



**MINISTÈRE
DE LA CULTURE**

*Liberté
Égalité
Fraternité*

**Direction générale des patrimoines
Service des musées de France**

**JOURNÉE D'ETUDE « BASES DE DONNÉES DOCUMENTAIRES : ETAT DES
LIEUX ET PERSPECTIVES »
Paris, 22 mai 2007**



Mise en ligne : fin 2007

**Présentation de "Collections", guichet unique d'accès en ligne aux données patrimoniales du
ministère de la culture et de la communication**

Jack Meurisse, Haut fonctionnaire chargé des systèmes d'information, Secrétariat général du ministère de la culture et de la communication

"Collections", le guichet unique d'accès en ligne aux données patrimoniales du ministère de la culture et de la communication est accessible sur Internet depuis avril 2007 grâce à un onglet spécifique du portail www.culture.fr.

Dressons l'historique de ce projet :

- dès janvier 2002 ; le rapport de Bruno Ory-Lavollée "La diffusion numérique du patrimoine, dimension de la politique culturelle" faisait déjà mention du défi que représentait l'interopérabilité des différentes bases de données du ministère de la culture et de la communication, permettant leur interrogation simultanée par une seule requête nécessitant un chaînage et la mise en commun d'un puissant moteur de recherche.

Cette ambition paraissait il y a cinq ans de cela quelque peu utopique. Elle partait du constat que le ministère de la culture et de la communication gère depuis trente ans une multiplicité de bases de données documentaires et patrimoniales fonctionnant grâce à des technologies variées : logiciel Mistral, systèmes libres, format d'échange XML, plate-forme SDX mise en service par la mission de la recherche et de la technologie.

Ces sources d'information, disparates et hétérogènes, pâtissaient d'un manque de visibilité.

Malgré la diffusion sur Internet entrée assez vite en jeu au sein du ministère et notamment pour la mise en ligne des bases, le mode d'accès à ces dernières, non fédérées, que ce soit via le site institutionnel du ministère ou via le portail événementiel, demeure toujours compliqué.

Mutualiser ces bases de données, tout en respectant l'intégrité de leur contenu et la diversité de leurs supports techniques, devient alors une nécessité afin de valoriser auprès d'un plus large public international le potentiel numérique du ministère de la culture et de la communication et la qualité du travail collectif de ses équipes.

En 2005, l'élaboration du schéma directeur des systèmes d'information du ministère, conjuguée aux réflexions de la LOLF, de l'administration électronique, introduit les bases de données comme l'un des premiers éléments du schéma informatique.

La première phase de ce schéma - DO-1 (Documentaire 1) - prévoit la mise en commun des bases de données documentaires ; recherche commune et simultanée sur l'ensemble des bases.

Dès septembre 2005, un groupe de travail interdirectionnel se réunit pour définir le périmètre de ce projet et déterminer les bases concernées.

En janvier 2006 : le projet est présenté au Comité stratégique des systèmes d'information (qui remplace l'ancien COMSI). Le directeur du cabinet du ministre et les directeurs présents donnent leur accord. Le projet est lancé officiellement. Le groupe de travail, officialisé, se consacre jusqu'en juillet 2006 à la rédaction du cahier des charges avec le soutien d'une assistance à maîtrise d'ouvrage extérieure. L'appel d'offre est lancé le 15 juillet 2006. Le rythme soutenu du projet est notamment dicté par l'échéance des élections présidentielles.

Le groupe de travail interdirectionnel est consolidé par la mise en place de maîtrises d'ouvrages transverses au ministère : communication et documentation, directement impliquées par la connexion de leurs bases Evénements et Capadoce.

Le guichet unique est l'un des projets de la maîtrise d'ouvrage documentaire transverse (le DO-2 abordera le problème des schémas de données du ministère afin d'améliorer leurs procédures de production et leur diffusion ; le DO-3 traitera des systèmes d'information géographique à partir de l'application Atlas du Patrimoine).

En septembre 2006 intervient le choix de la société commercialisant le moteur permettant l'accès en ligne des données patrimoniales des quatorze bases retenues dans le premier cercle. Il s'agit des bases de données jugées par les directions comme les plus significatives. Le fait qu'elles fonctionnent sur des supports techniques différents doit pouvoir prouver qu'un seul moteur peut les interroger, restituer des résultats communs tout en donnant accès aux documents sources. Les premières bases retenues pour la DMF sont Joconde, Atlas (œuvres exposées du musée du Louvre), agence photographique de la RMN, établissement public), notamment en raison de leur grand nombre d'images en ligne.

"Collections" doit s'intégrer dans l'interface du portail www.culture.fr ; un onglet spécifique est donc prévu pour proposer un accès unique en recherche plein-texte, transversal, simple et convivial pour un public non spécialiste. La deuxième version du guichet unique (en cours de réalisation) offrira des fonctionnalités de recherche avancée qu'un public plus spécialisé a l'habitude d'utiliser.

Il faut saluer l'effort collectif qui a permis de tenir les très brefs délais de mise en service. Les différentes directions étaient partie prenante du projet, en bon accord avec la société retenue.

"Collections" couvrira à terme divers domaines de connaissance dont certains transverses : musées, architecture, archéologie, objets mobiliers, arts du spectacle, événements culturels, généalogie, arts décoratifs, peinture, sculpture, plans et cartes, archives, enluminures...

Différents types de ressources documentaires sont fédérées : bases de données, sites web statiques (Lascaux, célébrations nationales) ou dynamiques (RMN, Atlas), publications électroniques (relevant de la direction du patrimoine et de l'architecture).

Des technologies différentes sont mises en relation : mistral, sdx, xml, sites web, html.

Le tout représente actuellement deux millions de documents en ligne, la plupart du temps illustrés.

La société choisie (parmi quatre) est SINEQUA. Il s'agit d'une PME française, commercialisant le moteur linguistique et sémantique nommé Intuition.

Le groupe projet a fait appel également à une autre société d'assistance à maîtrise d'ouvrage pour l'étape des tests fonctionnels linguistiques.

La réalisation de la version 1 s'échelonne d'octobre 2006 à avril 2007. Le temps de travail de la maîtrise d'ouvrage (groupe projet d'une dizaine de personnes et renfort ponctuel des spécialistes travaillant dans les directions) s'élève à 600 jours/homme.

Quelques mots sur le fonctionnement du moteur Intuition :

- Il s'agit d'un outil de recherche plein texte faisant de l'analyse morpho- syntaxique. Le principe est de retrouver automatiquement les mots sous toutes leurs formes en évitant les ambiguïtés : ainsi à partir de la racine du mot sont envisagées toutes ses évolutions (singulier, pluriel, féminin, déclinaisons verbales et formes conjuguées). A noter que les corrections phonétiques et orthographiques ne fonctionnent pas si les documents diffusés par les bases comportent de telles erreurs. Le moteur est insensible à la casse, à l'accentuation : ce qui est particulièrement important pour un portail international.

- Le principe d'analyse sémantique permet l'amélioration de l'ordre de présentation des réponses et de proposer des élargissements par le sens. C'est une nouveauté pour des bases structurées en champs et recourant à des vocabulaires contrôlés et hiérarchisés.

- Des expansions sont possibles : synonymie, reformulation (thesaurus allégé). Il est en effet possible d'élargir et de compléter les termes du dictionnaire central du moteur par des vocabulaires "métier", listes d'autorités ou faiblement hiérarchisés (un niveau).

Le classement par pertinence des résultats de recherches est déterminé par des statistiques calculées par le moteur.

Les résultats s'affichent sous forme d'une liste signalétique, avec un bref résumé du contenu de la notice et possibilité d'accès par lien cliquable à la notice source au sein même de la base concernée.

Afin d'affiner ou d'élargir la recherche, des renvois vers d'autres termes sont proposés dans l'espace à gauche des résultats.

[S'ensuit une démonstration de l'onglet "Collections" du portail www.culture.fr.]

Quelques statistiques de consultation :

En moins d'un mois de mise en ligne, l'onglet "Collections" a reçu des visites d'internautes des cinq continents (90 % Europe, 3 % Etats-Unis, 2% Asie ; 2% d'Afrique majoritairement francophone ; 2% d'Amérique du Sud).

A l'heure actuelle, les questions posées proviennent de navigateurs Internet français à 75 % et 25 % d'autres langues (anglo-saxon, espagnol, allemand, italien, portugais et notamment brésilien).

La deuxième version (été - automne 2007) proposera une fonction de "cross language" : des questions pourront être posées directement en anglais ou en espagnol. Elle proposera surtout l'accès à 17 bases de données patrimoniales supplémentaires, ainsi qu'une fonction de recherche avancée (recherche linguistique plus fine, par intervalles de date et synonymie élargie).

La possibilité d'ouvrir un compte permettra à l'internaute de conserver le résultat de ses recherches, de les transférer sur son poste de travail en plusieurs formats (HTML, CSV, fil RSS) et d'envoyer à des partenaires, des sélections via la messagerie. La généalogie fera l'objet d'un traitement particulier, avec son propre onglet dans le portail www.culture.fr.