

**Direction des archives de France**

**Instruction sur la publication électronique  
des instruments de recherche**

**Octobre 2005**

## **Sommaire**

### **1- Principes pour la bonne visibilité d'un site Internet**

- 1.1- Référencement par les moteurs de recherche. Intégration dans des portails**
- 1.2- Ergonomie générale du site internet**

### **2- Modes de consultation des instruments de recherche en ligne**

- 2.1- L'écran d'accueil**
- 2.2- La distinction possible entre état des fonds et instruments de recherche détaillés**
- 2.3- L'interrogation : navigation hiérarchique et recherche par critères**
  - 2.3.1- Navigation hiérarchique dans la structure des fonds**
  - 2.3.2- Recherche par critères**
    - Recherche générale/recherche spécialisée**
    - Recherche simple/recherche avancée**
    - Plein texte/indexation**
  - 2.3.3- Prise en compte des combinaisons de critères**
- 2.4- L'affichage des résultats**
  - 2.4.1- Ordre d'affichage des résultats**
  - 2.4.2- Liste sommaire des unités de description**
  - 2.4.3- Descriptions détaillées**
- 2.5- Les fonctions destinées à l'impression**
- 2.6- L'évaluation des fonctionnalités proposées**
- 2.7- Exemple d'un site récemment refondu**

### **3- Des modes techniques à adapter à l'existant**

- 3.1- Inventaires manuscrits ou anciennement dactylographiés ou imprimés**
- 3.2- Inventaires récemment dactylographiés ou imprimés**
- 3.3- Inventaires produits à partir d'un outil de traitement de texte**
- 3.4- Encodage en XML/EAD**
- 3.5- Notices saisies dans une base de données**
- 3.6- Tableaux synthétiques**

### **Annexe 1- Formats PDF et HTML dynamiques**

### **Annexe 2- Prise en compte, lors de la publication en ligne, des éléments propres aux instruments de recherche**

Des services d'archives de plus en plus nombreux choisissent de proposer un accès à distance sous forme électronique, notamment par le réseau Internet, à leurs instruments de recherche.

Cette démarche, qui rejoint la politique d'accès à l'information encouragée par la direction des archives de France et qui répond à une demande constante des lecteurs<sup>1</sup>, est entreprise de diverses manières :

- numérisation d'instruments de recherche papier pour publication sur le site Internet, au format PDF ou au format HTML (avec possibilité d'accéder directement à une partie de l'inventaire par l'intermédiaire d'un sommaire cliquable) ;
- publication sur le site Internet de données exportées à partir d'un logiciel documentaire (Clara, Arkhéia...) ;
- publication sur le site Internet d'instruments de recherche encodés suivant la DTD EAD (Description archivistique encodée), soit directement (avec un éditeur XML comme XMetaL), soit par conversion rétrospective à partir d'un instrument de recherche papier ou sous un autre format électronique non structuré (traitement de texte par exemple), soit encore par export au format EAD à partir d'un logiciel documentaire.

Afin de faire le point sur les modes de diffusion électronique des instruments de recherche, un groupe de travail<sup>2</sup> a été créé en 2004 à l'initiative de la direction des archives de France (département de l'innovation technologique et de la normalisation).

Les travaux du groupe ont abouti à la rédaction des présentes recommandations pour la publication électronique des instruments de recherche, quels que soient les formats natifs des documents (papier, outil de traitement de texte, PDF, HTML, XML/EAD, bases de données...). Ces recommandations pourront être reprises, en cas de prestation externe, dans un cahier des charges pour une interface d'édition et de consultation d'instruments de recherche sur le Web.

Au-delà des moyens techniques de production et de diffusion à mettre en œuvre, qui doivent être choisis en fonction d'un environnement propre à chaque service, certains grands principes de recherche, de consultation et d'affichage de l'information doivent être respectés, en vue de contribuer, par une ergonomie adaptée et par des similitudes fonctionnelles entre sites, au meilleur service à l'utilisateur.

Les recommandations envisagent également l'amélioration de l'accès aux instruments de

---

<sup>1</sup> Enquête nationale sur les publics des services d'archives (<http://www.culture.gouv.fr/culture/editions/r-devc/dc137.pdf>), enquêtes ponctuelles (Archives départementales de la Haute-Vienne, de la Lozère, du Morbihan...).

<sup>2</sup> Groupe composé de Françoise Banat-Berger (DAF/DITN), André Brochier (CAOM), Robert Chanaud (Archives départementales de la Haute-Vienne), Bruno Galland (CHAN), Henri Hours (Archives départementales du Puy-de-Dôme), Jean Luquet (Archives départementales de la Savoie), François Merlin (CHAN), Christine Nougaret (CHAN), Aude Roelly (Archives départementales de la Côte d'Or), Xavier de la Selle (Archives départementales de l'Aube), Claire Sibille (DAF/DITN), Olivier de Solan (DAF/DITN), Pascal Tanésie (CHAN).

recherche des services d'archives depuis des moteurs de recherche généralistes ou les portails spécialisés.

En effet, de plus en plus d'internautes parviennent sur les sites Internet des services d'archives non plus à la suite d'une démarche volontaire, mais par une requête sur un moteur de recherche, qui les amène sur une page quelconque du site.

Sur le plan fonctionnel, les présentes recommandations constituent un complément du guide *Les instruments de recherche dans les archives* publié par la direction des archives de France en 1999.

Sur le plan technique, elles élargissent le champ de la note DITN/RES/2002/001 du 21 juin 2002 sur la description archivistique informatisée, XML et la DTD EAD.

Par ailleurs, la direction des archives de France a entrepris une étude sur les fonctionnalités de description documentaire offertes par les logiciels de gestion d'archives.

La première partie des présentes recommandations expose quelques règles générales nécessaires à la bonne visibilité et à la facilité d'accès des sites Internet.

La deuxième partie décrit les fonctionnalités attendues de la consultation d'instruments de recherche en ligne.

La troisième partie présente diverses méthodes techniques pour la mise en ligne d'instruments de recherche en fonction de l'état d'origine de ces derniers.

## 1- Principes pour la bonne visibilité d'un site Internet

### 1.1-Référencement par les moteurs de recherche. Intégration dans des portails

De plus en plus, les pratiques des internautes les amènent sur les sites des services d'archives « par hasard », à la suite de requêtes sur des moteurs de recherche, des annuaires de liens ou des portails. Ils se trouvent alors sur une des pages du site sans être passés par la page d'accueil et sans avoir pris connaissance de la structure d'ensemble du site, qu'ils ne découvriront éventuellement que dans un second temps.

Un **moteur de recherche** est un outil de recherche qui référence automatiquement les pages Web se trouvant sur le réseau à l'aide d'un programme appelé robot (exemples : Voila, Google, Altavista...). Chaque moteur a son propre système d'indexation des pages Web : texte intégral sur tout le document, uniquement sur les premiers mots d'un paragraphe, uniquement sur les titres, etc.

Le référencement sur un moteur de recherche ne nécessite pas en principe de démarche particulière. Ceci étant, il est conseillé d'effectuer un certain nombre d'actions si on souhaite réellement optimiser son référencement, sur Google par exemple: par exemple, avoir un grand nombre de liens sur son site, bien choisir le titre des pages, ne pas utiliser de "frame", être inscrit dans les grands annuaires<sup>3</sup>.

Un **annuaire** est une base de données constituée manuellement décrivant une sélection de sites et les indexant à l'aide d'une liste arborescente de sujets (parfois appelés catégories) jouant à la fois le rôle d'une classification et d'une indexation matière (exemple : Yahoo). Le référencement dans un annuaire relève d'une démarche volontaire auprès de l'éditeur.

Un **portail**, dans son sens le plus courant, est un outil de recherche permettant un accès approfondi à des ressources dans un domaine donné. Pour cela, il exploite de manière spécifique les métadonnées des sites Internet qu'il fédère (exemple : portail généalogique Nomina).

Le référencement dans un portail résulte d'une démarche volontaire de la part des fournisseurs de données et suppose une mise en conformité avec le modèle de métadonnées retenu<sup>4</sup>.

---

<sup>3</sup>Pour plus de renseignements, consulter le site <http://www.webrankinfo.com/referencement/conseils/conseils.php>.

<sup>4</sup> Ainsi le protocole *Open Archive Initiative Protocol for Metadata Harvesting* (OAI-PMH) ou « Protocole de collecte de métadonnées de l'Initiative Archives Ouvertes » permet de centraliser les métadonnées référençant diverses ressources mais laisse ces ressources à leur emplacement initial, ce qui rend les temps de réponses très rapides. Il est alors possible d'accéder à ces ressources en architecture distribuée, en rassemblant les métadonnées et en les exploitant pour les besoins spécifiques d'un service : un fournisseur de services moissonne à intervalles réguliers, des métadonnées (au format Dublin Core) stockées dans un entrepôt OAI à partir d'extractions de données issues d'une base de données (les fournisseurs de données). Dans ce cas, bien évidemment, celui qui veut être moissonné doit le faire savoir au fournisseur de services et constituer un entrepôt OAI. C'est sur la base de ce protocole que fonctionne l'application NOMINA sur le portail France-Généalogie.

Plusieurs outils permettent d'évaluer la visibilité d'un site Internet à travers les moteurs de recherche. D'une part, le moteur de recherche Google : il s'agit de l'outil le plus utilisé et il constitue un objectif à lui seul. Les résultats obtenus par Google sont donc révélateurs. D'autre part, il est possible de recourir à un outil comme *www.referencement-2000.com/quiest-referencement.html*, qui analyse, pour un mot clé ou une combinaison de mots clés, le positionnement global du site concerné dans les moteurs francophones. Il faut cependant prendre garde au fait que les résultats obtenus sont une " photographie " de la situation au jour où les tests ont été menés. Ils évoluent sans cesse, en fonction de l'apparition de nouveaux sites et des variations des règles d'évaluation propres à chaque outil de recherche.

Plus généralement, il importe de maintenir une veille permettant de vérifier que le site est bien référencé sur d'autres sites spécialisés (sites des universités de la région, des bibliothèques, des centres et organismes de recherche, sites et portails de généalogie..).

Ceci étant, le référencement par les moteurs de recherche n'est pas forcément une fin en soi et certains services préféreront que l'accès aux internautes se fasse plutôt à partir de la page d'accueil du site et suivant la démarche correspondant à la structure et à l'ergonomie du site.

Dans certains cas, les archivistes pourront choisir que les moteurs de recherche ne puissent pas accéder directement à l'ensemble des données : ainsi, les **Archives départementales du Puy-de-Dôme** ont pris la décision de ne rendre accessible aux moteurs de recherche que les informations portées au niveau de la description des fonds, non dans les instruments de recherche détaillés. Dans ces derniers, en effet, se trouvent de nombreuses informations nominatives, qu'il peut être préférable de ne pas pouvoir retrouver par simple requête sur un moteur.

Il convient d'être conscient des conséquences de la mise en ligne des instruments de recherche, a fortiori s'ils sont indexés par des moteurs de recherche, en raison du changement d'échelle qui s'opère alors et de la présence de nouveaux types de publics, plus nombreux, plus hétérogènes. Cela est encore plus évident lorsque les instruments de recherche permettent l'accès à des images ou à des lots d'images. Cette prise de conscience ne doit en aucun cas dissuader de l'opportunité de la mise en ligne, l'élargissement du lectorat témoignant de l'adaptation des services d'archives aux possibilités offertes par les nouveaux outils.

Ainsi, les **Archives départementales de la Savoie** enregistrent plus de 25 000 consultations en ligne par mois de leurs instruments de recherche (très majoritairement des instruments de recherche anciennement sur support papier, qui ont été numérisés). Aux **Archives départementales de la Mayenne**, les chiffres sont identiques : 25 000 connexions par mois, au cours desquelles plus d'un million d'images sont téléchargées. Certains jours, plus de 1 000 connexions sont enregistrées. Ouverte en permanence, nuit et jour, cette salle de lecture virtuelle reçoit à certaines heures de la journée plus d'une centaine de visiteurs simultanément.

De même, depuis que la base des dossiers de légion d'honneur (LEONORE) du **Centre historique des Archives nationales (CHAN)** est indexée directement par Google, le nombre de demandes de reproduction de documents a augmenté très sensiblement, ce qui pose aujourd'hui la question de l'opportunité de numériser les dossiers eux-mêmes.

## *1.2 –Ergonomie générale du site Internet*

Par ailleurs, un soin particulier doit être apporté à l'ergonomie du site Internet en vue d'en permettre l'accès par le plus grand nombre. Il convient pour cela de se reporter au **référentiel « Accessibilité des services Internet de l'administration française, version 2004 »** élaboré

par l'Agence pour le développement de l'administration électronique (ADAE)<sup>5</sup> ainsi qu'à la recommandation de la direction des Archives de France qui sera publiée, en commentaire de ce référentiel, dans les semaines à venir. Le référentiel repose sur le principe suivant : « mettre le Web et ses services à la disposition de tous les individus quel que soit leur matériel et leur logiciel, leur infrastructure réseau, leur langue maternelle, leur culture, leur localisation géographique, ou leurs aptitudes physiques ou mentales »<sup>6</sup>.

De même, on veillera à ce que le site puisse être ouvert par tous les navigateurs standards existant. Si des restrictions doivent apparaître à cet égard, il convient de les signaler aux internautes.

---

<sup>5</sup> [http://www.adae.gouv.fr/IMG/pdf/referentiel\\_accessibilite.pdf](http://www.adae.gouv.fr/IMG/pdf/referentiel_accessibilite.pdf)

<sup>6</sup> Définition adoptée par le World Wide Web Consortium (W3C)

## 2 – Modes de consultation des instruments de recherche en ligne

### 2.1. L'écran d'accueil

La consultation peut se faire à partir d'un écran d'accueil qui doit proposer une aide à la recherche.

Cet accueil doit comprendre des informations générales sur les ressources en ligne : ce qui est présent, ce qui n'est pas présent, ce qui est présent ailleurs sur d'autres sites (ces informations sont éventuellement à décliner aux niveaux pertinents) et sur les limites éventuelles du cadre de classement (coupures chronologiques non strictes).

Il est indispensable de donner à l'internaute une information synthétique sur le contenu du site et les dernières dates de mises à jour.

Sur l'écran d'accueil peuvent être signalés les instruments de recherche nouvellement mis en ligne et/ou les nouveaux fonds communicables. Un pictogramme particulier peut également être placé à côté des titres des instruments de recherche nouvellement diffusés.

Certains services ont jugé intéressant d'indiquer le nombre total de fonds décrits sur le site, ainsi que le nombre total d'instruments de recherche détaillés mis en ligne (avec la date de la mise en ligne).

### 2.2. La distinction possible entre état des fonds et instruments de recherche détaillés

Il est possible de distinguer deux niveaux de recherche : un niveau de recherche général (état des fonds, état des inventaires, guide) et un niveau de recherche détaillé (répertoires, inventaires).

Cette séparation permet de disposer d'une recherche d'orientation générale, non " polluée " par des résultats de bas niveau et souvent plus exhaustive que la recherche dans les instruments de recherche détaillés. En revanche, cette démarche peut déconcerter le public familier des grands moteurs de recherche généralistes sur Internet, où il n'est jamais nécessaire de poser deux fois sa question pour obtenir une réponse complète.

On peut également proposer de sélectionner le niveau de recherche souhaité.

Ainsi sur le site des **Archives départementales de la Savoie**, l'internaute peut préciser si sa recherche porte uniquement sur les fonds d'archives ou aussi sur les dossiers.



☐ Fonds d'archives    ☒ Dossiers

Rechercher :

Lieu :  GUIDE

Effacer le lieu

Dates : choisissez ...

... un intervalle ...    ... ou une année exacte

☐ Après     ☐ En

☐ Avant

CHERCHER

C'est également le principe retenu sur le site des **Archives départementales du Puy-de-Dôme** :

**Recherche libre**

Tapez les mots significatifs de votre recherche :

☒ un des mots    ☐ tous les mots

Précisez les niveaux de votre recherche :

☒ état des fonds    ☒ inventaires

effacer    rechercher

**Les Archives départementales de la Haute-Vienne** ont même trois niveaux, une sorte de « chemin obligé », auxquels l'utilisateur accède par simple clic : le cadre de classement, un guide décrivant les séries et sous-séries de manière très générale (dates extrêmes, volumes) avec leurs limites (exemple : l'essentiel des archives des sous-préfectures se trouvant en M, dans la partie sur la série Z, on fait un renvoi à la série M) et les instruments de recherche détaillés.

**Le Centre des Archives d'Outre-Mer (CAOM)** a deux niveaux. On accède soit à l'état général des fonds soit aux instruments de recherche détaillés (progressivement mis en ligne en 2005 : bases de données comme l'état civil d'Algérie, liens vers d'autres sites comme Archives Canada-France, inventaires en EAD). L'onglet change de couleur selon que l'on se trouve dans l'état des fonds ou dans les instruments de recherche détaillés.

Il est enfin possible de n'avoir qu'un seul mode de recherche et de ne présenter le niveau de description (sommaire/détaillé) qu'à l'affichage des résultats.

### *2.3- L'interrogation : navigation hiérarchique et recherche par critères*

Différents modes d'accès peuvent être offerts aux internautes : soit par navigation dans des arborescences, à partir du cadre de classement et de l'état de fonds/producteurs, soit par des interrogations sur divers critères (analyse, cote, indexation, etc.). Il est souhaitable que ces deux modes de recherche soient proposés concurremment.

#### 2.3.1. Navigation hiérarchique dans la structure des fonds

La navigation hiérarchique dans la structure des fonds consiste à présenter à l'utilisateur une liste des fonds / producteurs, dans laquelle il pourra cliquer pour obtenir la liste des séries organiques subordonnées au fonds choisi et continuer ainsi son exploration dans les niveaux inférieurs au moyen de liens hypertexte.

Pour faciliter la lecture, il est conseillé de regrouper les fonds par grandes catégories : cadre de classement, thèmes, chronologie, etc.

#### **Accès types de producteurs/fonds**

Le projet des **Archives départementales de la Loire** consiste à organiser la base de données par type de fonds, les internautes étant moins familiarisés avec le cadre de classement réglementaire (présentation par liste des fonds classés par type de fonds puis par ordre alphabétique et par périodes). L'approche « par type de producteur/fonds » combine en fait des regroupements juridiques (notaires), des typologies documentaires (état civil), des périodes, des domaines d'intervention et des types de producteurs (mode intuitif). Ont été ainsi recensés 300 fonds environ, dont 200 privés (en ne prenant pas en compte les communes et les notaires) pour toutes les périodes.

Les **Archives départementales de l'Aube** ont, quant à elles, procédé à une indexation des producteurs pour pallier les insuffisances du cadre de classement.

#### **Accès par le cadre de classement**

Les **Archives départementales de la Haute-Vienne** ont privilégié l'approche par le cadre de classement réglementaire, afin de ne pas trop dérouter les utilisateurs connaissant déjà les archives. Mais afin de rendre le cadre davantage compréhensible du grand public, il a été remanié en opérant un compromis entre regroupements par grandes périodes chronologiques et par typologie : archives antérieures à 1800, archives administratives de 1800 à 1940, archives administratives postérieures à 1940, archives communales et hospitalières, état civil et notaires, archives privées, autres séries (Fi, Mi, NUM, bibliothèque).

1. ARCHIVES AVANT 1800	
▶ Juridictions d'Ancien Régime avant 1790 (série B)	▶ Collège de Limoges avant 1790 (série D)
▶ Administrations provinciales (série C)	▶ Fonds ecclésiastiques divers (série I)
▶ Clergé séculier avant 1790 (série G)	▶ Documents de l'époque révolutionnaire 1790-1800 (Série L)
▶ Clergé régulier avant 1790 (série H)	
2. ARCHIVES DE 1800 A 1940	
▶ Administration générale et économie, 1800-1940 (série M)	▶ Enseignement, culture, sports, 1800-1940 (Série T)
▶ Administration départementale, 1800-1940 (série N)	▶ Justice de 1800 à 1940 (série U)
▶ Administration communale, 1800-1940 (série O)	▶ Cultes de 1800 à 1940 (série V)
▶ Finances, cadastre, poste, 1800-1940 (série P)	▶ Assistance et prévoyance sociales, 1800-1940 (série X)
▶ Domaine, enregistrement, hypothèques (série Q)	▶ Établissements de répression de 1800 à 1940 (série Y)
▶ Affaires militaires, guerre, 1800-1940 (série R)	▶ Sous-préfectures de 1800 à 1940 (série Z)
▶ Travaux publics et transports, 1800-1940 (série S)	
3. ARCHIVES DEPUIS 1940	
▶ Archives dites contemporaines (Série W)	
4. ARCHIVES COMMUNALES ET HOSPITALIÈRES	
▶ Archives communales	▶ Archives hospitalières
5. ÉTAT CIVIL, NOTAIRES	
▶ Microfilms d'état civil	▶ Notaires (4 E)
6. ARCHIVES PRIVÉES	
▶ Archives privées (séries F et J)	
7. AUTRES SÉRIES	

Pour le **CHAN**, si l'entrée par le cadre de classement réglementaire reste incontournable, malgré ses défauts (il peut être non explicite et trop générique ou au contraire trop précis ; les coupures chronologiques ne sont pas strictes), une approche conjointe est également proposée, qui consiste à opérer des regroupements par type de producteur ou par type de fonds. L'accès aux fonds privés du CHAN peut s'effectuer grâce à une liste alphabétique des intitulés de fonds (producteurs et intitulés de fonds se rejoignent). L'accès aux fonds d'Ancien Régime privilégie une reconstitution intellectuelle des fonds éclatés entre les séries du cadre de classement. On arrive ainsi à une soixantaine de rubriques, présentées non pas par ordre alphabétique mais dans une liste méthodique, en distinguant les types de producteurs.

Pour les services d'archives départementales, une double entrée, qui serait d'une part le cadre de classement et d'autre part une entrée par fonds/producteurs, risque d'être redondante et a contrario, de troubler l'internaute. Il convient de se positionner en privilégiant soit le cadre de classement (remanié si nécessaire) soit l'approche fonds/producteurs.

### 2.3.2. Recherche par critères

La recherche par critères est la possibilité offerte à l'utilisateur de s'abstraire de la structure par fonds pour rechercher toutes les unités de description contenant certains mots ou termes d'indexation.

Ce mode de recherche se matérialise souvent par un formulaire d'interrogation avec des champs à remplir.

#### *Recherche générale / recherche spécialisée*

Il est possible de prévoir un formulaire de recherche générale et des formulaires de recherche spécifiques par types de documents (archives / bibliothèque / images ; état civil / notaires / cadastre, etc.).

Il est recommandé de prévoir au minimum un formulaire de recherche général, portant sur

l'ensemble des unités documentaires.

L'accès doit pouvoir se faire directement (d'emblée, depuis l'écran d'accueil) pour certains types de fonds particulièrement consultés : pour les archives départementales, accès particuliers pour l'état civil (registres paroissiaux et d'état civil / tables décennales), fonds notariaux, fonds du cadastre, qui à eux seuls contentent plus des deux tiers des lecteurs ou internautes ; ou des fonds emblématiques (ainsi le fonds de Clairvaux pour les Archives départementales de l'Aube, les dossiers de légion d'honneur pour le CHAN ou encore la cartothèque...).

### *Recherche simple / recherche avancée*

Il est conseillé d'offrir aux utilisateurs un formulaire de recherche général très simple, limité à quelques champs, voire à un seul champ (comme dans un moteur de recherche).

Pour l'accès par un formulaire de recherche, il convient de prévoir un paramétrage le plus souple possible des champs d'interrogation (dates/fourchettes de dates, lieux, personnes, objet, type de document, cote, etc.).

Comme cela a été indiqué au 2.2, soit le formulaire de recherche porte sur l'ensemble des descriptions, quel que soit le niveau de description, soit l'internaute peut être invité à préciser si sa recherche porte sur l'état des fonds ou sur les instruments détaillés (par défaut, elle portera sur l'ensemble).

Pour autant, le type d'instrument de recherche (inventaire, répertoire numérique...) ne semble par pertinent comme champ de recherche, les internautes ne connaissant généralement pas la typologie des instruments de recherche.

Si des images numérisées sont présentes sur le site (images ou lots d'images), il faut laisser la possibilité à l'internaute, s'il le souhaite, de restreindre sa recherche aux éléments pour lesquels il y a une image ou un lot d'images associées, et lui permettre d'accéder directement à la description donnant accès à l'image, sans passer par tous les niveaux supérieurs de la hiérarchie.

Un formulaire de recherche avancée, comportant plus de champs et destiné à des utilisateurs plus avertis, peut être proposé dans un second temps, pour tout ou partie des instruments de recherche.

La recherche avancée est surtout utile au service d'archives, pour mieux renseigner le public.

### *Plein texte / indexation*

La possibilité d'interroger en texte libre (comme dans un moteur de recherche) est un minimum

La présence de champs d'interrogation portant sur des indexations implique que les instruments de recherche publiés aient été indexés de manière homogène. Ainsi, il ne serait pas pertinent de proposer un champ " type de document ", si, par ailleurs, on a saisi très rarement la typologie.

En tout état de cause, il est indispensable, si l'ensemble des ressources publiées sur le site ne

sont pas indexées, de préciser pour l'internaute quelles sont les limites de l'indexation et sur quelles ressources elle porte.

Les **Archives départementales de la Haute-Vienne** ont construit un parcours de « pistes de recherche » permettant aux lecteurs d'accéder aux instruments de recherche, à partir de mots-clés correspondant à des thèmes récurrents (exemple : « enseignement »). Cent mots-clés avaient été prévus à l'origine, mais un nombre plus restreint suffit (une vingtaine), afin précisément de ne pas « perdre » l'internaute dans un grand nombre de concepts : il ne s'agit pas, dans ce cas, de faire une indexation-matière mais d'avoir une approche pragmatique qui corresponde aux demandes des lecteurs. Chacune de ces fiches indique au lecteur dans quelles séries, sous-séries ou fonds se trouvent les documents pertinents. Des liens lui permettent d'y accéder directement.

**L'intérêt de l'indexation par rapport au plein texte** doit être soigneusement pesé. Si l'indexation est trop générale (ne portant que sur des descriptions de haut niveau), les résultats risquent d'être trop pauvres. Son positionnement dès l'écran d'accueil suppose d'avoir tout indexé ou au moins des ensembles homogènes (toutes les archives contemporaines, toutes les archives privées).

On peut par ailleurs observer que, dès lors qu'un instrument de recherche est fortement structuré et les analyses suffisamment précises, l'indexation est moins nécessaire. Si, en revanche, l'index apporte une valeur ajoutée aux analyses, il sera intéressant de le récupérer<sup>7</sup>.

Une perspective prometteuse sera l'utilisation de thesaurus métier (évolutions du Web sémantique) permettant de rechercher automatiquement à partir d'un terme, les termes associés, voisins, appartenant au même univers lexical... et ce afin d'éviter des saisies d'entrées d'index normalisées beaucoup trop lourdes. Dans l'état actuel des choses, il semble que si des efforts en matière d'indexation et de base de connaissances peuvent être faits, ils doivent porter en priorité sur les lieux : indexation par noms de lieux, s'appuyant sur des dictionnaires de noms de lieux (découpages administratifs, géographie historique) permettant automatiquement d'améliorer les résultats de la recherche en créant des liens entre un lieu et ses différents rattachements. De plus en plus, de nouvelles interfaces permettront à l'internaute de formuler sa requête à partir d'une carte géographique.

Le **CAOM** a par exemple présenté un seul champ indexé : le nom de lieu (les noms de lieux sont présentés sous forme de liste structurée).

---

<sup>7</sup> Ainsi, dans le cas d'instruments de recherche anciens, il est arrivé que l'identification des noms de lieux se soit faite uniquement dans l'index, voire que certains termes qui ne figurent pas dans les analyses soient retenus dans l'index.

INSTRUMENTS DE RECHERCHE EN LIGNE

caom.aix@culture.gouv.fr

Etat général des fonds

Inventaires détaillés

Etat général des fonds

Recherche par le sommaire

AFFICHER LE SOMMAIRE

Recherche par critères

Plein

texte :

?

Lieu :

Année :

ou

entre

et

LANCER LA RECHERCHE

EFFACER

Contexte de la recherche

ÉTAT DES FONDS

FONDS MINISTÉRIELS

FONDS TERRITORIAUX

FONDS DÉPOSÉS

ARCHIVES PRIVÉES

ICONOTHÈQUE

CARTOTHÈQUE

BIBLIOTHÈQUE

DERNIÈRE MISE À JOUR : 23/11/2004

Le thesaurus géographique utilisé par les **Archives départementales de la Savoie** est constitué des niveaux hiérarchiques suivants : département, arrondissement, canton, commune, lieu-dit. Il associe les termes autorisés (dénomination actuelle) et les termes synonymes (dénominations anciennes). Un historique permet de préciser les dates, les re-découpages administratifs. Le moteur de recherche utilise les termes « vedettes » mais également les synonymes. Il possède une table automatique d'équivalences (St = Saint, Saint-Jean = Saint Jean...). La recherche est exécutée sur le champ « Lieux » mais aussi sur le texte intégral des notices. La complexité de la navigation a conduit à ne pas autoriser la frappe directe du terme lieu recherché, mais la sélection dans une liste que le système « auto-poste » dans le champ de recherche.

Si le champ de recherche correspond à une liste fermée, il est conseillé d'assister l'utilisateur, au moyen de listes déroulantes ou de thesaurus hiérarchisés par exemple. La liste déroulante doit présenter seulement les termes utilisés réellement, avec si possible le nombre d'occurrences concernées. Dans le cas d'un thesaurus, l'outil doit permettre d'afficher également les termes non utilisés pour l'indexation (cas du terme générique qui couvrirait des termes spécifiques qui, eux, ont été utilisés).

### *Prise en compte des combinaisons de critères*

Si plusieurs mots peuvent être saisis dans un champ, l'opérateur booléen ET doit être

implicite.

Le moteur de recherche doit permettre, quand on demande un intervalle de dates, de sélectionner toutes les notices qui entrent même partiellement dans cet intervalle de dates.

Ainsi, pour une recherche du type « 1900-1950 », les notices avec les dates 1930-1970, doivent pouvoir être retrouvées.

La recherche plein texte peut se faire à l'aide d'opérateurs booléens et permettre les fonctions avancées d'un moteur de recherche documentaire : gérer les masculins/féminins, les singuliers/pluriels, les espaces blancs, les traits d'union, les apostrophes....

La recherche doit dans tous les cas tenir compte de la hiérarchie des descriptions : la recherche au niveau inférieur doit récupérer toutes les informations saisies dans les niveaux supérieurs.

Par exemple, dans le fonds « Pré-inventaire des richesses artistiques de la Loire », il existe une série par commune (le champ lieu = « Lupé » par exemple) ; dans chacune de ces séries, il existe un dossier par objet ou monument.

Pré-inventaires des richesses artistiques de la Loire  
Série communale...  
Ville de ...  
Ville de ...  
Ville de Lupé  
Château  
Encensoir  
...

Si la recherche est « Lupé » ET « Encensoir », la réponse doit être le dossier « encensoir » en question, même si le terme Lupé ne figure pas dans la description de ce dossier : le résultat au niveau le plus bas (encensoir) doit en effet récupérer les informations des niveaux supérieurs (Lupé).

Par contre, si la recherche est « Lupé », la réponse doit seulement être la série, et non les dossiers contenus à l'intérieur.

## *2.4. L'affichage des résultats*

Comme cela a été rappelé dans le 1.1, l'internaute peut se retrouver, à la suite d'une requête à partir d'un moteur de recherche, sur n'importe quelle page du site. Cela entraîne un certain nombre de conséquences : ainsi, chaque page doit comporter la référence du service d'archives, de la série concernée, avec une possibilité d'accès direct à l'accueil et/ou à la recherche dans les instruments de recherche. C'est également la raison pour laquelle les documents PDF publiés devront comporter des titres courants.

Les conseils qui suivent s'appliquent au cas des instruments de recherche électroniques structurés (base de données, documents au format XML).

### 2.4.1 Ordre d'affichage des résultats

Le principe de rendre compte de la hiérarchie des niveaux de description doit également être

central pour l'affichage des résultats.

Il est indispensable que les réponses puissent être affichées dans leur contexte de description : affichage des niveaux supérieurs et possibilité de consulter les éléments subordonnés s'il y en a.

Les réponses qui s'affichent peuvent être de niveaux très différents (du fonds à la pièce). Elles doivent par conséquent s'afficher dans un ordre rendant compte de cette hiérarchie : classement par types de fonds/périodes, par fonds, par ordre des séries.

#### 2.4.2. Liste sommaire des unités de description

On doit pouvoir obtenir dans un premier temps la liste sommaire des unités de description répondant à la question. Si l'internaute le souhaite, il doit ensuite pouvoir sélectionner un élément de réponse pour obtenir une description détaillée et la liste des éléments subordonnés.

Plusieurs présentations sont envisageables : présentation hiérarchique avec des indentations sur un même écran (réponse cliquable et accompagnée d'un signe + par exemple), présentation hiérarchique en fenêtres successives, etc.

Il convient de tester des maquettes pour se rendre compte de l'affichage qui sera le plus familier et simple d'utilisation à l'internaute, avec le double souci de ne pas surcharger l'écran et, tout en conservant l'affichage du contexte, de ne pas compliquer exagérément la présentation, afin de ne pas perdre l'internaute dans un enchevêtrement d'informations finalement peu lisibles.

Il est recommandé que la liste de réponses indique le nombre d'occurrences qu'elle comporte.

Les notices sommaires pourraient comporter les indications suivantes :

- éléments ISAD(G) :

- niveau de description
- référence : dans certains cas, il ne sera pas pertinent de la retenir, notamment pour les instruments de recherche méthodiques, ainsi que pour les versements contemporains. Pour un article, il est conseillé de faire apparaître la cote de communication, exactement telle que devront l'indiquer les lecteurs pour obtenir la communication du document.
- intitulé/analyse suivant le niveau de description
- dates
- éventuellement, liens vers des images numérisées.

Il est recommandé de mettre en surbrillance les termes qui correspondent à la question posée.

On ne doit pas empêcher a priori l'internaute d'afficher autant de réponses qu'il le souhaite, même lorsque le nombre de réponses semble très élevé. Si la surcharge du moteur ainsi provoquée est susceptible de pénaliser d'autres utilisateurs, on peut prévoir l'affichage d'un message invitant l'internaute à affiner sa recherche. Une autre possibilité consistera à offrir la possibilité de télécharger le résultat, par exemple dans un format compressé. Dans tous les



cas, les résultats doivent s'afficher par lots de 20 ou 30 réponses au maximum.

A tout moment, un affinement de la recherche doit être possible : l'internaute est alors renvoyé au dernier formulaire rempli et peut compléter sa recherche.

Pour passer des notices sommaires aux notices détaillées, l'internaute doit pouvoir cliquer sur la réponse qui l'intéresse. S'il souhaite afficher plusieurs réponses, il doit pouvoir les cocher ; une fonction doit également lui permettre de sélectionner ou désélectionner en une seule opération l'ensemble des réponses.

#### 2.4.3. Descriptions détaillées

Dans la description détaillée, les vingt-six éléments prévus dans la norme ISAD(G) sont susceptibles d'être affichés, quel que soit le niveau de description.

#### *2.5. Les fonctions destinées à l'impression*

Dans le cas des instruments de recherche électroniques structurés (base de données, XML), deux niveaux de lecture peuvent être prévus : l'un adapté à l'affichage à l'écran, l'autre destiné à l'impression. Ainsi, dans les instruments de recherche en ligne du CAOM, ont été rédigés pour affichage à l'écran des textes de présentation courts, ne dépassant pas deux pages d'écran et renvoyant à l'introduction complète, ou des annexes, renvoyant dans un format imprimable type PDF à l'ensemble de l'instrument de recherche.

Il faut donner la possibilité à l'utilisateur d'obtenir une version imprimable de l'instrument de recherche (document PDF généré à l'avance à partir du document XML, enrichi d'éléments variés, d'annexes, etc.), ou à défaut des résultats de sa requête dans son contexte<sup>8</sup>.

#### *2.6- L'évaluation des fonctionnalités proposées*

Il convient d'évaluer, auprès du public, la pertinence des fonctionnalités proposées pour la consultation en ligne des instruments de recherche.

Pour l'évaluation d'un site déjà existant, la technique du questionnaire (enquête directe ou en ligne) semble la plus évidente, mais elle peut parfois donner des résultats décevants.

Il est possible d'associer des représentants du public à des séances de tests.

Une autre façon d'évaluer le site est d'analyser les messages électroniques envoyés par les internautes. Les questions posées permettent dans certains cas de se rendre compte que certaines informations, pourtant présentes sur le site, n'ont pas été repérées par les internautes et qu'elles doivent donc être mieux mises en évidence.

Les statistiques des recherches infructueuses dans le moteur de recherche peuvent également être très instructives. Elles peuvent permettre d'analyser des difficultés et éventuellement de corriger certaines imperfections du site ou des outils de recherche et de consultation. Plus

---

<sup>8</sup> Plusieurs projets en cours visent à permettre l'élaboration de feuilles de style adaptées à chaque grand type d'instrument de recherche.

généralement, il est recommandé de prévoir un certain nombre de statistiques sur le site, permettant de calculer l'impact de la mise en ligne d'instruments de recherche (et de documents d'archives) sur le nombre de connexions, voire d'analyser plus finement ce qui a été consulté (types de fonds, de séries, nombre de dossiers, de pièces), type de champs les plus fréquemment interrogés, nombre de recherches simples/avancées. Ces renseignements devraient permettre de faire évoluer les publications électroniques au plus près des besoins.

Il est conseillé en tout état de cause de prévoir, dans l'écran d'accueil, une adresse électronique permettant à l'internaute de poser des questions ; le service d'archives doit alors s'engager à répondre systématiquement et dans des délais raisonnables.

## *2.7 Deux exemples de sites récents (les Archives départementales du Puy-de-Dôme<sup>9</sup>, CAOM)*

### **Archives départementales du Puy-de-Dôme**

En amont, l'application utilise les exports au format XML conformes à la DTD EAD 1.0 générés par le logiciel d'archives Arkhéia.

L'internaute accède à deux types d'instruments de recherche : état général des fonds (1091 fonds décrits à ce jour) et inventaires détaillés (56 en ligne). Le rappel dans les bandeaux des termes "Fonds" et "Inventaire" lui permet de se repérer dans les niveaux de description. La publication des instruments de recherche s'effectuera progressivement au fur et à mesure de l'état d'avancement des classements et des reprises d'inventaires anciens sous Arkhéia.

Tout comme sur l'ancien site, les internautes effectuant des recherches sur l'histoire des familles peuvent accéder directement à l'inventaire des archives de l'état civil et à l'état des fonds des notaires.

Chaque fiche de description de l'état des fonds (symbolisé par une loupe sur le site) comporte plusieurs champs : identification, cotes, dates extrêmes, métrage, présentation du producteur, modalités d'entrée d'accès et de reproduction, présentation des instruments de recherche, sources complémentaires extérieures et dans le service, bibliographie et indexation matière, géographique et personne. Ces fiches peuvent être renseignées à minima pour des fonds non classés, dans d'autres cas, un lien direct permet d'accéder à l'inventaire détaillé (symbolisé par une page d'écriture) qui s'affiche dans une nouvelle fenêtre.

L'affichage de l'inventaire détaillé en pleine page permet une consultation plus confortable. Le sommaire affiché dans la partie gauche peut être déployé à la demande. L'affichage des intitulés et cotes des articles dépendant d'une même hiérarchie s'effectue dans la partie droite de l'écran. La notice complète d'un article peut être déployée en cliquant sur la cote lorsqu'elle apparaît en caractères bleu gras : ainsi s'affichent des informations supplémentaires sur le contenu, des données techniques, des modalités d'accès... De même, des informations complémentaires concernant des groupes d'articles peuvent être affichées à la demande lorsqu'un intitulé est suivi du symbole >>.

L'internaute dispose de plusieurs possibilités de recherche :

- une recherche arborescente dans l'état des fonds présenté selon le cadre de classement des archives. Il est possible également d'afficher la liste de tous les inventaires en ligne et d'y accéder directement ;

- une recherche guidée limitée à une interrogation sur le champ d'indexation matières de l'état des fonds. Le vocabulaire retenu est celui qui a été mis au point par la direction des archives de France pour l'élaboration de la « base d'orientation et de recherche dans les archives » (BORA). Une liste présentée sous forme alphabétique ou hiérarchique permet de sélectionner les termes et de valider la recherche. Le résultat de la recherche s'affiche dans la zone centrale en suivant le plan de classement (le nombre d'occurrences étant reporté entre parenthèses) ;

---

<sup>9</sup> Le site vient de faire l'objet d'une refonte.

- une recherche libre en plein texte. Elle peut se faire à l'aide d'opérateurs booléens (ou = « un des mots », et = « tous les mots »). La recherche s'effectue sur l'état des fonds, les inventaires ou les deux à la fois. Le résultat s'affiche comme précédemment en suivant le plan de classement, le nombre d'occurrences étant reporté derrière chaque symbole d'affichage (exemples : l'intitulé du fonds est suivi du symbole de la loupe avec (1) : le mot recherché est au moins présent une fois dans la fiche de description du fonds, l'intitulé est suivi du symbole page écrite avec (12) : le mot recherché est présent au moins une fois dans 12 intitulés d'articles de l'inventaire). Au niveau de l'inventaire l'affichage des occurrences (mots surlignés) se fait en tenant compte du contexte. Le moteur de recherche gère les singuliers/pluriels, les traits d'union.

### **CAOM (application Instruments de recherche en ligne ou IREL)**

L'accent a été mis dans un premier temps sur l'état des fonds (le nombre d'inventaires détaillés publiés est encore limité). Les résultats s'affichent de façon hiérarchisée, en fonction du cadre de classement. Le choix d'une seule fenêtre pour le résultat et pour le contexte des unités documentaires pertinentes a été retenu. La table des matières a été supprimée dans la mesure où, dans la partie gauche de l'écran, le contexte de la recherche permet de remonter dans les niveaux de description supérieurs. Existe également la possibilité de naviguer dans l'arborescence de l'état général sans passer par le formulaire de recherche.

L'état des fonds comprend un niveau « minimal » par fonds (nom du fonds, ses cotes, ses dates) , qui peut aller jusqu'à une description à l'unité documentaire dans certains cas (description de groupes d'articles ou d'articles individuels). Dans certains cas (présence d'un inventaire détaillé), l'internaute peut cliquer sur celui-ci : il se trouve alors dans un autre contexte (couleurs différentes) qui lui permet de se repérer.

La date de la dernière mise à jour est visible sur l'écran d'accueil.

Quand on descend dans les niveaux de description, on quitte le fragment de la réponse. L'historique des recherches a été abandonné volontairement, c'est le contexte de la recherche qui a été considéré comme important. Un lien permet de remonter à la liste des résultats, on peut aussi relancer très facilement la recherche. Il n'y a par contre pas de possibilité de stocker les résultats dans un « caddie » ou dans un « panier ». Il est possible de naviguer vers le résultat précédent ou suivant.

### 3- Des choix techniques à adapter à l'existant

Il existe plusieurs moyens techniques pour assurer la publication en ligne d'instruments de recherche.

Chaque service devra choisir le mode le plus adapté selon l'état initial des instruments de recherche et selon les moyens dont il dispose, en examinant les aspects positifs et négatifs de chaque solution. Très souvent, le choix ne sera pas univoque : par exemple, on pourra d'une part publier en ligne après une simple numérisation tous les instruments de recherche qui existent sur support papier, en salle de lecture (manuscrits, dactylographiés), tels qu'ils sont (à l'exception de quelques aménagements mineurs), et, d'autre part, des données extraites des bases de données utilisées dans le service (par exemple pour les fonds contemporains). Toutes les combinaisons peuvent être envisagées. Une telle approche permet de mettre en ligne plus rapidement un plus grand nombre de ressources, sans attendre par exemple une ressaisie de tout l'arriéré dans une seule base. Le site Internet devra alors être conçu de telle manière que les internautes appréhendent bien les différents accès proposés. Certains services, au contraire, préféreront offrir un accès homogène à l'ensemble de leurs instruments de recherche (par exemple, présenter tous les instruments de recherche en PDF au travers d'une navigation par le cadre de classement ou bien publier tous les inventaires au travers d'une seule base de données). D'autres choisiront plusieurs types d'accès.

#### *3.1- Inventaires manuscrits ou anciennement dactylographiés ou imprimés*

Si on dispose d'inventaires anciens manuscrits, dactylographiés ou imprimés de mauvaise qualité, il sera possible de générer un fichier PDF image (en utilisant un des logiciels de la suite Acrobat) et de le publier tel quel. On pourra décider de s'arrêter à cette étape, l'avantage étant bien évidemment la facilité et la rapidité de l'opération, ainsi que la facilité pour l'internaute d'imprimer l'inventaire sur un support papier. En revanche, il sera impossible d'effectuer des recherches dans cet inventaire et celui-ci ne sera pas accessible par les moteurs de recherche<sup>10</sup>. En outre, le fichier sera volumineux et long à télécharger<sup>11</sup>.

Si on décide par la suite d'aller plus loin, il sera nécessaire de ressaisir l'inventaire pour pouvoir générer un fichier PDF texte ou un fichier au format HTML ou encore un fichier au format XML EAD, ou pour pouvoir importer les données issues de ce fichier dans une base de données. La ressaisie est une opération lourde à réaliser, mais elle permet d'obtenir des résultats appréciables pour la recherche (voir les parties sur XML/EAD et les bases de données). La rétroconversion au format XML/EAD sera d'autant plus facile que l'inventaire initial sera conforme à la norme de description ISAD/G<sup>12</sup>.

#### *3.2- Inventaires récemment dactylographiés ou imprimés*

Si on dispose d'inventaires récemment dactylographiés ou imprimés, il sera possible de procéder à l'océrisation<sup>13</sup> pour obtenir par exemple un fichier PDF. Cette océrisation permettra de générer, à côté du fichier image, un fichier texte, dont on pourra choisir de

---

<sup>10</sup> Sauf à élaborer un sommaire en HTML qui pourra être alors indexé par les moteurs de recherche.

<sup>11</sup> Voir annexe technique sur le format PDF.

<sup>12</sup> En effet, dans ce cas, sur la base de concordances entre les différentes parties de l'inventaire et les balises EAD données par l'archiviste, le prestataire pourra réaliser une rétroconversion relativement automatiquement.

<sup>13</sup> Ou OCR (reconnaissance optique de caractères).

corriger ou non les inévitables erreurs de reconnaissance automatique. Les avantages de cette solution par rapport au format image sont nombreux : le fichier PDF texte sera directement accessible par les moteurs de recherche, la recherche en texte intégral plus aisée, le poids du fichier bien moins important. L'opération est par ailleurs relativement facile et économiquement intéressante à réaliser. L'impression papier reste toujours aussi aisée. En outre, les mises à jour sont simples à effectuer. En revanche, le mode de recherche ne sera pas structuré (impossible de rechercher sur un champ précis et de restituer le résultat dans son contexte de description). Par ailleurs, la gestion des liens peut se révéler problématique<sup>14</sup>.

A partir de l'océrisation, on peut également choisir de générer non pas un fichier PDF texte, mais un fichier au format HTML<sup>15</sup>, format reconnu par les moteurs de recherche, qui permet une navigation particulièrement aisée avec l'utilisation de liens hypertexte. La publication dans ce cas est facile et relativement rapide à mettre en œuvre, sans avoir à passer par un outil supplémentaire<sup>16</sup>. En revanche, comme avec le PDF texte, la recherche structurée ne sera pas possible. Pour l'affichage à l'écran, on prendra soin de découper l'inventaire en autant de pages correspondant à la taille de l'écran. Si la page HTML dépasse la hauteur d'un écran, on prévoira des liens pour le retour en « haut de page ». Par ailleurs, il n'existe pas de solution pour commander en une seule fois l'impression de plusieurs pages HTML : si l'instrument de recherche comporte plusieurs pages, il convient donc d'offrir à l'utilisateur une option de téléchargement de tout l'instrument de recherche, faute de quoi il devrait imprimer page à page<sup>17</sup>.

### *3.3- Inventaires produits à partir d'un outil de traitement de texte*

Si on dispose des instruments de recherche sous la forme de fichiers produits par traitement de texte, il est possible de produire directement un fichier au format PDF texte ou, moyennant un redécoupage et l'établissement de lien, des fichiers HTML (voir 3.1.2). On peut également opter pour une conversion en XML/EAD ou pour un import dans une base de données (voir ci-dessous).

### *3.4- Encodage en XML/EAD*

Un instrument de recherche au format XML/EAD peut être obtenu de plusieurs manières : soit par un encodage direct en passant par un éditeur XML (du type XMetaL ou XML Spy), soit par un export à partir d'un logiciel documentaire (par exemple Arkhéia ou Gaïa), soit par une conversion à partir d'un inventaire en format texte non structuré (PDF texte, éditeur de traitement de texte).

Dès lors, les options pour la publication sont multiples : production d'un fichier PDF ou

---

<sup>14</sup> Voir annexe sur le format PDF.

<sup>15</sup> Concernant ce format, on parle de pages statiques : les pages existent en tant que telles (elles préexistent aux requêtes).

<sup>16</sup> Si on produit du HTML natif, il est recommandé d'utiliser des feuilles de style plutôt que de marquer les polices et mise en forme dans le texte. En effet, si on change la maquette des pages, il suffira de changer les feuilles de style. Il est recommandé dans ce cas d'alimenter les pages avec des titres et des mots-clés.

<sup>17</sup> Il existe enfin d'autres solutions, comme le téléchargement dans un autre format comme le format RTF qui présente l'avantage d'être accessible à l'ensemble des internautes ou encore le téléchargement par réseau FTP qui peut être tout à fait adapté pour de gros volumes.

RTF<sup>18</sup>, production de fichiers HTML<sup>19</sup>, production de fichiers HTML dynamiques<sup>20</sup>. Seule la dernière option permet d'avoir les avantages liés aux pages dynamiques : facilité de mise à jour des informations, ajustement beaucoup plus fin de l'information fournie à l'internaute en fonction de sa question. Elle permet également de conserver tous les avantages de la recherche structurée et contextualisée à partir d'un fichier XML/EAD : richesse de la recherche et pertinence forte des résultats, utilisation d'un format normalisé qui pourra faciliter l'intégration dans des portails à venir, pérennité sur le long terme des fichiers XML/EAD. En revanche, il est nécessaire dans ce cas de disposer d'une application pour la publication. Il convient également de prendre des précautions si on souhaite que les données soient facilement reconnues par les moteurs de recherche<sup>21</sup>. De même, l'impression des résultats peut être plus complexe<sup>22</sup>.

### 3.5- Notices saisies dans une base de données

Si les descriptions archivistiques sont saisies dans une base de données, on aura là encore pour la publication, plusieurs choix possibles : utiliser un export au format HTML ou au format XML/EAD, si la base de production le permet<sup>23</sup> ; recourir à une base de données en ligne<sup>24</sup>. Les mêmes précautions que celles mentionnées au 3.1.4 doivent être prises pour la reconnaissance par les moteurs de recherche ou l'impression des résultats. A l'inverse, la recherche peut être particulièrement riche, mais les résultats seront d'autant plus pertinents et contextualisés que la base de production aura été conforme à la description ISAD/G.

### 3.6. Tableaux synthétiques

#### Processus de publication envisageables en fonction de l'état initial des instruments de recherche

\_ : impossible directement

+ : possible

++ : pertinent

	PDF image	PDF texte	HTML	(XML/EAD)	Bases de données
<b>IR manuscrits, non océrisables</b>	++	—	—	—	—
<b>IR océrisables</b>	+	++	+	—	—
<b>IR traitement</b>	—	++	++	—	—

<sup>18</sup> Ce fichier sert alors pour faciliter pour les internautes l'impression papier. Pour obtenir ce fichier, on appliquera une feuille de style à travers l'exécution d'un programme de transformation. Le format RTF est fréquemment proposé aux internautes en raison de sa relative portabilité.

<sup>19</sup> Si on produit du HTML à partir d'une structure XML, il faut paramétrer une feuille de style relativement sophistiquée. Les outils qui le permettent ne sont pas très ergonomiques et on se heurte vite à des problèmes de programmation. Il est nécessaire de faire appel à un informaticien.

<sup>20</sup> Par exemple, en utilisant la plate-forme SDX et son application Pleade.

<sup>21</sup> Voir annexe sur les pages dynamiques.

<sup>22</sup> Voir annexe sur les pages dynamiques.

<sup>23</sup> C'est le cas d'Arkheia par exemple.

<sup>24</sup> Par exemple, MySQL.

de texte					
IR XML/EAD	—	+	+	++	+
IR Bases de données	—	+	+	++	++

### Evaluation de l'ergonomie des divers modes de publication

- : moyen ou médiocre

+: assez favorable

++ : très favorable

	PDF image	PDF Texte	HTML	XML/EAD	Base de données
<b>Facilité de mise en œuvre</b> ( par rapport aux outils à utiliser)	++	+	+	—	+
<b>Rapidité de mise en œuvre</b> (par rapport au temps d'intervention)	++	+	+	+	+
<b>Facilité d'affichage à l' écran</b>	—	—	+	+	+
<b>Facilité impression papier</b>	++	++	—	—	—
<b>Pérennité des données</b>	+	+	+	++	—
<b>Rapidité d'affichage</b> (en raison du volume)	—	+	+	+	+
<b>Recherche plein texte</b>	—	+	+	+	+
<b>Recherche structurée et contextua— lisée</b>	—	—	—	++	++ <sup>25</sup>
<b>Reconnais- sance par les moteurs de recherche</b>	—	+	++	+	+
<b>TOTAL</b>	<b>7+</b> <b>4 _</b>	<b>8+</b> <b>2 _</b>	<b>8+</b> <b>2 _</b>	<b>9+</b> <b>2 _</b>	<b>8+</b> <b>2 _</b>

<sup>25</sup> A condition bien évidemment que la production dans la base de données en amont soit structurée.

--	--	--	--	--	--

## Annexe 1. Formats PDF et HTML dynamiques

### *Le format PDF*

Les fichiers PDF ne sont pas visibles si Acrobat Reader n'est pas installé sur le poste client. Il faut donc offrir un lien de téléchargement vers le site d'Adobe. Pour des raisons de propriété intellectuelle, même si le logiciel est gratuit, il n'est pas légal de permettre le chargement direct sur un site, ni de l'installer sur un CD/DVD.

Les fichiers PDF peuvent être produits avec de multiples outils, par exemple Acrobat ou OpenOffice. Les fichiers produits sous Acrobat Distiller (option à choisir lors de l'édition en PDF d'un texte) conservent un enrichissement de polices et formats utilisables chez un imprimeur. Les fichiers produits avec OpenOffice ou Acrobat Writer (si on choisit cette option) sont des formats d'impression « plats » (comme des images en pratique)<sup>26</sup>; en revanche les fichiers sont beaucoup moins volumineux.

Il convient par ailleurs de rappeler que deux modes sont possibles avec le format PDF : le format texte et le format image.

Un document au format texte (exemple d'un instrument de recherche produit à partir d'un outil de traitement de texte) est converti pour publication au format PDF mode texte (par conséquent « atteignable » par les moteurs de recherche).

Si le document à publier est dans un format image (exemple d'un inventaire papier dactylographié), il est possible, à partir de cette image, d'effectuer une océrisation (grâce à un logiciel de reconnaissance de caractères) qui permettra, à côté du format image, de générer un fichier au format texte. Le format PDF, en effet, permet de garder le format image pour afficher le texte non reconnu conjointement au texte reconnu.

Dans ce cas, les moteurs de recherche pourront reconnaître certains mots, voire tous si on a choisi de corriger manuellement les caractères mal reconnus.

Si on se contente d'un format PDF image (exemple des inventaires manuscrits), il est recommandé, pour pallier l'impossibilité d'accès par les moteurs de recherche, d'établir un sommaire de l'inventaire en HTML, qui sera reconnu par les moteurs et permettra à l'internaute de naviguer aisément d'une partie à l'autre.

La plupart des moteurs de recherche sont désormais capables de rechercher dans les pages PDF. Toutefois, en règle générale, les documents PDF figurent moins facilement que les pages HTML en première page d'un résultat de recherche. De même, au-delà d'un certain nombre de pages, il faut s'assurer que l'intégralité du fichier est bien repérée.

Une attention particulière devra être portée aux liens : en effet, le format PDF permet de générer des liens, par exemple entre une table des matières et un chapitre, un index et une occurrence du terme, et même entre plusieurs documents ou encore des liens externes entre la

---

<sup>26</sup> Dans sa version actuelle.



page de titre de l'instrument de recherche et la page d'accueil générale ou celle des instruments de recherche. Il s'agit cependant de liens « positionnels » et non structurels : si on change la mise en page, le lien n'est pas conservé ou, pire, devient erroné.

Par ailleurs, des liens externes générés à l'intérieur d'un fichier au format traitement de texte ne sont pas conservés en cas de conversion au format PDF. Pour des liens qui auraient été générés en interne (interne au fichier d'origine), il convient de s'assurer que la version d'Acrobat que l'on utilise permet de conserver ces liens.

Les fichiers PDF peuvent être très volumineux, donc demander des temps de chargement considérables. Il est donc recommandé d'éviter les illustrations. Il convient par ailleurs de proposer un « téléchargement continu » : la première page demandée s'affiche pendant que les suivantes se téléchargent en tâche de fond. L'option peut être sélectionnée pour chaque fichier PDF. Il convient également que les serveurs de l'hébergeur soient paramétrés pour laisser passer cette fonction, ce qui n'est pas systématique (ainsi, sur les serveurs Linux, l'option doit être activée). Une autre solution peut consister à tronçonner l'inventaire en différents fichiers auxquels on accède grâce à un sommaire en HTML.

### *Les pages dynamiques*

On appelle pages dynamiques (format HTML dynamique), des pages HTML générées au fur et à mesure des requêtes des utilisateurs.

Afin d'améliorer la reconnaissance par les moteurs de recherche, il convient d'offrir une présentation des fonds « étendue » en pages XML/HTML traditionnelles, par exemple en éditant les historiques et introductions des fonds. Ceci en effet permet de créer des pages statiques balisées par des titres et mots-clés sur lesquels les moteurs de recherche pourront opérer. Il n'est pas nécessaire que soient visibles ces pages statiques : il suffit qu'elles existent sur le site et qu'elles renvoient aux pages dynamiques.

Par ailleurs, il est recommandé de choisir un format de stockage des données reconnu par les moteurs de recherche : c'est le cas des bases de données en logiciels libres (type MySQL) et théoriquement XML. Il convient par ailleurs de prendre garde, dans le cas de données stockées dans une base de données XML, aux cadres qui peuvent empêcher d'accéder à ces données : il s'agit d'un élément à prendre en compte lors de la conception de l'architecture du site.

Il faut enfin tenir compte du problème des impressions de résultats. Un accès dynamique aux instruments de recherche peut être un format d'interrogation donnant un grand nombre de réponses (plusieurs milliers). Dans ce cas, il faut prévoir une présentation sur plusieurs pages avec navigation d'une page à l'autre et une méthode d'impression ou de téléchargement du résultat.



## **Annexe 2- Prise en compte, lors de la publication en ligne, des éléments propres aux instruments de recherche papier**

### *Comment présenter le sommaire ?*

S'il existe en tant que tel, le sommaire doit être synthétique (correspond au plan de classement). S'il doit être toujours visible, le sommaire doit aussi avoir un lien dynamique avec l'instrument de recherche concerné. Il faut également éviter de répéter l'arborescence du sommaire dans l'écran de visualisation.

Dans Gaïa, le nombre de niveaux de description affichés à chaque étape de la navigation est limité (premier niveau subordonné). Le sommaire est en fait constitué par la liste des éléments subordonnés (non dépliés), au-dessous de la description du fonds.

Aux Archives départementales de la Côte-d'Or s'affiche à gauche la liste des instruments de recherche comportant des résultats, et quand on sélectionne un de ces instruments de recherche, on a dans un premier temps les résultats correspondant à la requête présentés dans leur contexte hiérarchique ("affichage filtré"). Quand on veut visualiser ces résultats dans le contexte général de l'ensemble de l'instrument de recherche, alors la frame droite permet de visualiser le plan de classement.

Au CHAN, avec PLEADE, la fragmentation du document EAD peut être différente pour le sommaire (dans la partie gauche de l'écran partagé) et pour l'instrument de recherche.

Au CAOM, le parti pris a été de ne pas afficher le sommaire, mais de faire apparaître sur l'écran de gauche la contextualisation de la recherche (cela permet de descendre dans les niveaux inférieurs).

### *Les annexes des instruments de recherche*

On distinguera les annexes qui ont leur place naturelle dans les niveaux de description les plus élevés (introduction) et les annexes non obligatoires qui peuvent être appelées à la demande par un clic. Un premier principe est par conséquent à retenir : celui d'un affichage optionnel des annexes par l'internaute.

### Bibliographie

Si on souhaite présenter une bibliographie, il convient de la dater, avec le problème des mises à jour. On peut également éviter des mentions de ressources bibliographiques trop générales, que l'internaute ira de toute façon chercher sur des sites spécialisés, et ne retenir que les références les plus précises, difficilement repérables ou encore locales. On peut également choisir de rattacher ces ressources aux niveaux de description inférieurs, mais là encore, on risque de « parasiter » les informations essentielles pour la bonne compréhension de l'unité de description. On recommandera dans tous les cas de faire des liens vers des ressources bibliographiques extérieures à l'instrument de recherche électronique (soit internes au service d'archives, soit externes).

### Sources complémentaires

Les sources complémentaires de même provenance sont à signaler obligatoirement. En revanche, pour les sources complémentaires sans lien de provenance, on peut se contenter de renvoyer à un autre service par un lien (la description des documents sera toujours plus à jour sur le site de l'autre service).

### Récupération des index papier

La notion d'index (listes alphabétiques) comme annexe de l'instrument de recherche papier n'a plus de raison d'être pour un instrument de recherche électronique. Le contenu des index est le même, c'est la forme qui est différente : soit affichée dans le contexte de la réponse à une requête, soit sur des formulaires de recherche, sous forme de listes déroulantes. Concernant la récupération des index d'instruments de recherche papier, il est recommandé de ne l'envisager que dans le cas où l'index papier est plus riche que le corps de l'instrument de recherche. Même dans ce cas, il convient de mettre en place une stratégie qui prenne en compte les gains de l'indexation au regard des charges et, dans tous les cas, d'imaginer une récupération sous la forme la plus simple qui soit.

Les instruments de recherche papier du CAOM ne sont pas publiés sans index. Mais pour la mise en ligne de l'état général des fonds, le choix a été fait de ne pas reprendre les index dans un premier temps, dans l'attente des réactions des chercheurs. Le seul index au niveau de l'état des fonds est l'index des noms de lieux (équivalent des tableaux méthodiques de noms de lieux dans les instruments de recherche papier), qui permet une économie au chercheur et donne par conséquent une véritable valeur ajoutée. Les résultats de la requête « Martinique » seront plus précis par l'index géographique qu'en mode recherche plein texte. La décision de ne pas reprendre l'indexation n'est pas définitive : l'insertion ultérieure de balises d'indexation dans les documents XML natifs est toujours possible.