La numérisation des caractères non latins et ses contraintes

Bibliothèque nationale de France

DocAsie 2013

Valérie Louison-Oudot

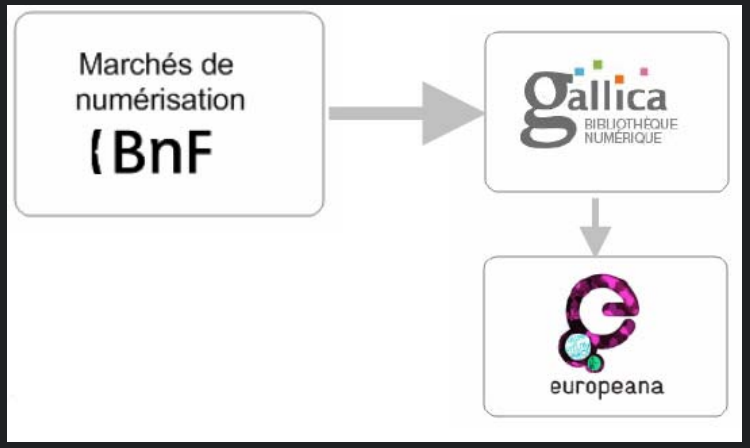
valerie.louison@bnf.fr

Gallica : les dates clés

- Mars 1992 : début de la sélection documentaire pour la numérisation
- 1997 : mise en service de Gallica 1 (documents du XIX^{e} siècle)
- 2000 : mise en ligne de Gallica (documents du domaine public)
- 2005 : charte documentaire Gallica
- 2007 : numérisation à grande échelle
- 2007 : lancement de Gallica 2 : offre de services plus interactifs et actualisation des modalités de recherche et de navigation au sein des documents numériques.
- Mars 2008 : Gallica et l'édition contemporaine : Expérimentation de l'offre payante avec les éditeurs (E-Distributeurs)
- Mars 2009 : le Blog Gallica
- 10 Février 2010 : Gallica atteint le cap symbolique du million de documents en ligne (dont près de 400.000 en mode texte) et adopte une nouvelle interface graphique

Chaîne de numérisation

- Politique documentaire
- Sélection physique et constitution des lots d'envoi
- Numérisation des documents par un prestataire
- Contrôle qualité de la production
- Ajout des documents numériques dans Gallica

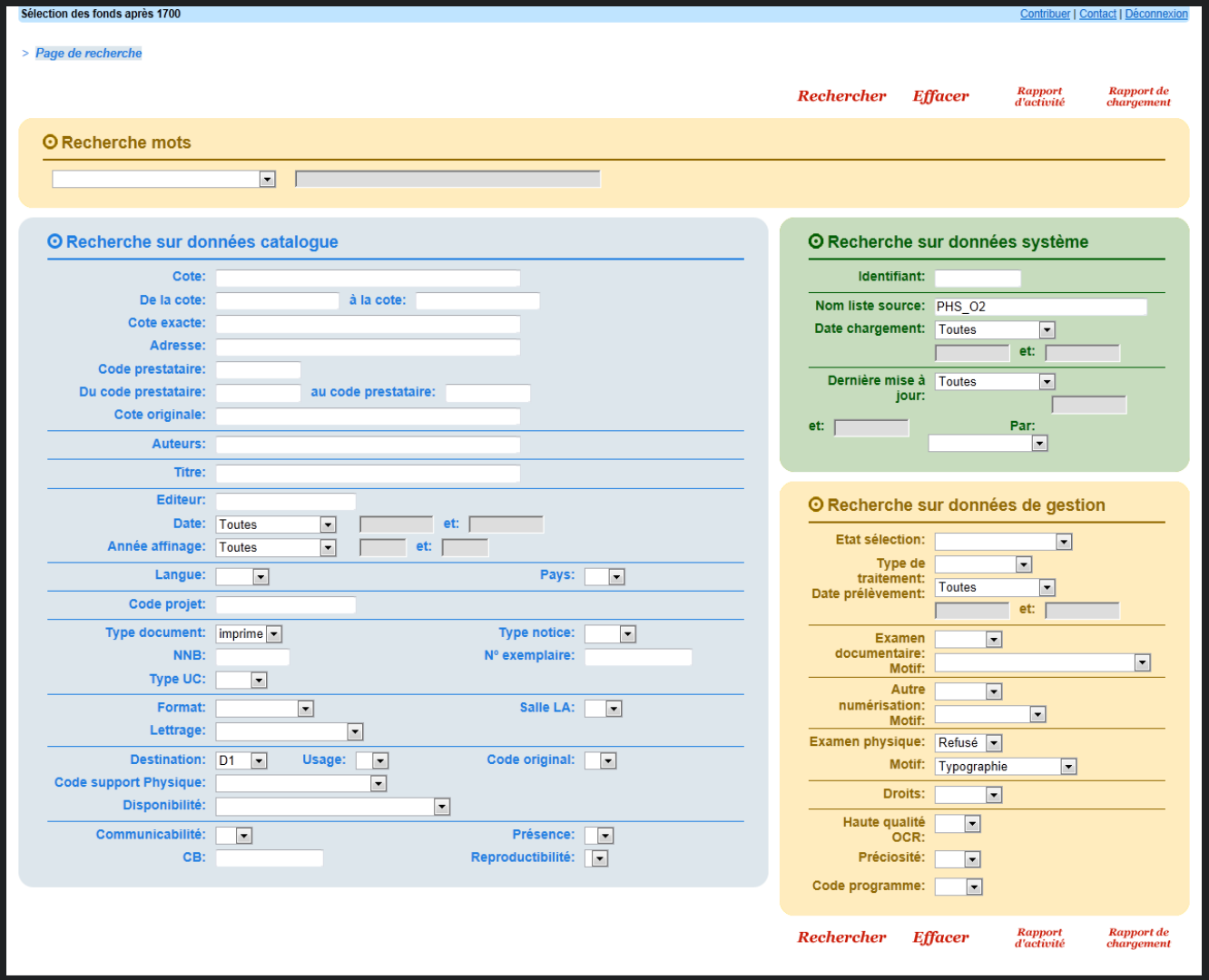


PHS en quelques chiffres

- 2007-2010 Marché SAFIG
- PHS réalise 45% du marché des imprimés
- 46 200 UC sélectionnées
- 6 036 284 pages
- Rythme des envois : 75 000 page par semaine
- une montée en charge en cours de marché qui a porté le taux des lots d'envois à 90 000 pages par semaine

- 2011-2013 Marché JOUVE /SAFIG
- PHS réalise 40% du marché des imrpimés
- au 27 juin 2013 PHS à sélectionné et envoyé en numérisation:
- 28 437 UC soit 4 646 678 pages
- Rythme des envois : 49 321 pages par semaine

Logiciel ADCAT-15



Critères de sélection documentaire

A	B	C	D	E	F
Etat de Sélection	**Examen documentaire**	**motif**	**Commentaire Documentaire**	**Commentaire Général**	
VALIDÉ	Accepté (**Microformes comprises**) (Sous réserve de validation des trois autres critères de sélection)				
	Documents en langue française publiés à l'étranger				
	Documents en langue étrangère ou en caractères non latin (20%) publiés en France				
	Série en plusieurs vol. : "Validé" les tomes non défectueux et pour les tomes posant problèmes ou non éligibles → les mentionner dans le Fichier de complétude DCP	Fichier de complétude: Base des séries lacunaires suite aux sélections PHS pour envoi en numérisation_ Maché Jouve/Safig 2011-2014			Haute qualité OCR: Oui Préciosité: Code programme:
	Série en plusieurs vol. : pour les tomes microformés **sélectionner le papier** si bon état.				
	Pour toute cote sélectionnée pour un traitement en OCR HQ	Haute qualité OCR: **Oui**			
REFUSÉ	Document présentant des anomalies	Anomalie du catalogue			
	Document ant. 1750	Pertinence et /ou Non numérisable		Proquest	
	Ré-édition non significative	Autre ed., reprint, extrait, tiré à part			
	Double d'un document	Autre exemplaire	Double de la cote 8-Z Le Senne-...		Refusé
	Document dans une langue étrangère non édité en France (sauf anciennes colonies- France 19ème-20ème)	Langues, pays			Anomalie du catalogue
	Indisponible Soit: " absence constatée" ou "manque en place" dans le catalogue	Non numérisable			Anomalie du catalogue Autre exemplaire Autre éd., reprint, extrait, tiré à part Langues, pays Non numérisable Pertinence Reproduit Tome isolé
	Document non pertinent	Pertinence			
	Ensemble incomplet, document isolé	Tome isolé			
	Accepté			Sauvegarde	
Á REVOIR	Recueil de pièces	Non numérisable	Recueil de pièces de 4-Z Larrey-45 à 56	Recueil de pièces - Reliées ensembles de ... cote à...cote	
NOTA BENE	**Tout document conforme à l'ensemble des critères de sélections, physique et documentaire est à passer à l'état de sélection "VALIDÉ"**				

Critères de sélection physique

Etat de Sélection	Examen physique	motif	Commentaire examen physique	Commentaire Général
REFUSÉ	Absence constatée, manque en place, en traitement	indisponible	Manque en place ou **en traitement**	
	Cartonnage romantique	façonnage	Cartonnage romantique	
	Plein cuir: reliure disloquée	Reliure		
	Papier acide et cassant (des particules se détachent, papier effrité)	papier		Sauvegarde (si doc non déjà reproduit)
	Couture cassée ou brochage cassé et cahiers et feuillets détachés	façonnage		
	Typographie illisible, fortes transparences, taches gênant la lecture	Typographie		
	Plus de 20% de caractères non latins	Typographie	signaler si gothique	
	Dépliant supérieur au A4 et ouverture < 80°	format	Dépliant	
	Marge < à 3mm, ET doc s'ouvre mal et/ou épais	marge		
	Partie de texte, feuillets, ou cahiers manquants	Lacune		
NB: tome de série/pério refusé	un ou plusieurs tomes d'une série sont HU non réparables ou MP	*Mettre le motif de refus*	*Noter la cote dans le tableau des séries lacunaires*	série
Á REVOIR	Partie de reliure (1/2 cuir, toile) qui se détache (plat, coiffe, dos…) mais couture OK	Reliure	Reliure	*Prélever le document, mettre un fantôme petite rep. et aller le déposer sur l'étagère du L3.001 + signet explicatif*
	Feuillets ou cahiers qui se détachent mais couture OK	façonnage	façonnage	*Prélever le document, mettre un fantôme petite rep. et aller le déposer sur l'étagère du L3.001 + signet explicatif*
	Onglet fragile en début ou fin d'ouvrage	façonnage	onglet	*Prélever le document, mettre un fantôme petite rep. et aller le déposer sur l'étagère du L3.001 + signet explicatif*

Examen physique: Refusé

Motif:

Commentaire:

Haute qualité OCR:

Nb pages:

Nb fascicules:

Préciosité:

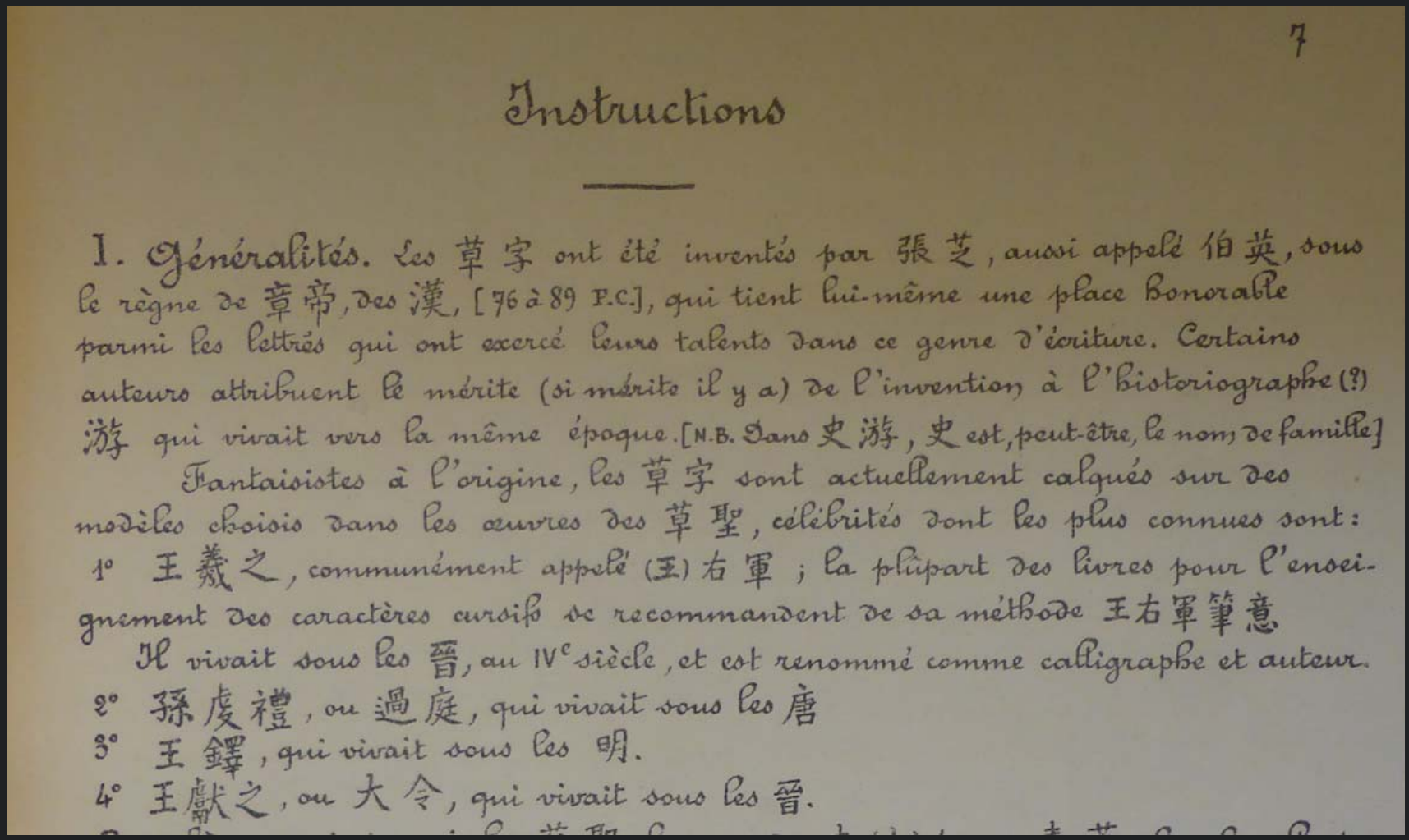
Code programme:

Commentaire:

Thématique:

- Façonnage
- Format
- Indisponible
- Lacune
- Marge
- Non numérisable
- Original HU sur micro
- Papier
- Petits formats
- Plein cuir
- Reliure
- Typographie

Exemple de refus



7

Instructions

I. Généralités. Les 草字 ont été inventés par 張芝, aussi appelé 伯英, sous le règne de 章帝, des 漢, [76 à 89 P.C.], qui tient lui-même une place honorable parmi les lettrés qui ont exercé leurs talents dans ce genre d'écriture. Certains auteurs attribuent le mérite (si mérite il y a) de l'invention à l'historiographe (?) 游 qui vivait vers la même époque. [N.B. Dans 史游, 史 est, peut-être, le nom de famille]

Fantaisistes à l'origine, les 草字 sont actuellement calqués sur des modèles choisis dans les œuvres des 草聖, célébrités dont les plus connues sont:

1° 王羲之, communément appelé (王) 右軍 ; la plupart des livres pour l'enseignement des caractères cursifs se recommandent de sa méthode 王右軍筆意

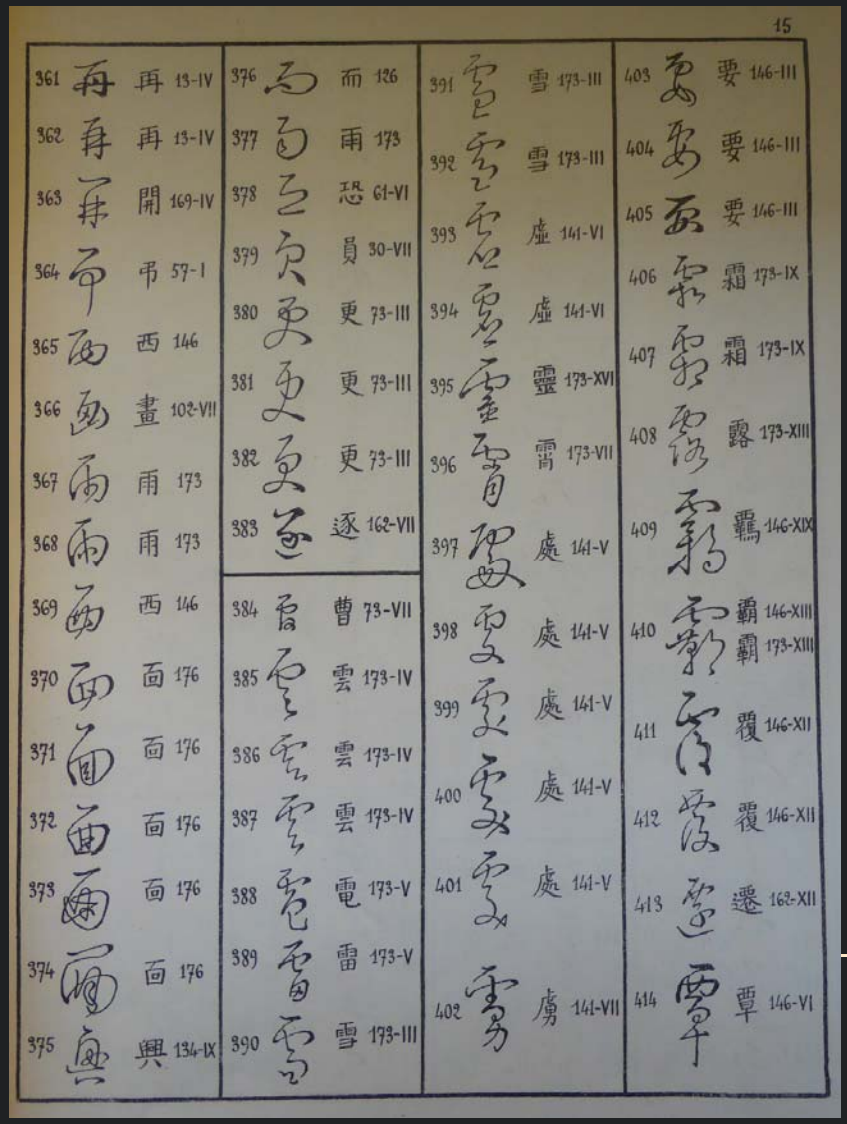
Il vivait sous les 晉, au IV^{e} siècle, et est renommé comme calligraphe et auteur.

2° 孫虔禮, ou 過庭, qui vivait sous les 唐

3° 王鐸, qui vivait sous les 明.

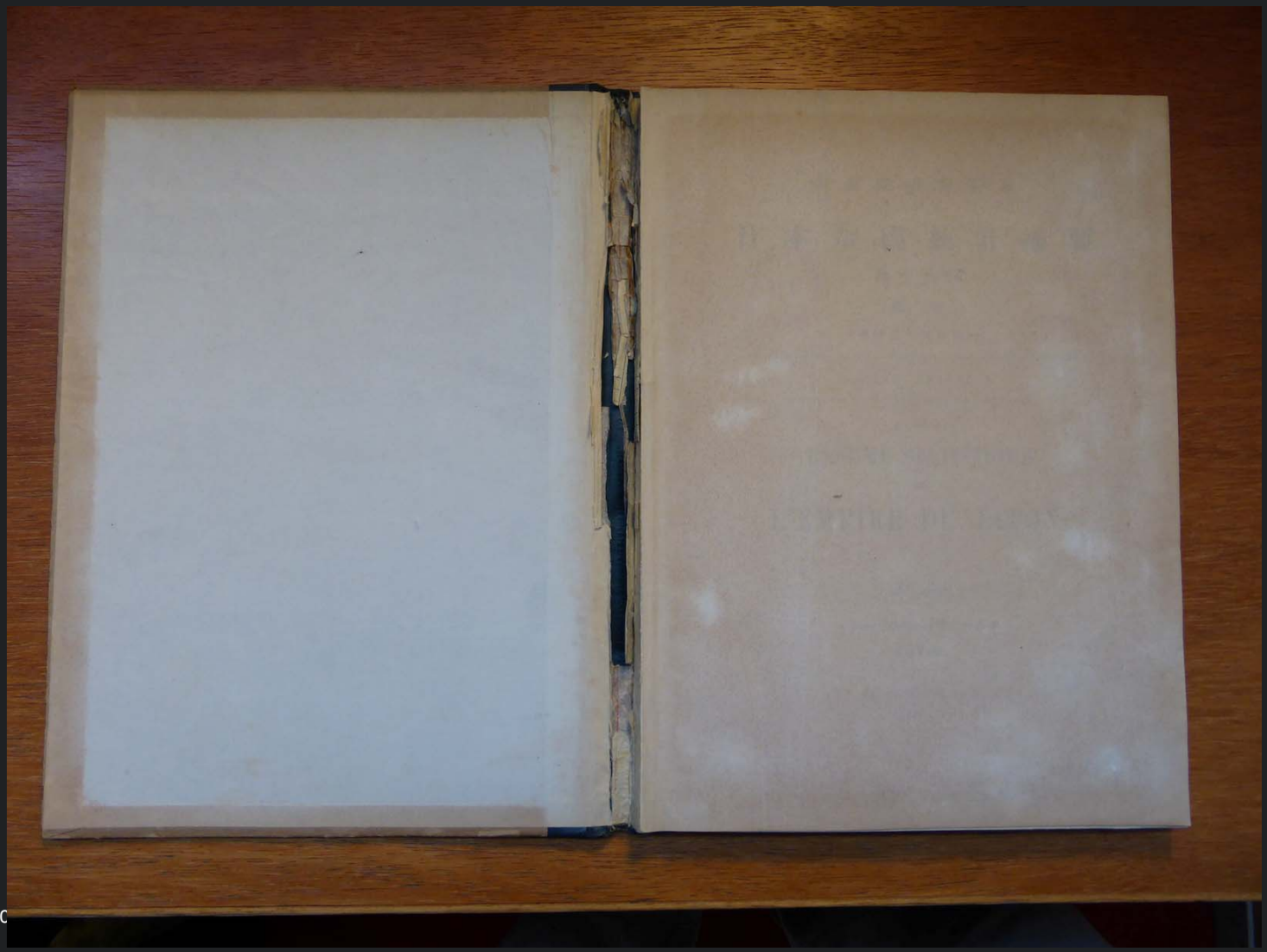
4° 王獻之, ou 大令, qui vivait sous les 晉.

Exemple de refus



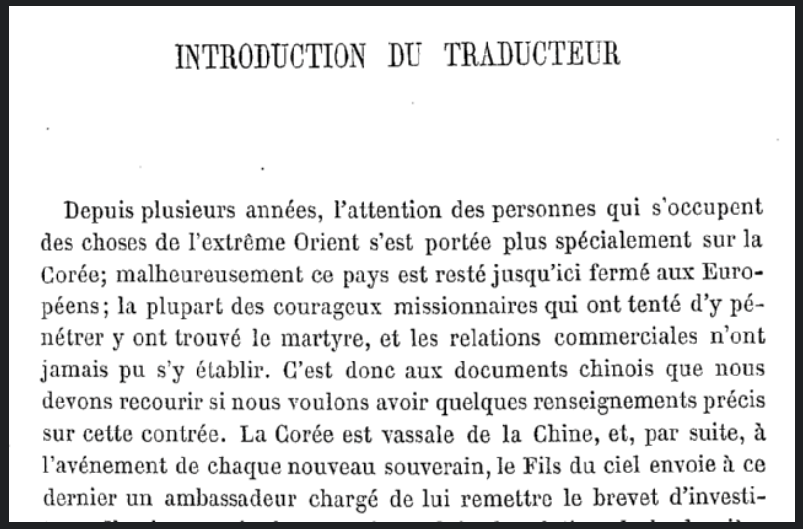
Exemple de refus

Exemple de refus



Modalités de consultation des documents

- Mode image



INTRODUCTION DU TRADUCTEUR

Depuis plusieurs années, l'attention des personnes qui s'occupent des choses de l'extrême Orient s'est portée plus spécialement sur la Corée; malheureusement ce pays est resté jusqu'ici fermé aux Européens; la plupart des courageux missionnaires qui ont tenté d'y pénétrer y ont trouvé le martyre, et les relations commerciales n'ont jamais pu s'y établir. C'est donc aux documents chinois que nous devons recourir si nous voulons avoir quelques renseignements précis sur cette contrée. La Corée est vassale de la Chine, et, par suite, à l'avénement de chaque nouveau souverain, le Fils du ciel envoie à ce dernier un ambassadeur chargé de lui remettre le brevet d'investi-

- Mode texte

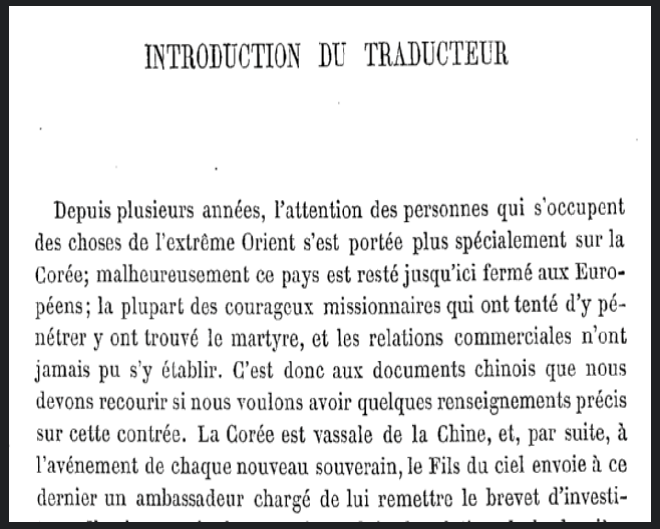
INTRODUCTION DU TRADUCTEUR

Depuis plusieurs années, l'attention des personnes qui s'occupent des choses de l'extrême Orient s'est portée plus spécialement sur la Corée; malheureusement ce pays est resté jusqu'ici fermé aux Européens; la plupart des courageux missionnaires qui ont tenté d'y pénétrer y ont trouvé le martyre, et les relations commerciales n'ont jamais pu s'y établir. C'est donc aux documents chinois que nous devons recourir si nous voulons avoir quelques renseignements précis

sur cette contrée. La Corée est vassale de la Chine, et, par suite, à l'avènement de chaque nouveau souverain, le Fils du ciel envoie à ce dernier un ambassadeur chargé de lui remettre le brevet d'investi-

OCR (Optical Character Recognition)
Reconnaissance optique de caractères

- Mode image



INTRODUCTION DU TRADUCTEUR

Depuis plusieurs années, l'attention des personnes qui s'occupent des choses de l'extrême Orient s'est portée plus spécialement sur la Corée; malheureusement ce pays est resté jusqu'ici fermé aux Européens; la plupart des courageux missionnaires qui ont tenté d'y pénétrer y ont trouvé le martyre, et les relations commerciales n'ont jamais pu s'y établir. C'est donc aux documents chinois que nous devons recourir si nous voulons avoir quelques renseignements précis sur cette contrée. La Corée est vassale de la Chine, et, par suite, à l'avénement de chaque nouveau souverain, le Fils du ciel envoie à ce dernier un ambassadeur chargé de lui remettre le brevet d'investi-

océrisation

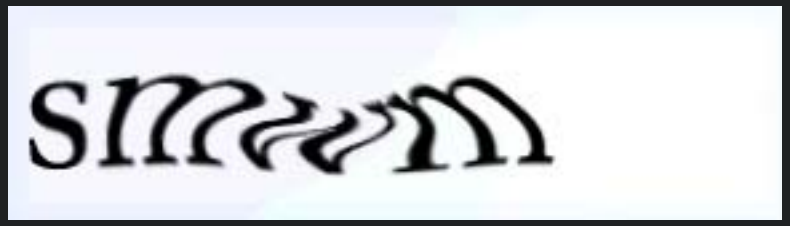
- Mode texte

INTRODUCTION DU TRADUCTEUR

Depuis plusieurs années, l'attention des personnes qui s'occupent des choses de l'extrême Orient s'est portée plus spécialement sur la Corée; malheureusement ce pays est resté jusqu'ici fermé aux Européens; la plupart des courageux missionnaires qui ont tenté d'y pénétrer y ont trouvé le martyre, et les relations commerciales n'ont jamais pu s'y établir. C'est donc aux documents chinois que nous devons recourir si nous voulons avoir quelques renseignements précis
sur cette contrée. La Corée est vassale de la Chine, et, par suite, à l'avènement de chaque nouveau souverain, le Fils du ciel envoie à ce dernier un ambassadeur chargé de lui remettre le brevet d'investi-

CAPTCHA

- Un CAPTCHA est une forme de test permettant de différencier de manière automatisée un utilisateur humain d'un ordinateur.



smwm

reCAPTCHA

- reCAPTCHA est un système mettant à profit les capacités de reconnaissance des utilisateurs humains mobilisées par les tests Captcha, pour améliorer par la même occasion le processus de numérisation de livre.

Document numérique

- Un fichier de métadonnées (refNum)
- Un répertoire d'images (TIFF ou JPEG 2000)
- Un répertoire d'OCRisation (fichier Alto)
- Un fichier table des matières (TEI)