articles (marque-page ; bibliothèque personnelle d'articles ; etc.)

Questions posées : quelles fonctionnalités seraient les plus intéressantes à développer pour ce projet ? Et comment aller encore plus loin par rapport à ce qui existe déjà ?

Anne-Laurence Hostin met en avant l'indexation à l'article et la création d'un index pour la navigation. En effet il s'agit de retrouver facilement les informations, notamment les informations régionales. Les articles traitent largement d'aspects locaux (le village, la campagne) et il faut être en mesure de les identifier rapidement.
Anne-Laurence Hostin insiste également sur l'importance des noms propres. Actuellement les utilisateurs effectuent en grande majorité des recherches de noms propres (nom de famille de personnalités, nom de lieux entre autres).
Pour la presse du 18e siècle, Denis Reynaud précise que tous les noms propres sont en italiques, ce qui peut faciliter leur reconnaissance.
Anne Meyer indique qu'actuellement des listes topographiques sont déjà accessibles via les notices MADS (*Metadata Authority Description Schema* ; modèle pour les données d'autorité développé par la Bibliothèque du Congrès aux Etats-Unis).

Recommandations :

- développer la création d'index dans lesquels il est possible de naviguer et d'accéder rapidement aux spécificités régionales et locales des titres
- travailler sur la structuration et la segmentation de la pages jusqu'à l'article
- reconnaître le type de nom propre : lieu ou personne, via le TEI
- dresser une liste des noms propres les plus utilisés dans la presse ancienne régionale et la transmettre au prestataire pour améliorer la qualité de l'OCR (termes à reconnaître)

Au 18e siècle la presse est structurée de manière fixe (rubriquage identique d'un titre à l'autre). Au contraire la presse du 19e siècle se caractérise par une très grande diversité dans l'organisation des pages (nombre de colonnes qui évolue, type de rubriques très varié, présence d'illustrations, etc.)

Recommandation :

- pour la presse du 19e siècle, identifier les rubriques les plus usitées et/ou les plus pertinentes à mettre en avant

Luc Maumet indique que le format de navigation au paragraphe est intéressant. Il faut savoir ce qui se passe lorsque du XML est transformé en Daisy (*Digital accessible information system*) de manière automatique. Selon lui la perte de structuration d'une page de presse lors d'un export dans d'autres formats accessibles n'est pas un problème (par exemple ne pas savoir si un article se trouve en haut ou en bas de la page, s'il occupe l'espace d'une page entière, etc.). La priorité reste avant tout l'accès au contenu ; rendre compte de la structuration d'une page est donc secondaire. Il serait éventuellement envisageable d'ajouter des métadonnées sur cette structuration, mais elles devront rester minimales.

Recommandation :

- tester une transformation en DAISY d'un article segmenté en XML/ALTO > que se passe-t-il au niveau audio ?

Le module de correction de l'OCR proposé par *Trove* est très intéressant. Néanmoins il s'agit d'un développement complexe à mettre en place et surtout à maintenir (nécessité d'avoir des serveurs suffisamment volumineux pour recevoir en permanence de nouveaux fichiers corrigés et une équipe dédiée pour administrer le module). De fait ce type de correction pourrait aussi être envisagé sur des échantillons et non sur la masse totale des titres disponibles. On sait également que la communauté la plus sensible aux corrections des textes sont les personnes directement concernées par les fonds (bibliothécaires, généalogistes, etc.)

Recommandation :

- réfléchir à la possibilité de rendre l'OCR corrigeable par les internautes ou tout au moins donner la possibilité aux utilisateurs de faire remonter des erreurs (via un formulaire dédié par exemple)

3. La question de l'accessibilité

L'accessibilité doit prendre en compte tous les publics, et par extension les différentes formes de handicap qui existent. Dans le cadre de l'expérimentation de l'Arald, les membres du groupe de travail sont d'accord pour développer en priorité des outils en direction des publics déficients visuels ; d'autant plus que ces travaux serviront également pour d'autres types de handicaps dont les handicaps moteurs ou les troubles de l'acquisition de la lecture (dyslexie). Pour prendre en compte au mieux les besoins et les spécificités de ce public il est crucial de travailler en étroite collaboration avec des professionnels du handicap et des utilisateurs empêchés.

L'accessibilité doit aussi être envisagée dans sa dimension sociale. Sans l'intégration et la participation concrète du tissu local, les contenus ne sauraient être pleinement accessibles. Les associations de soutien aux personnes déficientes visuelles sont des acteurs essentiels capables de proposer d'autres solutions et de mettre en réseau les projets et les utilisateurs. La lecture en voix humaine constitue une piste intéressante. Les publics déficients visuels utilisent en effet massivement le synthèse vocale ; pour autant, lorsqu'ils sont interrogés, ils mettent toujours en avant leur attachement à la voix humaine. D'autre part, la synthèse vocale est limitée. Luc Maumet indique que dès qu'on arrive à plus de 10 % d'erreurs de lecture, la synthèse vocale devient insupportable.

Recommandation :

- associer les associations locales et les communautés empêchées dans le développement d'outils plus accessibles. Mettre en place de nouveaux projets d'accessibilité (sessions d'enregistrements en voix humaines par exemple)

Luc Maumet met en avant le service pour les publics empêchés de la *National Library of Congress* aux Etats-Unis : *Service for the Blind and Physically Handicapped (NLS)* ; qui pourrait être une entrée intéressante pour l'expérimentation menée par l'Arald.

Recommandations :

- prendre contact avec le service dédié aux déficients visuels et moteurs de la Librairie du Congrès (USA)
- à ce stade de l'expérimentation et lors de tests utilisateurs, faire d'abord appel aux professionnels du handicap capables de répondre et d'apporter leur expertise

Denis Reynaud demande si l'idée d'accessibilité pourrait également concerner les publics scolaires. Selon lui la page journal pourrait être trop intimidante et freiner l'intérêt de certains apprenants. Une réflexion pourrait être menée sur les moyens de rendre la presse ancienne plus attractive.

Recommandation :

- réfléchir à une éditorialisation des contenus pour les rendre plus accessibles. Travailler en priorité sur de la presse illustrée et sur des corpus limités

Pour les publics empêchés visuellement l'accès aux illustrations est rendue possibles de deux manières :
- via une présentation en relief
- via une description sonore
Néanmoins il s'agit toujours de procédés très coûteux. Dans le cas de la presse illustrée il faudra nécessairement inclure un accès aux légendes des images. Certaines légendes peuvent d'ailleurs être très complètes (parfois jusqu'à dix lignes de texte).

Il est également important de développer un outil respectant les normes du RGAA 3.0 (application française des WCAG mises en place par le W3C (*World Wide Web Consortium*)).

Recommandation :

- effectuer régulièrement des audits pour le bon respect des normes par un prestataire extérieur spécialisé

4. Organisation de l'expérimentation

> Définition du corpus :

Le corpus sera constitué de trente journaux représentant dix titres de presse ancienne régionale.
Les critères de choix retenus pour la définition du corpus sont les suivants :

1. **critère de valorisation** : privilégier des titres qui ne sont pas encore présents sur le portail *Mémoire et Actualité en Rhône-Alpes*
2. **critère géographique** : faire état de la diversité géographique de la région Rhône-Alpes et de sa future réunion avec l'Auvergne. Inclure au moins un titre de chaque département et un titre auvergnat
3. **critère graphique** : montrer l'évolution visuelle et structurelle des titres (variation dans le nombre de colonnes, la présence d'illustrations, les polices utilisées, etc.)
4. **critère de contenu** : prendre en compte la multiplicité des contenus traités et des orientations/tendances politiques et religieuses
5. **critère symbolique** : créer une cohérence dans le choix des numéros (date anniversaire,

événement important, etc.)

(Cf. document annexe : « Corpus_expérimentation_presse_ancienne)

Denis Reynaud demande si des titres de presse féminine lyonnaise 19e font partie du corpus (par exemple *Le Papillon*). L'Arald répond que ces titres n'ont pas été numérisés dans le cadre de la campagne régionale de sauvegarde de la presse ancienne.

L'Arald propose d'intégrer un titre de la presse du 18e siècle issue des collections du groupe LIRE. Denis Reynaud consultera Pascale Ferrand pour sélectionner un titre adapté pour le corpus de l'expérimentation.
Les Archives municipales de Saint-Etienne proposent également de mettre à disposition des fichiers XML ALTO issus de leur toute récente numérisation.

> La rédaction d'un cahier des charges

L'Arald présente une première trame du cahier des charges qui sera réalisé dans le cadre de l'expérimentation (cf. document annexe). Ce document sera complété et présenté dans une version plus aboutie lors de la prochaine réunion du groupe de travail.

Conclusion

Les deux prochaines réunions du groupe de travail sont fixées le :
- **11 décembre 2015** de 10h à 12h, au sein de la Bibliothèque municipale de Lyon (salle de la Découverte)
- **8 janvier 2016** de 10h à 12h, au sein de la Bibliothèque municipale de Lyon (salle de la Découverte)

La réunion de décembre portera sur les aspects techniques de l'expérimentation et permettra notamment de présenter et de travailler sur le cahier des charges.

L'Arald précise qu'elle rédigera le compte rendu de cette réunion et le transmettra aux membres du groupe de travail et du comité d'experts associés.

Annexe 1 : Corpus_expérimentation_presse_ancienne

> Proposition :

Choisir 3 numéros de 10 titres différents (1 par département + 1 en Auvergne + 1 d'un partenaire associé)
Pour chaque titre toujours choisir :

- le premier numéro du titre (ou de la période numérisée)
- le dernier numéro du titre (ou de la période numérisée)
- un numéro intermédiaire

→ mettre en avant l'évolution du titre, les changements qui ont pu avoir lieu

AIN :

Le Réveil de l'Ain (1843-1847)
Non présent sur M&A ; fichiers numériques retraités en 2011.
Format :PDF, TIF (200/1)
Type : journal politique et généraliste
Rythme de parution : hebdomadaire

Numéros :

N°2, 1è année : 10/12/1843, 4 pages, 3 colonnes
N°48, 2è année : 26/10/1845, 4 pages, 3 colonnes
N°41, 4è année : 15/04/1847, 4 pages, 3 colonnes. ATTENTION : journal annoté en noir : problème avec la reconnaissance des caractères ?

→ attention à la qualité de la numérisation et du journal. Pages souvent numérisées avec saturation de noir, lettres indéchiffrables.

ARDECHE :

La Croix de l'Ardèche (1891-1944)
Non présent sur M&A. Présence de contenus dans d'autres langues (latin)
Format : 1891-1931 : TIF (190-200/1) ; 1932-1944 :JPG (200/8),
Contrainte : traitement à faire (indexation, ocr)
Type : journal religieux
Rythme de parution : hebdomadaire

Numéros :

N°1, 1è année : 28/03/1891, 4 pages, 4 colonnes
N°463, 10è année : 28/01/1900, 3 pages, 4 colonnes
N° ?, 53è année : 19/03/1944, 2 pages, 5 colonnes

DROME :

Le Courrier de la Drôme et de l'Ardèche (1832-1871)
Format : 1832-1851 = JPG (96/8) ; 1852-1871 = JPG (300/8) et TIF (300/8)
Contrainte : utiliser période 1852-1871 car la résolution est meilleure ?
Type : journal politique et généraliste
Rythme de parution : trois fois par semaine

Numéros :

N°1 : 01/05/1832, 3 pages avec 3 colonnes + 1 page en format horizontal consacrée aux « annonces judiciaires et avis divers »
N°7, 19è année : 09/01/1850, 4 pages, 3 colonnes
N°75, 40è année : 28/03/1871, 3 pages + 1 page d'annonces ; 5 colonnes

ISERE :

Le Moniteur Viennois (1842-1944)
Diversité du graphisme (nombreuses polices, présence de mots en gras dans le corps des articles ; nombre de colonnes variable, etc.)
Format : PDF
Type : journal judiciaire
Rythme de parution : hebdomadaire ou bi-hebdomadaire

Numéros :

N°28, 32è année : 14/07/1842, 4 pages, 3 colonnes
N°1, 124è année : 01/01/1916, 2 pages, 5 colonnes
N°33, 153è année : 19/08/1944, 2 pages, 5 colonnes

LOIRE:

Proposition A :

La Tribune Républicaine (1899-1944)
Non présent sur M&A + impératif suite à la Convention non réalisée avec la ville de Roanne
Format : microfilms
Contrainte : titre à numériser depuis les microfilms (qualité des microfilms inconnu)
Type : ?
Rythme de parution : ?

Proposition B :

Le Petit Stéphanois (1881-1916)
Format : PDF et TIF (300/8)
Type : journal républicaine d'informations généralistes régionales.
Rythme de parution : quotidien

N°1, 1è année : 14/07/1881. 4 pages, 4 colonnes. Dernière page consacrée aux annonces.
N°891, 4è année : 01/01/1884. 4 pages, 4 colonnes. Dernière page consacrée aux annonces.
N°1974, 6è année : 31/12/1886. 4 pages, 5 colonnes. ATTENTION : 2 dernières avec partie manquante (arrachée).

RHONE :

Le Salut Public (1848-1942)
Non visible sur M&A et sur Numelyo (BM Lyon)
Format : JPG, TIF (300/8)
Type : « politique, commercial et littéraire »
Rythme de parution : quotidien

Numéros :

N° 1: 13/03/1848, 4 pages, 4 colonnes
N°18, 53è année : 18/01/1900, 4 pages, 7 colonnes
N°1, 97è année : 31/12/1943-01-02/01/1944, 2 pages, 5 colonnes

SAVOIE :

Les Alpes Illustrées (1892-1902)
Non présent sur M&A
Format: JPG (96/1)
Contrainte : résolution basse (96dpi) > retraitement nécessaire ?; journal en couleur numérisé en N&B. Renommage à faire.
Type : journal d'informations à tendance culturelle et artistique. Nombreuses illustrations.
Rythme de parution : hebdomadaire

Numéros :

N°1, 6è année : 09/01/1892, 12 pages, 3 colonnes. Page 5 : illustration sur la demi-page supérieure ; page 7 : illustration sur toute la page ; page 9 : BD sur toute la page.
N°22-23, 9è année : 6-13/06/1895, 12 pages, 3 colonnes. Pages 4 à 7 : partition musicale
N°5, 16è année : 13/03/1902, 8 pages, 3 colonnes. Page 6 : poème sur toute la page

HAUTE-SAVOIE :

L'Avenir Savoyard (1904-1914)
Non présent sur M&A

Format : PDF et TIF (300/8)
Type : journal d'informations régionales
Numéros :

N°1, 1è année : 09/01/1904, 4 pages (dernière page consacrée aux annonces), 6 colonnes.
N°209, 5è année : 02/01/1908, 4 pages (dernière page consacrée aux annonces), 6 colonnes.
N°553, 11è année : 30/07/1914, 4 pages (dernière page consacrée aux annonces), 6 colonnes.

AUVERGNE :

Le Moniteur du Puy-de-Dôme (1856-1944)
Format : PDF, TIF, JPG, XML ALTO
Type : journal d'informations régionales

Numéros :

N°1, : 03-05-1856, 4 pages
N° ??:
N° ?? : 31/07/1944, 2 pages

PARTENAIRE ASSOCIE : **Musée de l'imprimerie** (don du collectionneur Bernard Gelin)

La Gazette de Lyon (XVII°-XVIII° siècles)
Journal d'une autre époque : graphisme différent.
Format : ?
Contrainte : numérisation ou retraitement numérique à effectuer ?
Type : gazette

> Numérisation de masse :

Les Alpes Illustrées (1892-1902)
Format: JPG (96/1)
Nombre de fichiers : 3 496
Contrainte : résolution basse (96dpi) > retraitement nécessaire ?; journal en couleur numérisé en N&B. Renommage à faire.
Type : journal d'informations à tendance culturelle et artistique. Nombreuses illustrations.
Rythme de parution : hebdomadaire

Les Alpes Pittoresques (1901-1914)
Format: JPG (96/1)
Nombre de fichiers : 6 934
Contrainte : résolution basse (96dpi) > retraitement nécessaire ?; journal en couleur numérisé en N&B. Renommage à faire.
Type : revue illustrée historique, artistique et littéraire. Nombreuses illustrations.
Rythme de parution : hebdomadaire

Annexe 2 : Proposition de Cahier des charges (1e trame)

Présentation et but du projet

Expérimentation > travail sur un échantillon diversifié (type de graphisme, de police, d'organisation stylistique, de format de fichier, etc.)
Traitement de masse (10 430 fichiers) pour les *Alpes Illustrées* et *Alpes Pittoresques*
Tester plusieurs formats (et plusieurs structuration de ces formats ?)

Lot à traiter (en fonction du corpus validé)

- 1 titre à numériser depuis des supports microfilms (*La Tribune Républicaine,* Roanne)
- 10 titres disponibles en fichiers numériques (différents formats et qualité de numérisation)

Documents applicables de référence

BnF :
- référentiel OCR, décembre 2013
- référentiel d'enrichissements des métadonnées – version METS, avril 2015

Numérisation

Originaux microfilms
Manipulation des masters de conservation

Retraitement des originaux numériques

Formats des fichiers disponibles :
PDF
TIFF
JPG

OCR

Cf. référentiel OCR de la BnF (décembre 2013)
Précision sur le type de collection
Polices variées et parfois fantaisistes
Langues : français, latin
Taux d'acceptation

Nommage des fichiers

Formats et structures

> Structuration des journaux

Modélisation des pages et ordre de lecture des blocs de texte
Segmentation voulue et règles à appliquer
Types d'éléments à détecter : illustration, légende, photographie, etc.

> Fichiers souhaités

METS
ALTO
DAISY XML
EPUB

> Outils de restitution

Contrôle qualité et validations
Validation des fichiers METS et ALTO
Contrôle de la qualité de la structuration
Intégrité des données
Vérification et corrections manuelles de l'OCR
Critère de refus des lots