Bibliothèques Virtuelles Humanistes
http://www.bvh.univ-tours.fr

Assemblée générale 2011
le 5 décembre 2011, CESR, Tours

Les BVH : programme de recherche en « humanités numériques »

- **Région Centre**

BLOT Anne-Françoise (Médiathèque BMVR Orléans), CHASSEGUET Véronique (SCD Tours), PUEL Stéphanie (SCD Tours), PUYNEGE-BATARD Colette (BM Bourges), SCHNEZ Lise (Archives municipales de Tours), TISSERAND Sylvie (Prytanée national militaire), TRANCHIDA Robert (BM Bourges), YARDIN Olga (AD 37)

- **Autres Partenaires**

BERTET Karell (L3i, La Rochelle), BOUJU Alain (L3i, La Rochelle), CLAERR Thierry (MCC, Service du livre et de la lecture), COUSTATY Mickaël (L3i, La Rochelle), GUILLEMINOT Geneviève (BnF, Réserve), HULVEY Monique (BM de Lyon), LETRICOT Rosemonde (TGE Adonis), LOLLIEROU Franck (Supersoniks), MAUREL Denis (LI, Tours), PIOCH Alexandra (Centre de recherche du château de Versailles), RAYAR Frédéric (LI, Tours), SORDET Yann (Bibliothèque Mazarine), TSOPZE Norbert (L3i, La Rochelle)

- **Université François-Rabelais & MSH de Tours**

BOUILLON Maud (Crévilles, MSH), CHAREILLE Pascal (CERMAHVA, UFR Arts & Sc. Hum.), CORREALE Francesco (CITERES, MSH), LALLIER Thomas (Crévilles, MSH), LAURENTI Jean-Noël, LORET Stéphane (Crévilles, MSH), LYAET Marie-Christine (Crévilles, MSH), QUENTIN Jacques (Centre de documentation, MSH), THOMAS Guillaume (CERMAHVA, UFR Arts & Sc. Hum.)

- **CESR & IRHT**

AQUILON Pierre, ARDESI Denise, BÉNÉVENT Christine, BERTRAND Lauranne, BLIN-DAVID Claire, BOUTEILLER Jean-Louis, BOUTROUE Marie-Elisabeth (IRHT), BREUIL Sandrine, BUSSON Sébastien, DEMONET Marie-Luce, FAUQUET Hélène, FINS Jorge, GEONGET Stéphan, GREIS Yvone, JIMENES Rémi, JOURNET Agnès, MASQUILIER Marie-Laure, NUE Alice, UETANI Toshinori

- **Excusés**

BOBIN Martine (Médiathèque François-Mitterrand, Poitiers), BOISSEUIL Didier (MC Histoire, université François-Rabelais), BONY Catherine (Bibliothèques de Blois-Agglopolys), BURGHART Marjorie (CIHAM, Lyon), BURNARD Lou (MEET Adonis), CORON Antoine (BnF), CRESPO Miguel (Digiscrib), CRON Geneviève (BnF), DELHAYE Anne-Sophie (BnF), DOSSO Diane (IHPST - UMR 8590), DRAELANTS Isabelle (Centre de médiévistique Jean-Schneider, MSH Lorraine, Université de Nancy 2), DUBOIS Alain (BM, Vendôme), DUBOULOZ Nicolas (Conseil régional du Centre), GALLIOT Nathalie (Archives municipales, Bourges), GOLDMAN Philippe (Conseil général du Cher), GUIGNARD Bruno (Bibliothèques de Blois-Agglopolys), ILLIANO Marie-Odile (MCC), JESTAZ Juliette (ENSBA), LAY Marie-Hélène (Ecrit.com, FORELL, Poitiers), LEGROS Alain (CESR), Le MORT Françoise (ISHS), LE ROLLE Vanessa (Arkhenum), MARTIN Frédéric (BnF), NAAS Laurent (BM Sélestat), NEVEU Valérie (Angers), PALLUAULT Florent (MCC, Service de la lecture), PORT Anita (BM Vendôme), POTARD Dominique (Médiathèque Châteauroux), RAMEL Jean-Yves (LI Tours), TRAINEAU-DUROZOY Anne-Sophie (BU Poitiers), VENE Magali (BnF), WALSBY Malcolm (School of History, University of St Andrews), WHITELOCK Jill (Cambridge University Library)

Les BVH dans le contexte des « Digital Humanities »

Les BVH dans les programmes de recherche au CESR

• Les programmes de numérisation en cours – Axe « Renaissance » du CPER, projet PADOVA (jusqu'en 2013)
• Labellisation par le CNRS de la Fédération de recherche constituée avec le CESCM
• Les programmes numériques déposés :
  • Labex PATRIMAR
  • Personae (prosopographie) APR régional obtenu : Ricercar, IRHT, BVH, CESCM
  • Equipex Biblissima : IRHT, EPHE, EnC, BnF, CIHAM (Lyon), BVH [*note* : la sélection de Biblissima a été annoncée le 20 décembre suivant l'AG]

Les BVH dans les TGIR (Très Grandes Infrastructures de Recherche) et dans DARIAH : le consortium CAHIER « Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche »

Les BVH adoptent les grands principes des « digital humanities », par : les problématiques liées à la propriété intellectuelle, la mise à disposition de données réutilisables-partageables-transmissibles (Open data) et l'archivage durable, à l'intention des communautés de chercheurs. BVH-TIPo est le centre de ressources numériques (CRN) pour les Textes Imprimés PatrimOniaux labellisé depuis 2010 par le TGE Adonis. Les CRN sont en cours de redéfinition depuis juin 2011 en articulation avec les consortiums CORPUS, dont le consortium CAHIER porté par le CESR. Les BVH participeront notamment au VCC3 (Virtual Center of Competence, Scholarship content) de DARIAH pour les données textuelles.

Coopérations

• **avec la Bibliothèque nationale de France :** La convention de « pôle associé » (signée en 2006) est prolongée "de fait" jusqu'à aujourd'hui ; un nouveau modèle intégrant les partenariats de numérisation et de recherche sera proposé au début 2012 ; une nouvelle convention devrait être mise en place avec la Réserve des imprimés dans le cadre du projet MONLOE.
• **avec Europeana :** une nouvelle convention sur la disponibilité des métadonnées a été signée en novembre 2011, suivant les recommandations de la Communauté européenne sur la numérisation (28 octobre 2011).
• **avec La Flèche, Bibliothèque du Prytanée militaire** : la convention est en cours de signature.

Téléchargements aux BVH

302 192 téléchargements de fichiers pdf /4 615 000 téléchargements tous fichiers confondus (page web, images, pdf, ...) pour les 11 mois de 2011.

Corpus Fac-similés

Corpus : 962 fac-similés numérisés, 607 titres en ligne au 5 décembre 2011 – 400 en déc. 2009 (383 Tours, 7 Châteaudun, 19 Vendôme, 38 Châteauroux, 108 Blois, 29 Poitiers, 6 Cambridge, 1 Budapest, Romorantin, Paris, Saumur, Nogent-le-Rotrou, Bologne… ; 21/272 Orléans), soit 351 610 images numérisées depuis 2003 et plus de 30 000 éléments extraits (bases de lettrines, portraits, matériel typographique).

2012 : Projet de Bourges : environs 150 documents imprimés et manuscrits (soit 30.000 images) seront numérisés, en réponse à l'appel à projet du ministère de la culture déposé en novembre 2011 par la ville de Bourges en partenariat avec le CESR ; l'objectif étant la valorisation du patrimoine berruyer par son accessibilité en ligne à tout public. [*note* : la notification de ce projet a été connue à la fin du mois de décembre]

Métadonnées : Les ressources numériques des BVH sont référencées par Claire Blin (bibliothèque du CESR) dans les notices du SUDOC pour lesquelles une « collection BVH » a été créée (liste des ressources UNIMARC, liens vers les notices, date de numérisation). Les BVH s'intègrent toujours aux réseaux de références : Europeana, Gallica, Isidore (via l'entrepôt OAI-PMH), ISTC (British Library, Incunables), GLN 15-16 (Genève), Edit16 (ICCU), etc. ; USTC (St Andrews), VD16, VD17 à venir. Le transfert des notices vers XML/TEI et Marc/XML est en cours de réalisation.

Corpus textuel Epistemon

Corpus : 46 textes en français de la Renaissance incluant des transcriptions inédites. En 2011, le corpus antérieur en html se résorbe (10 en 2011 contre 19 en 2010). Depuis 2007, 36 textes ont été encodés en XML-TEI (10 en 2009, 21 en 2010) et sont dorénavant téléchargeables aux formats PDF, TEI et HTML.

- Joachim Du Bellay, *L'Olive de Joachim Du Bellay*, Paris, 1549
- François Rabelais, *Le Quart Livre, Lyon, 1548*
- François Rabelais, *La Brève Déclaration de François Rabelais*, Paris, 1552
- Pierre de Ronsard, *Ode de la paix de Pierre de Ronsard*, Paris, 1550
- Pierre de Ronsard, *Élégie de P. de Ronsard Vandomois*, Paris, 1563

Evolution du modèle XML-TEI Renaissance : Le modèle a été mis à jour suivant les évolutions de la TEI et les derniers points non valides résolus : les descriptions bibliographiques de l'exemplaire et de l'édition sont désormais plus cohérentes et détaillées, les responsabilités sont gérées (pour la création d'index et la gestion des variantes), les états de transcription sont catégorisés (dissimilé, désabrégé, etc.).

Etude de la migration des bases de données des fac-similés en XML-TEI : Actuellement, pour le corpus Fac-similés, seuls les textes océrisés (PDF MRC) sont indexés en plein texte mais interrogeables sur quelques métadonnées sous le moteur de recherche XTF tandis que la consultation et la navigation dans les fac-similés s'effectue grâce aux bases de données des sommaires et notices avec une recherche bibliographique. Un modèle d'encodage en XML-TEI pour les fac-similés est en cours de réalisation et d'harmonisation avec le modèle du corpus Epistemon, afin de récupérer dans un fichier unique pour chaque ouvrage les données issues des différentes bases de données : le sommaire structuré et transcrit, les métadonnées de la notice, les liens vers les images, le texte océrisé (ou la transcription si elle est disponible).

Adaptation de la plateforme de publication : XTF est désormais paramétré de façon à distinguer les corpus BVH en préparation de l'import en XML-TEI du corpus Fac-similés. XTF dispose actuellement d'options de navigations avancées pour la consultation du corpus Epistemon.

- Nouvelles fonctionnalités : Indexation affinée, recherche textuelle par élément XML/TEI, ajout d'index des champs bibliographiques, ajout de facettes, ajout des notices bibliographiques <TeiHeader> des transcriptions XML/TEI.
- A venir : affichage page à page, affichage du texte et de l'image alignés, harmonisation avec le corpus Fac-similés (export des bases de données en XML/TEI), etc.

Outils de traitement du corpus Epistemon

- Edit-TEI, en partenariat avec la société Digiscrib : nouvelle fonctionnalité de dissimilation et de détildage en ligne avec Franscriptor (projet Google 2a).
- Analog, Dissimilog et Varialog : en partenariat avec Ecrit.com et l'université de Poitiers, Marie-Hélène Lay (Projet Google 2b)
- Philologic, en collaboration avec l'Université de Chicago, Marc Olsen : accessible en ligne depuis juillet 2011, via le menu Epistemon/Recherche par mot. PhiloLogic permet de constituer son corpus et de générer des statistiques à partir des textes Renaissance du corpus Epistemon.
- TXM, collaboration avec S. Heiden et D. Vigier, ICAR, ENS-Lyon (et labex numérique lyonnais) : traitement linguistique expert, plateforme pour construire et analyser le corpus Renaissance.

Corpus Manuscrits

Corpus

- La base « De minute en minute » compte 12 000 actes notariés, dont 4 500 documents d'archives du XV^e^ siècle des Archives départementales d'Indre-et-Loire – Transcriptions ou analyses ;
- Manuscrits du fonds Frotté (comptes de Marguerite de Navarre) : en cours de transcription et d'encodage XML-TEI en vue de leur publication.

Projet Actes Notariés

Dans le cadre du CRN TIPo, un modèle d'encodage en XML-TEI de ce type de manuscrits a été établi (en collaboration avec l'École Nationale des Chartes) avec pour objectifs la conversion de la base existante

« De minute en minute » et la création d'un formulaire de saisie (en collaboration avec TELMA, IRHT) accessible en ligne, afin que les étudiants et chercheurs puissent transcrire, commenter et alimenter la base de données XML-TEI. Un nombre restreint de balises TEI pourront être intégrées directement par l'utilisateur en fonction des compétences et droits d'accès accordés. La typologie proposée via le formulaire et validée par Pierre Aquilon pour la caractérisation des minutes transcrites sera soumise à la communauté des spécialistes. Les transcriptions des minutes tourangelles du XVe seront bientôt alignées avec leur fac-similé numérisé, en partenariat avec les Archives départementales d'Indre-et-Loire.

Iconographie & typographie

Evolution de la base BaTyR

Le corpus est constitué de 26 202 images extraites soit, après dédoublonnage automatique (à reprendre manuellement) : 7 133 lettrines, 426 bandeaux, 32 éléments d'encadrements, 420 fleurons, 245 marques typographiques. Jusqu'à présent, les bases de lettrines et de marques typographiques étaient distinctes et la recherche globale sur les éléments d'ornements n'était pas possible. On procède actuellement à la fusion de ces bases pour optimiser la recherche via un formulaire unique de la base BaTyR ou via des notices « imprimeurs » qui constitueront des répertoires des matériels de chaque atelier typographique. À court terme, on souhaite intégrer les marques typographiques répertoriées par Renouard (1 142 marques) et Silvestre (1 310 marques), et, à long terme, intégrer la base de polices de caractères typographiques (en réflexion).

L'OCR RETRO et le classement des caractères – Frédéric Rayar, pour Jean-Yves Ramel (LI, Tours) : Pattern Redundancy Analysis for Document Image Indexation and Transcription (PaRADIIT, Google Award 1)

Pour la transcription de textes de la Renaissance, le LI a développé plusieurs outils destinés à faciliter l'OCR des documents patrimoniaux de qualité : (1) AGORA : permet la segmentation de l'image en fac-similé afin d'isoler les éléments illustrés des blocs de texte (formats de sortie : ALTO, METS, XML-TEI), (2) Module de clustering : regroupe les formes extraites en groupes homogènes ; on cherche à identifier les contours de chaque caractère (*ex.* cas des caractères qui se chevauchent), (3) RETRO : une base de caractères typographiques va être constituée afin d'améliorer la qualité de l'OCR, ce qui, à terme, permettra l'étude de la circulation et de l'évolution des matériels typographiques.

Projets en cours

Google Award 2

- **Outil Franscriptor**, *Google 2a :* Développement d'un outil d'automatisation de la transformation en version patrimoniale, puis régularisée (par méthode n-grammes et dictionnaires), avec la société Digiscrib (Tours – La Riche)
- **Outil Varialog**, *Google 2b :* Développement d'un outil de requête à partir de la variation graphique (application de règles linguistiques et de dictionnaires), avec le laboratoire FORELL (Poitiers, Marie-Hélène Lay)

Informatisation des *Catalogues régionaux des incunables* (MCC/CESR)

L'objectif du projet est la diffusion en ligne, gratuite, d'un catalogue interopérable, rassemblant les données des *CRI* et enrichi de liens vers d'autres catalogues ou fac-similés d'incunables en ligne, grâce à nos collaborations européennes. En particulier avec l'ISTC (Incunabula Short Title Catalogue) de la British Library afin de récupérer automatiquement certaines métadonnées. Le format choisi est UNIMARC-TEI en fonction duquel le SIGB open-source KOHA est en cours de test de paramétrage afin de déterminer dans quelle mesure il faciliterait la rétroconversion des notices.

Nouveaux programmes de recherche

« Montaigne à l'œuvre » (MONLOE) – Projet ANR Corpus déposé (Oct. 2011)

Le projet MONLOE a pour objectif de reconstituer virtuellement la bibliothèque de Montaigne en partenariat avec les institutions détentrices des ouvrages originaux (33 exemplaires de la réserve de la BnF sélectionnés par G. Guilleminot, 30 exemplaires de la BM de Bordeaux, 11 de Cambridge – 6 déjà en ligne, dont 2 avec les nouvelles transcriptions par Alain Legros (CESR, chercheur associé) des notes manuscrites :

- Lucretius Carus, Titus, De rerum natura libri sex, Paris, 1563 [Annoté par Michel de Montaigne]
- Montaigne, Michel de, *Les Essais*, Paris, 1652 [Annoté par Jean-Jacques Rousseau]

MONLOE rendra possible la découverte de ce corpus entièrement numérisé et la mise en valeur des processus de création de l'œuvre de Montaigne par la transcription intégrale de « l'Exemplaire de Bordeaux ».

- **« Renom » – Projet régional obtenu (APR 2011), Denis Maurel (Laboratoire d'informatique, Tours) :** Le projet ReNom vise à permettre une meilleure exploitation des ouvrages – modes image et texte – par une recherche des entités nommées (principalement les noms de personnes et les noms de lieu) et par leur indexation, souvent absente des éditions. Ce projet, orienté vers la valorisation et le tourisme par le financement de la région Centre, rendra ces données accessibles au grand public via le web, les smartphones et des bornes interactives, la navigation à partir des noms de personne ou de lieu, tout en l'invitant à visiter la région (par exemple le musée de La Devinière s'il s'agit de Rabelais, mais aussi de châteaux qu'il mentionne).

Dissémination : documentation, publications, formations

Documentation en accès libre ou à la demande

- Un *Manuel d'encodage TEI "Renaissance"*, en ligne depuis 2008 (V2 en 2009), version 3 en préparation
- des descripteurs d'images (Iconclass, OLDB - lettrines)
- des cahiers des charges (numérisation, interface web, moteur de recherche, normes de saisie)
- des modèles de métadonnées (description d'imprimés, d'incunables, de manuscrits, d'archives)

Publications et, interventions

- Novembre 2010, Tours, Journée « Signalement et numérisation des incunables » (en ligne)
- Décembre 2010, Lyon-Valpré, Université d'hiver Adonis
- Janvier 2011, Journée édition génétique avec l'ITEM, Tours
- Juin 2011, Journée Patrimoine écrit, La Rochelle
- Juin 2011, congrès des Digital Humanities, Stanford (présentation du projet Montaigne et des premiers résultats obtenus avec la bourse Google)
- Juillet 2011, journées de travail avec l'ENC et l'ATILF ; avec ICAR-Lyon
- Octobre 2011, Séminaire de Caen
- Décembre 2011, Council of Content Providers, Europeana, Rotterdam
- Décembre 2011, Présentation des travaux réalisés avec les bourses Google (Centre culturel, Paris)
- *En projet* : participation au congrès des Digital Humanities, Hambourg, juillet 2012 – à la session plénière d'Europeana, Londres, juin 2012.

Formations

- Octobre 2011, ANGD (Aussois), gestion de projet (encadrement)
- Mai 2011, Journée formation (suivie) sur XSLT, Lyon
- Novembre 2011, Stage TEI formation continue (SUFCO), Tours, CESR
- Décembre 2011, Organisation de l'atelier du consortium CAHIER, Paris-Sorbonne
- Janvier-février 2012, Option « Bibliothèques virtuelles » du master 2 professionnel « Patrimoine écrit et édition numérique » (CESR, Tours) : traitement de l'image, initiation aux OCR et stage XML-TEI
- *En projet* : Printemps 2012, Formation sur les entrepôts OAI-PMH (Adonis et Consortium CAHIER)

Lancement du technopôle ValCoNum – **Problématiques de numérisation de documents historiques en masse** – Mickaël Coustaty (L3i, La Rochelle)

ValCoNum a vocation depuis septembre 2011 à devenir l'Institut de Valorisation des Contenus Numériques français intégrant des partenariats ponctuels et bilatéraux entre les acteurs de la recherche académique et ceux issus du développement industriel privé, avec pour but de constituer une filière numérique nationale et pérenne.