PANDOR : Portail d'archives numériques et de données de la recherche

Laurent Gautier, Céline Alazard, Agnès Viola, Hédi Maazaoui, Arnaud Millereux

PANDOR : Portail d'archives numériques et de données de la recherche

[ressource] Puissant outil d'interrogation et de valorisation des ressources numériques, Pandor est issu d'une réflexion menée conjointement entre chercheurs et personnels techniques de la Maison des sciences de l'homme (MSH) de Dijon et intégrant le vaste mouvement des humanités numériques.

Inauguré en octobre 2014[1], Pandor, puissant outil d'interrogation et de valorisation des ressources numériques, permet de localiser et d'accéder à un ensemble de données, le plus souvent inédites, issues de programmes de recherche pluridisciplinaires. Il couvre tous les champs thématiques des sciences humaines et sociales représentés à la MSH et intègre tous les types de données multimédias, qu'elles soient natives ou le fruit d'une numérisation. Répondant aux standards internationaux en matière de traitement de données, Pandor permet, grâce à une fine description des contenus, de repérer des documents difficiles d'accès. Il inclut aussi des archives et des productions de chercheurs constitués dans le cadre de programmes de recherche.

Données techniques

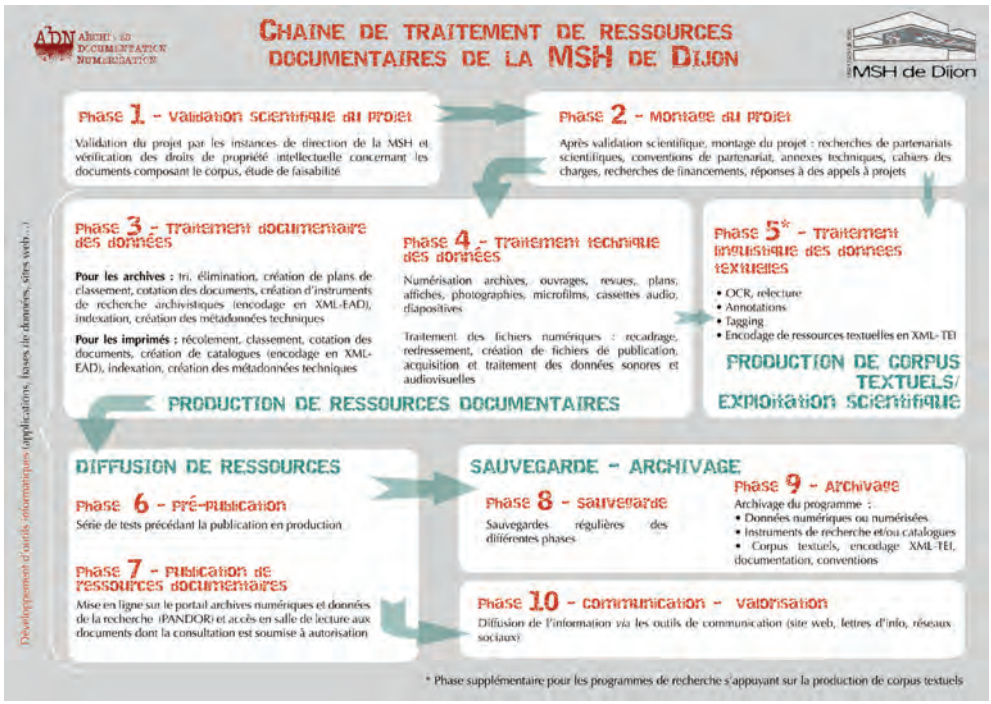
L'application développée en JAVA a été déployée sur un serveur d'application de type Tomcat sous Linux. Les données sont stockées dans une base de données MySQL afin d'accroître les performances et les capacités de l'application. La gestion de l'affichage et des traitements est déléguée à des routines réalisées par transformation xslt pour une portabilité et une évolutivité accrues.

Pandor permet la mise en ligne d'instruments de recherche et de catalogues créés au format XML/EAD et d'objets numérisés ou nativement numériques. Il s'appuie sur les standards du Web pour le traitement et la diffusion des données. Ainsi, les documents sont consultables *via* la visionneuse intégrée et compatibles avec les équipements de type smartphones, tablettes, ordinateurs.

Plusieurs procédés de traitement permettent de tirer le meilleur parti des documents numériques textuels. La recherche en texte intégral est rendue possible par l'application en amont de la reconnaissance optique de caractères et de la technologie XML Mets/Alto. L'interopérabilité de Pandor est assurée par la présence d'un entrepôt utilisant le protocole OAI-PMH[2]. Ainsi, Pandor échange déjà ses données avec des moteurs nationaux et européens tels qu'Isidore ou Europeana. ////

Illustration

Chaîne de traitement de ressources documentaires de la MSH de Dijon



//// Cette démarche s'appuie sur les préconisations et les bonnes pratiques des grandes institutions de la recherche et de la culture.

Des fonds inédits

Les fonds d'archives ou d'imprimés publiés sur Pandor s'insèrent dans les thématiques de recherche de la MSH. Ainsi, par exemple, elle a numérisé et mis en ligne des archives produites par l'entreprise Schneider au moment de la Grande Guerre[3], représentant quelques 100 000 pages de dossiers, 400 plans et 800 photos.

On y trouve aussi un échantillon de 3 000 brochures antérieures à 1940 issues du fonds de la Bibliothèque marxiste de Paris. Cette action de numérisation, de catalogage et de diffusion sur Internet permet d'accéder à un fonds de documents imprimés rares (car non totalement répertoriés par la BnF) et précieux (leur fragilité n'autorisant plus leur consultation par le public). Leur mise en ligne rend possible leur exploitation par tout type de public, dans des domaines qui dépassent le cadre de la recherche scientifique (produits éditoriaux, film documentaire).

L'avenir

Le devenir du portail s'appuie sur l'acquisition de nouvelles compétences, l'enrichissement de l'outil Pandor lui-même l'inscrivant dans le mouvement des humanités numériques. Il s'agit, par exemple, de transformer des données de corpus « statiques » en données dynamiques et intelligentes. Ainsi, la MSH[4] a développé, à partir d'un corpus textuel original et inédit du *Bulletin de l'Organisation Internationale de la Vigne et du Vin*, plus précisément « des notes de dégustation œnologique », un prototype d'indexation utilisable pour la fouille de données et pour une exploitation lexicale et sémantique permettant d'automatiser leur analyse.

Ce projet appliquera à ce corpus les prérequis de l'analyse de sentiment avec pour objectif de parvenir à l'extraction automatique du profil positif/négatif des évaluations de vins. Il s'agit d'intégrer à l'indexation et à l'extraction des données des savoirs experts propres à la filière viticole de référence, en particulier d'une analyse serrée du lexique employé. Les résultats du projet seront transférables à d'autres types de données plus hétérogènes comme les blogs ou forums d'amateurs dont le rôle prescriptif, avec l'essor du Web 2.0, ne doit pas être négligé dans les décisions d'achat des consommateurs.

Dans ce cadre, une réflexion a été engagée sur l'adoption de la Text Encoding Initiative (TEI) déjà utilisée par des institutions précurseurs en la matière[5]. Parallèlement, l'acquisition des compétences liées au traitement automatique des langues sera transférée à la communauté scientifique par des formations élaborées par l'équipe de la plateforme.

L'évolution technologique de l'outil Pandor passera par la mise en place d'outils de diffusion des contenus des corpus au moyen des réseaux sociaux simultanément au déploiement de l'outil de diffusion DTD-TEI. Les nouveaux programmes de recherche sur le traitement des corpus oraux évalueront la faisabilité de la mise à disposition en ligne sur le portail de retranscriptions (alignement son et texte). ■

> **L'équipe de la plateforme Archives Documentation Numérisation (ADN) de la MSH**

L. Gautier (responsable de la plateforme ADN et du projet Pandor), **C. Alzazard, A. Viola, H. Maazoui, A. Milleureux** et les personnels contractuels intervenant aux différentes étapes de production des données.

Contact : Laurent.gautier@u-bourgogne.fr

1. pandor.u-bourgogne.fr. Financé par le Plan d'action régional pour l'innovation (Pari) de la région Bourgogne et le Fonds européen pour de développement régional (Feder), Pandor s'appuie sur la solution libre multiplateformes Pleade.

2. Open Archives Initiative Protocol for Metadata Harvesting

3. On y découvre non seulement la participation de cette grande entreprise à l'effort de guerre avec la fabrication massive d'armement, mais aussi la vie des salariés.

4. *Via* sa plateforme Archives-Documentation-Numérisation (ADN)

5. Telles que le consortium international TEI, le consortium Écrits ou la MSH Val de Loire, etc.