Date : 10/07/2008

PERSÉE, UN OUTIL AU SERVICE DE LA COMMUNICATION SCIENTIFIQUE FRANCOPHONE

Nathalie FARGIER, Programme Persée
Responsable de la documentation et des relations revues

Valérie NEOUZE, Ministère de l'enseignement supérieur et de la recherche, Direction générale de l'enseignement supérieur.(C3-3).
Responsable du patrimoine et de la numérisation des collections

Meeting: **148. Social Science Libraries** with **Division II** & **Special Libraries** and **Geography and Map Libraries**

Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

World Library and Information Congress: 74th IFLA General Conference and Council
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

PERSÉE est un programme national de numérisation et de publication électronique de revues scientifiques francophones en sciences humaines et sociales (SHS). Il repose sur une chaîne de traitement intégrée permettant la conversion au format électronique d'articles scientifiques existant sous forme imprimée et sur un portail de diffusion permettant un accès libre et gratuit à ces contenus scientifiques, fruit d'une concertation étroite avec la communauté scientifique.

Lancé en 2003 à l'initiative du Ministère chargé de l'enseignement supérieur et de la recherche, ouvert au public en 2005, le portail PERSÉE continue trois ans plus tard à répondre aux besoins de son public cible et à l'accompagner dans les transformations profondes qu'une coopération quotidienne avec les directeurs de publications, les comités de rédaction et les chercheurs a su initier. Prospectif, ce programme s'est donné également pour ambition de donner à l'archivage pérenne des documents numérisés une place centrale, en nouant un partenariat privilégié avec le Centre informatique national de l'enseignement supérieur (CINES), afin de garantir l'intégrité et l'accessibilité à long terme de ce riche patrimoine scientifique national numérisé.

1. GENESE ET FONDEMENTS DU PROGRAMME PERSÉE

1.1. Contexte de la communication scientifique en sciences humaines et sociales

Reflet de la richesse du patrimoine national en sciences humaines et sociales, les revues scientifiques françaises représentaient au début des années 2000 un corpus de portée internationale très peu valorisé, à la différence des revues anglo-saxonnes. Pourtant la diffusion de ce patrimoine constituait déjà un enjeu stratégique évident, dont les acteurs tels que *JSTOR*[1] ou *Proques*[2]*t* avaient pris depuis longtemps la mesure. Comment expliquer cette très nette avance anglo-saxonne dans la diffusion en ligne des revues SHS?

Si l'utilisation de la langue anglaise constitue un premier élément de réponse, elle n'explique pas à elle seule la forte présence des revues scientifiques anglo-saxonnes en SHS sur les réseaux. En réalité, celles-ci bénéficiaient déjà d'infrastructures de diffusion performantes, conçues à l'origine pour le vaste et rentable domaine des revues en sciences et techniques, dont la « *marchandisation* », très supérieure, justifiait de tels investissements. Modèle économique rôdé et intégration de l'offre SHS dans les collections pluridisciplinaires, gage d'une visibilité minimale par défaut, sont autant d'atouts dont profitaient d'emblée les revues anglo-saxonnes en SHS. Il faudrait y ajouter les capacités financières et commerciales des maisons d'édition et des institutions scientifiques sans commune mesure avec nos moyens nationaux et le dynamisme des presses universitaires, qui prouvaient l'existence d'un espace public performant et non lucratif pour la diffusion des revues scientifiques en SHS.
En France, point d'oligopoles ou de maisons d'édition à vocation internationale aux infrastructures éprouvées. Mais une atomisation des acteurs, tant privés que publics, qui affaiblissait encore les capacités financières disponibles et contribuait à la frilosité générale, premier frein à la révolution numérique dans l'édition scientifique en sciences humaines et sociales. Sur les 200 revues françaises dont les abonnements laissaient supposer une audience internationale, moins d'une dizaine proposaient systématiquement leurs articles en ligne en 2002.

De ce constat et des attentes de la communauté universitaire et de recherche est né le programme public PERSÉE, lancé à l'initiative du MESR en 2003 pour préserver ce riche patrimoine et améliorer la visibilité de la recherche scientifique en langue française dans le domaine des sciences humaines et sociales.

1.2. L'appel d'offre pour le programme PERSÉE (2003)

Menée en collaboration avec l'Institut des sciences du document numérique (ISDN), une étude menée en 2002 a permis de recenser et d'analyser les modèles économiques et éditoriaux existants pour définir les contours d'un portail public francophone de diffusion qui concilierait accès ouvert, exigence de visibilité et performance de la diffusion, dans le respect du droit des auteurs.
Parallèlement, une seconde étude réalisée par la société AJLSM, dédiée à la faisabilité technique du projet et à sa nécessaire interopérabilité avec les plateformes en ligne, a conduit

[1] http://www.jstor.org
[2] http://www.proquest.com/products_pq/descriptions/periodicals_archive.shtml

à un partenariat privilégié avec le portail canadien ERUDIT, notamment sur le partage des modèles de données.

A l'issue de ces travaux fondateurs été lancé le 7 mars 2003 par le Ministère en charge de l'enseignement supérieur et de la recherche un appel d'offres auprès des établissements publics pour le développement d'une chaîne de traitement numérique en *open source*, permettant la numérisation, la diffusion et l'archivage pérenne des revues en SHS. Concrètement, il s'agissait pour les équipes candidates de développer une plateforme de production automatisée et normalisée et un portail de diffusion alimentant directement un dépôt d'archivage numérique au CINES, et de lancer la production, sur ces outils émergents, d'un corpus de revues qualifiées alors de « pionnières ».

La proposition du consortium regroupant l'Université Lumière Lyon 2, la Maison de l'Orient Méditerranée-Jean Pouilloux (MOM-CNRS), l'Université de Nice-Sophia Antipolis et l'Ecole Normale supérieure de lettres de Lyon été retenue le 6 juin 2003 avec, comme objectif, une réalisation dans un délai d'un an à dater de la signature d'une convention associant les partenaires du projet (octobre 2003).

L'appel d'offre stipulait également la collaboration nécessaire avec le CINES désigné par le Ministère comme l'opérateur de l'archivage pérenne des documents produits dans le cadre de ce programme.

1.3. Les principes fondateurs

En 2005 fut donc ouvert au public le portail PERSÉE, partie visible du programme dédiée à la diffusion des revues française en SHS, avec 7 revues pionnières[3]. Fort de 54 revues aujourd'hui, soit plus de 170 000 articles, auxquelles il faut ajouter 36 revues en cours de production et 16 nouvelles revues cooptées en juin 2008[4], le portail PERSÉE a rencontré son public cible et poursuit sa montée en charge dans le respect de quatre principes fondateurs qui demeurent d'actualité depuis 2005.

1.3.1. Accès libre et gratuit

L'exigence de visibilité d'un corpus en langue française, produit par une communauté de chercheurs sur fonds publics, explique le choix initial d'une diffusion libre et gratuite des collections rétrospectives de ces revues, dans le respect d'une barrière mobile définie conjointement avec le directeur de publication et l'éditeur de la revue, tous deux signataires d'une convention spécifique. Il s'agit en effet pour le programme PERSÉE de contribuer à la visibilité de ce patrimoine scientifique national sans nuire à la commercialisation des numéros récents, qui assurent encore une partie des recettes nécessaires à la poursuite de la publication papier. La situation économique des revues étant inégale, la durée de la barrière mobile est laissée à l'appréciation des revues et de leurs éditeurs. Elle se situe généralement entre 3 et 5 ans.

[3] *Annales, la Revue de l'art, Vingtième siècle, L'homme, Revue française de science politique, Matériaux pour l'histoire de notre temps, Bibliothèque de l'Ecole des chartes, Revue économique.*

[4] Le programme PERSÉE a été doté dès l'origine d'un comité de suivi chargé, deux fois par an, d'expertiser les demandes de partenariat et de statuer sur les évolutions techiques et docuementaires envisagées par l'équipe PERSÉE.

1.3.2. Le respect du droit d'auteur

La question juridique demeure l'une des plus difficiles à traiter dans le cadre de ce programme. Dès 2003, le Ministère s'est attaché les services d'un laboratoire du CNRS, le Centre d'Etudes sur la COopération Juridique Internationale (CECOJI). Spécialisé dans le droit de la propriété intellectuelle et le droit des TICs, ce laboratoire s'est vu confier l'accompagnement juridique du programme, en lien avec les directeurs de publication et les éditeurs publics ou privés concernés : en l'absence de contrats entre les auteurs et la revue, et devant l'interprétation difficile des anciens contrats d'édition signés à une époque où le numérique n'existait pas, il a été décidé que l'accord des auteurs des articles à numériser dans le cadre du programme PERSÉE serait individuellement requis et qu'une convention spécifique serait systématiquement signée entre le programme PERSÉE, le directeur de publication, l'éditeur et, le cas échéant, le diffuseur afin de garantir une situation juridique claire pour tous les acteurs concernés. Ce dispositif, lourd et onéreux, est néanmoins actuellement la seule solution qui permette l'accessibilité en ligne d'articles sous droit nécessaires à l'avancée de la recherche en SHS, dans le respect de la législation française en vigueur.

1.3.3. Choix de technologies ouvertes et standardisées

Le programme PERSÉE s'est inscrit dès l'origine dans une démarche *open source*, exigée dans l'appel d'offre. Point de doctrine du Ministère, ce choix s'explique également par la volonté à moyen terme de remettre à disposition des établissements publics une chaîne de traitement numérique innovante et conforme aux normes techniques et documentaires dont l'emploi est préconisé au niveau international.

L'évolutivité de ces outils, et donc la disponibilité des sources, ainsi que le recours aux normes pour garantir l'intéropérabilité de ce portail avec les portails similaires ou complémentaires, en France et à l'étranger, participent d'une réelle stratégie de diffusion nécessaire à la réussite du projet et à l'intégration de ce corpus francophone dans les circuits internationaux de communication scientifique.

L'ambition de pérennité et le souhait d'offrir une information structurée de qualité ont conduit naturellement au choix de XML et d'outils reconnus dans le monde des bibliothèques numériques. Le schema METS[5] développé par la bibliothèque du Congrès est utilisé pour restituer toute la complexité d'une collection de revue (issn, volume, numéro, sommaire de fascicule) et pour gérer l'ensemble des éléments numériques créés tout au long du processus de dématérialisation (fichiers images, texte issu de la reconnaissance optique de caractères, métadonnées). Dans un souci constant d'ouverture et de compatibilité, deux schémas complémentaires ont été retenus pour décrire le contenu même des articles (TEI[6] et schéma ERUDIT article) ainsi que plusieurs jeux de métadonnées, répondant à des usages et des publics différents (Dublin Core[7], Mods[8], MarcXML[9]). Dès son ouverture, le portail PERSÉE était compatible avec le protocole OAI-PMH pour assurer une large dissémination des métadonnées liées aux articles.

1.3.4. Un outil conçu pour et par la communauté scientifique

Le quatrième principe fondateur réside dans la place nouvelle et centrale donnée à la communauté scientifique dans un projet qui lui est destiné. Afin de répondre avec pertinence

[5] http://www.loc.gov/standards/mets/
[6] http://www.tei-c.org/index.xml
[7] http://dublincore.org/
[8] http://www.loc.gov/standards/mods/
[9] http://www.loc.gov/standards/marcxml/

et efficacité à leurs besoins, accompagner un changement profond dans leurs pratiques informationnelles et anticiper de nouveaux usages, la méthode retenue, qui sera développée ci-dessous, fut celle d'une coopération étroite avec les directeurs de publication de revues emblématiques de la recherche en SHS, offrant un panel sinon exhaustif du moins représentatif des disciplines, dans le cadre de groupes de travail.

2. LE PORTAIL DE DIFFUSION PERSÉE

2.1. Une conception en collaboration étroite avec les chercheurs

Associant bibliothécaires, chercheurs, éditeurs et informaticiens, le programme PERSÉE fut l'occasion de créer un espace de dialogue fructueux entre professionnels aux compétences complémentaires. L'objectif de ce groupe de travail chargé de préfigurer le portail était de recenser les besoins et les pratiques des chercheurs en SHS et, prenant acte des avancées technologiques, de proposer des outils documentaires accessibles en ligne permettant de les satisfaire. Il s'agissait également d'accompagner toute une communauté scientifique au sein d'une révolution numérique qui avait laissé les SHS en marge du processus.

Les conclusions de ce groupe de travail peuvent être ainsi résumées :

- le respect de l'identité visuelle de la revue s'est imposé comme un élément clé, qui requérait donc le choix du mode image. Ce choix correspondait par ailleurs au principe français du droit patrimonial attaché à la mise en page originelle de l'article.
- la recherche dans le texte intégral était ensuite avancée comme une fonctionnalité indispensable. Les collections de revues devaient être traitées dans leur intégralité, du premier numéro paru au plus récent, sans aucune restriction liée au droit des auteurs, et un logiciel de reconnaissance optique de caractères (OCR) systématiquement utilisé.
- l'unité de sens pour les chercheurs est l'article, et non le numéro qui constitue certes le niveau pertinent de catalogage en bibliothèque mais qui présente un intérêt limité en termes de recherche. Ce choix de l'article comme unité documentaire de base fait sans doute la particularité de ce portail patrimonial en permettant, à partir d'une revue « papier », une articulation naturelle avec les portails diffusant les articles récents (via le protocole OAI-PMH) ou les bases bibliographiques.
- enfin, un article sous sa forme papier est organisé de manière structurée (différents titres, résumé, note de bas de page, références bibliographiques, annexe…). Un tel agencement devait faire l'objet d'une retranscription sous forme numérique pour permettre une exploitation fine du document et multiplier les accès à l'information. Cette demande sous-entendait l'utilisation de la norme XML et de modèles de données normalisées (METS et TEI).

Lieux d'échange, d'apprentissage et de discussion, ce groupe de travail a permis de déterminer les choix documentaires à intégrer dans le développement de la chaîne de traitement pour une diffusion ciblée sur le futur portail. Enfin, à ces contenus structurés devaient être associés des services en terme de navigation, de recherche, de consultation et d'exploitation innovants, reflétant a minima les pratiques des chercheurs avec les revues papier.

2.2. Des contenus associés à des services

2.2.1. De PERSÉE 1 à PERSÉE 2 : un portail en évolution permanente, à l'écoute des nouvelles pratiques

Ouvert en 2005, le portail PERSÉE a consacré ses deux premières années d'activité à la mise en ligne d'articles, l'effet de seuil étant un facteur de visibilité déterminant sur Internet, et à l'analyse constructive de l'outil de diffusion. Retours nombreux des chercheurs, études d'usage et recours à un cabinet d'ergonomes ont permis d'identifier les besoins satisfaits et les attentes à court terme et d'anticiper de nouvelles pratiques, souvent issues du web 2.0. Cette nouvelle version de PERSÉE a été inaugurée au Salon du Livre de Paris, en mars 2008.

Reflet des conclusions du groupe de travail, le portail PERSÉE présente des fonctionnalités de navigation et de recherche, traditionnelles dans leur approche, innovantes dans leur mise en œuvre[10]. Il ne s'agit pas ici de procéder à une description exhaustive du portail mais d'insister sur ses spécificités : article comme unité de base du portail, outils documentaires associés à chaque document pour optimiser la recherche et la consultation et inspiration de la philosophie du web 2.0 pour donner au chercheur une place toujours plus centrale dans le portail.

Si les outils permettent une navigation aisée au sein de chaque revue par le biais de ses sommaires numériques, réponse aux pratiques habituelles des chercheurs, la recherche porte sur une collection d'articles, toutes revues confondues, favorisant ainsi la transdisciplinarité inhérente aux sciences humaines et sociales qu'autorise une masse critique de 170 000 articles disponibles. Au-delà de la recherche sur les métadonnées et sur le texte intégral des articles, la recherche avancée permet à chaque usager de spécifier son champ d'application en sélectionnant les revues qui l'intéressent principalement et en construisant des requêtes complexes. Les résultats de recherche sont présentés par pertinence et peuvent être triés par date, par revue ou par titre d'article selon le souhait de l'internaute. Partant de ces résultats, il est proposé d'élargir ou, au contraire, d'affiner la requête, en précisant plus finement ce qui est recherché (type de document, date, titre de revue, disponibilité de la ressource). L'internaute a donc le choix de définir a priori une recherche bien ciblée ou de privilégier une approche autre reposant sur une restriction progressive du champ de la recherche. Au niveau le plus fin, les termes recherchés sont surlignés dans l'image numérisée de la page et permettent ainsi de repérer rapidement le contexte dans lequel ils s'insèrent.

A chaque article sont associés des éléments, issus d'un lourd travail documentaire en amont, tel que le plan interactif de l'article pour accéder directement aux chapitres, la liste des figures, les résumés en différentes langues, la référence bibliographique exacte et les mots clés indiqués par son auteur.

La consultation de l'article se fait ensuite en mode image, dans le respect du choix des chercheurs, mais avec une possibilité d'afficher le texte issu de la reconnaissance optique de caractères pour l'exploiter par le biais notamment de la fonction « copier-coller ». Le feuilletage est d'autant plus facilité que des outils de défilement et de zoom permettent de passer sans rupture d'une page à l'autre et de visualiser les détails d'une illustration. L'internaute peut également choisir d'écouter l'article, dont le texte a fait l'objet d'une

[10] Il est important de rappeler au préalable que l'un des défis du portail PERSÉE est de permettre l'exploitation des articles numérisés à partir du papier de manière aussi riche que s'il s'agissait de fichiers nativement numériques.

synthèse vocale. Disponible à titre expérimental sur une faible partie du corpus, cette fonctionnalité sera progressivement étendue à l'ensemble des documents, facilitant ainsi l'accès à ces collections pour les malvoyants.

Au-delà d'un accès facilité à l'information scientifique, Internet offre l'opportunité de s'approprier les contenus mis à disposition, de personnaliser l'interface de consultation et de constituer des réseaux sociaux. Inspiré de ces nouvelles pratiques du web, le portail PERSÉE a également choisi de proposer au chercheur une place non plus en amont mais au cœur du projet, en lui offrant des contenus mais aussi des espaces personnel ou collaboratif de travail en ligne. Gratuit mais soumis à authentification, l'espace personnel permet entre autre au chercheur de participer à l'indexation du corpus en associant aux articles des étiquettes, venant alimenter les nuages de « tags » associés à chaque article. Gérant le tri des étiquettes, l'historique des requêtes et des résultats, cet espace personnel est donc doté d'une diffusion sélective de l'information qui renforce l'existence des flux RSS désormais disponibles sur l'ensemble du site : l'information vient dorénavant à la rencontre du chercheur et non l'inverse. En outre, le chercheur peut ajouter des commentaires à un article, l'annoter et partager son travail au sein d'un groupe qu'il aura lui-même constitué ou auquel il aura été invité à participer.

2.2.2. Le portail PERSÉE dans les circuits internationaux de la communication scientifique

L'ambition première de PERSÉE est d'assurer une meilleure visibilité aux résultats de la recherche francophone en sciences humaines et sociales et de les inscrire dans un réseau international de connaissance. Afin de réaliser une telle entreprise, PERSÉE conduit une double action en participant à des projets internationaux reconnus d'une part, et en collaborant avec les intermédiaires traditionnels de l'information scientifique que sont les bibliothèques et les éditeurs de bases bibliographiques d'autre part.

PERSÉE est affilié à CrossRef et des DOIs sont attribués à tous les articles scientifiques diffusés par le biais du portail. Le mécanisme du referencement croisé permet aux utilisateurs de naviguer à partir des références bibliographiques qu'ils contiennent et ainsi de passer d'un article à l'autre qu'ils soient diffusés par PERSÉE ou présents sur un autre portail. Cette opportunité offerte aux éditeurs constitue une exception parmi les diffuseurs de revues francophones en sciences humaines et sociales et, en plus d'améliorer le parcours des utilisateurs au sein des collections, elle constitue une avancée dans la mise en place d'outils bibliométriques.

PERSÉE est également partie prenante de l'Open Knowledge Initiative (OKI) portée par le *Massachussetts Institute of Technology (MIT)* qui, centrée sur les attentes hétérogènes et propres à chaque utilisateur d'Internet, vise à développer des outils et des interfaces permettant de s'approprier des contenus distants et de les personnaliser. A l'exemple d'autres diffuseurs de contenus scientifiques comme *JSTOR* et *Muse*[11], PERSÉE a développé un connecteur qui permet à tout internaute, via son interface de travail, d'interroger différentes ressources distantes et de rapatrier les résultats de ses requêtes en fonction de ses besoins spécifiques.

Alors que la quantité d'information présente sur Internet connaît une croissance exponentielle, la question de la validité et de fait de la validation de cette masse de données se pose avec force. PERSÉE souhaite développer des partenariats avec les bibliothèques et les fournisseurs

[11] http://muse.jhu.edu

de bases bibliographiques pour favoriser un accès « médiatisé » aux articles de PERSÉE parallèlement à un accès direct à son contenu. Une première étape fructueuse a permis d'établir des liens entre les catalogues des bibliothèques, comme celui de la bibliothèque de Yale, et les articles diffusés par.PERSÉE. Des discussions en cours avec *Google Scholar* et *Proquest* augurent de collaborations avancées offrant aux internautes la qualité des contenus scientifiques de PERSÉE et les services propres à chaque plateforme.

2.2.3. La mutualisation avec les autres portails de revues scientifiques en sciences humaines et sociales

Avec plus de 170 000 articles en texte intégral et une croissance annuelle de 30 nouveaux titres, PERSÉE est le premier portail de diffusion de revues scientifiques francophones en sciences humaines et sociales. Les collections des revues sont traitées et diffusées dans leur intégralité, du premier numéro paru jusqu'au numéro le plus récent. Chaque éditeur détermine, revue par revue, si le corpus présent sur PERSÉE a vocation à s'enrichir en fonction de l'avancement de la barrière mobile ou s'il demeure en l'état. Dans le seconde hypothèse, les numéros de l'année en cours peuvent être disponibles *via* un autre portail et soumis à abonnement (Cairn, revues.org, Armand Colin). PERSÉE a anticipé dans ses choix techniques et documentaires la nécessaire interopérabilité avec les portails diffusant les numéros actuels d'une revue afin d'assurer aux internautes la continuité dans la consultation des corpus. Sans différenciation entre les revues, l'internaute peut accéder à l'ensemble des sommaires même les plus récents et effectuer des recherches dans les métadonnées et le texte intégral des articles diffusés par PERSÉE et d'autres portails. Dans un souci de transparence, tout internaute est informé des conditions d'accès au texte intégral et du portail vers lequel il sera dirigé s'il souhaite y accéder (consultation libre et gratuite sur PERSÉE, autre mode de consultation sur les portails partenaires).

L'interopérabilité repose sur l'utilisation commune du protocole OAI-PMH et de modèles de données largement utilisés par les éditeurs scientifiques. Selon les spécificités technologiques propres à chaque plateforme et l'utilisation des standards pré-cités, PERSÉE échange différents types de données, des seules métadonnées des articles (en Dublin Core) à des données plus complexes et riches comme les sommaires des numéros (en METS), jusqu'au texte intégral à des fins d'indexation (en TEI).

PERSÉE a également établi des partenariats avec des portails assurant la diffusion d'autres revues scientifiques en sciences humaines et sociales, comme le consortium canadien Erudit. De tels échanges permettent *via* une même interface d'interroger un autre corpus et d'accéder aux revues diffusées par Erudit en plus de celles proposées par PERSÉE.

3. LA CHAINE DE FABRICATION PERSÉE

3.1. Une chaîne de traitement automatisée

L'objectif premier de PERSÉE est la diffusion de contenus scientifiques de qualité et la mise à disposition d'outils permettant de rechercher dans ces contenus et de se les approprier. Ce programme se distingue par une écoute attentive de la communauté des chercheurs qui bénéficie de ces services et des éditeurs scientifiques qui y participent. Les échanges sont constants et les besoins, de l'une ou l'autre des parties, s'expriment directement auprès de l'équipe PERSÉE ou par le biais d'un forum ou d'études d'usages. La formalisation des

demandes des utilisateurs de PERSÉE complète un travail de veille et un suivi appliqué des évolutions technologiques. Les évolutions régulières de la chaîne de traitement numérique et du portail de diffusion témoignent de la réactivité de PERSÉE.

La chaîne de traitement PERSÉE est automatisée pour assurer la conversion au format électronique de documents pré-existant sous forme papier et pour intégrer des fichiers électroniques fournis par les éditeurs. Elle repose sur un serveur d'objets et sur un outil dénommé jGalith qui gère l'ensemble du workflow et assure l'agencement efficient des différentes étapes de traitement et de contrôle.

En premier lieu, les éditeurs fournissent les collections papier des revues et l'équipe PERSÉE procède au récollement et à un premier contrôle de l'état matériel des ouvrages. Chaque page est numérisée à haute résolution (400 dpi) et des images TIFF sont générées en niveau de gris ou en couleurs. Un logiciel de reconnaissance optique de caractères (OCR) est alors appliqué à l'ensemble des pages scannées et des applications automatisées permettent de lancer des traitements massifs (repagination des fichiers, amélioration de la qualité des images et des contrastes). En l'état, la chaîne PERSÉE est dimensionnée pour traiter un million de pages par an. Un des caractères distinctifs du programme PERSÉE est le travail conséquent de documentation qui fait suite à la numérisation et aux traitements massifs des corpus.

Pour chaque numéro, les tables des matières sont créées avec la saisie des titres des documents, de leurs auteurs, la pagination, la langue de rédaction du document et le type de document (article, compte rendu, note critique). A un niveau plus fin de granularité, la structure de chaque article est précisément décrite (intertitres permettant de générer le plan de l'article, résumés, mots clés, liste des illustrations, liens vers d'autres articles à partir des références bibliographiques). Chacune de ces étapes est suivie d'un contrôle qualité automatique ou humain qui permet de poursuivre le traitement et d'avancer dans le processus. Au terme de la chaîne, les données numériques créées sont à la fois transmises au CINES pour archivage et diffusées *via* le portail PERSÉE.

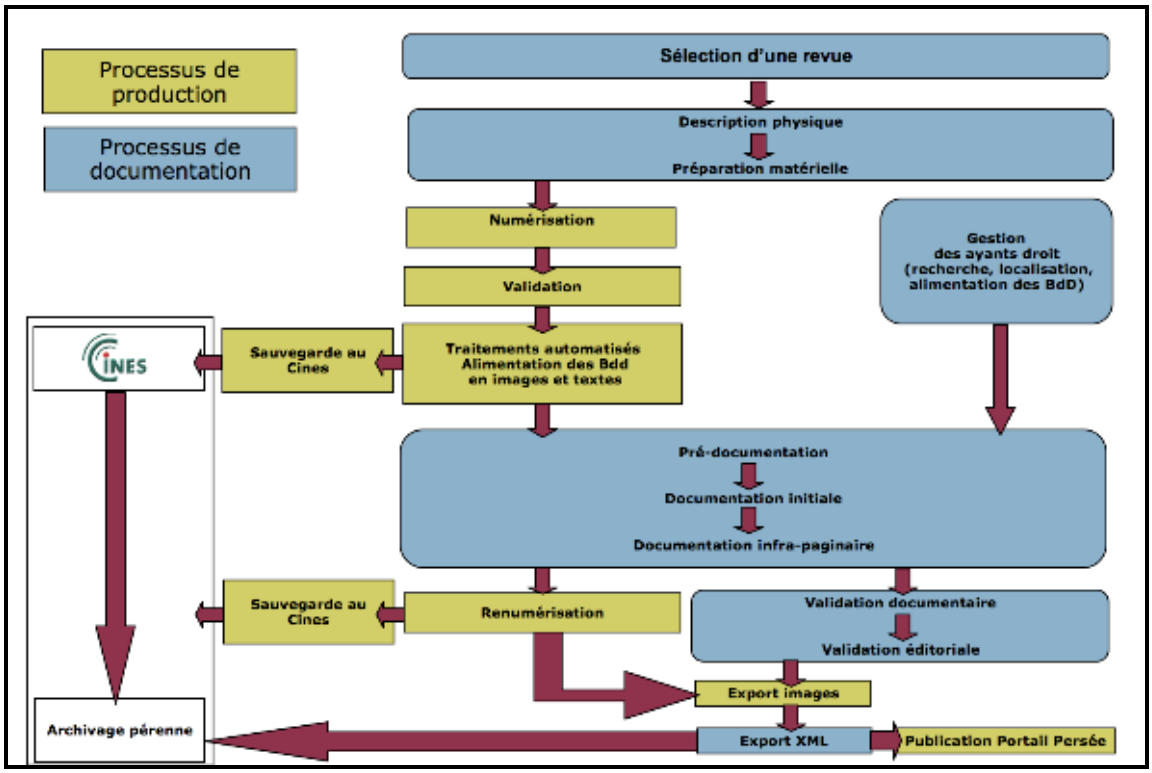

Fig.1- Schéma de la chaîne de traitement Persée

Répondant à la demande du Ministère de l'Enseignement supérieur et de la recherche, la chaîne PERSÉE a été développée en *open source* sous la double licence CeCILL et GPL. Ce choix de technologies libres permet un essaimage de la plateforme PERSÉE et la constitution à l'échelle nationale d'un réseau de sites pouvant conduire leurs propres projets de numérisation.

3.2. Essaimage de la chaîne de traitement

En 2008, la chaîne de traitement PERSÉE est parvenue à un état de développement stabilisé, qui permet d'envisager un premier essaimage technologique auprès d'établissements publics susceptibles de s'associer au programme afin de poursuivre en l'accélérant la montée en charge du portail de diffusion.

Réalisés en *open source*, les outils PERSÉE nécessitent néanmoins, dans un premier temps, un accompagnement pour en appréhender toute la complexité, en particulier sur les étapes de documentation. C'est pourquoi cet essaimage a pris la forme d'une « réplication » auprès d'un établissement universitaire, Université Paris Descartes. A la faveur d'une convention, celle-çi accueille au sein de l'une de ses composantes[12] les technologies PERSÉE et bénéficie pour cela d'une formation par l'équipe PERSÉE et d'une assistance à distance pour démarrer la production numérique parallèle de revues destinées à rejoindre l'unique portail de diffusion. La convention prévoit par ailleurs la possibilité pour le « pôle de réplication » d'utiliser à ses fins propres de numérisation tous les outils PERSÉE, fondant ainsi les bases d'une communauté d'utilisateurs aptes à en promouvoir et poursuivre le développement en *open source.*

[12] UMS 3036 : Unité Mixte de Service : service de l'université associé avec le CNRS

Cette expérimentation, lancée au printemps 2008, devrait permettre de mesurer la faisabilité d'un essaimage plus ouvert, à destination de tout établissement public français ou non intéressé par l'installation d'une chaîne intégrée et normalisée de traitement numérique,, favorisant ainsi la diffusion de bonnes pratiques en matière de production, de diffusion et d'archivage pérenne des données.

3.3. L'archivage pérenne des données

Partenaire historique du programme, le Centre informatique de l'enseignement supérieur (CINES) est un établissement public placé sous la tutelle du ministère de l'enseignement supérieur et de la recherche. Consacré au calcul intensif, il met aujourd'hui à disposition ses capacités informatiques et les compétences de ses équipes pour proposer un service d'archivage pérenne des données numériques. Choisir d'intégrer cette dimension de manière native dans le programme PERSÉE en fait l'une de ses spécificités..

Membre du groupe PIN (Préservation de l'information numérique) et de *l'Alliance for permanent access to the records of science,* le CINES a développé PAC[13], une plateforme d'archivage sur le modèle OAIS[14], retenu par les autres acteurs nationaux et internationaux comme la Bibliothèque nationale de France (avec SPAR) ou la Direction des archives de France et la Direction générale de la modernisation de l'Etat (avec PILAE). Lancé en 2007, un appel d'offre du CINES a permis l'installation de l'infrastructure matérielle et logicielle propre à évoluer de la sauvegarde à l'archivage pérenne des données nativement numériques ou des documents numérisés.

Une collaboration étroite entre les équipes du CINES et l'équipe PERSÉE a permis de définir, sur le mode de l'expérimentation, les modalités de versements, d'archivage et d'accès en dernier recours aux articles numérisés dans le cadre du programme, représentant dans un premier temps une volumétrie de 20 téraoctets. Etape finale du processus modélisé dans le programme PERSEE, l'archivage pérenne s'inscrit dorénavant naturellement et automatiquement dans la chaîne de traitement PERSÉE.

Initié pour répondre au défi que représentait la mise en ligne des revues scientifiques françaises en sciences humaines et sociales, peu visibles sur les réseaux, le programme PERSÉE a choisi d'associer étroitement une communauté scientifique, impatiente, fidèle et motivée depuis la conception, toutes disciplines confondues. Cette démarche, originale dans son écoute permanente des besoins des chercheurs, se traduit d'une part dans les évolutions technologiques de la chaîne de traitement numérique, d'autre part dans la cooptation des nouvelles revues qui rejoignent le portail de diffusion. Comptant aujourd'hui 54 titres en ligne représentant près de 170 000 articles en accès libre et gratuit, auxquels il faut ajouter 36revues en cours de production, le portail PERSÉE compte chaque mois 2,5 millions de connexions qui laissent espérer que l'objectif premier de visibilité de ces revues sur Internet est atteint. Chiffres sans doute satisfaisants qui ne doivent pas masquer qu'une intégration réussie dans les circuits internationaux de communication scientifique impliquent une veille technologique et documentaire quotidienne, une attention renforcée aux besoins du public que l'on sert et des partenariats à poursuivre ou construire avec tous les acteurs concernés, chercheurs, bibliothécaires, informaticiens, éditeurs et diffuseurs.

[13] Pour des informations techniques plus précises, consulter le site du CINES : http://www.cines.fr , rubrique *Archivage pérenne*.
[14] OAIS-ISO 14721 : *Reference model for an Open Archival information System*

Annexe 1 : Revues disponibles sur le portail Persée et en cours de traitement (juillet 2008)

Revues actuellement disponibles sur le portail Persée

1. Actes de la recherche en sciences sociales (Seuil)
2. Annales, Histoire, Sciences Sociales (Editions de l'EHESS)
3. Annales de géographie (Armand Colin)
4. Année psychologique (Armand Colin)
5. Archives de sciences sociales des religions (Editions de l'EHESS)
6. Bibliothèque de l'école des chartes (Société de l'Ecole des chartes)
7. Bulletin de correspondance hellénique (Editions de l'Ecole française d'Athènes)
8. Bulletin de l'école française d'extrême-Orient (Editions de l'EFEO)
9. Bulletin de la société préhistorique française (Société préhistorique française)
10. Bulletins et mémoires de la société d'anthropologie de Paris (Sté d'anthropologie de Paris)
11. Cahiers de l'AIEF (AIEF)
12. Cahiers de linguistique – Asie orientale (EHESS)
13. Cahiers d'études africaines (Editions de l'EHESS)
14. Dialogue d'histoire ancienne (Presses universitaires de Franche Comté)
15. Faits de langue (Ophrys)
16. Flux (Métropolis)
17. Genèses (Belin)
18. Géocarrefour (Association des amis de la revue géographique de Lyon)
19. Géomorphologie (GFG)
20. Histoire, Economie, Société (Armand Colin)
21. Histoire et mesure (Editions de l'EHESS)
22. L'Homme (Editions de l'EHESS)
23. Journal de la société des américanistes (Société des américanistes)
24. Journal des africanistes (Société des africanistes)
25. Langages (Armand Colin)
26. Langue française (Armand Colin)
27. Livraisons d'histoire de l'architecture (Association LHA)
28. Matériaux pour l'histoire de notre temps (Association des amis de la BDIC et du Musée)
29. Mélanges (Editions de l'Ecole française de Rome)
30. Mil neuf cent. Revue d'histoire intellectuelle (Société d'études soréliennes)
31. Mots. Les langages du politique (ENS Editions)
32. Paléorient (CNRS Editions)
33. Pôle Sud (OPPES)

34. Politique étrangère (IFRI / Armand Colin)
35. Politix (Armand Colin)
36. Population (INED)
37. Réseaux (Lavoisier)
38. Revue d'anthropologie du Portugal (GAP – MSH Paris)
39. Revue d'histoire des sciences (Armand Colin)
40. Revue de géographie alpine (Armand Colin)
41. Revue de la numismatique française (Société française de numismatique)
42. Revue de l'art (CNRS Périodiques jusqu'en 2004)
43. Revue de l'OFCE (Presses de Sciences Po)
44. Revue d'économie industrielle (Editions techniques et économiques)
45. Revue économique (Presses de Sciences Po)
46. Revue européenne de migrations internationales (AEMI)
47. Revue française de science politique (Presses de Sciences Po)
48. Revue française d'économie (Association française d'économie)
49. Revue française de sociologie (Ophrys)
50. Revue internationale de droit comparé (Société de législation comparée)
51. Revue Tiers Monde (Armand Colin)
52. Romantisme (Armand Colin)
53. Tiers Monde (Armand Colin)
54. Vingtième siècle. Revue d'histoire (Presses de Sciences Po)

Revues qui vont intégrer le portail Persée

1. Annales d'Ethiopie (La Rable Ronde)
2. Annuaire des collectivités locales (CNRS Editions)
3. Annales historiques de la révolution française (Armand Colin)
4. Annuaire français du droit international (CNRS Editions)
5. Archipel. Etudes interdisciplinaires sur le monde insulindien (Association Archipel)
6. Arts Asiatiques (Ecole Française d'Extrême-Orient)
7. Cahiers d'Extrême-Asie (Ecole Française d'Extrême-Orient)
8. Cahiers de linguistique hispanique médiévale (ENS Editions)
9. Cahiers du monde russe (Editions de l'EHESS)
10. Communication et Langages (Armand Colin)
11. Communications (EHESS)
12. Comptes rendus de l'Académie des Inscriptions et Belles-Lettres (AIBL)

13. Critique internationale (Presses de Sciences Po)
14. Déviance et société (Médecine et Hygiène)
15. Economie et statistique (INSEE)
16. Economie rurale (Société française d'économie rurale)
17. Journal de la société des océanistes (Société des océanistes)
18. Médiévales (Presses universitaires de Vincennes)
19. Mélanges de la Casa Velazquez (Ecole Française de Rome)
20. Mètis (Daedalus)
21. Paléo (SAMRA)
22. Quaderni (Sapientia)
23. Quaternaire (AFEQ)
24. Recherche sur Diderot et l'encyclopédie (Société Diderot)
25. Réforme, Humanisme, Renaissance (Association RHR)
26. Revue archéologique de Picardie (Société archéologique de Picardie)
27. Revue internationale de politique comparée (De Boeck)
28. Revue des mondes musulmans et de la Méditerranée (AESHAN)
29. Revue d'histoire de l'église de France (Société d'histoire religieuse de France)
30. Revue de l'histoire des religions (Armand Colin)
31. Revue du christianisme social (Mouvement du christianisme social)
32. Revue du Louvre (RMN)
33. Revue française de pédagogie (INRP)
34. Revue philosophique de Louvain (Peeters)
35. Sociétés contemporaines (Presses de Sciences Po)
36. Syria (IFPO)