



HAL
open science

Métadonnées dans un outil de gestion d'images

Stéphane Pouyllau, Lucie Secchiaroli, Julie Sastrada

► **To cite this version:**

Stéphane Pouyllau, Lucie Secchiaroli, Julie Sastrada. Métadonnées dans un outil de gestion d'images : Expériences de mise en ligne d'archives et de corpus numériques scientifiques par le pôle Histoire des Sciences et des Techniques en Ligne du Centre Alexandre Koyré-CRHST. 2005. halshs-00004902v2

HAL Id: halshs-00004902

<https://shs.hal.science/halshs-00004902v2>

Preprint submitted on 11 Oct 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Métadonnées dans un outil de gestion d'images

Expériences du pôle HSTL du CAK-CRHST

Stéphane Pouyllau, IE CNRS

Lucie Secchiaroli, documentaliste contractuelle au pôle HSTL



- Le pôle HSTL du CRHST
 - Plateforme d'informatisation des données de recherche au sein d'UMR
 - Création en 2001 par Pietro Corsi
 - Mise en place en 2002
 - Mission : assister les équipes de recherche en matière d'informatique et de web dans le domaine de l'Histoire des Sciences



- Equipe CNRS
 - 1 IE responsable du pôle et développeur
 - 2 techniciens d'édition web et de gestion de bases de données
 - 1 documentaliste contractuelle
 - 1 secrétaire gestionnaire (demi-poste)



- Le corpus scientifique numérique pour la recherche

Œuvres et rayonnement de Jean-Baptiste Lamarck (1744-1829) - Mozilla Firefox

http://www.crhst.cnrs.fr/j-lamarck/

Œuvres et rayonnement de Jean-Baptiste Lamarck

Jean-Baptiste LAMARCK (1744 - 1829)
ANIMAUX SANS VERTÈBRES

Vie de Lamarck <ul style="list-style-type: none">PrésentationChronologieTémoignages et biographiesHistoriographieDocuments	Œuvres de Lamarck en texte intégral <ul style="list-style-type: none">Ouvrages<ul style="list-style-type: none">LivresDiscoursArticles de revueArticles de dictionnaireMémoiresManuscritsCorrespondanceManuscritsHerbier	Lamarck en son temps <ul style="list-style-type: none">PrésentationContemporains et critiques<ul style="list-style-type: none">Textes en françaisTextes en anglaisTextes en italienBibliographies des études sur la vie et l'œuvre de Jean-Baptiste Lamarck	Auditeurs de Lamarck <ul style="list-style-type: none">PrésentationAccès à la banque de donnéesDocuments sur les auditeurs de LamarckNotes prises par des auditeurs au cours de Lamarck et d'autres professeurs du Muséum National d'Histoire Naturelle (Nouveau)
---	---	--	---

Rechercher dans le texte des œuvres : Lancer

Terminé

@. ampère : Autobiographie d'Ampère. Cahier manuscrit de 16 feuillets dont les 8 premiers feuillets sont autographes, les 8 suivants de la main d'...

http://www.ampere.cnrs.fr/ice/ice_page_detail.php?lang=fr&type=text&bdd=ampere&table=ar

CRHST (UMR n°8560... Centre Alexandre K... Novell WebAccess @. ampère : Ampère... Buffon : l'Œuvre co... CC-IN2P3: Bases de... histsciences.univ-pa...

Accueil Recherche libre dans les textes Lancer

Documents biographiques • AMPERE, Autobiographie d'Ampère, [1824].

Page 2 Aller à la page Autobiograph... p.1 voir

retour au texte

A. 21-22-1825

ampère - marie ampère naquit à Lyon le 20 janvier 1775 de Jean - jacques ampère négociant, et de jeanne - antoinette de futières - parey. son père qui n'avait jamais été de l'école de la littérature latine et française, ainsi que plusieurs branches des sciences, l'éleva lui-même dans une campagne voisine de la ville où il était né. jamais il n'écrivit de latin quoique ce soit mais il fut lui

Terminé



- Le corpus scientifique numérique pour la recherche
 - Ensemble des documents qui sont utiles aux chercheurs (archives, archives perso, documentation secondaire, etc.)
 - Accompagnement de l'équipe de chercheurs tout au long du projet
 - Développement d'outils dédiés



- Réalisations :
 - www.lamarck.science.gouv.fr
 - www.lavoisier.science.gouv.fr
 - www.ampere.cnrs.fr
 - www.buffon.cnrs.fr
 - www.histmap.net
 - ...



2) Numérisation photographique



- Numérisation des images et photos au pôle HSTL
 - Vocation directe du pôle
 - Numérisation respectant les chartes du Ministère de la Culture (OAI, DC, EAD, 600 dpi)





- Numérisation externe :
 - Grands formats et grosse quantité
 - Partenariat avec la société AZENTIS (Paris)
 - 6 contrats en 2005

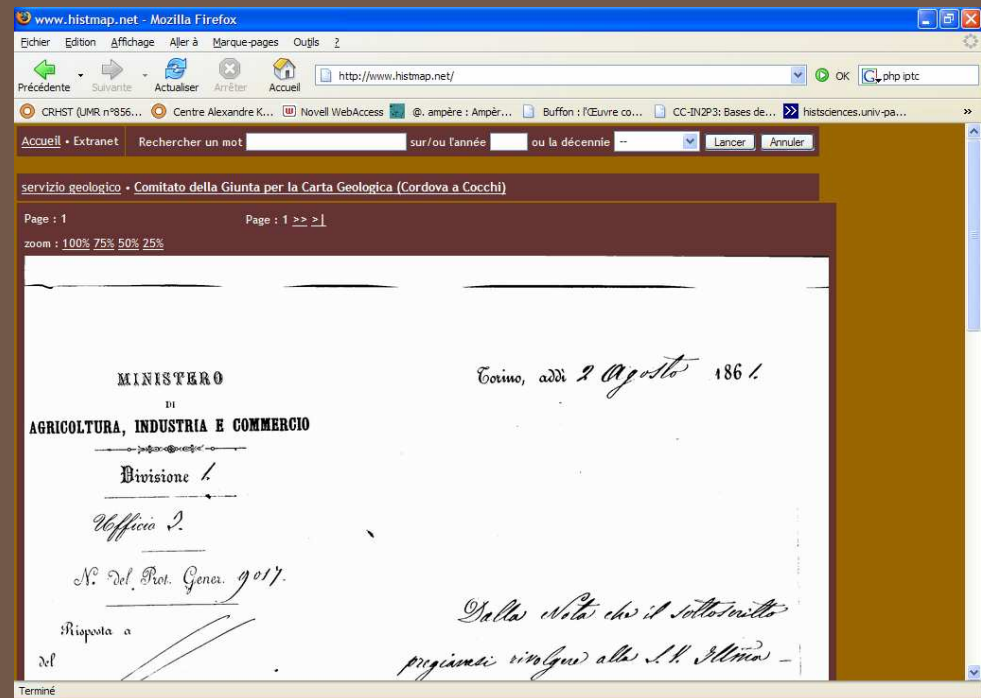


JumboScan azentis

- 30000 x 12000 pixels
- La plus grande surface de capteur au monde 78 x 195 mm
- Le plus grand scanner au monde 5 x 2 m
- Le scanner le plus rapide au monde 280 Mbits/sec.



- Numérisation « locale »
 - Documents transportables
 - Moins de 1000 documents
 - Scan A3, A4
 - OCR (Omnipage)





- Numérisation photo : la nummobile
 - Documents précieux
 - Mobilité
 - Coût faible
 - Qualité bonne à très bonne (400 dpi – 12 millions de pixels)
 - Pilotage via l'ordinateur
 - www.nummobile.com





International Press and Telecommunications Council (IPTC)

Historique

– 1965

- Organisation créée afin de définir des standards d'échanges de données pour le monde de la presse

– 1990

- Élaboration d'un modèle global de données, le IIM (Information Interchange Model) en association avec la NAA (Newspaper Association of America)

- Norme : IPTC-NAA IIM



- Adobe Photoshop (par ex.) : permet d'indexer une image selon différentes entrées IPTC :

- Légende
- Mots-clés
- Catégories
- Crédits
- Source
- Copyright

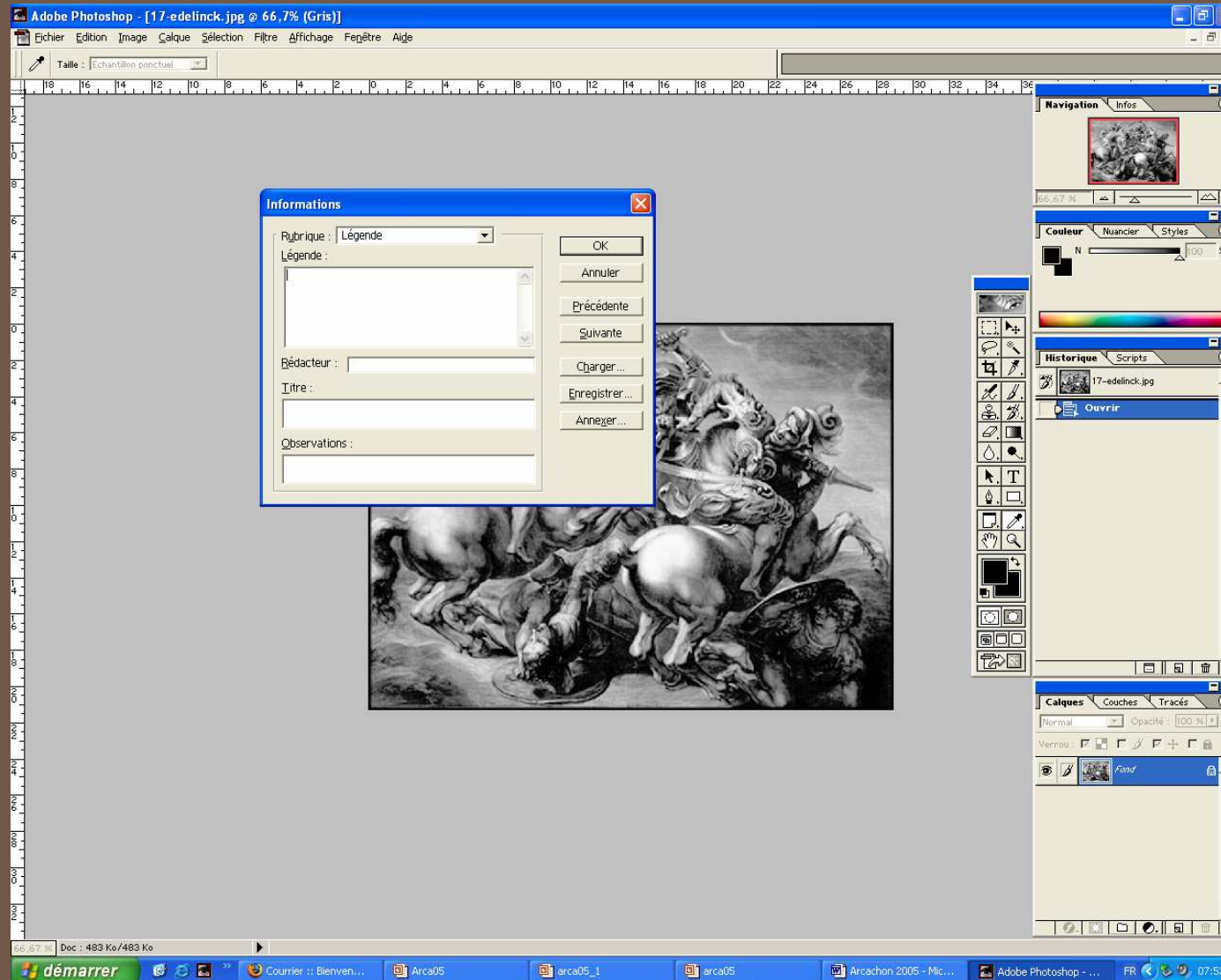
[Adobe Photoshop intègre cette fonction directement
– Fichier > Informations]



- Synthèse sur l'IPTC pour la documentation réalisée par le pôle HSTL =

<http://halshs.ccsd.cnrs.fr/halshs-00004902>

Ref : halshs-00004902



Exemple d'une entrée de données sous Photoshop

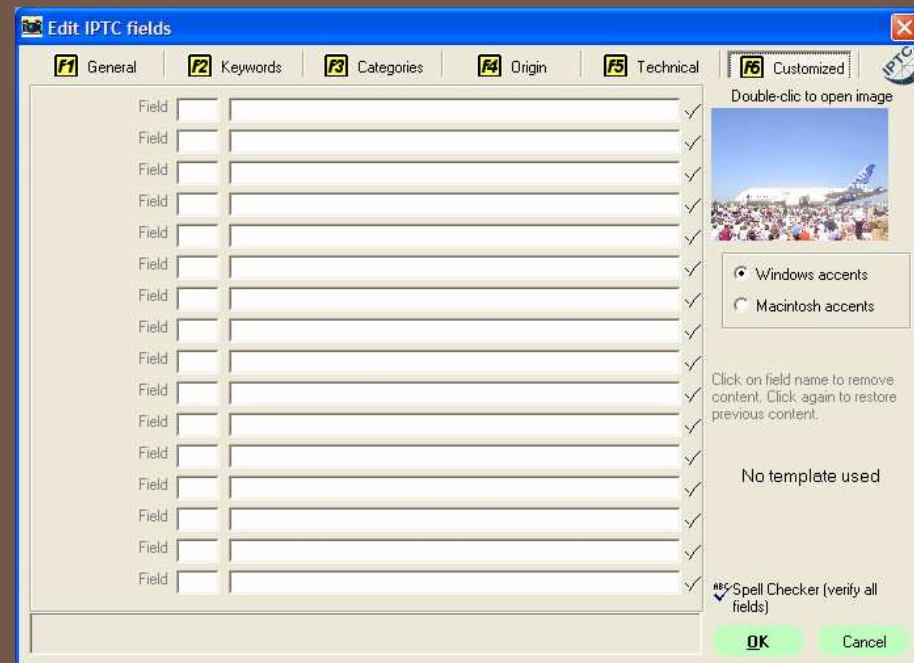


- Limites documentaires d'IPTC-NAA :
 - Cette norme est utilisée dans des cas très précis de traitement documentaire, principalement la photographie
 - Si l'on doit traiter d'autres types de ressources...




• La norme IPTC et les besoins en SHS

- Informations techniques
- Informations descriptives
- Informations juridiques
- Ajouts complexes
- Opération lourde
- Logiciels propriétaires
- Limite des formats (tif, jpg)





- Récupération de métadonnées IPTC à l'aide d'un script php 
 - Utilisation de la GD 2.0
 - iptcparse
 - PHP 3 >= 3.0.6, PHP 4, PHP 5
 - Programmation lourde
- **Démo...**



3) Traitements documentaires et métadonnées



- Numérisation textuelle
 - Saisie des grands textes – WordPro
 - OCR
 - Edit-MathML (développement pôle HSTL)
 - Moteurs de recherche

The screenshot shows a Mozilla Firefox browser window with the URL <http://histsciences.univ-paris1.fr/j-corpus/lavoisier/page-detail.php?bookId=54&pageNumber=580>. The page title is "Lavoisier, Antoine Laurent (1743-1794) : Mémoire sur la combustion du fer dans l'air vital, p.580". The website content includes a search bar, navigation links like "Accueil", "Les livres", "Les mémoires", and "La correspondance". The main text is a scanned document page with the title "Mémoire sur la combustion du fer dans l'air vital" and a small portrait of Antoine Lavoisier. The text describes an experiment involving oxygen and iron.



OCR : Utilisation d'OmniPage 12 (anglais, français, italien, espagnol)

The screenshot shows the OmniPage Pro software interface. The main window displays a document with a yellow background and a blue border. The document text is partially visible, including the title "Julien Bredin à peine âgé" and the subtitle "LETTRES INÉDITES". A verification dialog box is open over the document, showing the text "LETTRES INÉDITES" and a list of suggestions. The dialog box has buttons for "Remplacer", "Bypasser", "Ignorer tout", "Ajouter", and "Enlever".

Tableau de bord du document

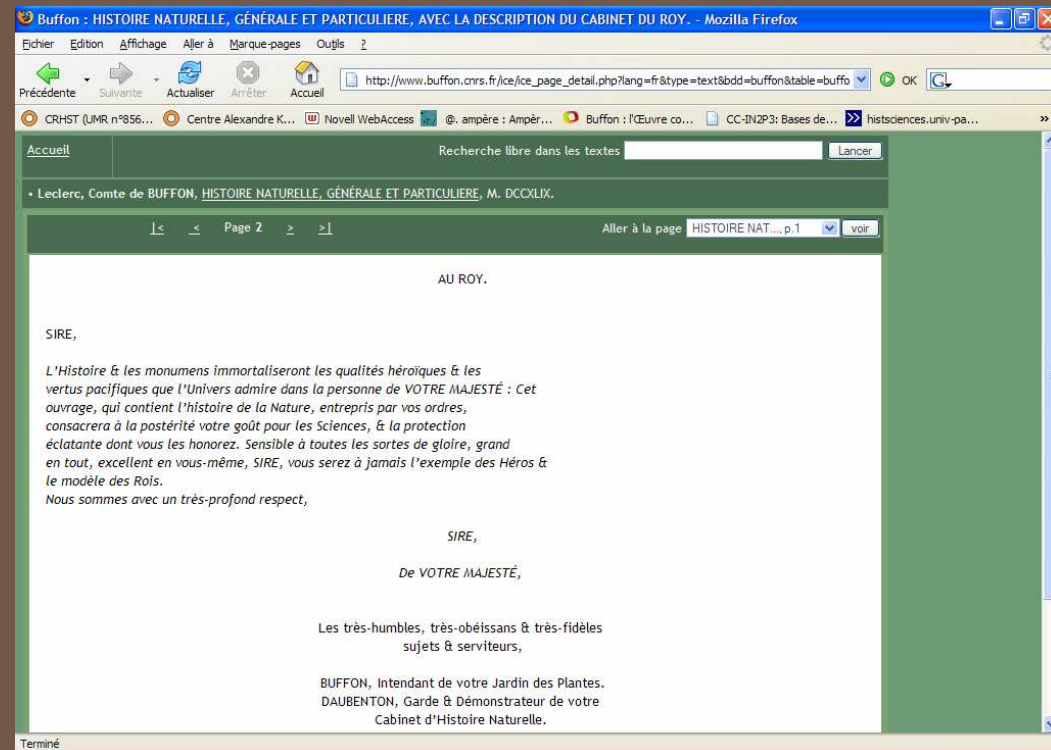
Page	Etat	Mots douteux	Car...	Mots	Préc...	f
001	Rec...	37	2075	391	100...	
002	Rec...	16	2129	432	100...	
003	Rec...	15	2039	405	100...	
004	Rec...	15	2109	415	100...	
005	Rec...	23	2148	396	100...	
T...		106	10500	2039	100...	

Tableau de bord de la page

Page	Etat	Mots douteux	Car...	Mots	Préc...	f
001	Rec...	37	2075	391	100...	
002	Rec...	16	2129	432	100...	
003	Rec...	15	2039	405	100...	
004	Rec...	15	2109	415	100...	
005	Rec...	23	2148	396	100...	
T...		106	10500	2039	100...	



- Re-saisie en double clavier (France, Azentis, WordPro, Inde) : sur les grands textes (œuvres complètes, etc.)





Dublin Core (DC)

- **Date de création : 1995**
 - Ensemble de 15 éléments regroupés selon 3 catégories
 - » Le Contenu (Title, Description, Relation,...)
 - » la Propriété intellectuelle (Creator, Contributor, Rights,...)
 - » la Version (Date, Format, Language,...)
- La syntaxe XHTML permet d'inclure des éléments du Dublin Core dans une page XHTML en utilisant les balises "META"




Exemple d'entête d'une page XHTML + DC

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Document sans nom</title>

<meta name="DC.Language" content="fr"/>
<meta name="DC.Title" content="Une page XHTML" />
<meta name="DC.Title.Subtitle" content="Le Dublin Core" />
<meta name="DC.Subject.Keywords" content="Métadonnées, Dublin Core;" />
<meta name="DC.Creator" content="pole HSTL" />

</head>
```



- Norme de description très générale :
 utile pour le domaine des SHS
puisque'elle permet de décrire une grande
variété de ressources hétérogènes



- **Mise en ligne : ICEberg et ICEberg-XML...**

ICEBERG permet :

- La gestion des archives de chercheurs
- La gestion de ressources texte intégral + images
- Gestion séparée du contenu et du contenant

= **Gestion de silos de documents numériques**

Démo !

