La chaîne documentaire

Cours « Document et Web Sémantique »

Nicolas Malandain, Nicolas Delestre

INSA ROUEN NORMANDIE

1. Le document
   - Description
   - Contenu & Forme
   - Diffusion
   - Interprétation
2. La chaîne de traitements
   - Problématique
   - Reconstruction document papier
   - Interprétation
3. Le sens du document
   - Interprétation
   - Indexation
4. Conclusion

Quelques définitions

Document

- Larousse
  « pièce écrite servant d'**information**, de preuve »
- Wikipedia
  « Un document renvoie à un ensemble formé par un **support** et une **information** (le contenu), celle-ci enregistrée de manière persistante. » (2021)

Données Suite de caractères représentant une valeur

Information Données contextualisées de manière formelle ou informelle

Connaissance Information qui permet de valider une information et/ou d'inférer une nouvelle information

Évolution des supports

À voir : « L'Odyssée de l'écriture »

La pierre les peintures rupestres

La tablette d'argile en Mésopotamie 4e millénaire avant JC
recto-verso, lecture haut-bas, droite-gauche (forme la plus ancienne d'écriture)

Le Papyrus Égypte et proche orient 3e millénaire avant JC

- craint le pliage, l'humidité et le feu
- apparition du *Volumen* : rouleau formé de deux axes verticaux, que l'on déroule.
- longueur jusqu'à 15 mètres
- nécessite les deux mains pour le dérouler et lire
- ne permet pas de relire quand on écrit, pas de marquage possible
- très utilisé pour la lecture à haute voix

Le parchemin Pergame (aujourd'hui Bergama, Turquie) 2^{e} millénaire avant JC

- fabriqué à partir de peau (mouton/chèvre)
- souple
- écriture des deux côtés
- apparition du *Codex* : assemblage de parchemins
- la bible est copiée par les chrétiens sur codex

Le papier Chine 2^{e} siècle avant JC

- monopole de 6 siècles
- se répand à partir de 751
- arrive en Europe au 12^{e} siècle
- support souple et léger, il se conserve mieux
- au 15ème siècle apparition de l'imprimerie avec Gutenberg

Dématérialisation de nos jours le document se dit électronique ou numérique et n'a plus de support spécifique.

Caractérisations

Un document se caractérise par :

- Son contenu
- Sa forme
- Sa diffusion
- Son utilisation
- Son interprétation

Contenu 1 / 2

Informel, semi-formel et formel

Alan Turing

Pour les articles homonymes, voir Turing.

Alan Mathison Turing, né le 23 juin 1912 à Londres et mort le 7 juin 1954 à Wilmslow, est un mathématicien et cryptologue britannique, auteur de travaux qui fondent scientifiquement l'informatique.

Pour résoudre le problème fondamental de la décidabilité en arithmétique, il présente en 1936 une expérience de pensée que l'on nommera ensuite machine de Turing et des concepts de programme et de programmation, qui prendront tout leur sens avec la diffusion des ordinateurs, dans la seconde moitié du xx^{e} siècle. Son modèle a contribué à établir la thèse de Church, qui définit le concept mathématique intuitif de fonction calculable.

Durant la Seconde Guerre mondiale, il joue un rôle majeur dans la cryptanalyse de la machine Enigma utilisée par les armées allemandes. Ce travail secret ne sera connu du public que dans les années 1970. Après la guerre, il travaille sur un des tout premiers ordinateurs, puis contribue au débat sur la possibilité de l'intelligence artificielle, en proposant le test de Turing. Vers la fin de sa vie, il s'intéresse à des modèles de morphogenèse du vivant conduisant aux « structures de Turing ».

Poursuivi en justice en 1952 pour homosexualité, il choisit, pour éviter la prison, la castration chimique par prise d'œstrogènes. Il est retrouvé mort par empoisonnement au cyanure le 8 juin 1954 dans la chambre de sa maison à Wilmslow. La reine Élisabeth II le reconnaît comme héros de guerre et le gracie à titre posthume en 2013.

230 A. M. TURING [Nov. 12,

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable *numbers*, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers for explicit treatment as involving the least cumbrous technique. I hope shortly to give an account of the relations of the computable numbers,

Naissance	23 juin 1912 Maida Vale (Londres) (Royaume-Uni)
Décès	7 juin 1954 (à 41 ans) Wilmslow (Cheshire) (Royaume-Uni)
Domicile	Wilmslow (Cheshire, Angleterre)
Nationalité	Britannique
Domaines	Informatique, mathématiques, logique, cryptanalyse
Institutions	Université de Manchester National Physical Laboratory Université de Cambridge
Diplôme	Université de Manchester Université de Princeton
Renommé pour	Problème de l'arrêt Machine de Turing Cryptanalyse d'Enigma

Contenu 2 / 2

Avant le document numérique

- contenu principalement unique : texte / image / vidéo / voix / ...
- accès séquentiel ou direct
- quelques mélanges : livre avec texte et image
- duplication pouvant entraîner des pertes

À l'ère du document numérique

- contenu totalement hétérogène : texte + image + vidéo + voix ...
- accès séquentiel / direct / hypertextuel
- duplication 100% conforme à l'original

Forme

Avant le document numérique

- support physique
- humainement déchiffrable
- stockage volumineux et relativement pérenne

À l'ère du document numérique

- le support est lié au stockage, mais non à l'utilisation
- stockage peu volumineux
- quasiment humainement indéchiffrable
- le support est pérenne mais les documents non (problème de format)

Accélération de la diffusion

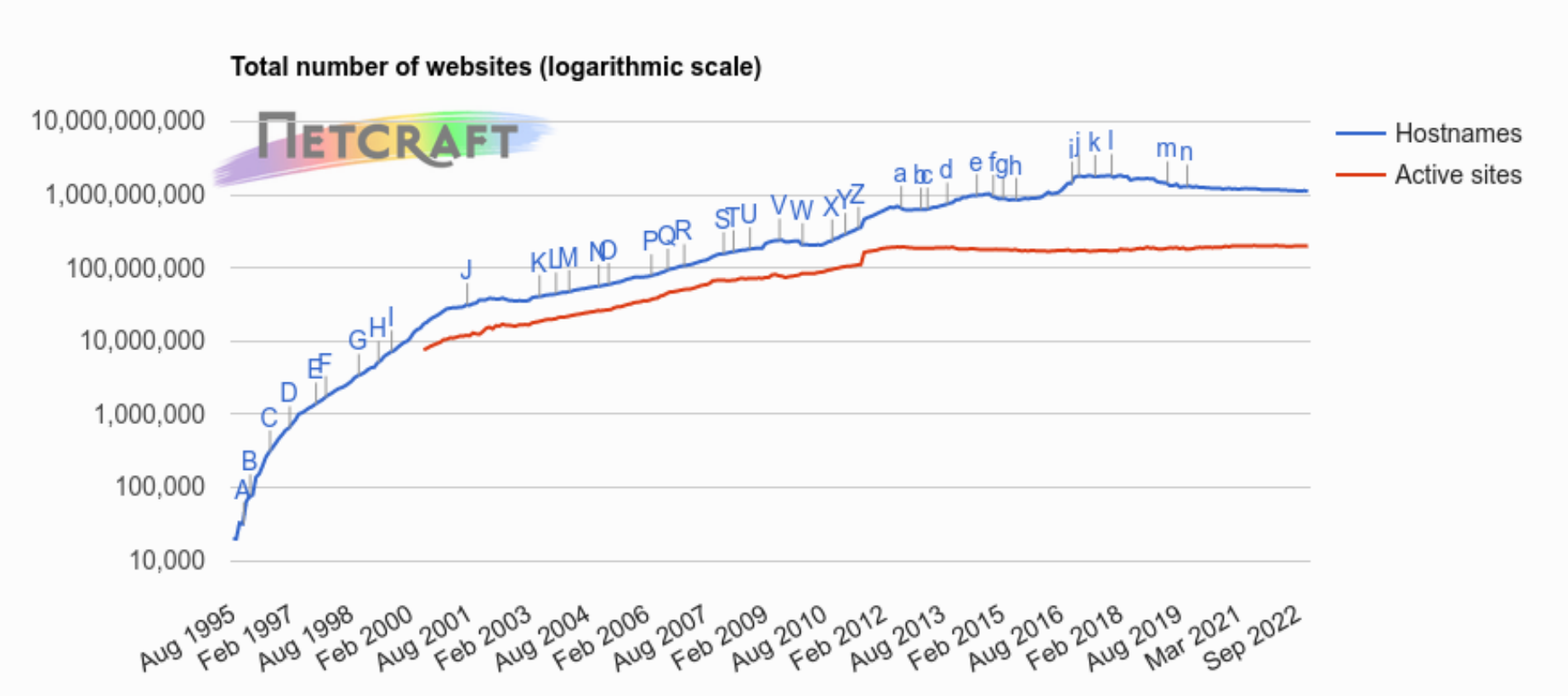
Changement du support

- poids
- transport
- copie
- disparition du support

Changement de la publication

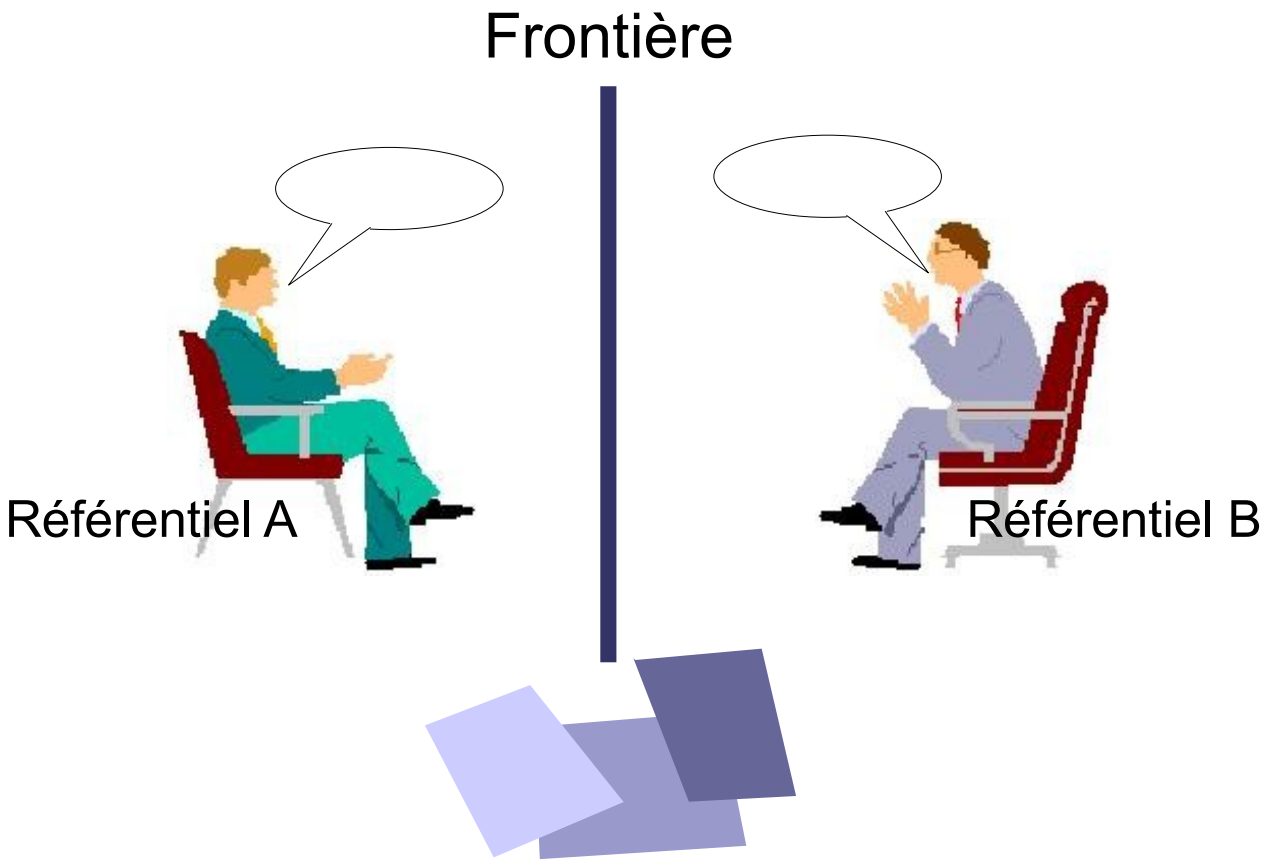
- rédaction : des “élites” au quidam
- publication : des éditeurs au quidam
- moins de sélection
- disponibilité immédiate

Nombre de sites web

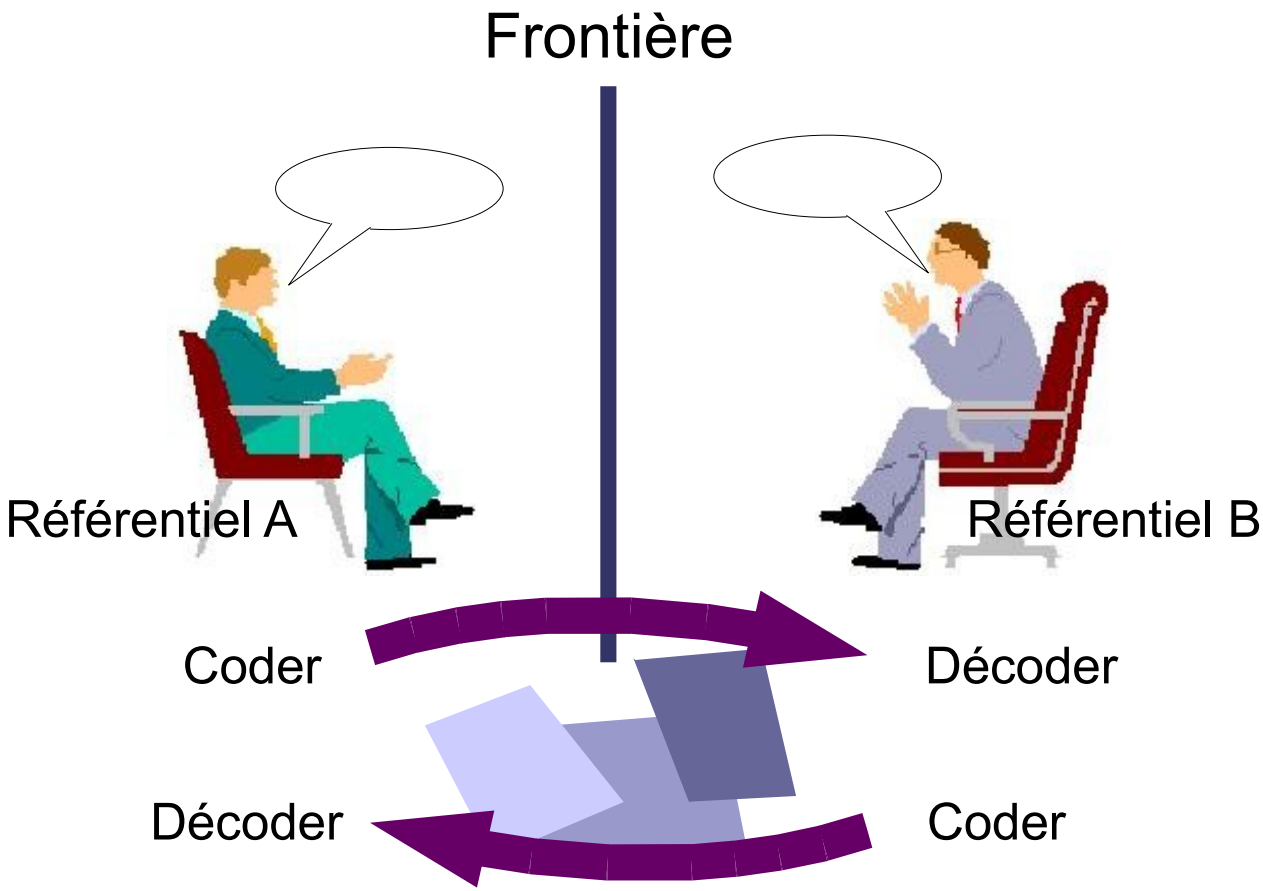


source : www.netcraft.com

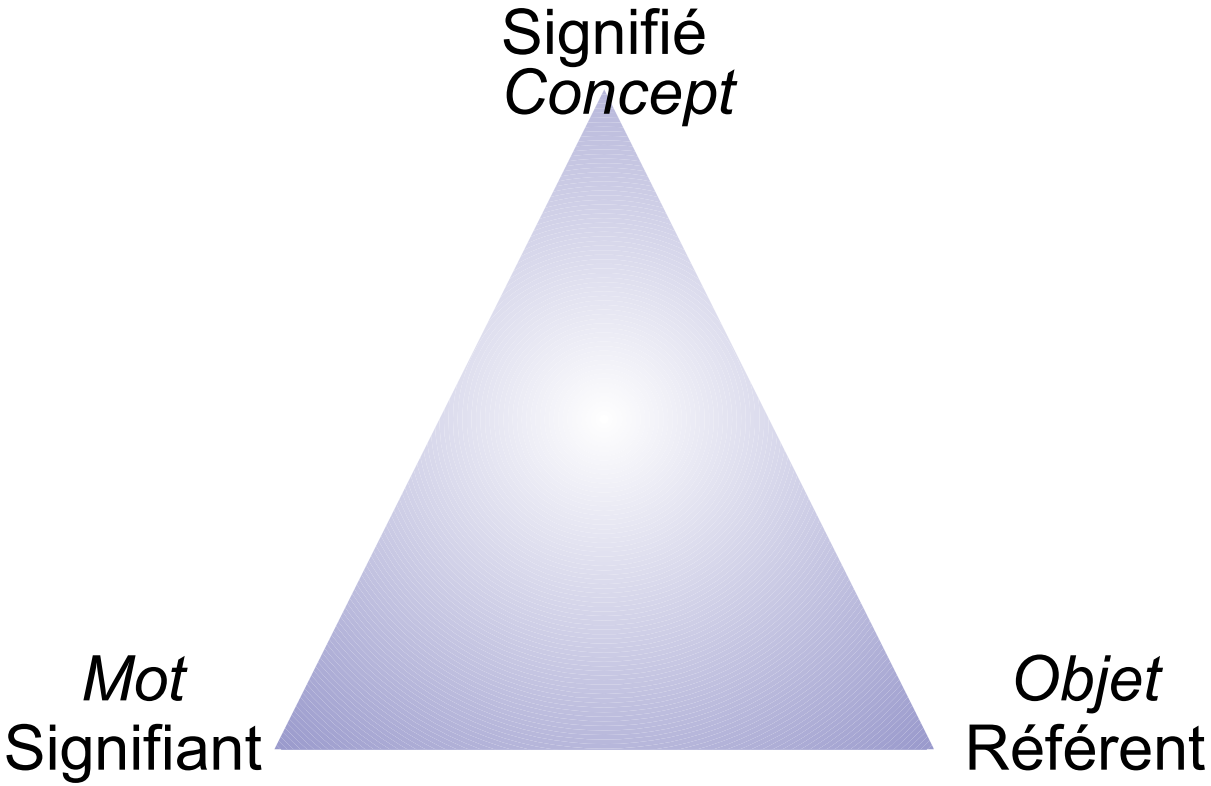
Le mécanisme de la communication



Le mécanisme de la communication



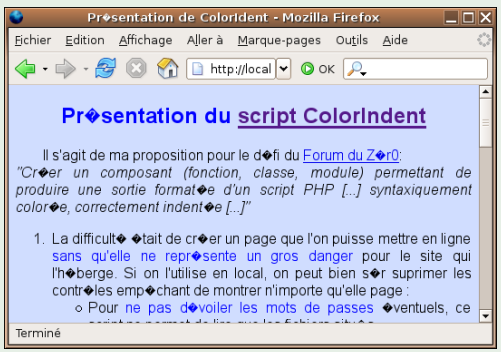
Le triangle sémiotique



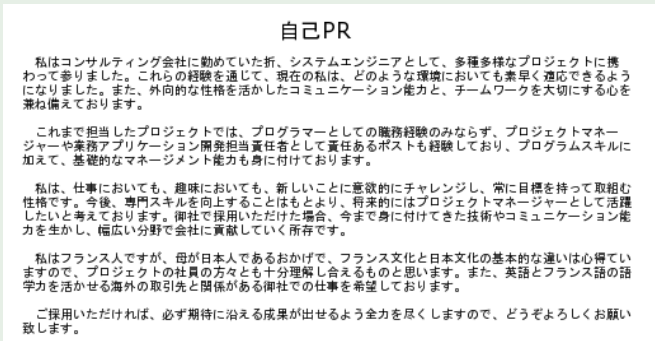
L'interprétation d'un document 1 / 2

Elle dépend

- écriture/lecture asynchrone
- écriture/lecture référentiels différents
  - homme ↔ homme / homme ↔ machine / machine ↔ machine
- codages multiples
- mises en forme



http://si2ra.ouvaton.org/si2ra/pub/mode_d039emp/installation/un_encodage_correcte.php



自己PR

私はコンサルティング会社に勤めていた折、システムエンジニアとして、多種多様なプロジェクトに携わって参りました。これらの経験を通じて、現在の私は、どのような環境においても素早く適応できるようになりました。また、外向的な性格を活かしたコミュニケーション能力と、チームワークを大切にする心を兼ね備えております。

これまで担当したプロジェクトでは、プログラマーとしての職務経験のみならず、プロジェクトマネージャーや業務アプリケーション開発担当責任者として責任あるポストも経験しており、プログラムスキルに加えて、基礎的なマネージメント能力も身に付けております。

私は、仕事においても、趣味においても、新しいことに意欲的にチャレンジし、常に目標を持って取組む性格です。今後、専門スキルを向上することはもとより、将来的にはプロジェクトマネージャーとして活躍したいと考えております。御社で採用いただけた場合、今まで身に付けてきた技術やコミュニケーション能力を生かし、幅広い分野で会社に貢献していく所存です。

私はフランス人ですが、母が日本人であるおかげで、フランス文化と日本文化の基本的な違いは心得ていますので、プロジェクトの社員の方々とも十分理解し合えるものと思います。また、英語とフランス語の語学力を活かせる海外の取引先と関係がある御社での仕事を希望しております。

ご採用いただければ、必ず期待に沿える成果が出せるよう全力を尽くしますので、どうぞよろしくお願い致します。

http://www.tartoinjapan.com/fr/blog_post-157-cv_lettre_motivation.html

Pierre Laurence

L'interprétation d'un document 2 / 2

Elle nécessite

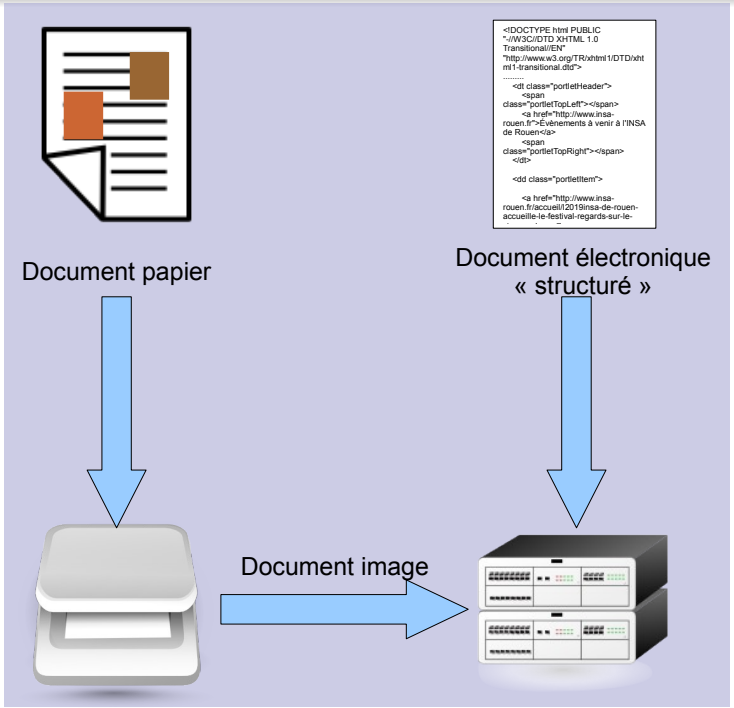
- L'ajout de données sur (ou au sein d') un document : métadonnées
  - Par exemple : l'auteur du document, la date de parution
- Ces métadonnées peuvent être intrinsèques ou externes

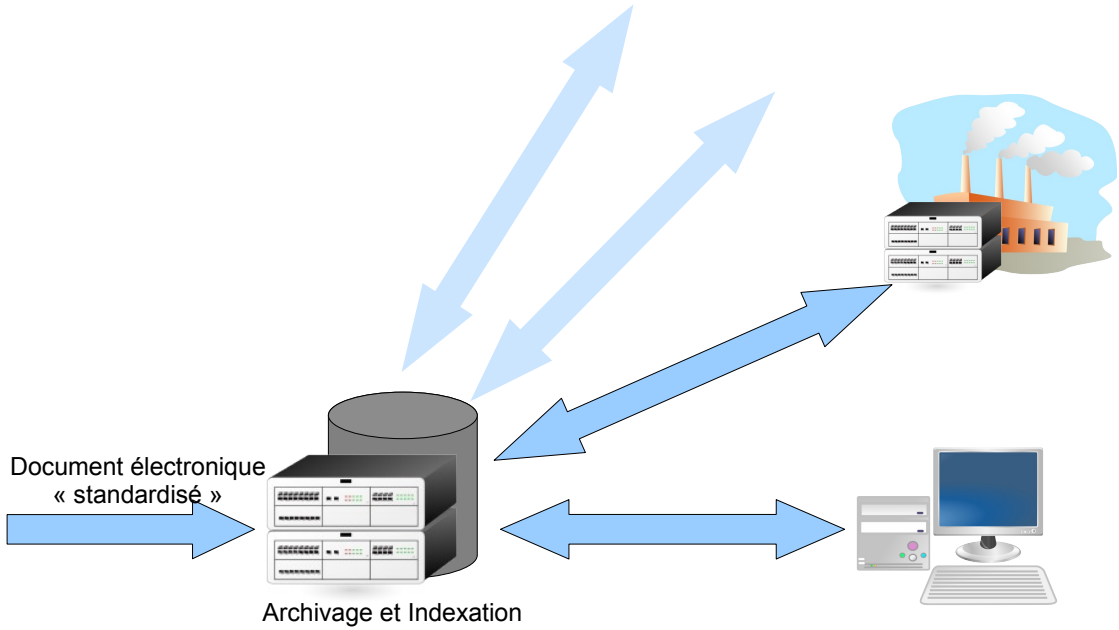
Une chaîne de traitements, pourquoi ?

Quels sont les problèmes liés aux documents ?

- Pérennité
  - la masse de documents papiers, audio et vidéo (analogiques) existante
  - la masse d'information papier/audio/vidéo en perpétuelle augmentation
  - la masse de documents électroniques en perpétuelle augmentation
  - la perte importante de documents
  - le stockage
- Accès à l'information
  - accès distant
  - mise en relation de documents
  - recherche d'informations : une aiguille dans une meule de foin
  - langues
  - mises à jour
  - format

Présentation d'une chaîne de traitements générale





Exemple de chaînes de traitements

- Projets à grande échelle : BnF, Google (Google books)
- GED pour les entreprises : Nuxeo
- GED pour les particuliers : Paperwork
  `https://github.com/jflesch/paperwork`

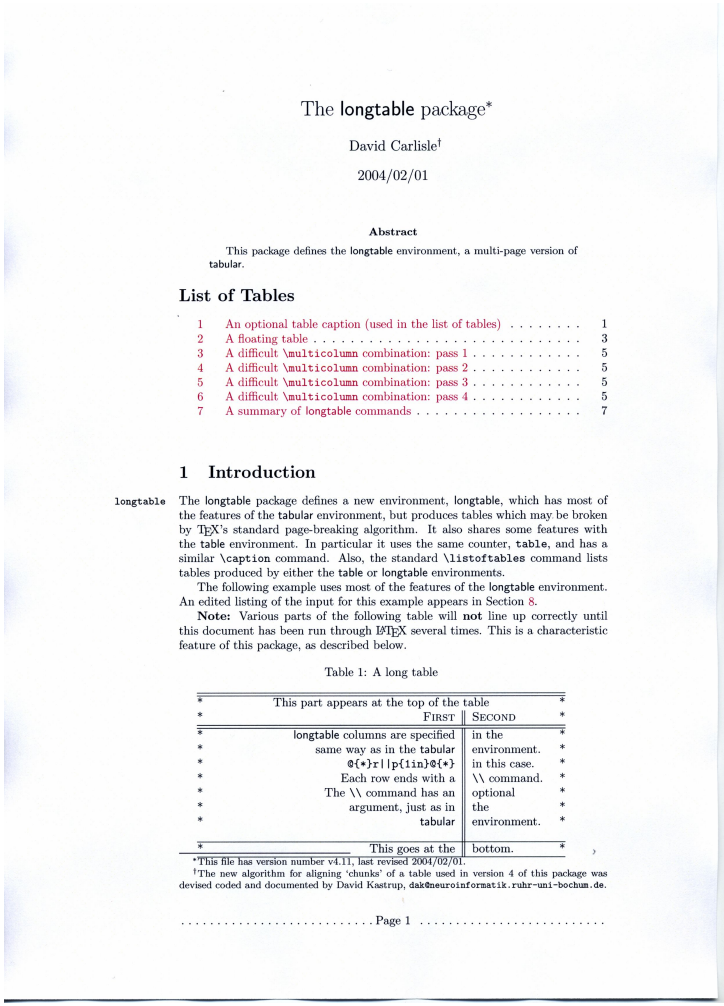
Numérisation

3 “formats”

- format papier
- format électronique “image”
- format électronique “structuré”

Le format image

issu de l'utilisation d'un scanner



The longtable package*

David Carlisle†

2004/02/01

Abstract

This package defines the longtable environment, a multi-page version of tabular.

List of Tables

1 An optional table caption (used in the list of tables) 1
2 A floating table 3
3 A difficult \multicolumn combination: pass 1 5
4 A difficult \multicolumn combination: pass 2 5
5 A difficult \multicolumn combination: pass 3 5
6 A difficult \multicolumn combination: pass 4 5
7 A summary of longtable commands 7

1 Introduction

longtable The longtable package defines a new environment, longtable, which has most of the features of the tabular environment, but produces tables which may be broken by TEX's standard page-breaking algorithm. It also shares some features with the table environment. In particular it uses the same counter, table, and has a similar \caption command. Also, the standard \listoftables command lists tables produced by either the table or longtable environments.

The following example uses most of the features of the longtable environment. An edited listing of the input for this example appears in Section 8.

Note: Various parts of the following table will **not** line up correctly until this document has been run through LATEX several times. This is a characteristic feature of this package, as described below.

Table 1: A long table

	First	Second
This part appears at the top of the table		
longtable columns are specified	in the	
same way as in the tabular	environment.	
@{*}r||p{1in}@{*}	in this case.	
Each row ends with a	\\ command.	
The \\ command has an	optional	
argument, just as in	the	
tabular	environment.	
This goes at the	bottom.	

*This file has version number v4.11, last revised 2004/02/01.

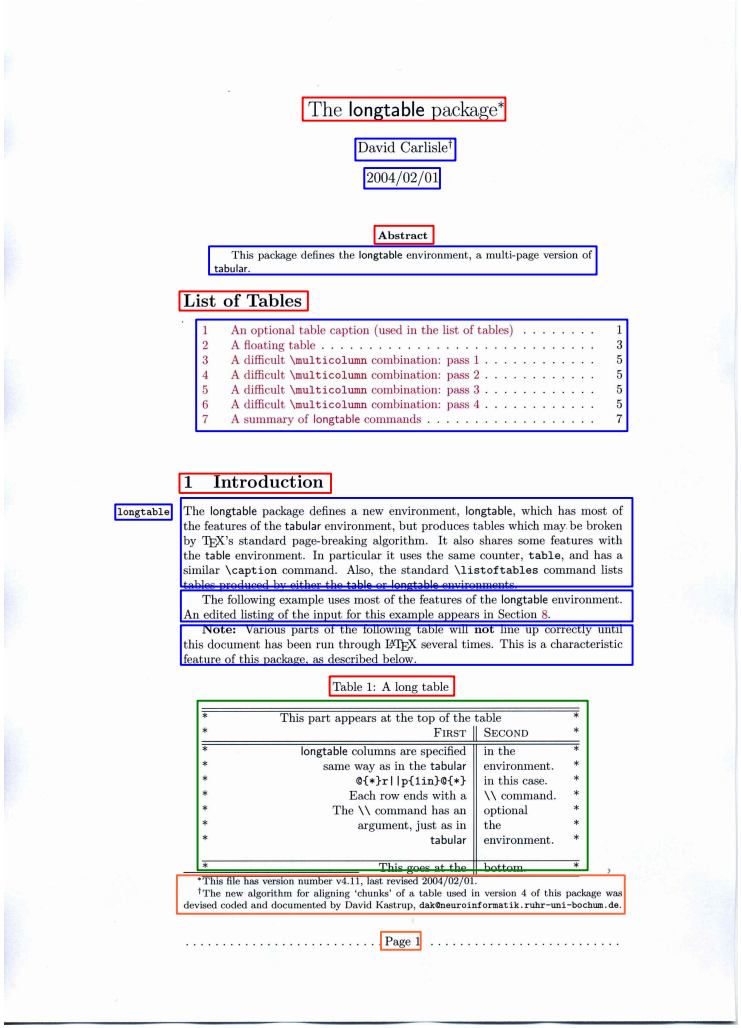
†The new algorithm for aligning 'chunks' of a table used in version 4 of this package was devised coded and documented by David Kastrup, dak@neuroinformatik.ruhr-uni-bochum.de.

. . . . . Page 1 . . . . . .

- information de plus bas niveau
- format bitmap décrivant chaque pixel
- seul le pixel est "connu" de l'ordinateur
- volumineux



Reconnaissance du document



The longtable package*

David Carlisle†

2004/02/01

Abstract

This package defines the longtable environment, a multi-page version of tabular.

List of Tables

1 An optional table caption (used in the list of tables) 1
2 A floating table 3
3 A difficult \multicolumn combination: pass 1 5
4 A difficult \multicolumn combination: pass 2 5
5 A difficult \multicolumn combination: pass 3 5
6 A difficult \multicolumn combination: pass 4 5
7 A summary of longtable commands 7

1 Introduction

longtable

The longtable package defines a new environment, longtable, which has most of the features of the tabular environment, but produces tables which may be broken by TEX's standard page-breaking algorithm. It also shares some features with the table environment. In particular it uses the same counter, table, and has a similar \caption command. Also, the standard \listoftables command lists tables produced by either the table or longtable environments.

The following example uses most of the features of the longtable environment. An edited listing of the input for this example appears in Section 8.

Note: Various parts of the following table will **not** line up correctly until this document has been run through LATEX several times. This is a characteristic feature of this package, as described below.

Table 1: A long table

This part appears at the top of the table	
	First / Second
longtable columns are specified	in the
same way as in the tabular	environment.
@{*}r||p{1in}@{*}	in this case.
Each row ends with a	\\ command.
The \\ command has an	optional
argument, just as in	the
tabular	environment.
This goes at the	bottom

*This file has version number v4.11, last revised 2004/02/01.

†The new algorithm for aligning 'chunks' of a table used in version 4 of this package was devised coded and documented by David Kastrup, dak@neuroinformatik.ruhr-uni-bochum.de.

Page 1

Avant : quelques étapes

- redressement
- binarisation et segmentation
- reconnaissance des caractères
- typage de la structure logique
- traduction dans un format

Aujourd'hui

- tout en même temps (cf. thèse de Denis Coquenet)

Difficultés

Variabilité du contexte

- mise en page : journal, magazine, livre, article scientifique
- culture : langue, sens de lecture
- ...

Variabilité de la présentation

- police de caractère
- taille, style
- alignement
- ...



CLUB S.V.M.

LES FONCTIONS CACHÉES DES LOGICIELS

NÉGOCIER UN RABAIS

VINCENT GAUVIN, DE Bazancourt, a fait ce qu'il est convenu d'appeler une affaire. Il nous raconte comment.

« Il y a peu, je décide de m'acheter un compatible IBM PC. Mais, disposant de peu de moyens, je me tourne vers le marché de l'occasion. Mon choix se porte vers un Amstrad PC 1512 DD monochrome, j'ai constaté avec surprise que les prix pratiqués étaient de 10 à 15 % inférieurs à ceux du matériel neuf. J'abandonne donc cette solution. Puis, sachant que certains hypermarchés vendent du PC, je me rends dans l'un d'eux (Continent de Reims) : guère de surprise, les prix y sont presque les mêmes que chez un distributeur classique. C'est alors que j'ai l'idée de demander au chef de rayon si je pourrais éventuellement acquérir un modèle de démonstration, qui est en triste état, en bénéficiant d'un rabais important. Il me répond qu'il n'est pas question pour lui de se séparer de celui-ci tant qu'il aura du neuf en stock. Mais il ajoute qu'étant donné la difficulté pour eux d'écouler ce matériel, en partie à cause de l'incompétence des vendeurs, il me consent un rabais de 3 690 F sur un modèle neuf.

C'est ainsi que j'ai acheté un PC 1512 DD, avec moniteur couleur, plusieurs jeux, Calcomat, Evolution Sunset et Superbase, pour 6 000 F avec une garantie d'un an ! Il me semble que les autres chaines d'hypermarché ont les mêmes problèmes, tout au moins en province. Peut-être cela intéressera-t-il aussi les lecteurs de SVM ?... »

UNE MACRO POUR SYMphony, de Lucien Bossoutrot de Taverny.

« Votre exemple d'appel à Print Graph à partir de la feuille de calcul de Symphony peut être amélioré pour permettre le basculement répété de la feuille à cet utilitaire d'impression. L'inconvénient de votre exemple vient de ce qu'après le rattachement de MS-DOS, un nouvel appel à la macro entraîne l'erreur : "DOS déjà rattaché". Voici la solution :

\G {CASERREUR PGRAPH|
...|SERVICES| ARDOS"
PGRAPH {SERVICES}
...AADOS" C :\SYM\PGRAPH"

En scindant la macro en deux champs nommés, l'erreur entraîne le saut sur la seconde ligne et permet le rappel de PGraph. La commande sortie de PGraph rappelle alors la feuille de calcul.

Quelques précisions sur le choix des polices de caractères sur Macintosh, d'un lecteur qui a oublié de signer.

« Pour alléger le système du Macintosh, vous conseillez aux utilisateurs d'éliminer les fontes de grande taille. Ceci n'est valable que si l'on ne se soucie pas trop de la qualité d'impression. En effet, lorsqu'on utilise une fonte de taille 12, par exemple, il est utile d'avoir la même fonte en 24, car lors de l'impression en qualité supérieure, le système du Macintosh recherche cette dernière ; si elle n'existe pas, l'ordinateur imprime avec la même densité de points que celle qui figure à l'écran. Dans ces conditions, le résultat peut être très médiocre, surtout dans le cas de polices de caractères au dessin complexe (cursives, gothiques, etc.). En revanche, si le corps 24 est disponible, le système calcule à partir de ce dernier le dessin des caractères à imprimer, et le résultat est nettement meilleur. Voici un exemple obtenu avec la police New York, en impression de qualité supérieure donc, sur Imagewriter 1. »

NEW YORK CORPS 12, avec corps 24 dans le système

NEW YORK CORPS 12, sans corps 24 dans le système

DÉBORDÉ PAR LE SUCCÈS

LE PROGRAMME LE CIEL, proposé gratuitement dans SVM n°47, est en train de faire une victime : son auteur, Jean-Jacques Nguyen, qui croule sous les demandes d'envoi.

« En un mois, nous écrit-il, j'ai reçu environ 500 demandes, dont une centaine sans les timbres pour le retour. Toutes les demandes accompagnées de timbres (ou de diverses contributions spontanées) ont été satisfaites. Mais je ne puis continuer indéfiniment à ce régime, sous peine de perdre la santé ! Aussi, je souhaite faire savoir que je ne suis plus en mesure de continuer à diffuser moi-même le programme. Cette diffusion sera reprise par un ou plusieurs organismes spécialisés dans la distribution des logiciels du domaine public (tractations en cours, surveiller la presse astronomique et informatique). Les personnes qui n'ont pas joint de timbres pour le retour de leur disquette sont priées de le faire (lent : 3,70 F ; rapide : 5,60 F ; étranger : coupons-réponse internationaux), et de rappeler leurs nom et adresse, afin que je retrouve leur demande. »

LA GUERRE DES SALONS

A LA SUITE DE NOTRE ARticle « La guerre des salons est déclarée » (SVM Actualités n°47), Max Hermieu, président du SICOB, nous écrit :

« Mme Poyen a effectivement passé quinze ans dans l'équipe du SICOB, mais le SICOB que j'ai fondé avec quelques collègues en 1950 existe depuis 38 ans. Il est donc prétentieux et fallacieux que Mme Poyen affirme être à son origine. Elle a par ailleurs été à l'origine des « positionnements successifs » de la manifestation MICRO de septembre dont elle dit aujourd'hui qu'elle n'était pas satisfaisante.

[Deux des] autres choix de 1987 concernaient la Convention informatique et l'Exposition S.S.I.I. qui effectivement ne peuvent être considérées comme des réussites : ces deux manifestations étaient placées sous la responsabilité de Mme Poyen. Cela a conduit le SICOB à cesser, en accord avec le SFIB et le SYNTEC, de soutenir la Convention informatique. Pour l'avenir, le SICOB continue à servir la profession pour laquelle et au nom de laquelle il est organisé, tirant parti de ses réussites, corrigeant ce qui doit être amélioré. S'adaptant sans cesse depuis 38 ans aux besoins de ses exposants, le SICOB n'a jamais failli à sa mission. »

166 SCIENCE & VIE MICRO N°49 · AVRIL 1988

Interpréter un document

Extraire du sens

Mais dans quel but ?

L'interprétation d'un document n'a de sens qu'en fonction du besoin

La "compréhension" en fonction de l'objectif

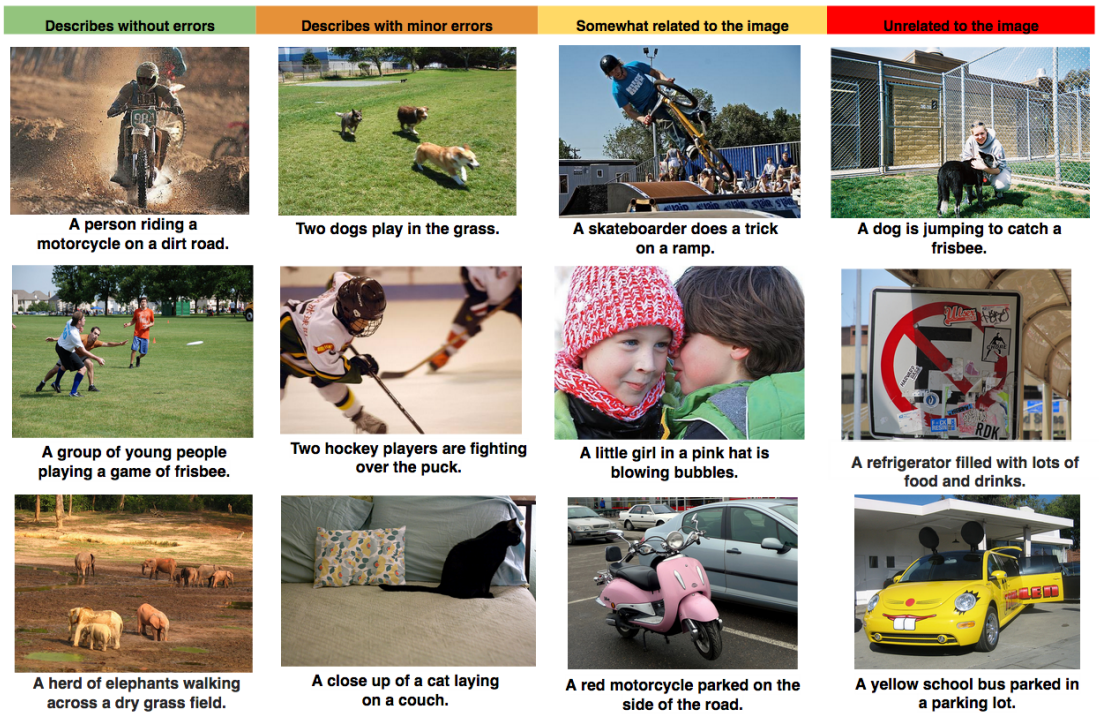
- **indexation**
- recherche d'information particulière (extraction)
- correction orthographique / grammaticale
- veille technologique
- résumé automatique
- traduction automatique
- ...

Documents à indexer

Contenu d'un document électronique

- texte
- image
- son
- video

une indexation principalement textuelle

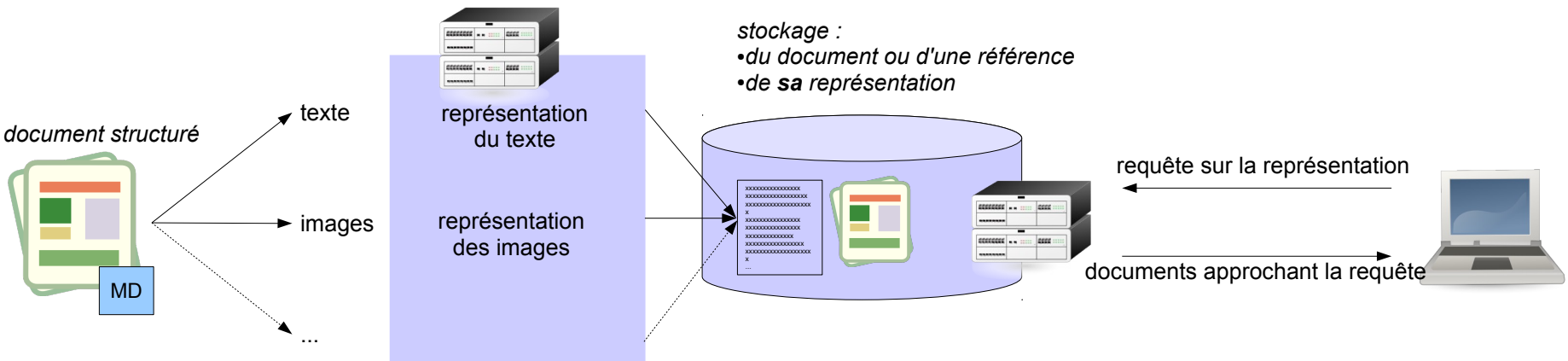


exemple de résultats de l'algorithme de Google

Indexation de document

Objectif

- sélectionner les informations caractéristiques dans le document
- extraire ces informations pour en faire une **représentation**
  - représentation structurée : métadonnées
  - représentation non structurée



Deux types d'indexation

1. manuelle
2. automatique

Indexation manuelle

Caractéristiques

- Elle permet de fixer des valeurs à la représentation structurée :
  - Ces valeurs peuvent être structurée (utilisation de thésaurus) ou pas (mots libres)
- Elle n'est réalisable que lorsque le nombre de documents est faible
- Elle peut varier en fonction de l'indexeur

Paramètres variables

- personne (connaissance, subjectivité, ...)
- temps
- technique
- objectif
- langue
- ...

Indexation automatique

Caractéristiques

- Elle permet d'indexer de gros volumes de documents
- Elle permet de renseigner la représentation structurée et la représentation non structurée
  - représentations structurée : à partir des métadonnées associées aux documents (facile si ce sont les mêmes, difficile sinon) ou à partir d'une analyse du texte
  - représentation non structurée : à partir d'une analyse du texte

Indexation plein texte

Principe

- Extraire automatiquement des informations caractérisant le texte

Analyse du texte

- segmentation (tokenization) : découpage
- analyse lexicale : association des tokens à des lemmes + propriétés
- analyse syntaxique : détermination de la classe morpho-syntaxique dans le corpus
- ...

Extraire les informations pertinentes

Qu'est ce qu'un mot "pertinent" :

- fréquence importante ?
- mot seul ? co-occurences ?
- structure logique : résumé, titre, ...
- ...

$\Longrightarrow$ pondération des mots

$\Longrightarrow$ association des mots de poids important au document

Conclusion 1 / 2

Plan du cours

- Du document papier au document numérique
- Prérequis :
  - XML le métalangage de représentation des documents
  - DTD le langage (non XML) de base de description de grammaires XML
- XPath le langage (non XML) d'identification/de désignation d'une partie d'un document XML
- XSLT le langage (XML) de transformation de documents XML
- XSD le langage (XML) de description de grammaires XML

Conclusion 2 / 2

Plan du cours

- La recherche d'information avec des représentations de documents non structurées
- Les métadonnées ou la recherche d'information avec des représentations structurées
- RDF et SPARQL où comment représenter et interroger des métadonnées
- RDFS et OWL où comment définir des schémas de métadonnées
- Prolog où comment raisonner/inférer sur des connaissances

