Quoi de neuf dans les bibliothèques numériques ?

Mathieu Andro

Quoi de neuf dans les bibliothèques numériques ?

Mathieu Andro

ENSSIB 4/11/2015

De l’histoire ancienne ?

Il est difficile de trouver des documents qui ne sont pas déjà numérisés et qui peuvent l'être du point de vue juridique.

Que faire au delà de la numérisation ?

- Solutions de diffusion
- Ebooks
- Print on Demand
- TEI, Text mining, bibliométrie
- Crowdsourcing
- Identification du plagiat

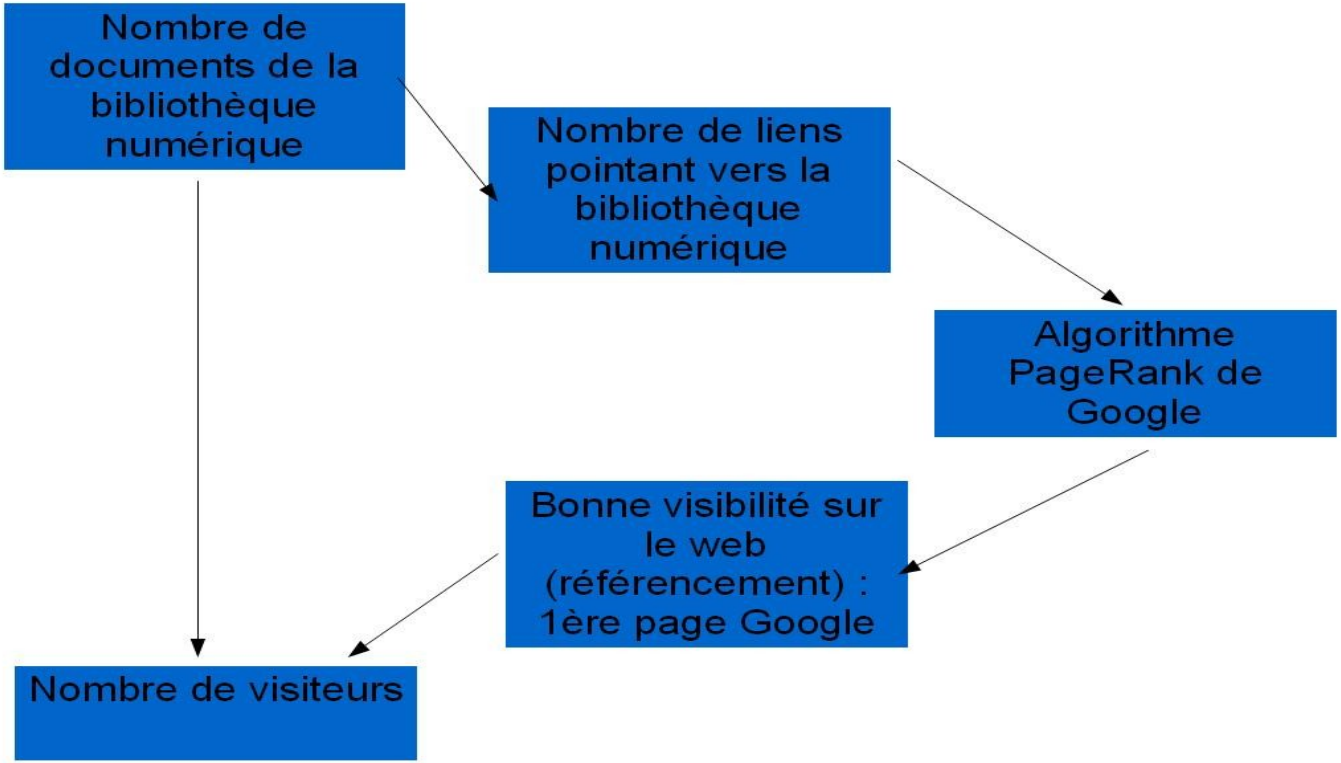
1. Comment développer une bibliothèque numérique gratuitement et simplement pour une visibilité, une pérennité et des fonctionnalités optimales ?

Les logiciels de bibliothèques numériques

- YooLib (Polinum)
- Invenio (CERN)
- ORI-OAI (Universités)
- DSpace (DuraSpace)
- DigiTool (Ex Libris)
- Mnesys (Naoned)
- ContentDM (OCLC)
- Eprint (Université de Southampton)
- Greenstone (Université de Waikato)
- Omeka (Université George Mason)

La majeure partie des documents numérisés par les Bibliothèques en France n'est pas en ligne et « dort » sur des DVD et disques durs dont la durée de vie est limitée car :

- Gallica ne peut les héberger (pas encore).
- Il n'existe pas de plateforme pour les Bibliothèques de l'Enseignement Supérieur.
- Le développement d'une plate-forme est coûteux pour un résultat rarement satisfaisant.
- Il y a un déficit d'information sur les solutions de diffusion.


Nombre de documents de la bibliothèque numérique
Nombre de liens pointant vers la bibliothèque numérique
Algorithme PageRank de Google
Bonne visibilité sur le web (référencement) : 1ère page Google
Nombre de visiteurs

	Nombre de documents	Nombre de visites ou de downloads en 2011	Nombre de visites ou de downloads par mois en 2011	Nombre de visites ou de downloads par mois et par document
Internet Archive[2]	3 678 804 livres (au 25 octobre 2012)	227 244 392 downloads	18 937 033 downloads par mois	5,15 downloads par livre et par mois minimum[3].
Gallica[4]	1,6 millions de documents dont 224 322 livres (fin 2011)	9 485 603 visites	790 467 visiteurs par mois	2,02 visites par document et par mois 3,52 visites par livre et par mois
e-corpus	30 743 notices dont 23 870 textes (au 1er janvier 2012)	418 215 visites	34 851 visites par mois	1,13 visites par notice et par mois 1,46 visites par texte et par mois minimum

Old School ?

Vieille école :
Chaque bibliothèque crée sa propre bibliothèque numérique et cherche à y attirer des internautes.

Nouvelle école :
Elle participe à des bibliothèques collectives déjà fréquentées par les internautes (Internet Archive, Flickr)

(Waibel, 2008)

Principaux acteurs

1- Google books (USA-International, plus de 30 millions de livres)

2- Europeana (Européen, 26 millions d'objets moissonnés)

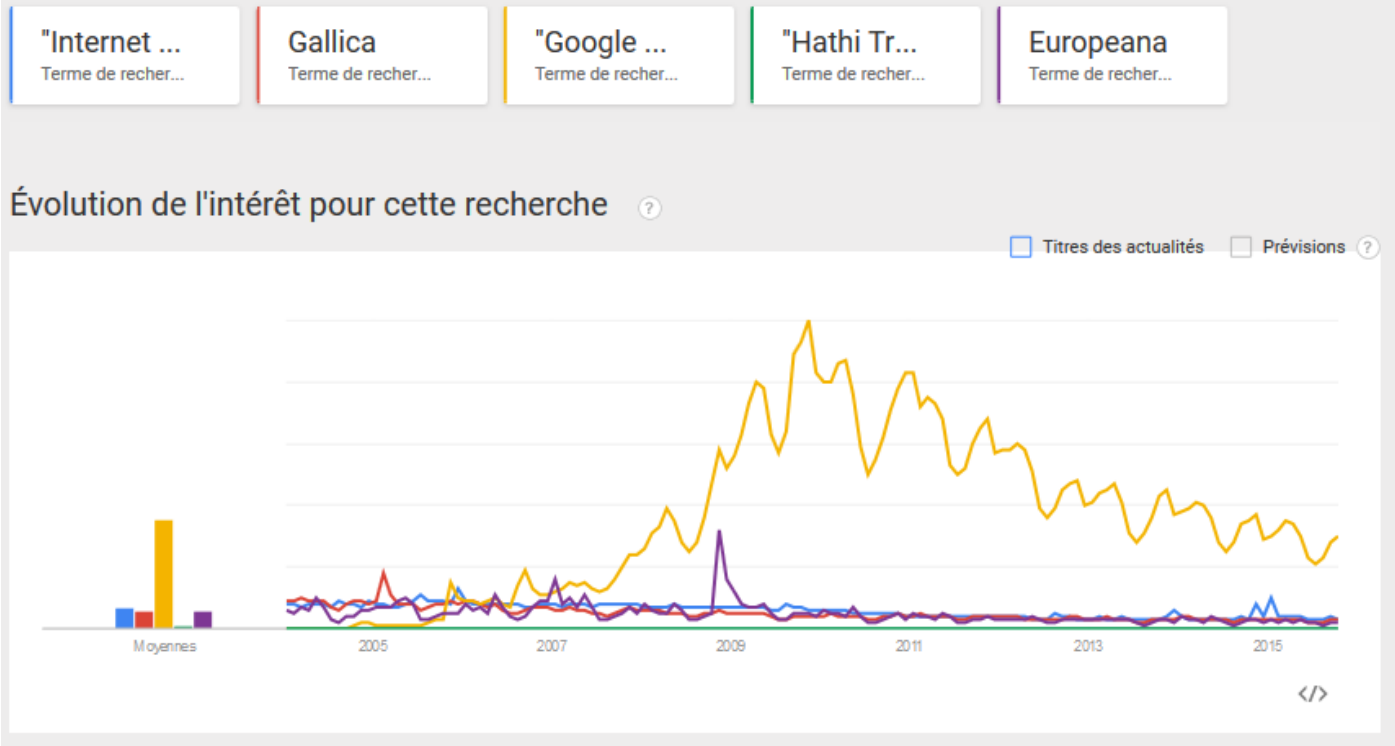
4- Internet archive (USA-International, 8,3 millions de livres en libre accès)

3- Hathi Trust (USA-International, 6,8 millions de livres dont 5,5 en libre accès)

5- Gallica (France, 3 millions de documents dont 550 000 livres)

Google évalue à 129 millions le nombre d'imprimés

Statistiques comparatives (monde)



Statistiques comparatives (France)

Gallica

- OAI-PMH
- Marchés BnF

- Gallica marque blanche
- Numistral 15 septembre 2013 (BNUS)

Hathi Trust

- Octobre 2008 par l'université d'Indiana et celle du Michigan avec le soutien de la Mellon Foundation afin de pérenniser l'accès à ce qui a été numérisé par Google
- Automne 2010 : adhésion de l'Université de Madrid
- Payant : partage de coûts
- Annonce 6,8 millions de livres numérisés (dont 5,5 million du domaine public et en libre accès).
- Importation possible des métadonnées METS depuis Internet Archive et Google Books
- Archivage pérenne aux standards PREMIS et OAIS.
- Poursuivi par l'Union des écrivains du Québec pour les 7 millions d'œuvres orphelines numérisées par Google et diffusées sur HathiTrust alors qu'elles sont sous droits.

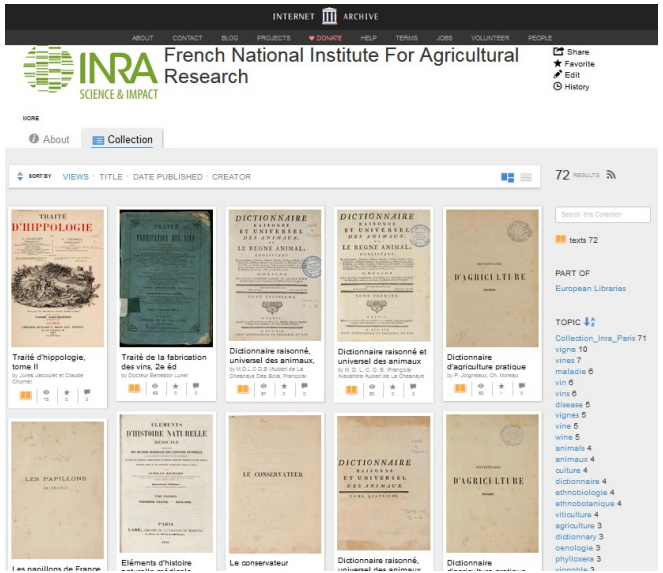
Internet Archive

- Organisation américaine non gouvernementale et à but non lucratif localisée à San Francisco.
- Fondé en 1996 par Brewster Kahle. Accessible en ligne depuis 2001. Membre de l'Open Content Alliance.
- 15 millions de dollars par an, 200 personnes dont 50 programmeurs, bibliothécaires et administrateurs.
- Data centers à San Francisco, Redwood City et Moutain View (Californie) et serveurs miroir à la Bibliothèque d'Alexandrie en Egypte.
- Plus de 8 millions de livres dont 300 000 Microsoft Live Book Search, 1 million de Google + projet Gutenberg.
- En France : BSG, Inra, Sciences Po...
- ARK, **EPUB, MOBI pour Kindle**, Daisy, licences juridiques, OAI-PMH, Zotero, RSS...
- OpenLibrary, vidéos, sons, wayback machine
- Gratuit

Très concrètement : Internet Archive

1. Créer un compte
2. Uploader un PDF ou un ensemble de JPEG
3. Saisir la notices dans les champs
4. Au delà de 50 versements, demander la création d'une collection et l'administrer

https://archive.org/details/inra



2. Au delà de la numérisation : text mining et crowdsourcing.

L’archivage pérenne

Le Centre Informatique National de l'Enseignement Supérieur (Cines) :

fichier avec format non reconnu
support détérioré (durée de conservation)
pas documenté (métadonnées)

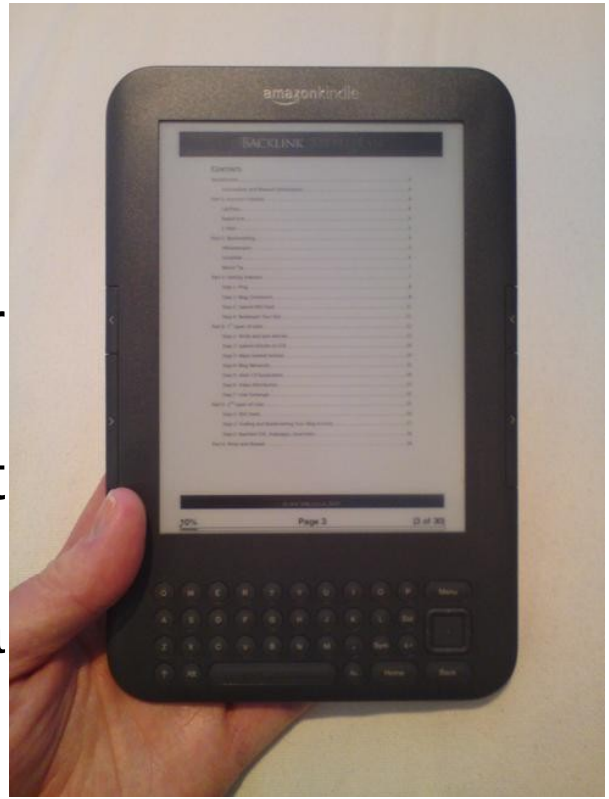
La Bibliothèque nationale de France : SPAR

Pérenniser la diffusion pour les générations futures

La lecture sur ebook

Chiffres datant de 2013 :

- 4 millions de EPUB sur Google Books
- 4 millions de EPUB et Mobi sur Internet Archive
- 1,8 million de Mobi sur Amazon (dont plusieurs milliers gratuits)
- environ 1000 EPUB et MOBI sur la Bibliothèque Sainte-Geneviève
- 159 EPUB sur Gallica



By JM1981 (http://www.donttouchmykindle.co.uk/kindle-wifi/) [CC BY-SA 3.0], via Wikimedia Commons

Identification du plagiat

Georges de La Tour [Public domain], via Wikimedia Commons

C’est quoi le text mining?

Au lieu de rechercher un mot dans un document...
rechercher des dictionnaires de mots dans des masses de documents…

et **identifier les relations** entre ces mots
et **extraire des données** de ces documents

Flickr, Kate Ter Haar, CC BY 2.0

La culturomique et les humanités numériques

Michel J.B. Shen Y.K. Aiden A.P. Veres A. Gray M.K. Pickett J.P. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 331(6014) : 176-82.

- Evolution de la langue (burned vs burnt)
- Trotsky disparaît de la littérature russe après la mort le Lénine

https://books.google.com/ngrams

C’est quoi le crowdsourcing ?

Juin 2006, Jeff Howe, Wired Magazine.

Crowd = foule (des internautes)

\+ (out)**sourcing** = externalisation

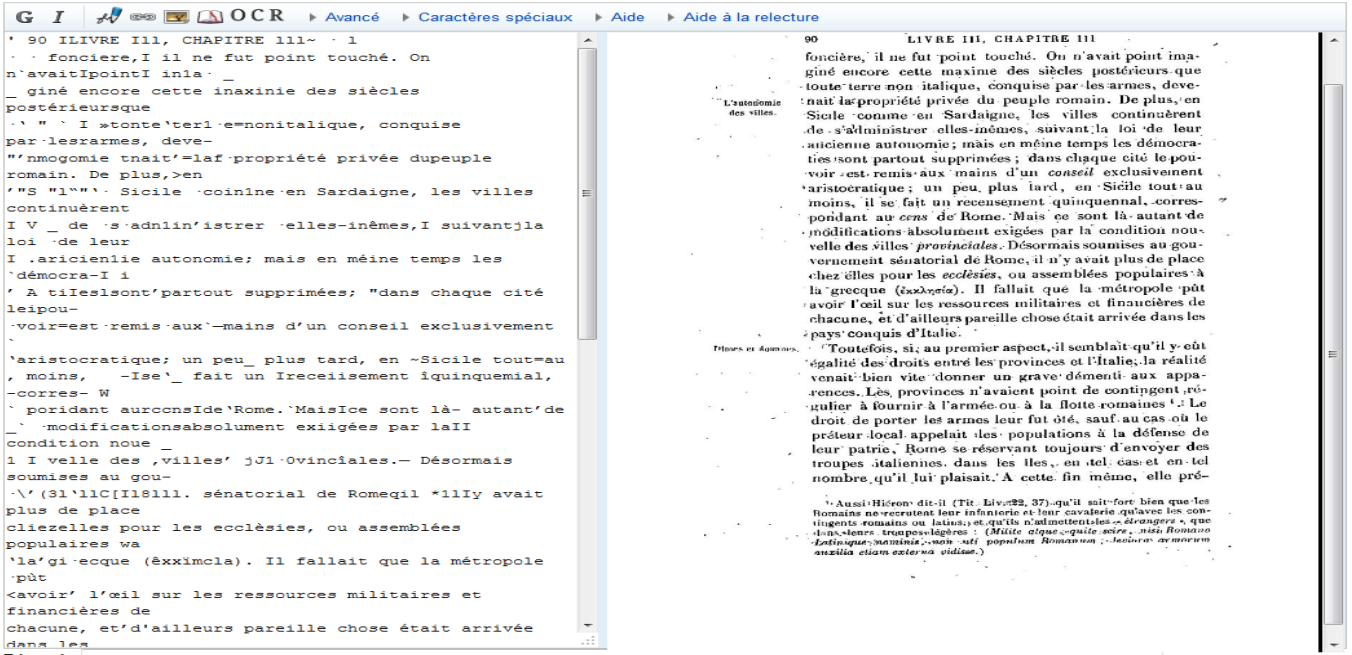
Au lieu d’externaliser vers des pays à main d’oeuvre à bas coût, pourquoi ne pas bénéficier des internautes ?

→ travail, compétences ou connaissances moins limitées.

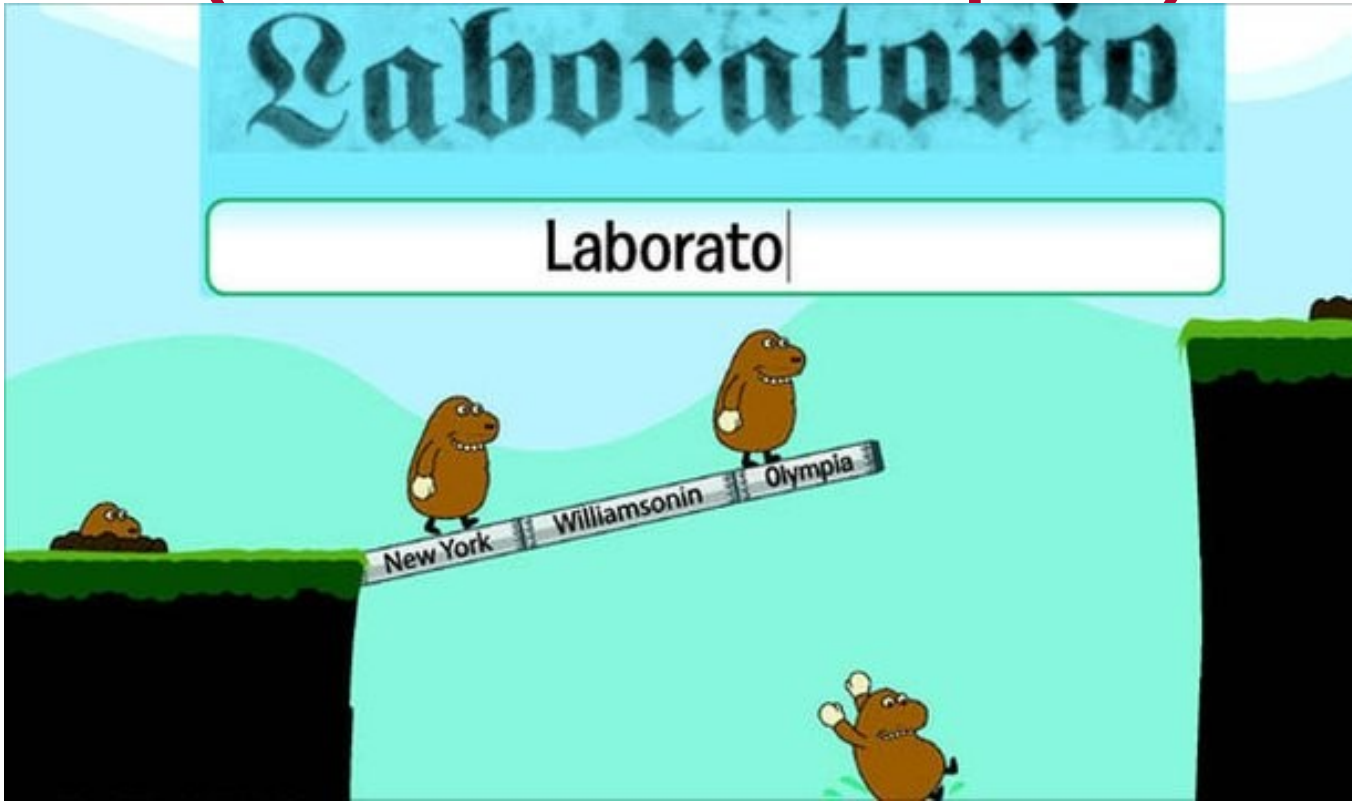
→ argent (crowdfunding).

Le crowdsourcing explicite classique

Modification de Page:Mommsen - Histoire romaine - Tome 3.djvu/94



La gamification (Games With A Purpose)



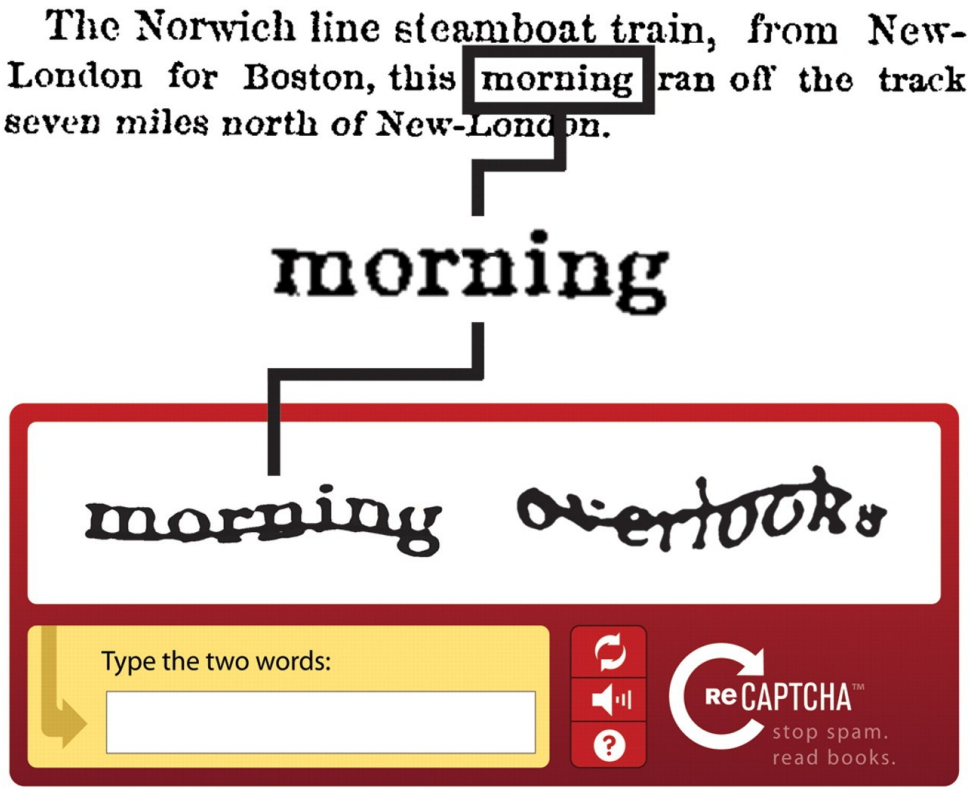
Le crowdsourcing implicite (involontaire)

2007 : reCAPTCHA.

Par jour :

- 200 millions de mots dont 88 validés
- 12 000 heures
- 1147 ebooks à 99,1%

Corrigés par des pays à bas coût, il faudrait environ 146 million d'euros / an.



Le crowdsourcing rémunéré

2005 : Amazon Mechanical Turk Marketplace.

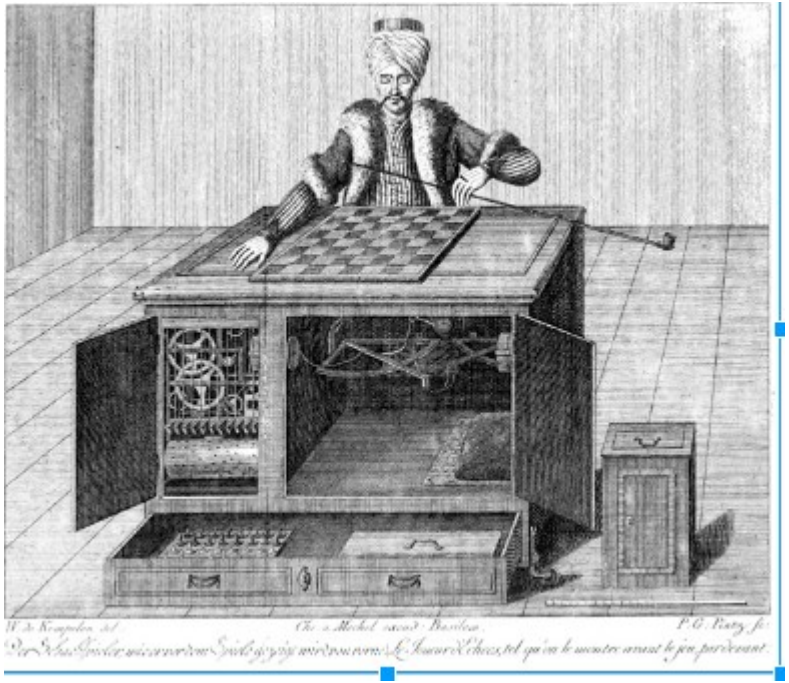
500 000 travailleurs de 190 pays

Jeunes américains + indiens

En avril 2010 : 6 701 406 HITs proposés par 9436 sociétés pour une valeur marchande de 529 259 $

700 000 HITs par semaine

Travailler où on veut, quand on veut, pour qui on veut, être tantôt employeur tantôt employé



« Tuerkischer schachspieler windisch4 » par Karl Gottlieb von Windisch. 1783. Public domain via Wikimedia Common

Le crowdfunding : numérisation à la demande

Ebooks on Demand, Numalire, Amis de la BnF

Proposer aux internautes d'en financer la numérisation

Partager notre politique d'acquisition

Se concentrer sur ce qui ne peut être numérisé par le privé

Délégation de service public :

- Usagers : service de reprographie
- Bibliothèques : renforcer leurs programmes
- Fondations : visibilité
- Investisseurs : trafic web (ROI)

L’impression à la demande

- Print on Demand (POD) : commander un fac-similé imprimé.
- Editeurs : produire en flux tendu, sans stocks ni invendus
- Amazon BookSurge, Jouve, Librissimo, UniBook...
- Exemple : Bibliothèque de l'Université du Michigan
- Délégation de service public (prestataire)
- Retour sur investissement
- Espresso Book Machine

Pour continuer à savoir ce qu'il y a de neuf dans les bibliothèques numériques

Surveiller l'environnement de son projet :

- veille juridique
- veille institutionnelle (tutelles, partenaires, concurrents, prestataires)
- veille "commerciale" : appels à projets
- veille technologique : nouvelles plateformes, nouvelles sociétés, opportunités de collaborations ou partenariats

Au lieu d'aller vers l'information, c'est elle qui vient à vous.
Le renseignement humain : tutelles, institutions et entreprises

User innovation

Amateurs = ruptures innovantes car ne cherchent pas à reproduire les modèles établis du métier

Von Hippel 46 % des entreprises américaines dans des domaines innovants ont pour origine un simple utilisateur

Conduite du changement

Merci pour votre attention

Contact : mathieu.andro@versailles.inra.fr

Diaporama : http://tinyurl.com/qaybz6y

Thèse : http://www.bibliotheque-numerique.fr