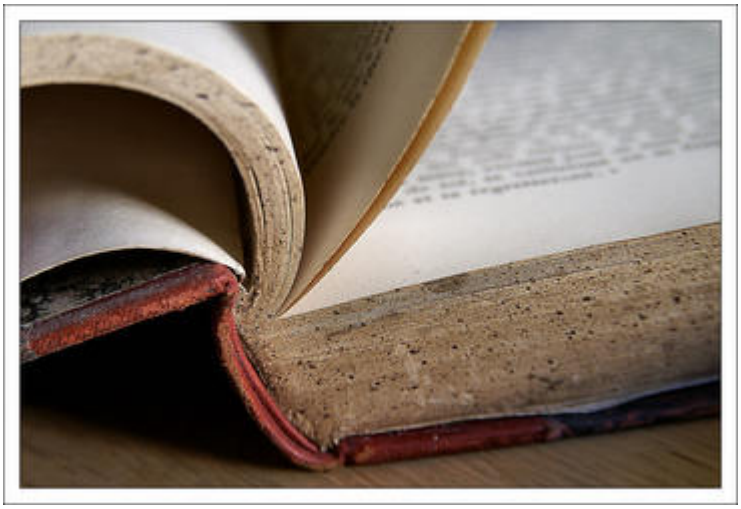
Clubic vous emmène dans une usine «tourne pages»

SCANNER

L'usine que nous avons visitée se trouve dans la campagne Alsacienne, et elle est souvent désignée sous le nom de « tourne pages ». De quoi s'agit-il ? Rien de moins que d'un pôle de numérisation d'ouvrages à grande échelle, mis en place pour un projet portant sur la numérisation de 80 000 ouvrages en trois ans, soit un projet parmi les plus ambitieux jamais menés. À l'heure où les Google, Yahoo et autres MSN occupent l'espace médiatique à coup d'annonces de lancement de vastes projets de numérisation, qu'en est-il des technologies et de la gestion de projet nécessaires pour mener à bien de telles entreprises ? Cette visite dans l'usine d'Infotechnique, à La Walk, nous permet de nous en faire une bonne idée et de découvrir toutes les étapes en détail. On y découvre un scanner tourne pages, unique en son genre… Pour bien apprécier la visite, et s'y retrouver dans les enjeux, nous vous proposons quelques pages d'introduction sur les différents acteurs déclarés de cette course (guerre ?) du contenu.



Visite d'une usine « tourne pages »

Est-ce la fin du papier ? Faut-il tout numériser ? Google fait-il main basse sur la culture ? Y a-t-il urgence à mettre en place une riposte européenne ? Ces différentes questions ressurgissent de plus en plus fréquemment dans l'actualité, qu'elle soit grand public, politique ou orientée nouvelles technologies. Il y a matière donc à faire un « point » sur la situation, tant les données et les enjeux nous dépassent – on parle numérisation sur 300 ans pour Google Print ! – et en même temps nous touchent de près : si nos archives sont perdues ou se dégradent au point de devenir illisibles, quelle connaissance garderons-nous de notre histoire ?

Les vastes projets de numérisation dont la presse débat prennent par ailleurs trop souvent l'allure d'un mauvais film : « Le méchant Google américain contre la gentille Bibliothèque Européenne ». Mais la question n'est pas aussi simpliste. Il y a deux mois, Yahoo faisait part de sa volonté de se jeter dans la bataille, et un consortium américain destiné à contrer l'initiative Google Print était mis en place. Plus récemment encore, c'est Microsoft qui, par l'intermédiaire de son moteur de recherches MSN, informe qu'il se lance dans la bataille. La culture n'a peut-être jamais suscité autant de passions, voire de conversions ! Mais au-delà des enjeux culturels et financiers, cet article se propose d'éclaircir les questions trop souvent laissées dans l'ombre : celle des moyens techniques qui sous-tendent ces projets de numérisation à grande échelle. Une usine unique en son genre nous a ouvert ses portes pour nous présenter un projet atypique, celui de la numérisation de 80 000 ouvrages. De la logistique à la technique, nous avons vu beaucoup, et posé tout autant de questions. Cette usine préfigure-t-elle celles qui seront implantées lorsque le projet de Bibliothèque numérique européenne sera officiellement lancé ? On le dirait bien, alors en route pour la visite !

Les livres : or noir du XXIe siècle ?

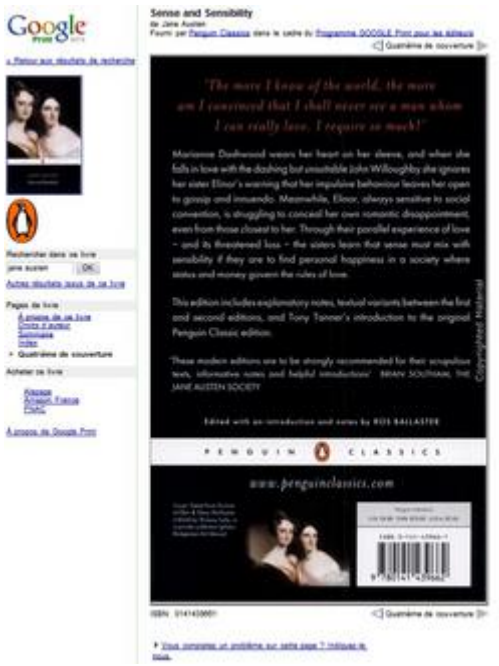
Quel est le point commun entre Google, Yahoo, MSN / Microsoft, la Bibliothèque Nationale de France (BNF) et le projet de Bibliothèque Numérique Européenne (BNE) ? Le contenu ! Chacun à sa façon s'intéresse de près – voire convoite -, les ouvrages de nos bibliothèques. Après avoir pris la poussière pendant des années dans l'indifférence la plus générale, après avoir été malmenés au fond de sacoches d'étudiants peu attentifs à les rapporter dans les délais, voilà que ces ouvrages parfois « mal-aimés » se retrouvent sous les projecteurs. D'où vient ce revirement ?

Il y a ceux qui conservent…

Comme nous autres, les livres ont une durée de vie limitée, susceptible d'être prolongée lorsque les conditions de stockage sont idéales (à l'abri de l'humidité et de la lumière notamment). Mais même les meilleures stratégies de conservation ne pourront empêcher le vieillissement naturel des ouvrages et leur dégradation. L'exemple le plus marquant est celui des journaux. Un journal, qu'il soit quotidien ou hebdomadaire, est fait pour être lu et sitôt « jeté » (dans le conteneur papier s'il vous plait), pas pour durer 100 ans ou plus. Ceux qui s'intéressent à la numérisation de ces ouvrages ne le font pas uniquement pour leur valeur documentaire, mais dans un premier temps pour assurer leur sauvegarde.

… et ceux qui exploitent

Qu'achète t'on lorsque l'on achète un ouvrage : du papier ou de l'information ? Et comment sont alimentés les moteurs de recherche ? Le contenu, qui semble presque être devenu l'or noir du XXIe siècle, est la réponse commune à ces deux questions. La problématique de Google notamment est bien celle-ci : « Acquérir » du contenu, pour le commercialiser d'une part (Programme Google Editor) et l'inclure dans les données accessibles depuis son moteur de recherches (Google Print). Le contenu devient une valeur marchande, et sa conservation par le biais de la numérisation, même si elle est bien réelle, est plutôt secondaire.



Résultat d'une recherche Google Print, avec des liens vers les sites marchands partenaires.

Cette partition un peu simpliste peut être affinée, mais elle permet en attendant de situer les enjeux et les motivations des différents acteurs de ces projets de numérisation.

Qui sont les acteurs de la numérisation ?

La « liste » que nous vous présentons ne vise pas l'exhaustivité ! Bien des projets de numérisation ont déjà été menés à bien ou sont en cours, loin des sirènes de la médiatisation. Les acteurs signalés ici sont ainsi ceux qui sont sous les feux de l'actualité.

Amazon

En août 2005, Amazon France a fait part de son projet baptisé Search Inside (« Chercher au Cœur »). Le but de son service est de permettre aux internautes d'effectuer une recherche dans le contenu numérisé, soit quelques 120 000 livres, dont 5 000 en français. Amazon recherche des partenariats avec des auteurs et des éditeurs. Il a dores à déjà conclu un partenariat avec 12 maisons d'édition françaises (Dargaud, La Découverte, Le Petit Futé, Ellipses, etc.). Cette fonction de recherche qu'il propose est déjà accessible depuis les sites d'Amazon aux États-unis, en France, en Allemagne, au Canada et au Royaume-Uni.
Le 4 novembre 2005, le groupe américain a présenté deux nouveaux programmes qui s'inscrivent dans ce projet :

- *Amazon Pages* qui prévoit de permettre d'acheter « simplement et à peu de frais » des parties d'ouvrages (pages, sections ou chapitres), sélectionnées d'après une recherche par mots-clés.
- *Amazon Upgrade* qui permet quant à lui, pour tout livre acheté, de bénéficier en supplément d'une version numérisée de l'ouvrage.

Un livre sur deux acheté sur Amazon.com aux États-unis le serait par le biais de ce programme. L'ambition d'Amazon est ainsi, selon Jeff Bezos son PDG, de collaborer avec les éditeurs pour « faire en sorte que les livres du monde entier soient accessibles instantanément n'importe quand, n'importe où».

Pour la recherche suivante.

Ce programme « Chercher au cœur » donne accès à de nombreuses informations : les phrases les plus récurrentes (tous ouvrages confondus), les termes récurrents, des statistiques (les 100 mots les plus utilisés, la lisibilité, la complexité…), etc. L'option la plus intéressante est celle de recherche par mot-clé (ici « heart »).

Les résultats de la recherche pour « heart ».

À partir de cette liste, il est possible d'afficher la page qui contient le terme en question, ainsi que les deux suivantes et les deux précédentes.

La Bibliothèque Nationale de France

La Bibliothèque Nationale de France (BNF) travaille depuis longtemps sur un projet de numérisation d'ouvrages de grande envergure : Gallica. Cette base de données contient quelques 70 000 ouvrages numérisés, 80 000 images et plusieurs heures de bandes audio. Toutes ces données sont accessibles gratuitement depuis cette page d'accueil.

L'ambition de Gallica est de constituer une bibliothèque patrimoniale et encyclopédique.

Les ouvrages contenus sont accessibles soit en mode image (conservation de l'aspect visuel du document original), soit en mode texte, permettant la recherche. Les ouvrages et illustrations que propose le moteur sont issus des fonds de la BNF ou de collections extérieures (livres illustrés de la Bibliothèque du Musée de l'Homme et de la Bibliothèque du Muséum national d'Histoire naturelle, fonds d'archives photographiques de la Médiathèque du patrimoine et de l'architecture, etc.). La BNF est dirigée par Jean-Noël Jeanneney.

La Bibliothèque Numérique Européenne

Pour le moment, ce dernier projet n'a guère dépassé le stade des déclarations d'intention. L'idée de Bibliothèque Numérique Européenne (BNE) a été proposée par Jacques Chirac à la suite des premières déclarations de Google. Le projet est dirigé par Viviane Reding, la Commissaire en charge du projet de la bibliothèque européenne virtuelle, à la tête d'un comité constitué de 40 personnes issues des secteurs de l'édition, des bibliothèques et de l'université. Vingt bibliothèques européennes ont dores et déjà adhéré au projet : Allemagne, Autriche, Belgique, Danemark, Espagne, Estonie, Finlande, Grèce, Hongrie, Irlande, Italie, Rome et Florence, Lituanie, Luxembourg, Pays-Bas, Pologne, République tchèque, Slovénie, Slovaquie et Suède.

Ce qui motive la « riposte » européenne au projet Google est bien résumé par cette phrase de Jean-Noël Jeanneney : « Les critères de sélection de l'éditeur numérique Google seront puissamment marqués car toute entreprise de ce genre implique des choix drastiques. S'affirme ainsi le risque d'une domination écrasante de l'Amérique dans la définition de l'idée que les prochaines générations se feront du monde ».

4digitalbooks

Au même titre que Kirtas Technologie, 4digitalbooks est un fabricant de scanner « tourne pages ». Le Ditizing Line qu'il propose a été réalisé avec la société I2S, chargée de la partie optique. Six de ces scanners sont utilisés dans le monde : quatre d'eux sont chez Infotechnique en Alsace, un autre à la bibliothèque universitaire de Stanford en Californie et le dernier dans celle de Southampton, en Angleterre. Les scanners « tourne pages » conviennent aussi bien pour les livres anciens que récents, quelle que soit leur taille et leur texture. La cadence est de 1 500 pages à l'heure et le Ditizing Line est commercialisé à environ 300 000 euros.

Kirtas Technologie

Kirtas Technologie n'est pas un énième concurrent ayant la volonté de se lancer dans la bataille de la numérisation. Kirtas est en fait uniquement un industriel fabriquant des scanners, et le premier à avoir mis au point des modèles qui tournent les pages automatiquement. Mais Kirtas s'est récemment fait connaître plus largement, en se posant de façon un peu exagérée comme un concurrent à Google Print. Ce qui est sûr, c'est que cet industriel vise le marché européen.

Les « Bookscan » qu'il propose ont comme atout d'être mobiles et plutôt légers (84 x 76 x 122 cm pour 77 kilos). Les bibliothèques pourront s'en équiper, ce qui n'est pas le cas des DigitalBook. Mais la contrepartie de cette compacité est que la machine est limitée aux formats compris entre le 11 x 17 et 27 x 35 cm. Pour la numérisation, l'ouvrage est placé dans un berceau. La machine se charge de tourner les pages tandis qu'un système de miroir permet de scanner les pages de droite et de gauche sans avoir à bouger la caméra digitale. Le Bookscan numérise en 300 ou 600 dpi, restitue

des fichiers Tiff, Jpeg ou PDF et dispose d'un système de reconnaissance optique pour 177 langues. Chaque Bookscan est commercialisé autour de 120 000 euros, et la cadence annoncée est de 1 200 pages numérisées à l'heure. Actuellement, plus de 20 Bookscan ont déjà été vendus.

Le Bookscan de Kirtas.

Google

Google a été un des premiers à se lancer dans la bataille de la numérisation. Ses premières déclarations remontent à 2003. Le projet Google nommé « Océan » se distingue par son ampleur et ses ambitions. Les annonces font ainsi état d'une volonté de numérisation de 14 millions d'ouvrages. Des accords ont été conclus avec des universités réputées comme Harvard, Stanford, celle du Michigan, d'Oxford et de New York.
On peut être pris de vertige devant l'ampleur du projet annoncé, dont la réalisation est évaluée à quelques 300 ans ! Le projet Google est double et s'articule autour de deux entités : Google Print et Google Editor.

Google Print
Ce service s'adresse aux bibliothèques. Les bibliothèques confient à Google les livres qu'elles souhaitent voir numériser. Elles reçoivent en retour un double numérique de leurs ouvrages. Elles peuvent ensuite l'utiliser librement, à condition de ne pas en faire d'utilisation commerciale.

Google Editor
Ce service s'adresse aux éditeurs détenteurs d'un numéro ISBN. Elle les invite à signer un contrat, puis à envoyer leurs livres pour qu'ils soient scannés et mis en ligne. Le moteur de recherche donne ensuite accès à un résumé (4e de couverture, biographie de l'auteur ou introduction) et à une portion de l'ouvrage comprise entre 20 et 100 % du contenu, en fonction de la volonté de l'éditeur. La page de recherche affiche également un lien vers le site de l'éditeur, et Google prévoit une rémunération liée aux revenus publicitaires générés par la présence d'encarts publicitaires. Pour tous ces ouvrages, Google bloque les fonctions « imprimer », « couper », « copier » et « enregistrer » pour les pages consultées.

De nombreuses erreurs et abus (numérisation et mise en ligne d'ouvrages sans accord préalable des ayant droits) ont conduit ces derniers mois Google devant les tribunaux. À présent, les éditeurs ont la possibilité de faire retirer (pour les ouvrages déjà numérisés) ou d'exclure (pour ceux qui ne sont pas déjà) du programme les ouvrages de leur choix. Les clauses du contrat que les lie à Google Print ainsi que les questions les plus courantes peuvent être consultées depuis cette page.

Même si Google se montre plutôt informatif concernant le cadre juridique de son programme, il reste malgré tout difficile de s'y retrouver ! La « machine Google Print » est complexe, et répond à un double objectif :

- Générer du trafic via le moteur de recherche en augmentant le nombre des pages disponibles.
- Générer des revenus grâces aux liens sponsorisés présents sur les pages de consultation des ouvrages. Ces revenus sont partagés avec les éditeurs, lorsque ceux-ci adhérent au programme Google Print. Mais il n'est pas évident d'aller plus loin dans le détail sur cette partie. Google ne divulgue pas la répartition exacte des gains, et les informations de type « rapport » (taux de clics et revenu total) communiquées aux éditeurs sont confidentielles.

Rappel chronologique

- *30 mai 2005*

Le projet Google Print en version Bêta est officiellement ouvert.

- *12 août 2005*

Google Print marque une pause. Google se heurte à plusieurs groupes d'auteurs et d'éditeurs lui reprochant de violer la loi sur le droit d'auteur en numérisant les pages de certains de leurs ouvrages sans véritable autorisation préalable.

- *21 septembre 2005*

Google Print fait l'objet de poursuites judiciaires de la part de « L'Authors Guild » (association qui représente plus de 8 000 auteurs aux États-Unis). D'après Nick Taylor, responsable de l'Authors Guild, « Google Print est une véritable et robuste violation des droits d'auteurs. Ce n'est ni à Google, ni à d'autres personnes de décider si tel ou tel ouvrage peut être copié. Ce droit revient aux auteurs et aux possesseurs des droits d'auteurs ». Une seconde plainte sera déposée quelques temps plus tard devant le tribunal fédéral de Manhattan, et pour les mêmes raisons, par « l'Association of American Publishers » (AAA).

- *19 octobre 2005*

À l'occasion de la foire du livre de Francfort, Google annonce l'ouverture de son service en Europe. Il existe désormais huit versions pour chacun des pays suivants : Italie, Allemagne, Hollande, Autriche, Suisse, Belgique, Espagne et France.

- *1er novembre 2005*

La numérisation reprend après presque de trois mois de pause.

MSN / Microsoft

Le moteur de recherche de Microsoft se lance à son tour dans la bataille. Le 25 septembre dernier, la firme de Redmond a dévoilé son projet **MSN Book Search**. Elle rejoint ainsi l'OCA (Open Content Alliance), dont fait déjà partie son rival Yahoo, en apportant quelques 5 millions de dollars susceptibles de servir à la numérisation de 150 000 livres. La version Bêta de MSN Book Search est attendue pour début 2006. Elle donnera accès aux contenus tombés dans le domaine public, dans leur intégralité et gratuitement. Concernant les livres protégés par le droit d'auteur, Microsoft réfléchit à la façon de facturer leur consultation.

Le 4 novembre dernier, Microsoft a indiqué avoir conclu un accord direct avec la British Library pour numériser l'équivalent de 100 000 ouvrages. Cet accord permet à cette dernière d'accélérer son programme initié en juin, et qui prévoit la mise à disposition de nombreux contenus : CD-ROM, journaux, livres numérisés, etc. Microsoft envisage cet accord qui implique un investissement de 2,5 millions de dollars comme un investissement à long terme. Il compte d'une part sur un partenariat durable avec la British Library, tandis qu'il envisage d'autre part de proposer à d'autres entreprises, les technologies développées pour ce projet (numérisation, classement, gestion de contenu, etc.).

Open Content Alliance

L'Open Content Alliance (OCA) est une association à but non lucratif basée à San Francisco, mise en place à l'initiative d'institutions culturelles, d'industriels et d'organismes gouvernementaux. L' OCA regroupe des bibliothèques, des archives et des maisons d'éditions autour d'un objectif commun :
la construction d'une immense bibliothèque numérique à même de contrer le projet Google Print. Cette bibliothèque fournira un accès libre aux ouvrages numérisés et permettra d'effectuer des recherches sur l'ensemble des contenus, et de les télécharger gratuitement. Au commencement du projet, le contenu (constitué aussi bien de livres que de sons ou de vidéos), sera fourni par les universités de Californie et de Toronto, le National Archives du Royaume-Uni, l'éditeur O'Reilly Media et l'European Archive. Adobe, HP et Yahoo prendront en charge l'aspect technique du projet et feront éventuellement des contributions financières.

Cette bibliothèque virtuelle contiendra des œuvres du domaine public mais aussi certaines soumises aux droits d'auteurs, numérisées avec l'accord des ayants-droit. Les 18 000 premiers ouvrages correspondant à la sélection de l'université de Californie devraient être mis en ligne vers la fin 2006. Ils seront hébergés par l'association Internet Archive dont la vocation est constituer une bibliothèque de sites Internet. Les projets MSN et Yahoo! s'inscrivent dans le cadre de l'OCA.

Yahoo!

Le projet de Yahoo! s'inscrit dans le cadre défini par l'Open Content Alliance (OCA). Yahoo assure la réalisation du moteur de recherche du site de l'OCA. Le contenu de cette bibliothèque numérique sera par ailleurs accessible par le moteur de recherche Yahoo!

Une démarche en 4 étapes

Numériser, c'est bien, mais cela ne suffit pas ! C'est même loin d'être la toute première étape d'un vrai projet de management de contenu. D'autre part, ce n'est pas la numérisation proprement dite qui cristallise les enjeux, mais l'indexation. Cet aspect est bien souligné par Patrick Bazin, directeur de la bibliothèque de Lyon, dans un entretien accordé à *Libération (12/11/05)* : « L'indexation représente un enjeu géostratégique. [...] C'est très bien que les Américains montrent la voie mais il ne faut pas qu'ils soient les seuls. La prospérité intellectuelle dans dix ou quinze ans dépendra de la capacité à maîtriser les connaissances. Or on prend du retard en France depuis dix ans. [...] L'enjeu est le mode texte, qui permet à tout un chacun d'avoir accès à un extrait. Or c'est là-dessus que les moteurs parient : arriver par mots clés à des séquences de textes. »

- 1 - L'inventaire

L'inventaire est l'étape préalable à celle de numérisation. Il est nécessaire d'informatiser tous les catalogues des différentes bibliothèques concernées par un même projet de numérisation. Cette étape permet de connaître le nombre précis d'ouvrages, et surtout de s'organiser afin d'éviter d'éventuelles numérisations en double.

- 2 – La numérisation

L'étape suivante est celle de la dématérialisation des documents textes et images. On quitte le support papier au moyen de scanners automatiques ou manuels, de saisie et de lecture automatique (OCR, ICR).

La numérisation n'est que la deuxième étape.

- 3 – L'organisation du contenu

Les données sont ensuite converties dans un format précis en fonction de l'utilisation qui en sera faite : PDF pour la transmission, HTML pour l'affichage web, XML et SMGL... Cette étape est également celle où l'on ajoute les méta-données documentaires.

- 4 – La diffusion et l'organisation

La dernière étape est celle de la diffusion et de la conservation, avec la création de communications multimédia (off-line et on-line), et l'hébergement de données et de copies de sauvegarde.

En route pour la campagne alsacienne

Pour parvenir jusqu'à l'usine « tourne-pages » d'Infotechnique, il faut quitter Strasbourg et emprunter quelques routes qui zigzaguent dans la campagne alsacienne. On s'arrête ensuite dans le petit village de La Walck, devant les bâtiments construits par le Conseil Général du Bas-Rhin et qui regroupe des entreprises relevant du pôle de compétences Technologies de l'information et la communication. Infotechnique est une filiale de Getronics. Une vingtaine de personnes travaillent dans l'usine de La Walck, sur le projet Amalfi (« Alsace-Moselle, Application pour un Livre Foncier Informatisé ») confié à Infotechnique par le Gilfam (Groupement pour l'Information du Livre Foncier d'Alsace-Moselle). Le commanditaire du projet est le Ministère de la Justice, et les bâtiments sont placés sous son contrôle. Vous le verrez au fil de la visite, les précieuses archives qui transitent par l'usine le temps de leur numérisation sont bien surveillées !

Nous voici arrivés.

80 000 ouvrages à numériser

Un marché de 60 000 000 euros, un investissement de 4 000 000 euros pour mettre au point l'usine de La Walck, un partenariat avec IBM pour la partie reprise de données : voilà ce qu'a nécessité ce projet portant sur la reprise des 40 000 ouvrages du Livre Foncier d'Alsace-Moselle, équivalent des Hypothèques dans les autres départements français. Vingt-trois des 60 000 000 euros du budget reviennent à Infotechnique pour la numérisation. Chaque ouvrage est numérisé en deux temps, ce qui porte à l'issue du projet à 80 000 le nombre d'ouvrages traités. Ces ouvrages sont consultés quotidiennement par de nombreux professionnels (notaires, géomètres, collectivités locales, etc.) et constamment mis à jour. Le projet de numérisation qui les concerne s'explique en partie par le fait que leur volume a « explosé » en même temps que le marché immobilier dans les années 80 et 90, et que leur gestion s'est compliquée. Derrière l'aspect numérisation et reprise de données structurées manuscrites, le projet Amalfi a pour but de proposer une base de données et des outils qui permettent la consultation des données à distance, et de permettre un gain en fiabilité et en performances. Plus généralement, ce projet est régi par une loi de 1924 modifiée en 2002 qui consacre la valeur informatique de ces documents. Ces versions numérisées auront la même valeur que leurs équivalents papier, et Amalfi sera ainsi un test grandeur nature d'une loi de mars 2000 sur la validité de la signature électronique.

Amalfi est un test grandeur nature de la loi sur la signature électronique

Suivons le cheminement des ouvrages

Pour rythmer cette visite, nous allons suivre le circuit qu'empruntent les ouvrages candidats à la numérisation. Nous commençons donc par l'entrepôt par lequel arrivent les camionnettes chargées des précieux registres. Deux d'entre elles, qui viennent de décharger les ouvrages, sont garées à l'extérieur tandis que d'autres, embarquant du matériel pour une numérisation « à domicile », sont entrées dans le bâtiment. Tant que le projet est en cours et que l'application destinée à la consultation des données n'est pas achevée, ce matériel mobile permet de se rendre dans les différents Bureaux afin de numériser les nouvelles pages. Des archives vivantes demandent en effet des mises à jour régulières !

Plusieurs camionnettes servent à l'acheminement des ouvrages.

Plus loin, à l'intérieur des locaux, une carte permet de suivre leur cheminement en temps réel au moyen d'un système GPS.

Quelques heures plus tôt, dans les Bureaux Fonciers…

Les ouvrages sont bien là, à l'abri dans leurs caissons, mais le travail a en fait commencé quelque peu en amont, au sein des quelques 46 Bureaux Fonciers géographiquement répartis dans trois départements. Équipées d'une mallette digne de James Bond, deux personnes se rendent dans les Bureaux Fonciers une semaine avant la sortie des ouvrages. Elles sont chargées d'effectuer un premier inventaire des archives et d'évaluer leur état (le degré de conservation est-il suffisant pour résister à un traitement automatique ?). Une étiquette portant un code unique est apposée à l'intérieur de l'ouvrage. Elle permet de l'identifier, de conserver des indications sur l'état de l'ouvrage et de le localiser. Un convoyeur passe ensuite pour amener les ouvrages jusqu'à l'usine de La Walck, où elles seront conservées pendant cinq jours.

Au moyen des outils contenus dans ce « nécessaire high-tech »,
on identifie chacun des volumes, et on leur appose une étiquette unique.

Les registres voyagent en caissons protecteurs…

Les registres sont transportés dans des caissons conçus spécialement pour le projet, et similaires à ceux utilisés pour le transport en avion. Ils répondent à un cahier des charges précis, exigeant la résistance au feu, à l'eau et aux chocs. Les volumes n'en sortent que le temps nécessaire à leur numérisation. Les sas dans lesquels sont rangés les caissons sont numérotés « 1, 2, 3, 4 et 5 » pour « Lundi, mardi, mercredi, jeudi et vendredi ». Une des contraintes du projet est d'assurer une numérisation rapide, permettant un prompt retour des registres dans leurs Bureaux d'origine où ils sont attendus pour le travail quotidien. Chaque volume emprunté doit être restitué au bout de cinq jours. Pendant ce laps de temps, les volumes sont numérisés, les données envoyées à Madagascar pour y être structurées et saisies tandis qu'un contrôle qualité pour la « partie métier » se charge ensuite de vérifier certains points.

De tels portails jalonnent les locaux pour suivre les registres. Sur la seconde vue, les documents sont stockés en attente de leur traitement et de leur retour.

lorsque la numérisation s'est passée sans problème, rouge lorsqu'une anomalie est survenue.

… et les données par satellite

Mais n'anticipons pas sur les étapes à venir ! Il suffit de savoir qu'une organisation rigoureuse a du être mise en place pour permettre un traitement et un renvoi aussi rapide des registres à leurs propriétaires. Là où l'entreprise a gagné un temps considérable, c'est en parvenant à compresser suffisamment les données pour permettre leur envoi par satellite. À l'origine, la solution envisagée pour acheminer les données numérisées vers les ateliers de ressaisie situés dans l'Océan Indien était celle du transport par avion, et elle nécessitait un minimum de deux jours.

Les données sont stockées sur place pendant la durée du projet…

… elles sont ensuite envoyées par satellite pour l'étape de structuration.

Les images issues de la numérisation sont aussitôt stockées sur les serveurs et la sécurité est assurée par des serveurs de backup. Les images fournies par l'atelier de numérisation pour l'envoi sont au format LDF (LuraDocumentFile) qui permet de diviser par dix la taille des fichiers. Une double page LDF du Livre Foncier (au format A2) pèse entre 70 et 150 Ko. Elle serait ainsi comprise entre 700 et 1,5 Mo au format standard Tiff. Le LDF permet de conserver la qualité requise pour une bonne lecture des documents et permet d'optimiser l'affichage de ceux qui intègrent texte et image.
Ce format sépare les données en trois couches :

- Le texte en bi-tons est traité selon le standard CCITT Groupe 4 (soit le format utilisé pour les télécopies).
- Le fond est une image au format Jpeg 2000.
- Le texte couleur est également traité en Jpeg 2000.

Ce tour d'horizon effectué, il ne nous reste plus qu'à pousser les portes suivantes pour nous retrouver au cœur de l'atelier de numérisation !

Dans l'atelier

La porte s'est refermée sur nous, et nous voici au cœur de l'usine tourne pages. Une salle blanche lumineuse est presque entièrement consacrée au projet Amalfi, avec quatre scanners tourne pages dans la partie centrale.

Une fois la porte refermée, on entre dans le vif du sujet !

Où on découvre quatre Ditizing Line à l'œuvre

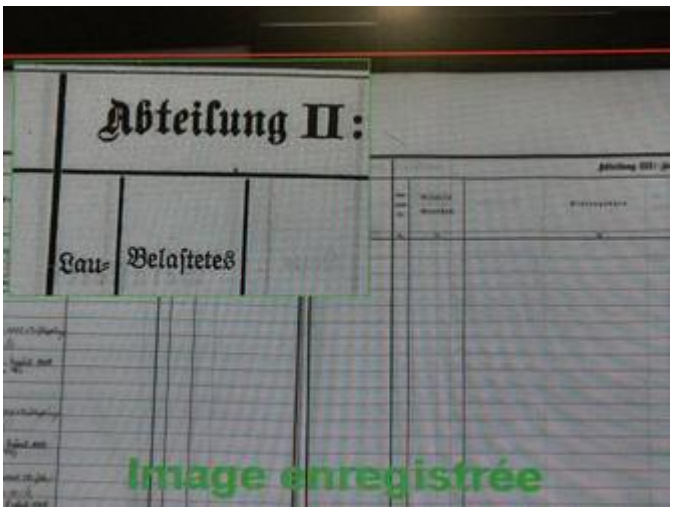
Il n'existe que six scanners de ce type au monde, et seuls ceux que possède Infotechnique sont utilisés de manière industrielle. Ce chiffre de quatre scanners permet en effet de numériser en optimisant les coûts de numérisation. Un unique opérateur est nécessaire pour la surveillance des quatre machines. L'intervention humaine est par ailleurs limitée au maximum. Dans cet atelier qui fonctionne en 3/8, le premier opérateur met en service les scanners un à un, en respectant un temps d'attente entre chaque machine. Dans la mesure où tous les registres ont un format et un nombre de pages identiques, cette précaution permet d'échelonner le moment où chacun des scanners arrive en fin de tâche.

Quatre scanners tourne pages sont utilisés de façon industrielle. Un unique opérateur suffit à assurer le bon fonctionnement des quatre scanners.

Préparation des ouvrages et numérisation

Avant d'être placés dans le sas des scanners, les ouvrages subissent un rapide nettoyage destiné à séparer les feuilles éventuellement collées. Il est en effet nécessaire de s'assurer que la numérisation se fera dans de bonnes conditions, chaque passage du bras du scanner étant programmé pour tourner une page après l'autre. Un contrôle est alors effectué, afin de vérifier qu'aucune page n'a été oubliée. Tous les volumes sont identiques en terme de format et de nombre de pages. Le programme de reconnaissance s'appuie sur cette caractéristique pour vérifier la numérotation des pages au fur et à mesure de leur numérisation. Les pages s'affichent sur le PC relié au scanner, et l'opérateur sera ainsi alerté en cas de nombre de pages différent de celui qui est attendu.

Ce pistolet à air permet de séparer délicatement les pages de ces ouvrages souvent anciens et fragilisés.



Après numérisation, les pages sont immédiatement visualisables sur le PC relié au scanner.

Les contrôles automatiques

Une fois le registre introduit dans le scanner, le travail de numérisation se fait de façon automatique. Trois diodes de couleur liées à chacun des scanners signalent l'avancement du travail et les éventuels blocages :

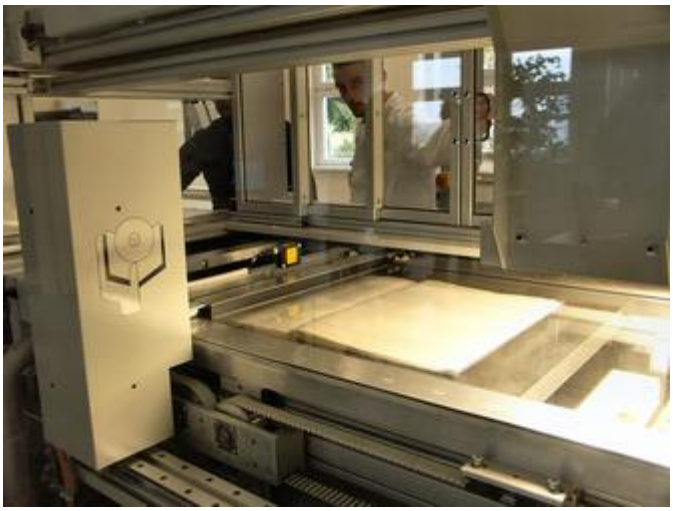
- Vert : tout va bien.
- Orange : numérisation du volume bientôt achevée.
- Rouge : erreur dans le processus (détection d'une page volante, nombre de page différent de celui attendu, présence de poussière en trop grand nombre…).

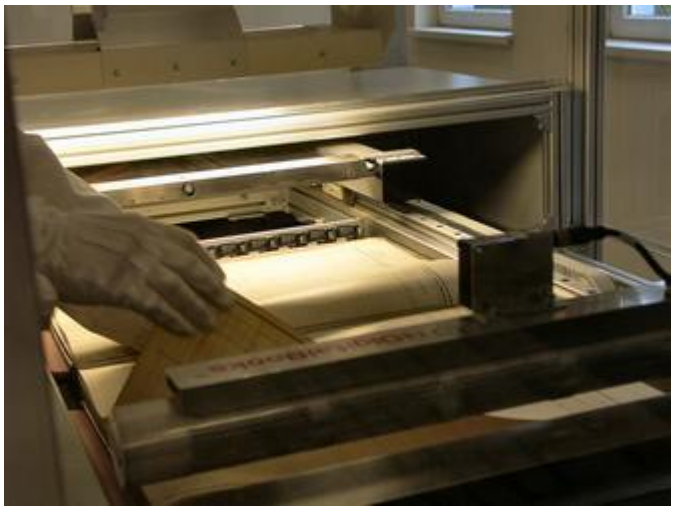
Les rares interventions humaines ont lieu à cette occasion.

La progression et l'état du travail de numérisation sont suivis au moyen de ces voyants lumineux.

Quelques cas d'interventions manuelles

L'introduction des documents dans le sas du scanner et l'intervention en cas de problème (feuille volante, poussières…) sont les seuls cas pendant lesquels les documents sont manipulés. Directement pris dans leurs caissons protecteurs, ils y sont placés à nouveau sitôt la numérisation achevée. Lors de ces rares étapes de manipulation, l'opérateur manipule délicatement les ouvrages en ayant pris la précaution de se munir de gants.





Les interventions humaines sont limitées au strict minimum et font l'objet de beaucoup de précautions.

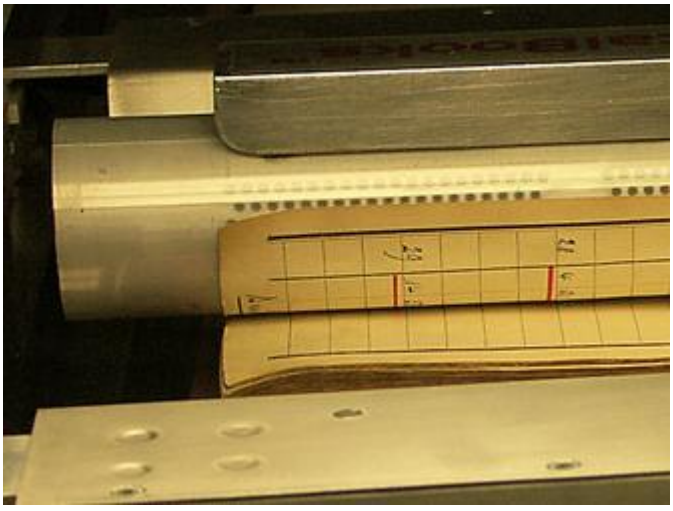
Il était une fois

On vous en parle depuis le début : voici donc enfin ce scanner pas comme les autres ! Nous vous proposons juste une anecdote avant de vous le montrer de plus près. À la lecture du cahier des charges du projet Amalfi, les équipes d'Infotechnique ont du réfléchir aux solutions à mettre en œuvre. L'histoire veut que tout ait débuté lors d'une pause café, avec la réflexion d'un développeur épris de robotique : « Et pourquoi pas un scanner tournant les pages automatiquement ? ». Piqué par cette idée, il aurait abrégé sa pause pour aller plus vite se consacrer à cette idée. Pendant ce temps, d'autres équipes cherchaient des traces d'éventuels scanners tourne pages existants. Il y avait bien le modèle Kirtas Technologies (voir cette partie), mais ses caractéristiques ne correspondaient pas aux contraintes du projet. La société I2S, fabricante de scanners et de caméras optiques, a été invitée à participer au brainstorming. Rapidement, un industriel suisse a été repéré, pour le travail qu'il avait déjà mené dans cette direction. Les deux entreprises ont ensuite travaillé ensemble, et dans des délais très courts, à réaliser un prototype fonctionnel. Le Digitizing Line était né.

Une société suisse (4DigitalBooks) et une société bordelaise (I2S) ont uni leurs compétences pour créer le Digitizing Line.

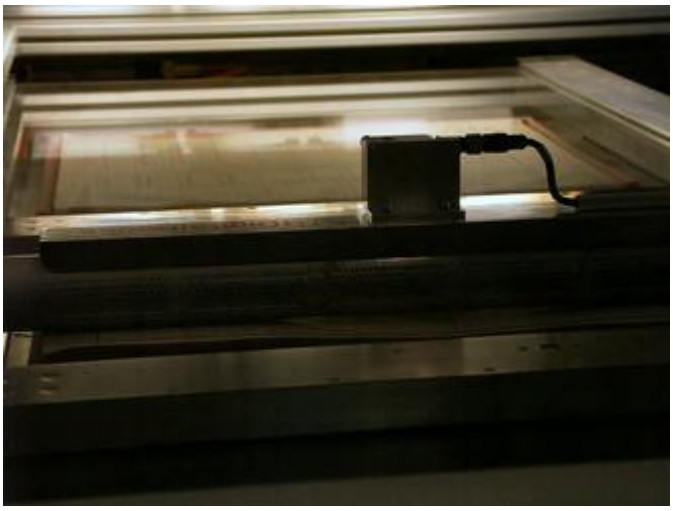
Un ingénieux système d'aspiration

Comment le Digitizing Line tourne t-il les pages ? Tout simplement en les aspirant pour les sélectionner une à une, et permettre à un bras automatique délicat de les rabattre pour les tourner. Une courte vidéo en parlera mieux que nous ! (Télécharger une vidéo du scanner en activité sous forme d'une archive .zip).



Aspirer la page pour la séparer des autres et pouvoir la tourner : il suffisait d'y penser !

Le procédé est non seulement délicat, mais précis. De nombreux tests ont été effectués pour s'assurer que le scanner ne prend bien qu'une page à la fois. Le bras qui aspire la page se déplace ensuite latéralement pour la tourner. La caméra du scanner passe alors pour numériser la double page qui se présente à elle.



Le bras tourne les pages une à une ; c'est ensuite au tour de la caméra optique des balayer de son faisceau.

Questions de cadences

Seuls deux autres scanners de ce type existent par ailleurs. L'un se trouve dans les locaux de l'université de Stanford (un des épicentres du projet Google Print, et université d'origine des fondateurs de la société du même nom), tandis que l'autre est en Angleterre à Southampton. Néanmoins, nous l'avons vu précédemment, une unique machine ne permet pas de numérisation de façon industrielle. Chaque scanner permet de traiter 800 pages au format A2 à l'heure. Ces modèles ont été donc réalisés sur mesure pour le projet Amalfi. Chacun d'entre eux a coûté quelques 300 000 euros. Au fil des jours, les équipes sont parvenues à affiner les techniques pour gagner en efficacité. De son côté, le fabricant du scanner travaille à des améliorations. Le matériel et l'organisation du travail sont donc optimisés tout au long du projet.

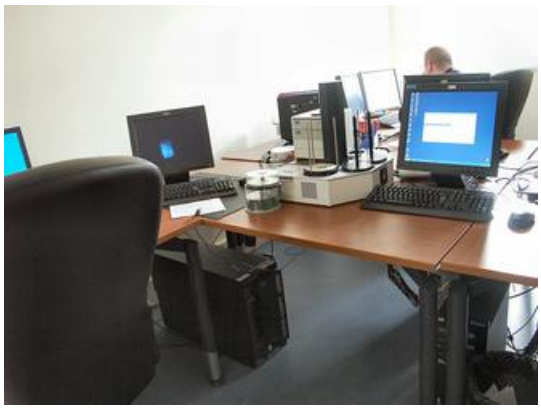
Chaque scanner traite huit cent pages à l'heure. Cinquante mille pages sont numérisées à la fin d'une journée.

Ça y est, on a quitté le support papier

Une fois numérisées, les pages sont transmises par satellite à l'atelier off-shore situé à Madagascar. Au milieu de l'océan Indien, des opérateurs sont chargés de structurer les données en XML au moyen d'une définition de document type (DTD) créée pour le projet. Les registres ont cette caractéristique propre aux livres de droit d'être des documents rigoureux, dont les informations sont structurées et se suivent dans un ordre immuable. Le contenu est typé au moyen de balises (« Date », « Nom », « Lieu » etc.) qui serviront ensuite à la recherche. Une autre action consiste à supprimer les pages dites « mortes » (actes de ventes correspondant par exemple à une propriété passée en d'autres mains). Une fois structurées et « nettoyées », les données reprennent le chemin de l'Alsace par cette même voie des ondes qu'elles avaient empruntée pour leur acheminement. Une dernière étape les attend dans les locaux d'Infotechnique : la vérification « métier ».

La vérification « métier »

Quelques anciens des Bureaux fonciers travaillent dans la dernière pièce que nous visitons. Cette petite équipe est en relation directe avec celles de Madagascar. Chacun des opérateurs en charge de la structuration peut ainsi questionner les experts dès qu'une anomalie est détectée (la rigueur des écrits juridiques est parfois prise en défaut) ou qu'une question d'ordre vraiment « métier » se pose. Un mail parvient jusqu'à l'équipe de La Walck ; il contient toutes les données du problème ainsi que les indications permettant d'identifier le volume et la partie concernés. Il semble que l'habilité des équipes chargées de la ressaisie à Madagascar soit telle qu'elle s'est presque transformée en expertise, à la grande surprise des spécialistes présents à La Walck. Une promotion d'experts fonciers serait-elle en train de se former sur le tas au milieu de l'océan Indien ?



On entre dans la partie « métier » du projet.

Place à IBM

Une fois la partie reprise des données achevées, c'est au tour d'IBM, le partenaire d'Infotechnique sur ce projet, d'intervenir. Les deux sociétés sont complémentaires sur ce projet. Tandis qu'Infotechnique assure la reprise des données, IBM se charge de l'aspect « signature électronique » et met en œuvre, dans ses bureaux de Strasbourg, l'application destinée aux Bureaux Fonciers. Cette même application sera ensuite utilisée par les notaires, les géomètres, les collectivités locales, etc.

Après traitement et vérifications, les données sont gravées sur CD pour être envoyées aux bureaux fonciers. IBM prend ensuite la relève.

Que retenir de tous ces projets d'envergure, et de cette visite ?

Concernant la bibliothèque numérique européenne, l'objectif affiché est clairement celui de sauvegarde, de mise à disposition et de valorisation de notre patrimoine culturel. La problématique de Google est en revanche bien différente. Il ne s'agit plus de préservation de notre patrimoine culturel, mais bien de sa « commercialisation ». Le choix des ouvrages numérisés est laissé à la discrétion des bibliothèques et des éditeurs participant aux projets. Ceux-ci peuvent confier des ouvrages rares ou menacés à Google pour assurer leur sauvegarde, mais il ne s'agit en rien d'une volonté expresse de Google. Par ailleurs, dans le cas des éditeurs, les ouvrages soumis ne seront pas retournés à leur propriétaire dans la mesure où la reliure est ôtée pour la numérisation.

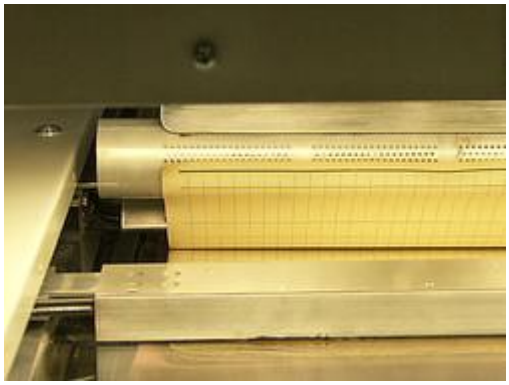
La numérisation est la seule façon de sauvegarder les ouvrages fragiles.

Une autre différence vient des étapes des différents projets. Celui de bibliothèque numérique européenne ne pourra sérieusement débuter sans un inventaire systématique des fonds des différentes bibliothèques. Sur un projet d'une telle envergure, on ne peut effectivement pas se permettre de numériser en double. Imaginez la perte de temps que représenterait la double voire triple numérisation d'un quotidien centenaire, de son tout premier numéro jusqu'à nos jours ? Les projets que l'on qualifiera plutôt de commerciaux n'ont quant à eux pas cette contrainte d'inventaire.

Quoi qu'il en soit de ces différences, la sauvegarde du patrimoine passe sans doute par des solutions mixtes, reposant sur des partenariats public / privé. Certaines bibliothèques européennes ont dores et déjà confié des fonds à des projets de type Google ou Yahoo, et reçu en contrepartie une version numérique de leurs ouvrages dont elles peuvent faire un usage libre tant qu'il n'est pas commercial. Pour les bibliothèques depuis trop longtemps dans l'attente de débloquement de fonds et confrontées à des difficultés de gestion de leurs ouvrages (manque d'espace de stockage, documents en fin de vie), le recours à des partenariats de type Google est presque une démarche de bon sens.

La dernière chose à retenir, est que la numérisation - et à plus forte raison la gestion des connaissances (knowledge management) dont il est de plus en plus question -, est un vrai métier industriel. La numérisation à grande échelle fait appel à des compétences, des techniques et des outils que les bibliothèques ne maîtrisent pas. La remarque vaut également pour Google Print et ses concurrents. Cette visite nous apprend que l'on ne s'improvise pas expert en numérisation et traitement des données, quel que soit le type de projet de sauvegarde ou de numérisation que l'on a. Bref, personne ne peut faire l'impasse sur les questions techniques si peu abordées, pour le moins publiquement, lorsqu'il est question de ces projets.

Qui sait réellement comment et au moyen de quelles technologies sont numérisés les ouvrages dans le cadre du projet Google ? Font-il appel à des scanners de type Kirtas, ou bien les pages sont-elles tournées manuellement avec tous les risques d'erreur que cela implique ?



Aujourd'hui les scanners tourne pages, et quoi demain ?

Aujourd'hui donc les scanners tourne pages, et quoi demain ? Sans doute des technologies similaires mais largement optimisées (cadence plus élevée, mobilité accrue, coûts réduits). La prochaine génération de scanners tourne pages pourra peut-être sortir des usines pour arriver jusqu'au cœur des bibliothèques. Ou bien ces scanners feront partie de solutions mixtes : tourne pages automatiques pour les ouvrages de grandes dimensions et calibrés, Bookscan pour les petits livres, scanners feuille à feuille pour les documents de moindre intérêt patrimonial, numérisation manuelle pour les documents les plus sensibles... Et sans doute d'autres machines inventives qui n'attendent que des idées un peu folles et un cahier des charges exigeant pour débarquer !

Article du Mardi 15 Novembre 2005 écrit par Anne

http://www.clubic.com/article-29049-1-visite-d-une-usine-tourne-pages-.html