Métadonnées dans un outil de gestion d'images

Stéphane Pouyllau, Lucie Secchiaroli, Julie Sastrada

Métadonnées dans un outil de gestion d'images

Expériences du pôle HSTL du CAK-CRHST

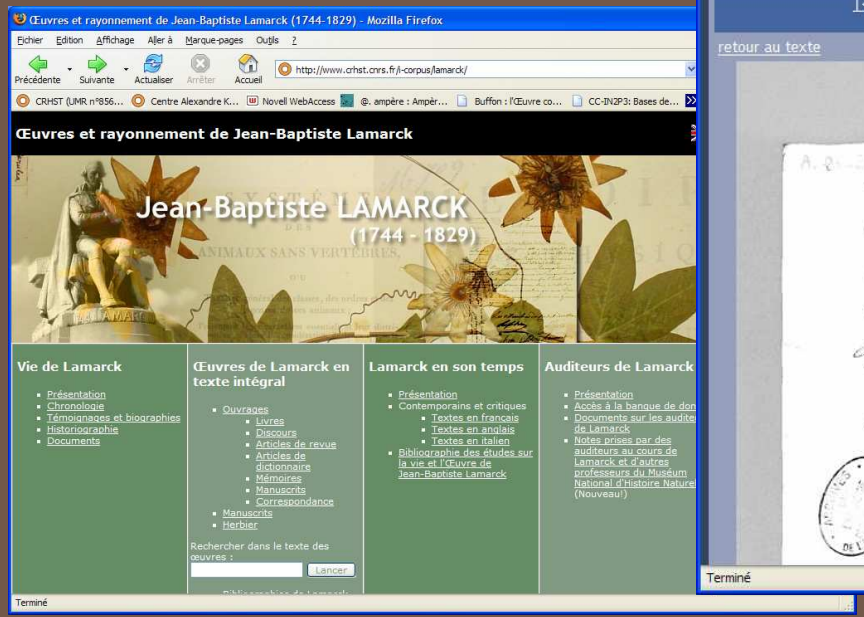
Stéphane Pouyllau, IE CNRS

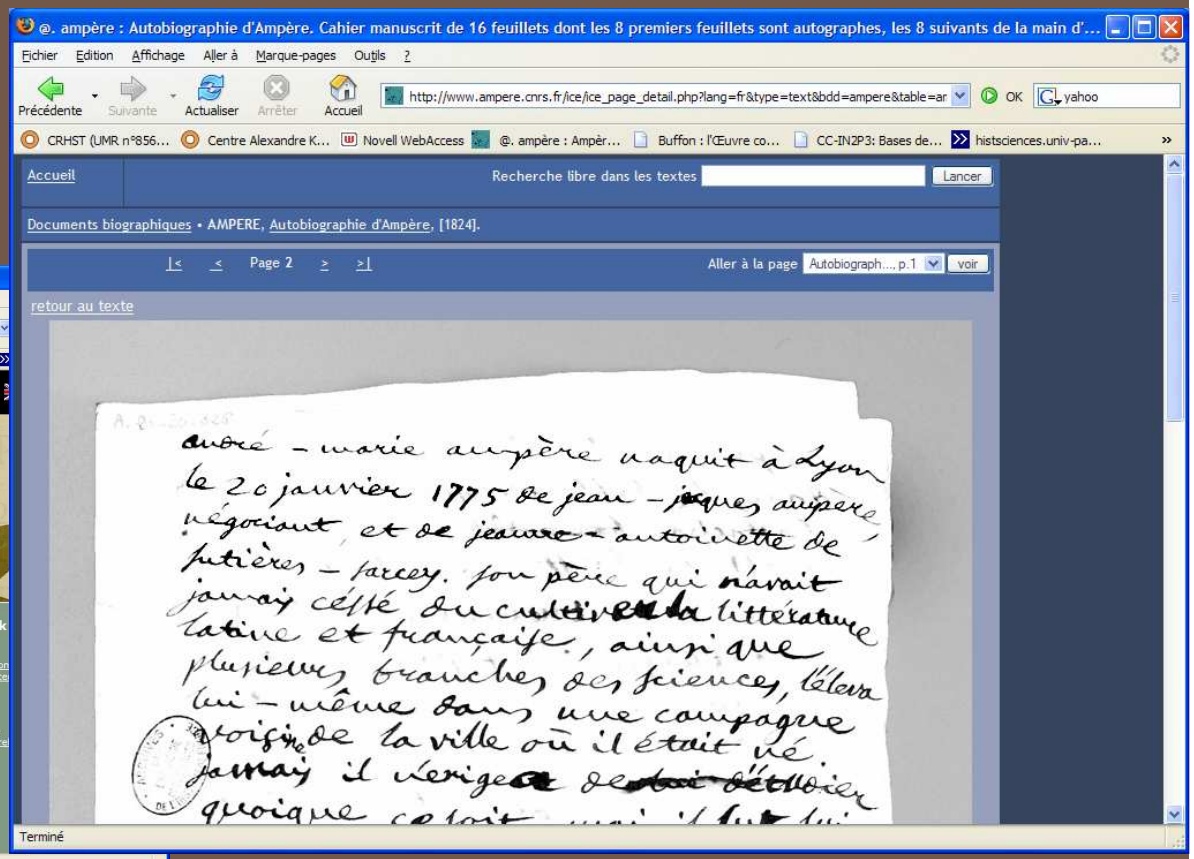
Lucie Secchiaroli, documentaliste contractuelle au pôle HSTL

- 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





- 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
  - …

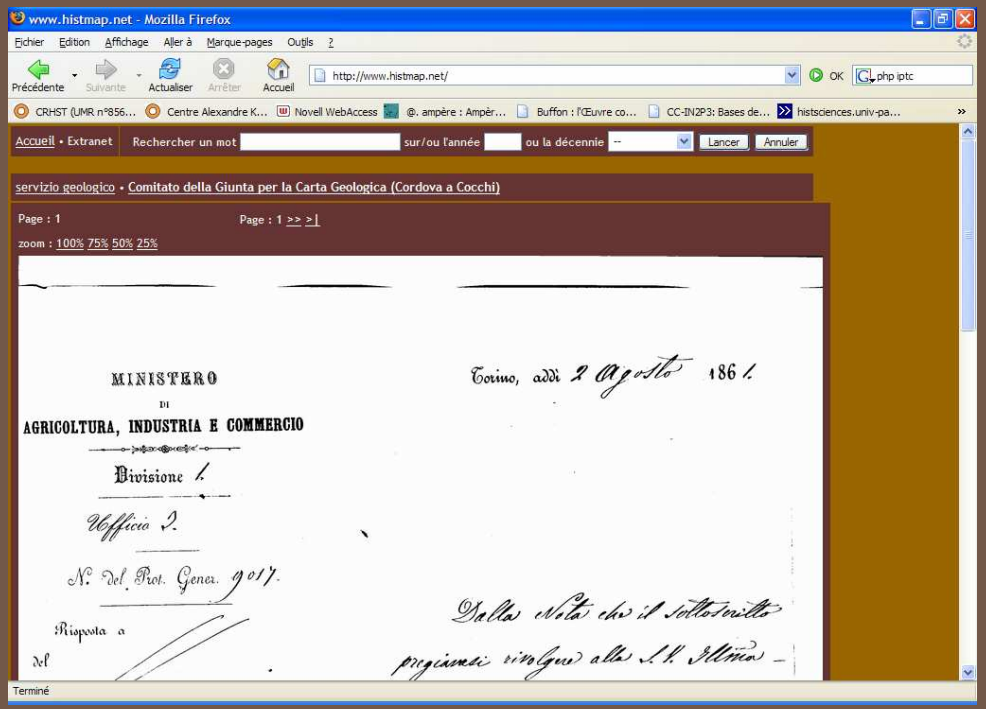
2) Numérisation photographique

- 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**

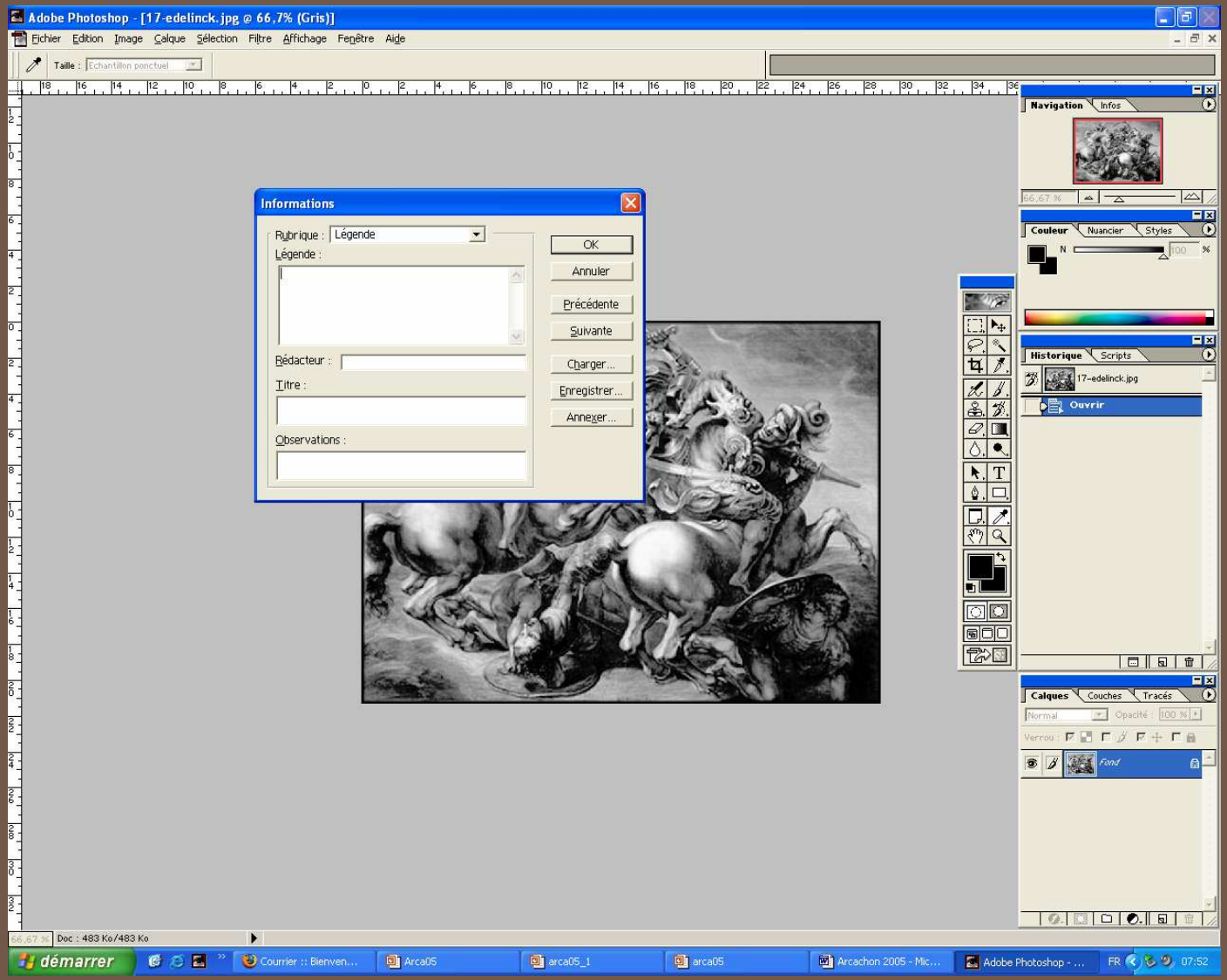
- Adobe Photoshop (par ex.) : permet d'indexer une image selon différentes entrées IPTC :
  - Légende
  - Mots-clés
  - Catégories
  - Crédits
  - Source
  - Copyright

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

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

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

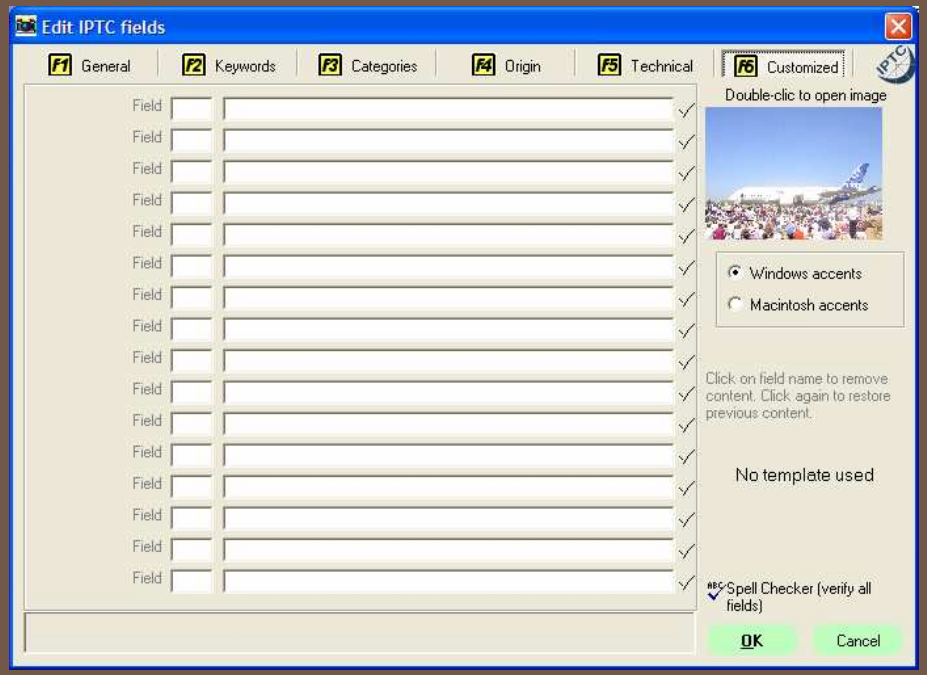
Ref : halshs-00004902



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)

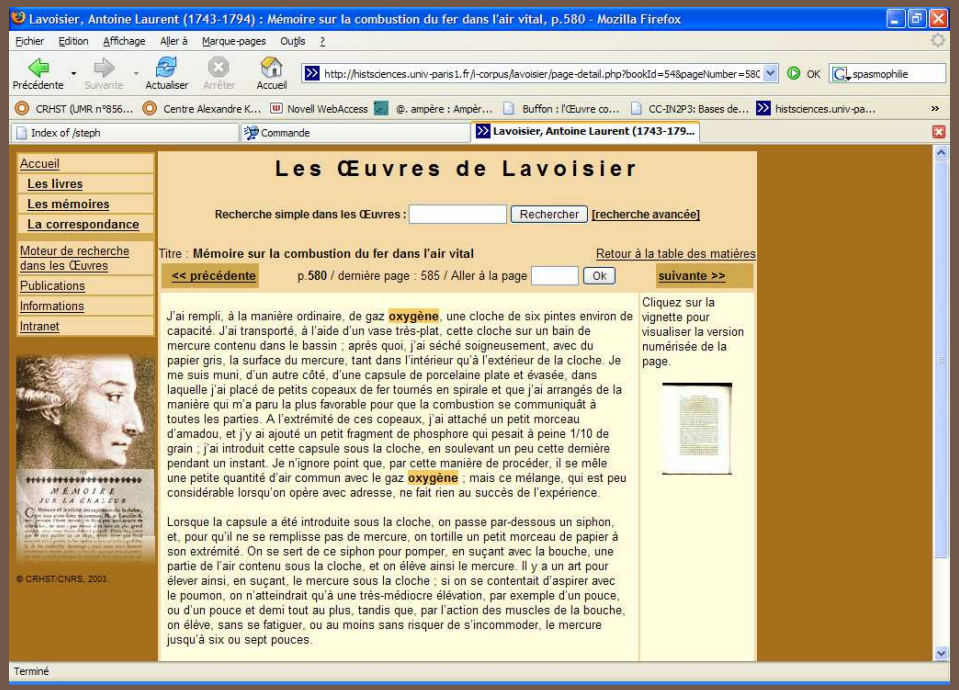


- Récupération de métadonnées IPTC à l'aide d'un script php
  - Utilisation de la GD 2.0
    - iptcparse
    - PHP 3 >= 3.0.6, PHP 4, PHP 5
    - Programmation lourde

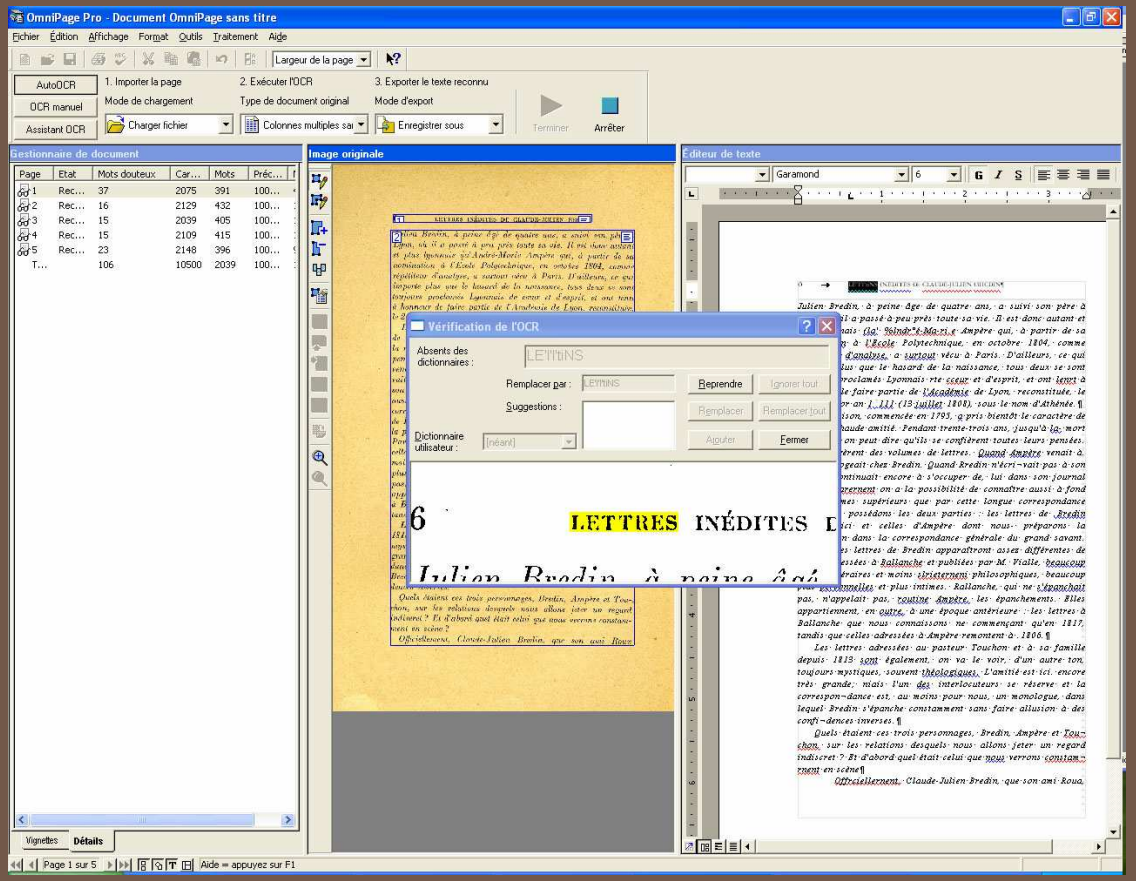
- **Démo…**

3) Traitements documentaires et métadonnés

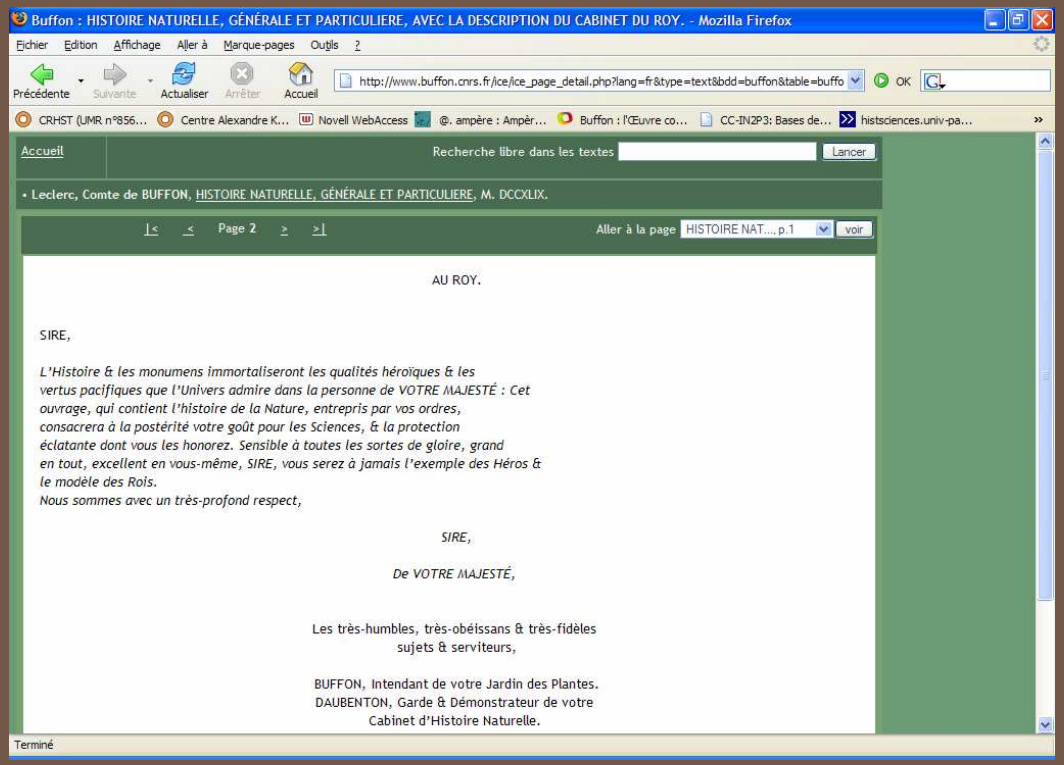
- Numérisation textuelle
  - Saisie des grands textes – WordPro
  - OCR
  - Edit-MathML (développement pôle HSTL)
  - Moteurs de recherche



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



- 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 puisqu'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 !