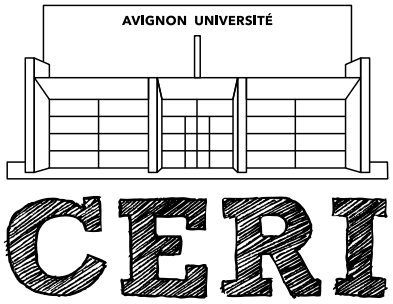
Partie 1
Introduction à la recherche d'information

Vincent Labatut

Laboratoire Informatique d'Avignon – LIA EA 4128
vincent.labatut@univ-avignon.fr

2019/20

M2 ILSEN
UE Ingénierie du document et de l'information
UCE3 Indexation & Recherche d'information



Centre d'Enseignement et de Recherche en Informatique

AVIGNON UNIVERSITÉ

Plan de la séance

1. Présentation de l'UCE
   - Organisation
   - Contenu

2. Introduction à la recherche d'information
   - Notion de recherche d'information
   - Historique de la RI
   - Principales notions relatives à la RI
   - Architecture d'un système de RI

Section 1

Présentation de l'UCE

AVIGNON
UNIVERSITÉ

Présentation de l'UCE
Organisation

- Modalités :
  - CM : 7 séances d'1h30 (dont 1 examen écrit)
  - TP : 7 séances d'1h30 (dont 1 examen en salle)
- Évaluation :
  - Examen écrit (50%) : 21/10/19 8h30-10h00
  - Examen de TP (50%) : 06/01/20 8h30-10h00
  - Sous réserve de modification : la dernière date donnée oralement prime
- Documents/communication : sur e-uapv
  - Supports de cours
  - Sujets et corrections de TP
  - Références bibliographiques [1]
  - Questions hors-séance → forum
- **Note :** contenu mis à jour par rapport aux années précédentes

1. Figure sans référence apparente → clic

Présentation de l'UCE
Contenu

1. Introduction
2. Recherche booléenne
3. Construction du lexique
4. Recherches positionnelle et approchée
5. ~~Construction d'index~~
6. ~~Compression d'index~~
7. Ordonnancement des documents
8. Évaluation des performances
9. ~~Approche probabiliste~~
10. ~~Recherche Web~~

11. Séance d'évaluation

Section 2

Introduction à la recherche d'information

AVIGNON
UNIVERSITÉ

Notion de recherche d'information
Définition

Recherche d'information (eng : Information retrieval)

La RI est le **domaine** traitant des méthodes permettant à un **utilisateur** d'identifier dans une **collection** de documents $\mathcal{C}$ ceux qui correspondent à ses **besoins**.

- Exemple : cabinet juridique
  - Accès à une collection de nombreux textes légaux
  - On veut obtenir tous (et seulement) ceux qui sont relatifs à des avis rendus par la cours des comptes (CC) à propos de la transparence financière des ONG
- Autres types de documents traités :
  - Pages Web, catalogues, articles, vidéos, chansons, images...

  → Données non-structurées

Notion de recherche d'information
Sous-problèmes et moyens mis en œuvre

- Sous-problèmes :
  - Représentation des documents
  - Stockage des documents
  - Organisation des documents
  - Accès aux documents
- Moyens mis en œuvre :
  - Indexation
  - Recherche
  - Modélisation
  - Recherche Web
  - Classification de textes
  - Interface utilisateur
  - Visualisation

Notion de recherche d'information
Tâches de RI et domaines proches

- Tâches :
  - Recherche *ad hoc* : tâche standard
  - Filtrage : collection en évolution
  - Navigation : raffinage itératif des besoins

- RI vs. Extraction d'information :
  - Obtenir et stocker des données structurées à partir de documents non-structurés
  - Peut être vue comme préparatoire à la RI
- RI vs. Fouille de données :
  - Calcul permettant d'extraire une information de haut niveau à partir de nombreuses données (informations de bas niveau)

Historique de la recherche d'information
Stockage et organisation de la connaissance

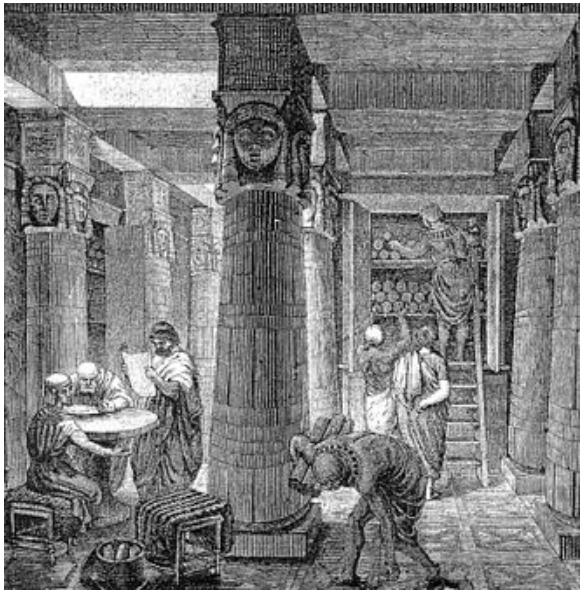
- $-5\,000$: apparition de l'écriture
  - → Comment organiser la connaissance ?
    - Problème fondamental pour le développement d'une civilisation
    - Tablettes → rouleaux → codex → livres
- Exemple : machine à vapeur
  - Ctésibios d'Alexandrie -270
  - Héron d'Alexandrie -150
  - Orgue de la cathédrale de Reims 1120
  - Léonard de Vinci 1502
  - Jacob Besson 1519
  - David Rivault 1606
  - Salomon de Caus 1615
  - Edward Somerset 1663
  - Samuel Morland 1683
  - Puis ère industrielle

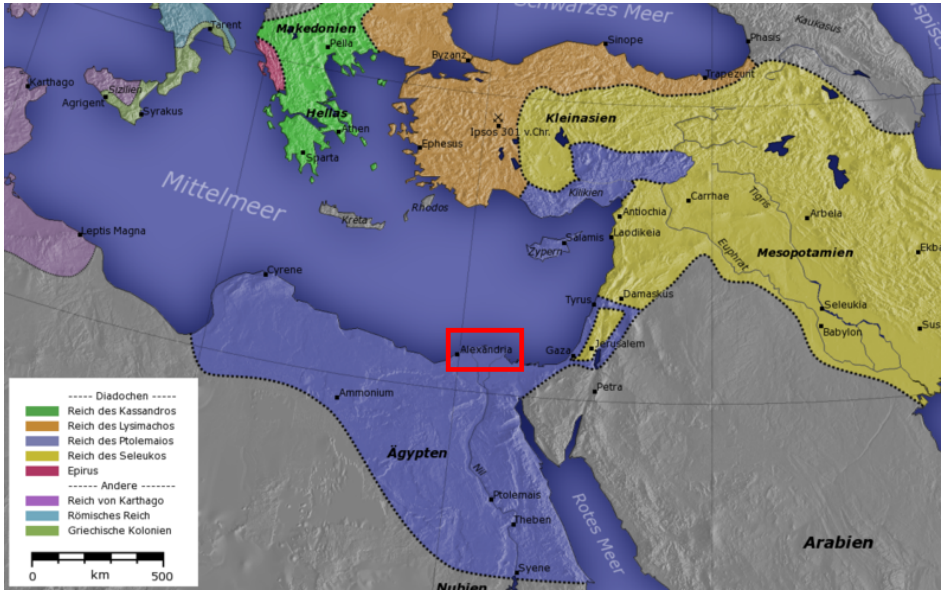
Historique de la recherche d'information
Concept de bibliothèque

- $-3\,000$: concept de bibliothèque
- -288 : bibliothèque d'Alexandrie (700 000 volumes)

→ Nécessité d'utiliser des index

- Catégories définies manuellement
- Sous-catégories pointant vers des documents spécifiques

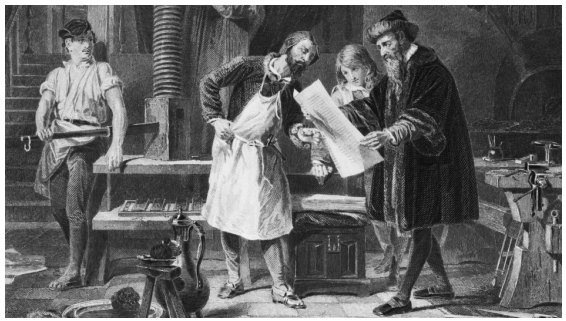




Historique de la recherche d'information
Développement technologique

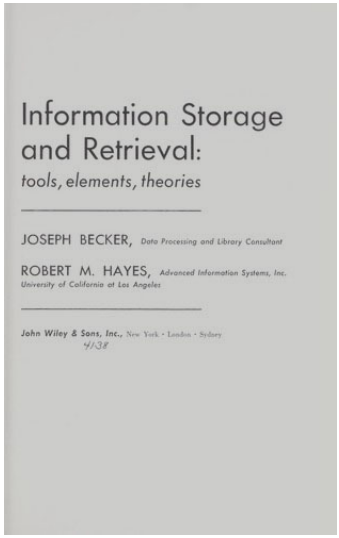
- -200 : codex (Rome), ancêtre du livre
- 100 : papier (Chine), puis 750 Arabes, puis 1000 Européens
- 600 : xylographie (Chine), impression par gravure
- 1000 : caractères mobiles en terre cuite (Chine) puis métal (Corée, 1200)
- 1300 : mécanisation de la production du papier
- 1440 : mécanisation de l'imprimerie (Gutenberg)

$\rightarrow$ 50 ans après l'invention de l'imprimerie en Europe, 20 millions de livres avaient été imprimés



Historique de la recherche d'information
Révolution computationnelle

- ≈ 1950 : création du terme Information Retrieval
- Automatisation des méthodes existantes
- 1962 : premier livre, par Becker & Hayes [BH62]
- 1978 : première conférence (ACM SIGIR)
- Nouvelles méthodes et fonctionnalités (i.e. GUI pour navigation)


Information Storage and Retrieval:
tools, elements, theories

JOSEPH BECKER, *Data Processing and Library Consultant*

ROBERT M. HAYES, *Advanced Information Systems, Inc.*
University of California at Los Angeles

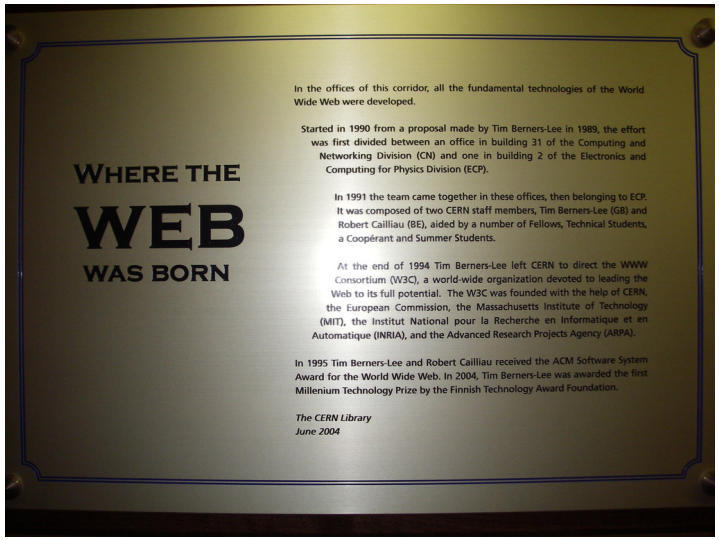
John Wiley & Sons, Inc., New York · London · Sydney

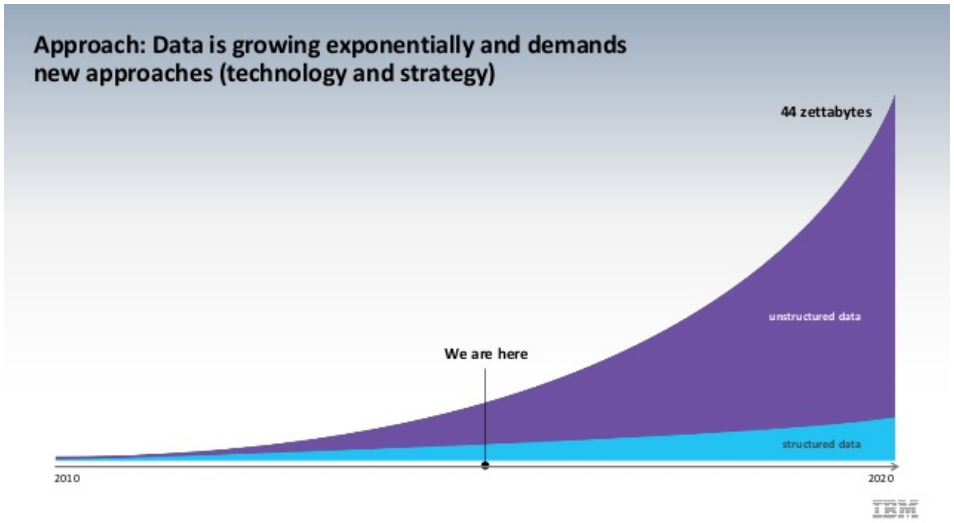
SIGIR
Special Interest Group
on Information Retrieval

Historique de la recherche d'information
Changement d'échelle

- 1990–92 : création du Web au CERN
- Explosion du nombre de documents accessibles : ≈ 50 milliards (2016)
- Accès répandu et en hausse : 47% (2016), 51% (2017), 55% (2018) de la population mondiale

→ Nécessité de développer de nouvelles méthodes de RI





Historique de la recherche d'information
Effets du Web

1. Étape supplémentaire de crawling
   - Documents composés de plusieurs pages Web
   - Connectées via des hyperliens
2. Traitement de gros volumes
   - Gros volumes de données
   - Gros volumes de requêtes
3. Évaluation difficile de la pertinence
   - Gros volume de données → traitement manuel plus possible
4. Utilisation généralisée
   - On ne cherche plus seulement une information bibliographique
   - Ex. : prix d'un service, n° de téléphone...
5. Données très bruitées
   - Contenu et forme non-contrôlés
   - Apparition de la notion de spam

Historique de la recherche d'information
Évolution légale et données ouvertes

- Notion de donnée ouverte
  - ≠ donnée accessible publiquement
  - Information numérique sous licence ouverte
  - Utilisable sans restriction légale, technique, ni financière
  - Caractéristiques : brute, complète, accessible, exploitable, permanente, gratuite
  - Exemples : génétique, cartes, loi, marchés publics
- Donnée ouverte publique
  - Produite dans le cadre d'un service public
  - Objectif : contrôle démocratique, efficacité de l'action publique, innovation économique et sociale
- Bref historique :
  - 1789 : droit d'accès aux informations publiques
  - 2002 : droit d'accès aux données juridiques
  - 2016 : loi pour une république numérique
  - 2018 : ouverture des données des collectivités locales

Principales notions relatives à la RI
Deux approches complémentaires

1. Initialement un sous-domaine des sciences de l'information
   - But : offrir les méthodes de recherches les plus adaptées à l'utilisateur
   - Approche orientée *humain*
2. Appartient maintenant aussi (automatisation) à l'informatique
   - But : améliorer les performances relatives au traitement d'une requête
   - Approche orientée *machine*

Besoin informationnel

Il s'agit du renseignement désiré par l'utilisateur, noté b. Ce besoin est **implicite**, ou exprimé mais de façon **informelle**.

Exemple : opinion de la CC à propos de la transparence financière des ONG

Corpus

La **collection de documents** $\mathcal{C} = \{d_1, ..., d_D\}$ faisant l'objet d'un traitement, où d_i dénote un document et D la taille de la collection ($1 \leq i \leq D$).

Principales notions relatives à la RI
Définitions II

Requête

Expression **formelle** du besoin informationnel, notée q (pour *query*) : elle est **explicite**, et compatible avec l'outil de RI.

Exemple : `auteur="cours des comptes" ET sujet ("finance" ET "ONG")`

Résultat de la recherche

Sous-ensemble $R \subset \mathcal{C}$ de documents issus du corpus $\mathcal{C}$, et noté $R = \{d_{i_1}, ..., d_{i_N}\}$, où N est le nombre de documents renvoyés ($0 \leq N \leq D$). Peut alternativement être une **séquence** $R = \langle d_{i_1}, ..., d_{i_N} \rangle$.

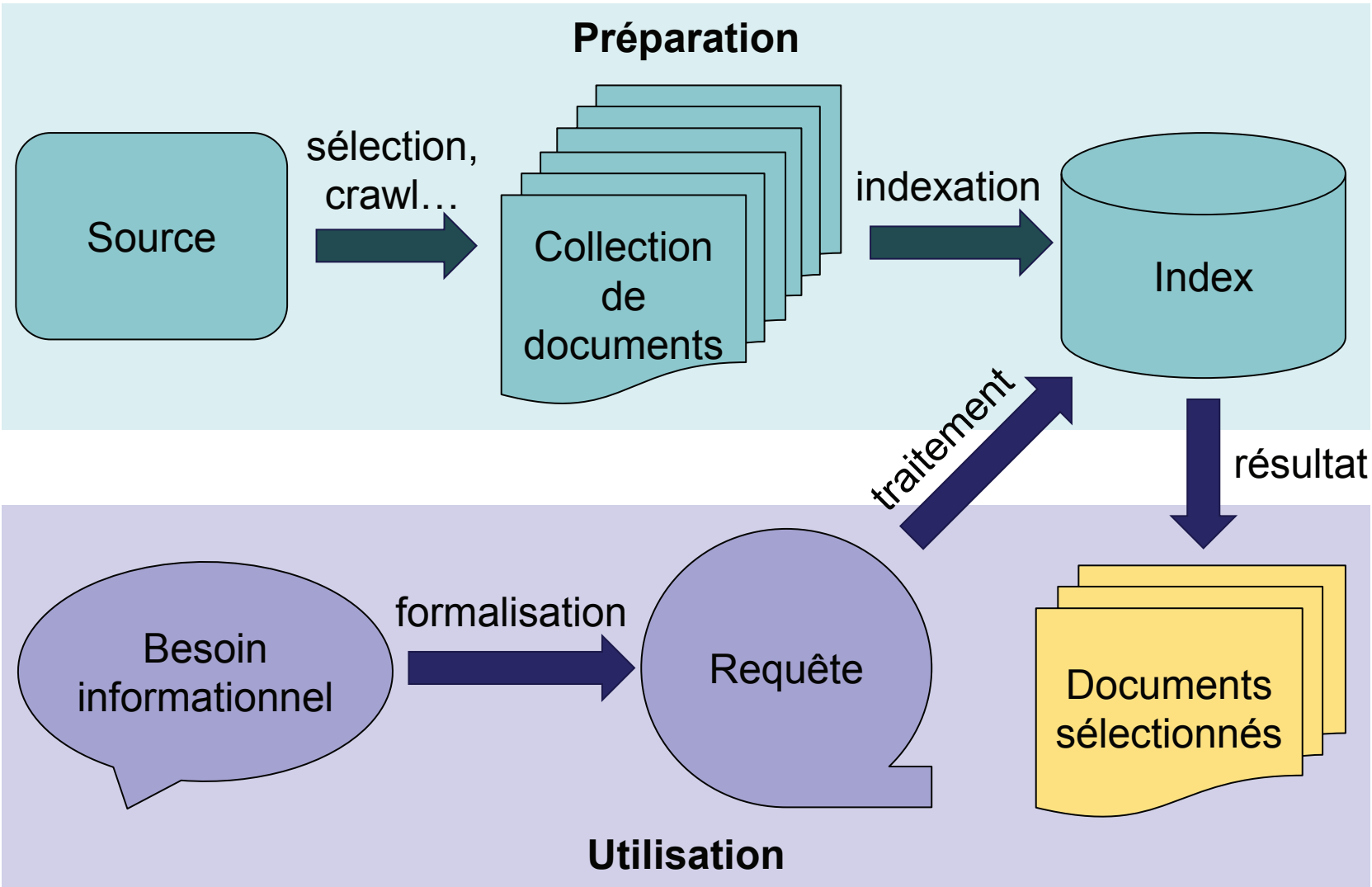
Séquence car le sous-ensemble peut être ordonné : pertinence, date, sujet...

Principales notions relatives à la RI
Objectif et performance

- Objectif : renvoyer
  - Tous les documents pertinents... (*rappel*)
  - ...et aucun document non-pertinent (*précision*)
- Pertinence : adéquation des documents renvoyés à la requête formulée
  - Notion *relative* : utilisateur, temps, espace...

  → difficile de définir un outil *toujours* pertinent
- Performance :
  - Précision : proportion de documents pertinents, parmi les documents retournés
  - Rappel : proportion de documents retournés, parmi les documents pertinents du corpus

Architecture d'un système de RI



Section 3

Conclusion

AVIGNON
UNIVERSITÉ

Concepts abordés dans cette partie

- Recherche d'information
- Tâche de filtrage
- Corpus
- Indexation
- Information structurée
- Besoin informationnel
- Tâche de navigation
- Index
- Architecture
- Information non-structurée

Lectures recommandées

[MRS08] *Introduction to Information Retrieval*, chapitre 1.

[BCC10] *Information Retrieval : Implementing and Evaluating Search Engines*, chapitre 1.

[BR11] *Modern Information Retrieval : The Concepts and Technology behind Search*, chapitre 1.

[AG13] *Recherche d'information - Applications, modèles et algorithmes*, chapitre 1.

[CMS15] *Search Engines : Information Retrieval in Practice*, chapitres 1 & 2.

Références bibliographiques I

[AG13] M.-R. Amini et É. Gaussier. *Recherche d'information - Applications, modèles et algorithmes.* Paris, FR : Eyrolles, 2013. url : https://www.eyrolles.com/Informatique/Livre/recherche-d-information-9782212673760/.

[BR11] R. Baeza-Yates et B. Ribeiro-Neto. *Modern Information Retrieval : The Concepts and Technology behind Search.* 2nd Edition. Boston, USA : Addison Wesley Longman, 2011. url : http://people.ischool.berkeley.edu/~hearst/irbook/.

[BH62] J. Becker et R. M. Hayes. *Information Storage and Retrieval : Tools, Elements, Theories.* Hoboken, USA : John Wiley & Sons, 1962.

[BCC10] S. Büttcher, C. L. A. Clarke et G. V. Cormack. *Information Retrieval : Implementing and Evaluating Search Engines.* Cambridge, USA : MIT Press, 2010. url : http://www.ir.uwaterloo.ca/book/.

Références bibliographiques II

[CMS15] W. B. Croft, D. Metzler et T. Strohman. *Search Engines : Information Retrieval in Practice*. Pearson, 2015. url : `http://www.search-engines-book.com/`.

[MRS08] C. D. Manning, P. Raghavan et H. Schütze. *Introduction to Information Retrieval*. New York, USA : Cambridge University Press, 2008. url : `http://www-nlp.stanford.edu/IR-book/`.