Département Informatique
Année 1998/1999

Rapport de Synthèse

CITHER : Consultation en texte intégral des thèses en réseau
Etude XML, intégration de LaTeX

Julien Tognazzi

Entreprise d'accueil : Doc'INSA
Adresse : INSA – bât. 220
20, av Albert Einstein
69621 Villeurbanne Cedex

Enseignant responsable : Jean-Marie PINON
Tuteur du projet : Jean-Michel MERMET

Résumé
La publication électronique des thèses est un domaine en forte croissance, la plupart des universités développent ou réfléchissent à un projet de publication sur internet. Différents formats de fichiers ont été adoptés : HTML, SGML, XML, PDF, PostScript, offrant différents conforts de recherche, de navigation et de visualisation. L'apport de tels projets est très important pour la communauté scientifique. Cela permet un accès plus facile à toute cette littérature (littérature grise), et lui permet d'être connue du grand public. Le projet CITHER en est à sa deuxième phase de développement, des problèmes de droit de diffusion ayant freiné son essor au début, il devrait connaître une montée en charge pour cette année. Ses principales caractéristiques sont une diffusion des fichiers au format Adobe PDF, une recherche en texte intégral, une navigation par liens hypertextes. Cette étude montre les nouveaux développements mis en place pour l'intégration des thèses rédigées sous LaTeX, et démarre une nouvelle réflexion sur les possibilités offertes par le langage XML.

Mots clefs
THESE, BIBLIOTHEQUE VIRTUELLE, DOCUMENT ELECTRONIQUE, SERVEUR, TEXTE INTEGRAL, INTERNET, MICROSOFT WORD, PDF, LATEX, XML

Abstract
Electronic thesis is an increasing research field. Most of the universities are currently developing a project to put their thesis on internet. Different file formats exist: HTML, SGML, XML, PDF, PostScript, providing different features to search, browse or view an electronic document. Such projects allow an easier access to scientific documents for the research community and the general public. The CITHER project (On-line full text thesis consultation) starts its 2nd phase. Copyright problems slowed down its development in the beginning, but it should reach a full production this year. The main characteristics are documents in the Acrobat PDF file format, full text search and hyperlinks browsing. This study shows the new developments with the LaTeX written thesis integration, and starts a reflection on the use of XML for project improvements.

Keywords
THESIS, VIRTUAL LIBRARY, ELECTRONIC DOCUMENT, SERVER, FULL TEXT, INTERNET, MICROSOFT WORD, PDF, LATEX, XML

I. Introduction

Avec le lancement du projet CITHER (**C**onsultation en texte **I**ntégral des **THè**se en **R**éseau), une première étude à été menée durant l'année 1997/1998, à la bibliothèque Doc'INSA, dépositaire des thèses soutenues à l'INSA de LYON. Elle a abouti à la mise en place d'une chaîne de conversion de documents électroniques pour la réalisation du serveur de thèses[1] [Huneau 98]

Ce projet constitue la suite de cette première étude, par l'extension des fonctionnalités de la chaîne de traitement, pour une montée en charge du serveur (conversion de fichiers sources en Latex, portabilité de la chaîne, etc.) et l'analyse de nouvelles technologies pouvant servir le projet CITHER. Une étude du langage XML a été menée sur ses possibilités en matière d'archivage, de publication et de sécurisation / authentification.

I. Présentation de l'existant

Le poste de conversion

Le poste de conversion se compose de l'ensemble Logiciels/Matériels suivant :

- Un PC sous Windows 95
- Un scanner
- Un graveur de CD-ROM pour l'archivage
- L'application Chaîne d'édition numérique (CEN)
- MS Office 97
- Adobe Acrobat 3

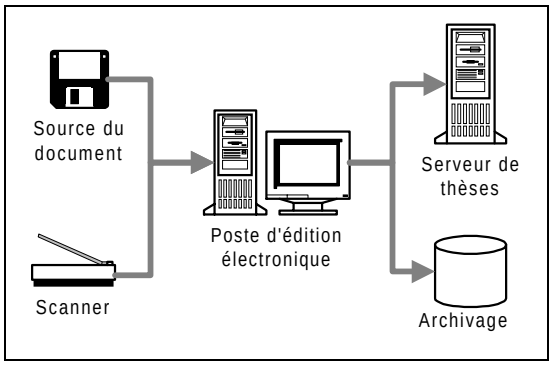

Figure 1 : Vue générale du dispositif

La chaîne d'édition numérique (CEN)

La chaîne d'édition numérique ou CEN est le logiciel développé lors de la précédente étude. Cette application, programmée sous Delphi 3 dans l'environnement Windows 32 bits, prend en charge le traitement des fichiers électroniques, du fichier source (au format Word 97) jusqu'à la publication sur le serveur.

Elle contrôle les autres applications via plusieurs mécanismes : MS Word et Acrobat Exchange sont pilotés via COM/OLE[2], alors que Acrobat Distiller est contrôlé par des messages Windows[3].

Le format de publication utilisé est le format propriétaire Adobe PDF[4].

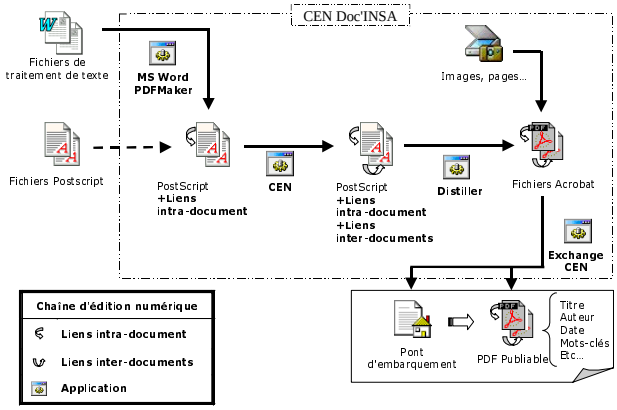

Figure 2 : Opérations de la chaîne d'édition

La conversion se déroule en quatre étapes :

- Tout d'abord, une macro-commande Word (Adobe PDFMaker [Adobe 98]) crée un fichier PostScript enrichi d'instructions *pdfmark*[5] [Adobe 97] à l'intention d'Acrobat Distiller. Cette macro-commande crée (le cas échéant) des liens à partir des champs 'note', 'table', etc. Elle crée également un repère Acrobat pour chaque titre (Liens intra-document).
- Les fichiers PostScript obtenus sont alors directement modifiés par l'application qui y ajoute des repères (toujours via *pdfmark*) désignant les autres fichiers (Liens inter-documents).
- Les fichiers PostScript sont ensuite convertis en PDF par Distiller.
- Enfin, les fichiers PDF sont 'retraités' à l'aide d'Exchange : leurs champs titre, sujet, auteur, etc. sont renseignés ; les miniatures de pages sont créées et les fichiers

[1] http://csidoc.insa-lyon.fr/these

[2] **C**ommon **O**bject **M**odel / **O**bject **L**ink **E**mbedding : modèle objet de Windows.

[3] Mécanisme de base de communication entre les entités de Windows

[4] PDF: **P**ortable **D**ocument **F**ormat. Format propriétaire développé par Adobe

[5] Opérateur du langage PostScript, destiné à Acrobat Distiller

optimisés pour une lecture en ligne (opération permettant au serveur d'envoyer le document page à page).

A ce point, le traitement par lot est terminé, et un rapport de conversion a été généré.

L'application génère en outre un "pont d'embarquement" vers la thèse, page HTML rassemblant la référence bibliographique du document et des liens vers tous les fichiers PDF. Enfin, elle peut préparer les fichiers à un archivage en les rassemblant dans un répertoire.

Le format PDF est un langage de représentation de page, impropre à l'archivage : Ne comprenant pas la notion de structure logique de document (paragraphes, titres, etc.), il ne peut efficacement servir de source à une éventuelle conversion vers un nouveau format. La solution actuelle d'archivage garde donc les fichiers PDF publiables et les documents sources (fournis par l'auteur et éventuellement retouchés sur le poste d'édition), pour permettre une évolution vers de nouveaux formats (SGML ou XML).

Par ailleurs, un guide de conversion sous forme de liste de contrôles permet à l'opérateur de se repérer dans les différentes phases de la conversion.

II. Les besoins du projet

De nouveaux besoins ont été définis par Doc'INSA avec l'arrivée au sein du projet d'autres universités (notamment Lyon I pour l'année 1999/2000) :

- Maintenance et évolutions du CEN, pour optimiser le temps de conversion d'un document, et corriger les problèmes existants.
- Etude de la portabilité de la chaîne d'édition numérique, pour permettre une installation facile sur de nouveaux postes de conversion.
- Extension des types de fichiers sources acceptés en entrée de chaîne, avec plus particulièrement l'intégration des fichiers sources en LaTeX.
- Réflexion sur les possibilités offertes par le langage XML comme format d'archivage ou de publication.

III. Maintenance de la chaîne d'édition

Plusieurs entretiens avec l'opérateur de conversion ont permis de définir les problèmes ou manques de l'application, notamment au niveau du guide opérateur.

Une mise à jour du guide a été effectuée, tenant compte de l'expérience acquise par l'opérateur et de ses astuces.

La correction et l'ajout de plusieurs fonctionnalités ont été implémentées :

- Fonction d'impression du rapport de conversion
- Définition de l'URL[6] en fonction du nom de l'auteur et de la date de soutenance
- Ajout automatique d'un nouveau lien dans les fichiers PDF pour revenir au pont d'embarquement.

IV. Portablilité de l'application

L'application CEN a été développée sous Delphi 3, en environnement Windows 95. Mais, jamais aucun test n'avait été effectué quant à sa portabilité sur d'autres machines, ou sur d'autres systèmes Windows 32 bits (Windows NT/98).

Une installation sur un poste Windows NT, et sur un nouveau poste de conversion équipé de Windows 98, mît en évidence certains problèmes :

- Clés manquantes dans la base de registre Windows pour l'interface COM/OLE des produits Acrobat.
- Fonctionnement perturbé par le déplacement des répertoires de travail

Une fois ces problèmes détectés, ils ont été résolus en modifiant la procédure d'installation et en corrigeant le code correspondant de l'application CEN.

V. Intégration de LaTeX

La part de thèses rédigées en LaTeX sur l'INSA est faible mais non négligeable[7], et avec l'arrivée de Lyon I dans le projet, elle va augmenter fortement.

Présentation de LaTeX

LaTeX est un traitement de texte particulièrement adapté à la rédaction de documents scientifiques et mathématiques, mais il sert aussi à écrire toutes sortes de documents, de la simple lettre, à des livres complets. Il est utilisé par beaucoup d'étudiants, de chercheurs et d'éditeurs à travers le monde.

LaTeX faisant partie du monde des logiciels libres, il est disponible sur la plupart des plates-

[6] URL à laquelle les fichiers seront transférés sur le serveur.

[7] 7% des thèses recensées lors d'une enquête de Novembre 1996 à Novembre 1997

formes informatiques, du PC au Mac, en passant par les systèmes Unix et VMS.

Dans le cadre du projet, il a été décidé d'étudier l'intégration du traitement des fichiers LaTeX à la chaîne d'édition numérique à partir d'une distribution Windows 32 bits.

Choix de la distribition

Plusieurs distributions existent pour Windows, proposant toutes un environnement complet (Miktex, Fptex, etc.). Fptex [Fptex 99] a été choisi pour les tests, pour son suivi des programmes en cours de développement, (notamment PdfTex, un programme de conversion de fichiers latex en PDF) et ses mises à jour régulières.

Une nouvelle chaîne de traitement

Une étude des différents programmes de conversion présents sous LaTeX à mis en évidence deux chaînes de traitement possibles :

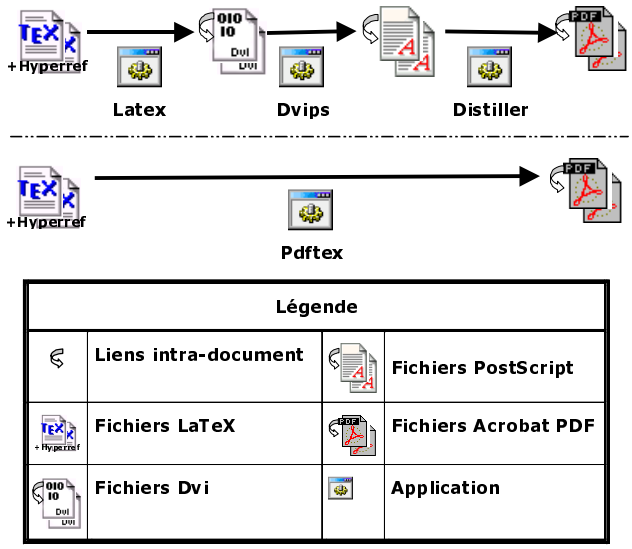

Figure 3 : Chaînes de traitement LaTeX

La première chaîne utilise le format de sortie traditionnel de LaTeX : le fichier dvi[8]. Ensuite, un premier programme, Dvips, convertit le fichier Dvi en fichier PostScript, et enfin le programme Distiller d'Adobe Acrobat, transforme le fichier PostScript en fichier PDF.

La deuxième chaîne est basée sur un nouveau programme, encore en cours de développement, PdfTex [Pdftex 99]. Pdftex remplace la compilation traditionnelle Latex, pour donner directement un fichier de sortie au format PDF, et non plus un fichier Dvi.

[8] DVI : Device Independent (indépendant du périphérique de sortie)

Dans les deux cas, l'intégration des liens intra-document, s'effectue par l'ajout du module Hyperref [Hyper 99]dans le préambule (en-tête) du fichier source Latex.

Ce module permet de définir au moyen de commandes Pdfmark les renvois aux notes, la table des matières dynamiques, etc., de la même manière que la macro-commande PDFMaker pour les fichiers Word. Il permet de plus une gestion des "back references" pour la bibliographie, en indiquant après chaque référence bibliographique les pages où elles ont été citées.

Ces commandes sont intégrées, pour la première chaîne, au fichier Dvi et PostScript puis interprétées par Distiller lors de la conversion au format PDF.

Pour la deuxième chaîne, ces commandes sont directement intégrées lors de la création du fichier PDF.

Comparaison des différentes chaînes

1ère chaîne	2ème chaîne
Utilisation de 3 programmes Latex, Dvips, Distiller	Un seul programme Pdftex
Programmes stabilisés, offrant un comportement sûr	Programme encore en phase de développement
Compatible avec tous les formats d'image utilisés sous LaTeX	Les fichiers d'image Eps ne sont encore pas reconnus

La chaîne de traitement basée sur Pdftex, permet une conversion plus simple et plus rapide, mais l'absence de reconnaissance du format Eps est un inconvénient majeur, ce type de fichier étant très utilisé par les utilisateurs Unix/Linux, principaux rédacteurs sous LaTeX.

Notre choix s'est donc porté sur la première chaîne de traitement présentée, comprenant l'utilisation successive des programmes latex, dvips, distiller.

Remarque à propos des références croisées

Pour une bonne gestion des références croisées sous Latex, il est nécessaire d'effectuer plusieurs passes (généralement deux). De plus, dans le cas d'un document contenant une bibliographie, comme c'est le cas pour une thèse, On doit faire appel à un autre

programme, Bibtex, pour les références à la bibliographie.

On obtient donc une chaîne de conversion faisant appel à 4 programmes différents (Latex, Bibtex, Dvips, Distiler) dont certain doivent être lancés plusieurs fois successivement (Latex, Bibtex).

L'utilisation d'un script Perl[9] Latexmk, résout ce problème, en s'assurant lui-même du bon enchaînement des programmes Latex, Bibtex et Dvips. Il ne reste plus qu'à lancer Distiller pour obtenir le fichier PDF.

Intégration au CEN

La chaîne de conversion LaTeX choisie, utilisant le programme Acrobat Distiler, comme la chaîne de conversion Word, elle peut être intégrée facilement à la chaîne existante.

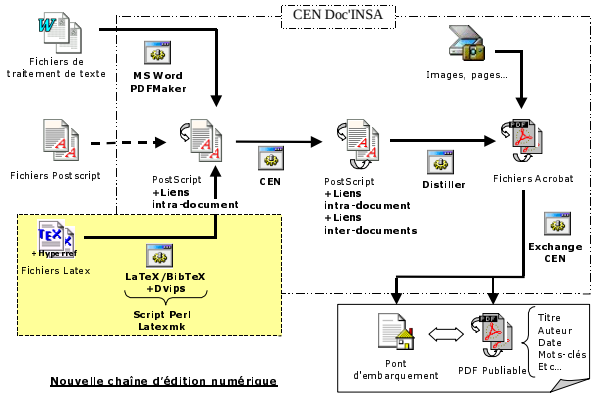

Figure 4 : Intégration à la chaîne existante

L'application CEN lance le script Perl qui s'occupe de la conversion des fichiers Latex en fichiers Postscript contenant les liens intra-document. A partir de là, on rejoint la chaîne existante qui poursuit la conversion par la création des liens inter-documents (dans le cas de plusieurs fichiers à traiter) puis des fichiers PDF par le Distiller et enfin, un retraitement et l'optimisation des fichiers avec Acrobat Exchange.

Remarque :

Le CEN prévoit le cas de thèses mixtes, où une partie du document serait développée sous LaTeX, et une autre sous MS Word (par exemple, la page de titre et certaines annexes en Word, et la thèse en Latex).

[9] Perl : Practical Extraction and Report Language. Langage de script très puissant développé par Larry Wall

VI. Evolution vers XML ?

Le langage XML

Le langage XML est un langage de balisage de document, comme HTML. Une recommandation du W3C[10] du 10 février 1998 définit XML dans sa version 1.0 [XML 98].

XML semble un format prometteur puisque tous les grands acteurs du monde informatique le soutiennent (Oracle, IBM, SUN, Microsoft, etc.).

Ses principales caractéristiques sont :

- Un ensemble extensible de balises (contrairement à HTML)
- Une séparation entre la présentation et les données.
- Un codage de caractère en UNICODE – ISO 10646
- Une bonne adaptation à la diffusion sur internet.

Un document XML s'accompagne généralement d'une DTD (Document Type Définition) qui permet de définir la structure du document XML, et de le valider.

Les développements liés

D'autre développements sont en cours, liés au langage XML.

XSL [XSL 99] est en langage pour définir des feuilles de style, il est composé de deux parties :

- un langage de transformation de documents XML permettant, à partir d'un document source XML, de produire un document cible XML composé de nouveaux éléments et / ou d'éléments présents dans le document source.
- un langage permettant de spécifier de manière très précise la présentation des données.

Une feuille de style XSL spécifie la présentation d'une classe de documents XML en décrivant comment une instance de cette classe est transformée en un autre document XML utilisant le langage de présentation des données.

Il est donc possible, à partir d'un fichier pivot, de dériver plusieurs versions adaptées au périphériques de sortie, en définissant les feuilles de style appropriées. (par exemple : impression sur papier A4, affichage à l'écran, etc.)

[10] W3C :World Wide Web Consortium

Enfin, XLL [XLL 99] (XML Linking Language) définit les liens hypertextes au sein du document XML. Une distinction est faite entre les liens externes, et les liens internes pointant sur des documents XML. Un lien est une relation explicite entre au moins deux données ou ensemble de données.

CITHER et XML

Une première étape consiste à vérifier l'existence de DTD permettant la définition de longs documents structurés. Plusieurs sont disponibles sur internet :

- The Book DTD – ISO 12083
- La DTD du TEI (Text Encoding Initiative) ou sa version simplifiée TEI lite [TEI 99]

Ces DTD ont été reprises de SGML, elles sont très complètes, et permettent la définition d'une thèse.

Il semble donc possible de définir un fichier au format XML, ne contenant que les données, la structure logique (Titre, auteur, texte, citation) et d'utiliser ce fichier pour l'archivage et la génération des différents formats pour la publication (HTML, PDF, XML).

Le problème réside dans l'obtention d'un tel fichier à partir des documents sources fournis par les doctorants (fichiers Word ou Latex)

Les outils de conversion (pour la génération du fichier pivot XML et ensuite sa dérivation en plusieurs autre formats de sortie) n'étant pas encore complètement disponibles, il a été jugé que l'on n'obtiendrait pas la qualité offerte par les fichiers PDF et qu'il valait mieux attendre l'aboutissement de toutes les normes liées à XML (XSL, XLL, etc.) et l'arrivée d'outils de conversion et de visualisation.

Remarque

Plusieurs d'universités, (l'université Laval au Québec, Les Presses universitaires de Montréal, l'université Lyon II, etc.) se sont associées pour le développent d'une chaîne de conversion de thèses autour des langages SGML[11]/XML [LyonII 99]. Dans le cadre de la réflexion sur l'évolution du projet CITHER, plusieurs réunions ont été tenues avec ce groupe d'universités pour la présentation de la chaîne, et son test avec une thèse scientifique de l'INSA. Les résultats sont encore insuffisants dans l'état actuel du projet, mais une collaboration est envisageable.

[11] Langage de balisage définit en 1986, à l'origine d'HTML et d'XML.

VII. Conclusion

Le projet CITHER offre maintenant une chaîne de traitement plus complète pour la publication électronique des thèses de l'INSA.

Plus d'une vingtaine de thèses ont déjà été converties, et la chaîne semble maintenant prête à une montée en charge.

Mais le projet ne s'arrête pas là, un groupe de travail à été formé sous l'impulsion de Doc'INSA, pour étudier la mise en place de feuilles de style propres aux thèses pour aider les étudiants dans leur rédaction et pour faciliter les traitements de conversion. Une formation à la rédaction de longs documents structurés va être instaurée.

Un rapprochement avec d'autres projets est aussi envisagé :

- Pour le test de nouvelles chaînes de conversion,
- Le partage de connaissances,
- La mise en commun des méta-données[12],
- La constitution d'un catalogue des thèses électroniques.

Avec l'arrivée de MS Office 2000 sur le marché et la sortie de la version 4.0 d'Adobe Acrobat, les prochains développements du projet vont consister à étudier leur intégration au CEN. La veille technologique autour de XML va être maintenue pour permettre l'évolution du projet vers ce langage dès que cela sera possible.

[12] Les meta-données sont les informations portant sur le document (nom de l'auteur, année, laboratoire, résumé, abstract, mots-clés, etc.)

Références bibliographiques

[Hyper 99], *Hypertext marks in LaTeX: the hyperref package*, [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://tug.org/applications/hyperref/manual.html>

[TIE 99], *Text Encoding Initiative* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www-tei.uic.edu/orgs/tei/>

[Pdftex 99] *PDFTeX support* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.tug.org/applications/pdftex/>

[LaTeX 99] *Une courte (?) introduction à LaTeX* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:ftp://ctan.tug.org/tex-archive/info/lshort/french/flshort-3.3.pdf>

[Fptex 99] fpTeX 0.3 User's manual [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:ftp://ftp.loria.fr/tex-archive/systems/win32/fptex/fptex.pdf>

[XML 98] Extensible Markup Language (XML) 1.0 [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.w3.org/TR/1998/REC-xml-19980210>

[XSL 99] Extensible Stylesheet Language (XSL) working draft [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.w3.org/TR/WD-xsl/>

[XLL 98] XML Linking Language (XLink) working draft [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.w3.org/TR/1998/WD-xlink-19980303>

[Huneau 98] Huneau M.E., "*Serveur de thèses en texte intégral : Rapport de Projet de Fin d'Etudes*" [On-line]. Villeurbanne (Fr.) : INSA – IF, 1998, 29 p. Available from internet : <URL: http://csidoc.insa-lyon.fr/these/doc/rapport_pfe.pdf>

[Adobe 97] **Adobe Developer Support**, *Acrobat Distiller Control Interface Specification*. Adobe, July 1997, Technical Note #5158

[Adobe 98] *Adobe PDFMaker 1.0 for Microsoft Word 97* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.adobe.com/supportservice/custsupport/LIBRARY/4d9e.htm>

[LyonII 99] Service des Nouvelles Technologies pour l'Information Et la Réalisation de Serveurs (SENTIERS) [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : http://phebus.univ-lyon2.fr/sentiers/