Le numérique et ses tendances

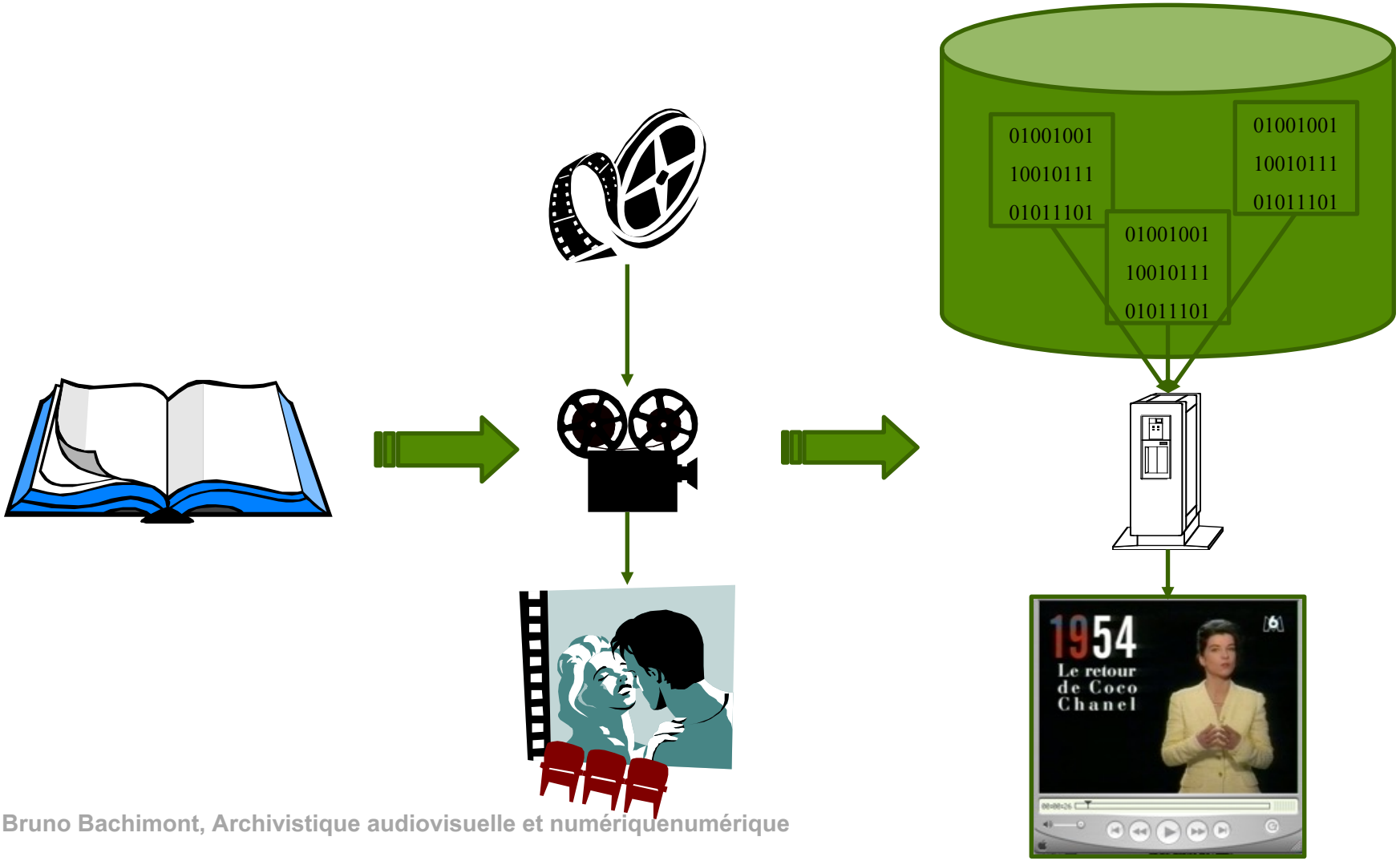
Bruno Bachimont, Université de technologie de Compiègne, France

Tendance documentaire /
Tendance archivistique /
Intégration des IA

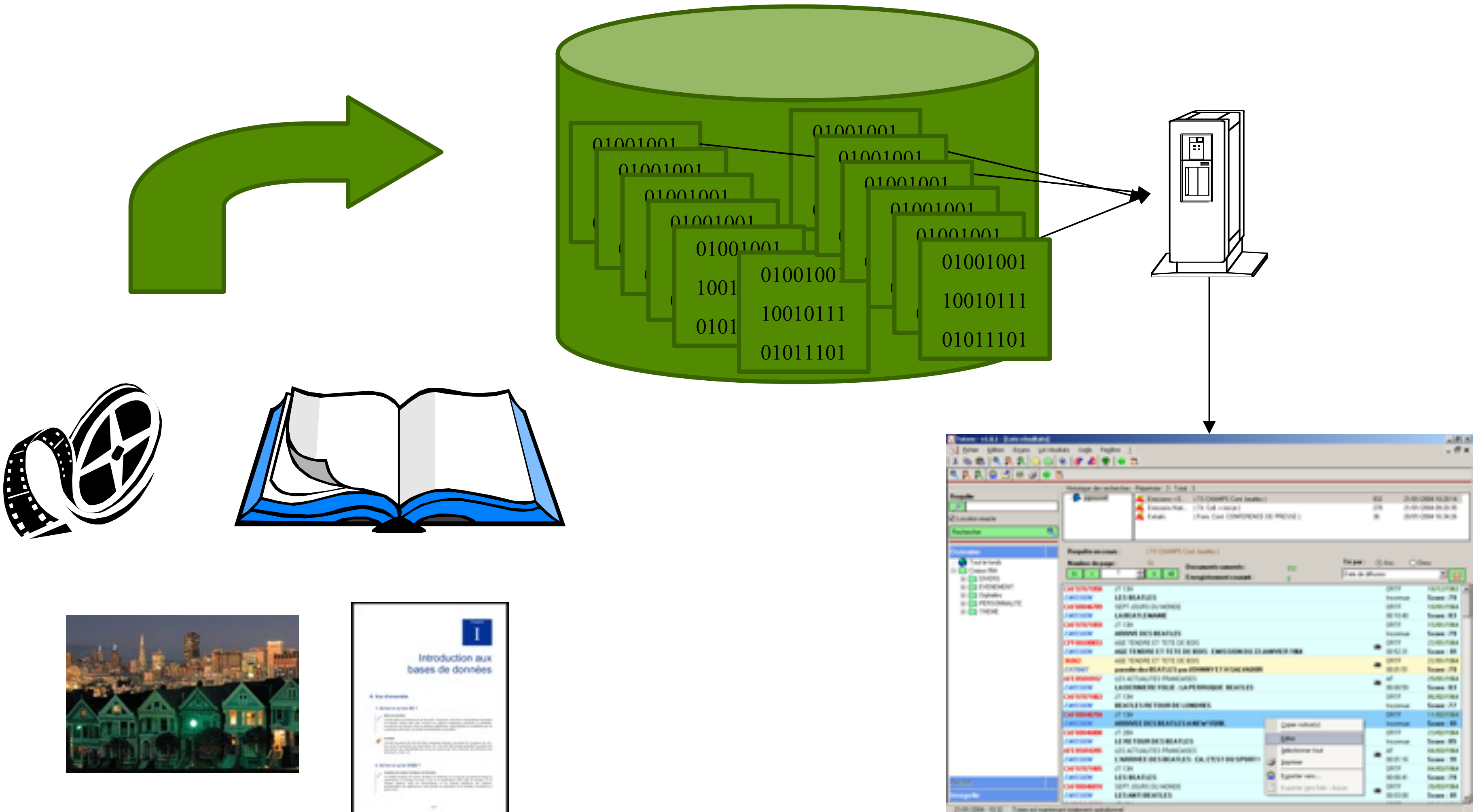
la tendance documentaire

Vers une éditorialisation des archives

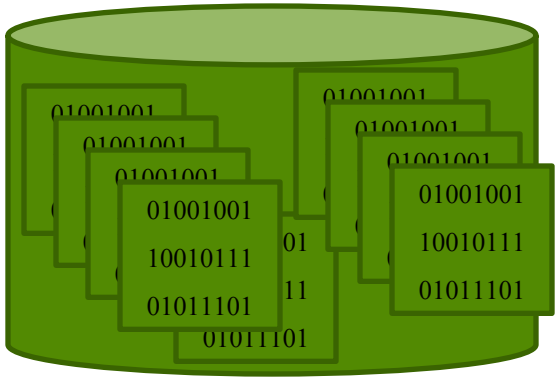
Du contenu numérique…

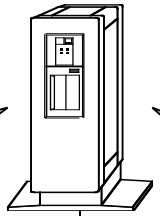


… aux bases numériques



Des supports diversifiés





Tendance documentaire

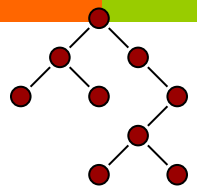
- L'indexation avait tendance à refléter la structure et le contenu du document initial;
- La fragmentation numérique rompt le lien avec le document initial;
- La documentation cherche alors à rendre possible les publications futures.
- La documentation rendait compte de l'origine dans les termes de l'usage, elle tend à reconfigurer l'origine pour l'usage.

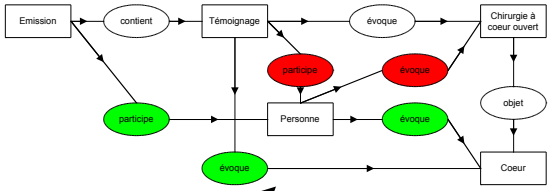
Il en résulte :

- En amont:
  - Virtualisation et dislocation des contenus:
    - Les repères habituels, liés au support physique disparaissent ;
    - On obtient un magma numérique sur disque dur.
- En aval :
  - Reconfiguration multi supports, multi formats, multi usages.
- Au milieu:
  - Possibilité d'adressage arbitraire et illimité ;
  - Recombinaisons illimitées à la discrétion des possibilités calculatoires.
- Par conséquent :
  - Passer d'une organisation physique des contenus à une modélisation logique ;
  - Surmonter l'arbitraire de la fragmentation numérique et de l'accès aléatoire.

Nécessité d'une indexation et d'un modèle d'usage

Du signal au contenu culturel



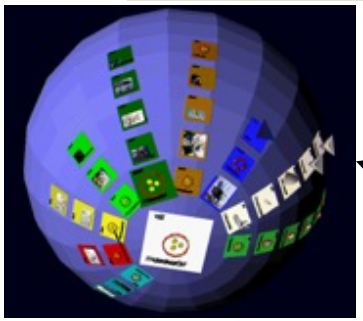


Raisonner
Représenter
Manipuler

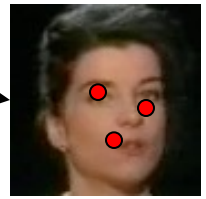
Bla bla bla Bla bla bla

Bla blaBla bla bla Bla bla bla Bla bla

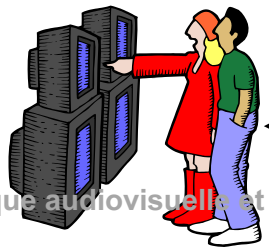
Bla bla Bla bla bla Bla bla bla



Visualiser



Caractériser (extraire)



Comprendre

Statut des index

- Traditionnellement:
  - Index:
    - pointer vers le document pertinent pour l'usage visé;
  - But:
    - Retrouver des documents, tels qu'ils sont;
- Tendance:
  - Métadonnées:
    - Information sur l'exploitation de l'information
  - Caractéristiques:
    - Pointe sur une partie arbitraire du contenu;
  - But:
    - Sélectionner un segment, le transformer pour l'exploiter.

Des mutations:

- Du document à la ressource
- De la recherche à la composition
- De l'indexation à l'éditorialisation

Du document à la ressource

- Ce qu'on appelle document tend à n'être plus que le « contenant » d'origine à partir duquel extraire des fragments;
- Ces fragments sont des ressources, dont le sens viendra des conditions d'exploitation et d'utilisation.
- La cohérence et cohésion documentaire d'origine n'est plus constitutive du contenu de la ressource.

De la recherche à la composition

- La principale motivation applicative de l'indexation est la recherche d'information:
  - Trouver le document ou contenu qui exprime l'information recherchée.
- La principale motivation applicative des métadonnées est la sélection de ressources pour créer de nouvelles informations:
  - Ce n'est pas la ressource en tant que telle qui aura une valeur, mais le contexte dans lequel elle sera intégrée.

La documentation rendait compte de l'origine dans les termes de l'usage,

elle tend à reconfigurer l'origine pour l'usage.

De l'indexation à l'éditorialisation

- Objectif:
  - Trouver des ressources pour créer de nouveaux contenus
- Contrainte:
  - Le fragment est décontextualisé de son contexte d'origine ;
  - Nécessité de le recontextualiser dans son nouvel environnement:
    - Support spécifique;
    - Harmonisation avec des fragments hétérogènes.

Créer de l'information et du contenu pour l'intégrer aux ressources sélectionnées

Plusieurs visions

- Posture « généalogique »:
  - La ressource sélectionnée est enrichie pour être resituée dans son contexte d'origine
  - Travail éditorial qui publie le travail documentaire.
- Posture « amnésique »:
  - La ressource sélectionnée est enrichie dans un nouveau contexte oublieux de l'ancien.
  - Le travail éditorial est une création en coupure avec le travail documentaire effectué sur la ressource.
- Posture « créative » :
  - La ressource est réutilisée indépendamment de son sens et de son origine : création esthétique/artistique (e.g. musique électro-acoustique).

France Libre Actualités -3 nov. 1944 - [0:02:50]

Reconstitution d'un parachutage d'armes pour la Résistance

Images de guerre

TRANSCRIPTION

France Libre Actualités - 3 nov. 1944
- Transcription du sujet 1 - [0:02:50]

Reconstitution d'un parachutage d'armes pour la Résistance

- (Journaliste) -

Dans tous les villages de France, les années de l'occupation ont été aussi celles de la Résistance. A la barbe des Allemands, on attendait des armes parachutées par avion. Aujourd'hui, sur les points que l'ennemi occupe encore, ces parachutages continuent.

(Musique)

- (Résistant) -

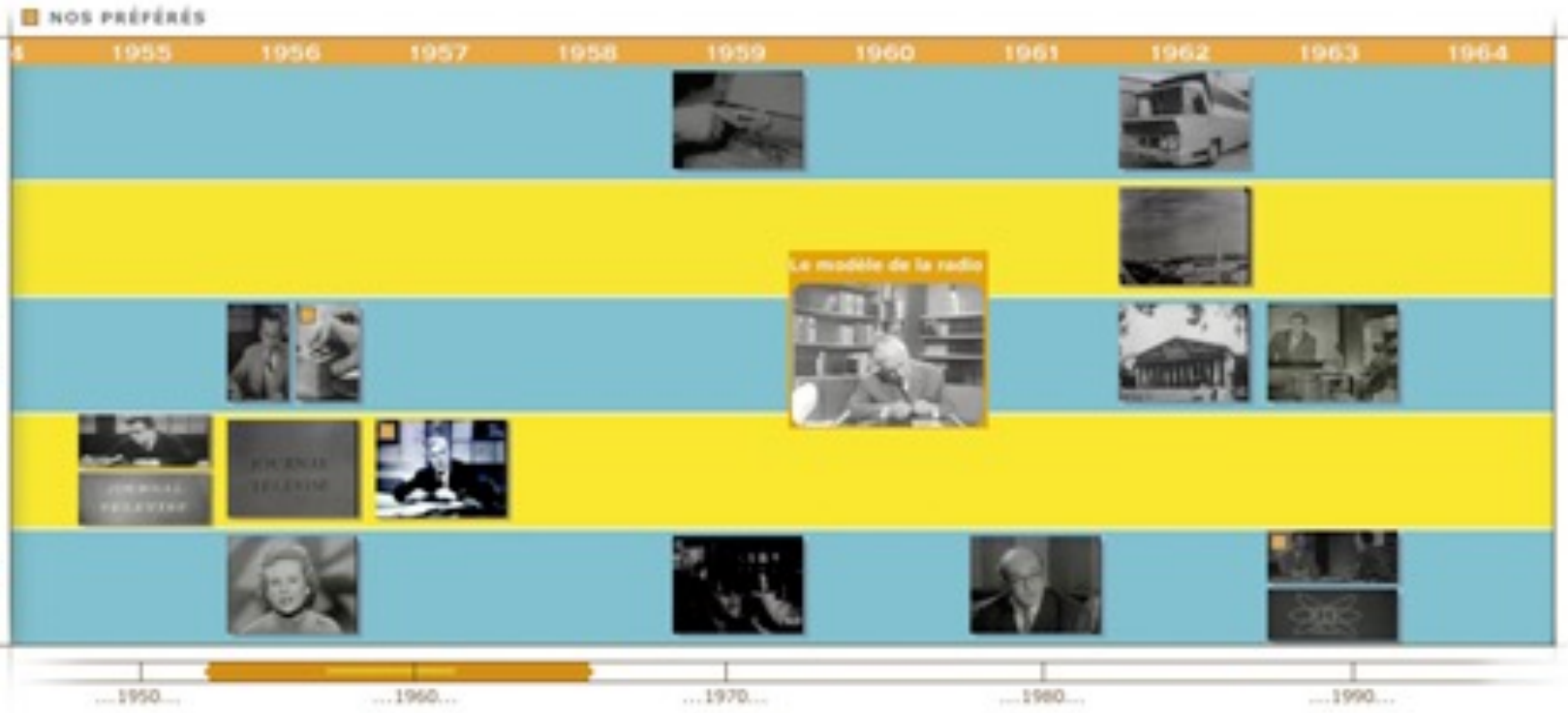
Ici, Londres. Veuillez écouter tout d'abord quelques messages personnels. L'étoile filante repassera. Le chien du jardinier pleure. La bibliothèque est en feu. Nous disons : «La bibliothèque est en feu», deux fois.

- (Journaliste) -

1960 : le modèle de la radio

JT, RTF, 24.11.1960

Claude Darget semble un journaliste particulièrement décontracté en faisant écouter une correspondance d'Alger sur la tentative d'insurrection menée par Lagaillarde et Ortiz. C'est qu'il vit et pense son travail sur le modèle de l'information parlée de la radio : l'image du direct est encore très rare et l'attitude même du journaliste n'est pas pensée pour capter l'attention du téléspectateur.



Ina.fr

Midi Québec

09 JUIL 2010

Partager l'article |

Publié par Rédaction Ina le 9 juillet 2010 à 16:39
Pas de commentaires

La Fondation Zellidja attribue des bourses de voyage pour permettre à des jeunes de 16 à 20 ans d'effectuer seuls un voyage d'étude sur le sujet de leur choix, dans le pays de leur choix. L'INA soutient l'opération et diffuse leur reportage de bord. Découvrez un extrait de "Midi Québec" commenté par sa réalisatrice Anaïs Le Berre :

"Été 2009, Zellidja m'accorde une deuxième bourse pour partir un mois seule au Québec tourner un film.Le concept est simple : filmer tous les jours à midi, les gens que je rencontre et les lieux par lesquels je passe, saisir des personnalités et des ambiances, traduire ma sensibilité, mon regard sur l'inconnu. Bon voyage... Au fait, pourquoi midi ? L'idée m'est venue suite à la vision d'un film de Jim Jarmusch, « Night on earth », qui relie cinq courts-métrages prenant place dans cinq grandes villes du monde grâce à l'intermédiaire d'horloges de gare indiquant chacune les heures respectives des cinq villes au même moment. L'heure est un moyen astucieux pour rattacher différentes ambiances de manière cohérente, tout en permettant un sentiment de simultanéité."

rechercher

CATÉGORIES
- Documentaires

ARTICLES RÉCENTS
- Midi Québec
- La chaleur des glaciers

facebook
Ina.fr
J'aime
Ina.fr a 1,297 fans

Mutation professionnelle

- La documentation ne consiste donc plus seulement à documenter mais à éditer des ressources qu'il faut enrichir.
- Plusieurs niveaux:
  - Enrichissement éditorial
    - Travail généraliste renvoyant à une compétence documentaire;
  - Enrichissement expert:
    - Travail spécialisé renvoyant à la compétence scientifique et l'autorité académique.

Système technique et éditorialisation

- La numérisation du système technique audiovisuel est achevé: il reste à le déployer.
- Le besoin technique se concentre sur les systèmes de repurposing :
  - Un même contenu doit être démultiplié sur des cibles multi-usages, multi-supports, multiformats.
  - Cette démutiplication n'est pas seulement technique, mais éditoriale.

Par exemple

Include Full text capability

show content according search / task

Allow video segmenting and triming

Speed-up production with pre-editing

Stream line the operation matching workflow

Handle complex content management

Aggregate Metadata for rich media repurposing

Facilitate operation with task based action

La tendance archivistique

Vers une économie de la variante

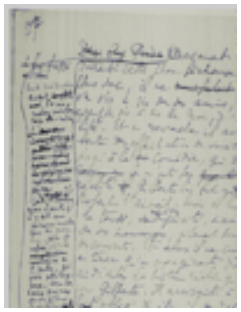
De la solution au problème

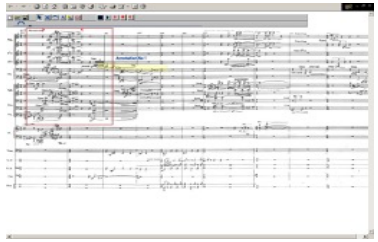
- Deux âges :
  - Le numérique comme solution à la conservation du patrimoine ;
    - Réponse à la
      - Corruption des supports,
      - Accessibilité des contenus.
  - Le numérique comme problème pour la conservation du patrimoine;
    - Obstacles dus à:
      - La prolifération des formats
      - La mutabilité des contenus
      - La complexité du numérique.

Ça partait pourtant bien…

- **Pour** : en théorie, le numérique permet:
  - Recopie parfaite entre les exemplaires;
  - Ubiquité : accès non concurrentiel au contenu ;
  - Universalité : tout contenu peut être numérisé;
  - Homogénéité : le cycle de vie est intégré dans un même système technique interopérable.
- **Contre :** en pratique, on est confronté à :
  - Nouveaux formats et obsolescence logique du contenu;
  - Prolifération de copies transformées, adaptées.
  - Environnements complexes et hétérogène de lecture.

De la théorie à la pratique





Plates-formes et OS : Mac, PC, Windows, MacOS, Linux…

Environnements : Word, WMP, Realplayer, VLC, EMACS, VI….

Formats (métadonnées) : XML, LaTeX, mpeg-7, mxf, rdf, TEI,…

Formats (codage) : unicode, ascii, iso-latin1, mpeg, jpeg,tiff,aiff, pdf…

0010111001110110001100100010010001111101000101

Copie : au bit près: c'est parfait et éternel

Ubiquité: on a tous accès à la même chose

Universel: tout est numérique

Homogène: tout est traitable numériquement

Avec le numérique…

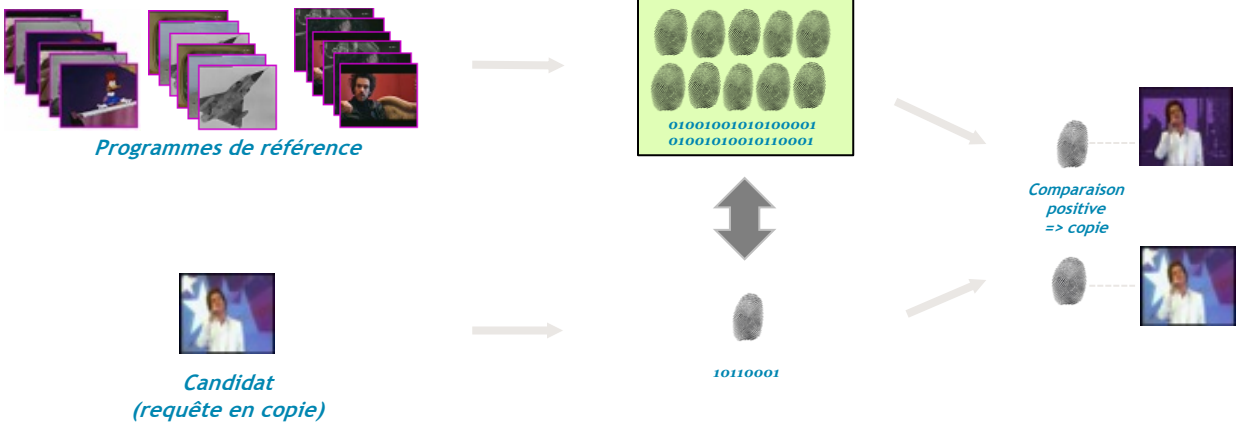
- Le contenu n'est pas préservé:
  - On ne conserve que les ressources et les outils ;
  - On reconstruit le contenu ;
  - Le contenu n'est accessible qu'à travers les fonctionnalités des outils.
- Conséquences:
  - La reconstruction est variable ;
  - Les outils d'accès conditionnent l'interprétation.

Interlude : projet Signature de l'Ina

Détection de copie automatisée à grande échelle

par génération d'empreintes compactes

puis comparaison robuste aux transformations



INA Signature | Principe général

INA Signature | Robustesse aux transformations

G : copie détectée - D : référence

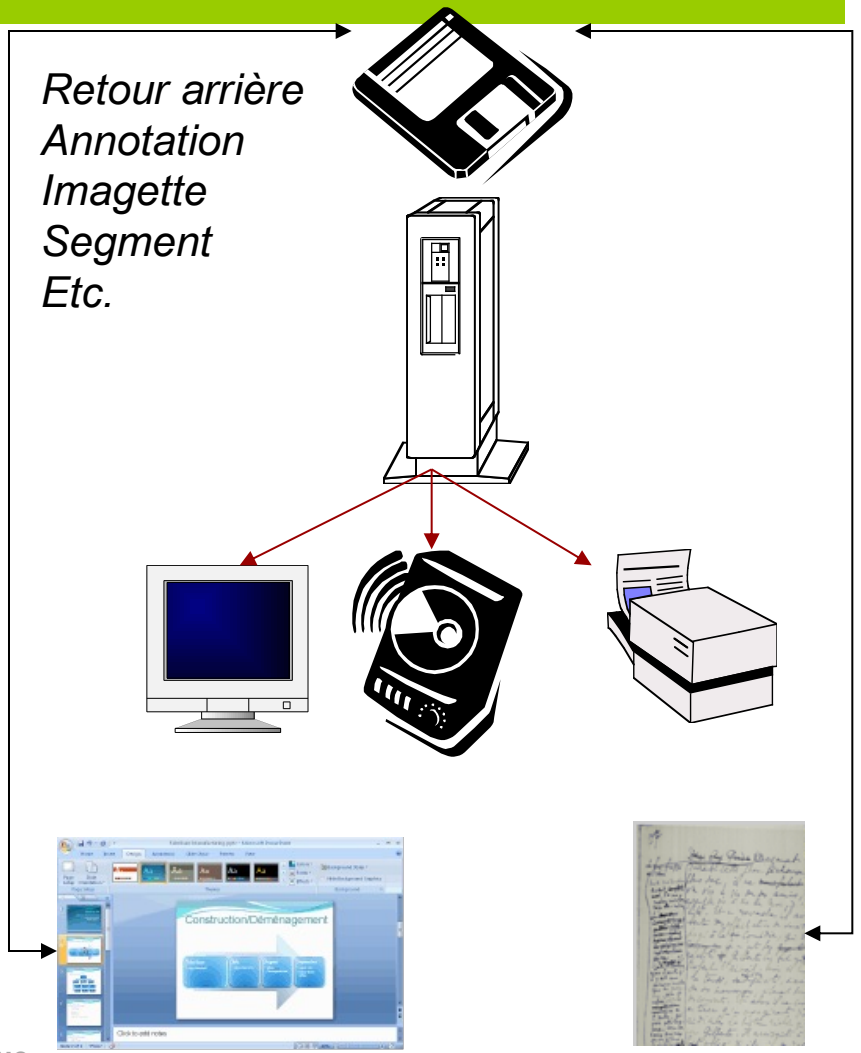
Prolifération des variantes

Economie de la variante

- De nombreuses versions d'un « même » contenu sont en circulation:
  - Altération technique (codage)
  - Altération éditoriale (habillage)
  - Altération sémantique (ce qui est représenté).
- La question se repose sur:
  - Ce qui fait l'identité d'un contenu
  - Les variations acceptables;
  - distinguer les variantes de/s l'original/aux.

Interprétation dépendant des outils

- Documents numériques:
  - Les ressources en mémoire permettent la publication de multiples vues différentes du contenu.
  - Les données stockées sont inaccessibles en tant que telles.
  - Le contenu n'est accessible qu'à partir des vues multiples publiées.
  - Les conditions d'interprétation sont définies par les fonctionnalités des outils de lecture.



Synthèse :

Mutabilité

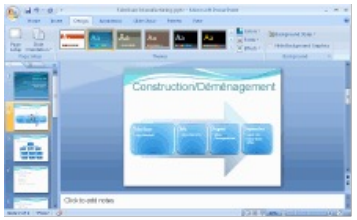
Hétérogénéité

Ubiquité

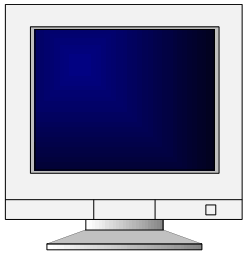
Codage

Éternité

00101110011
10110001100
10001001000
11111010001



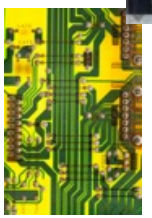
Plongement sémantique



Interprétation

Interaction

Implémentation



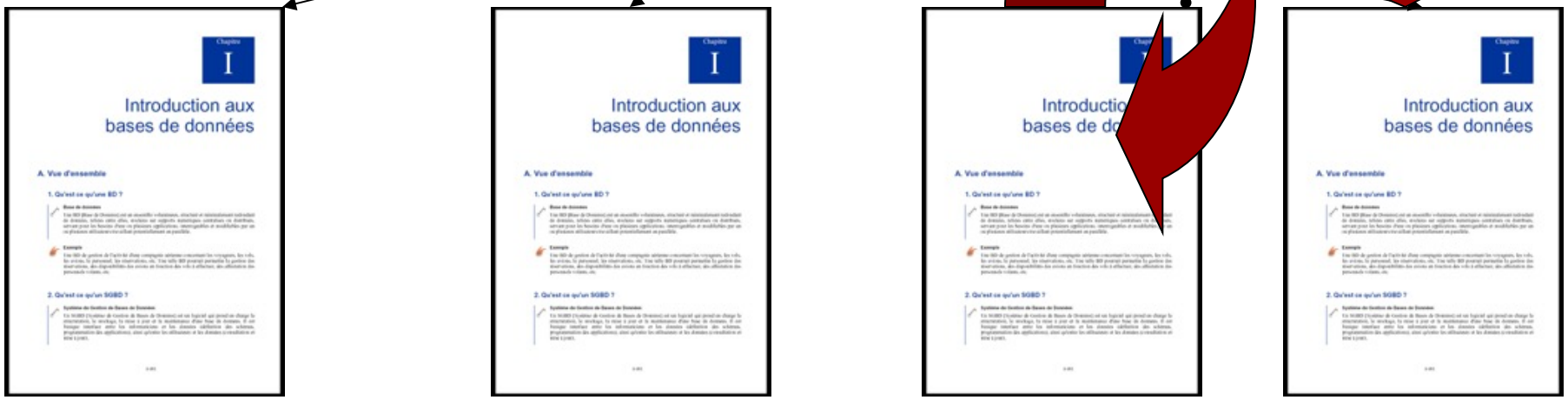
Calcul idéal ***Ancrage matériel*** ***Usage*** ***Variabilité***

Rendre compte de la multiplicité…

Ce qui est stocké: un contenu, en fait, des bits…

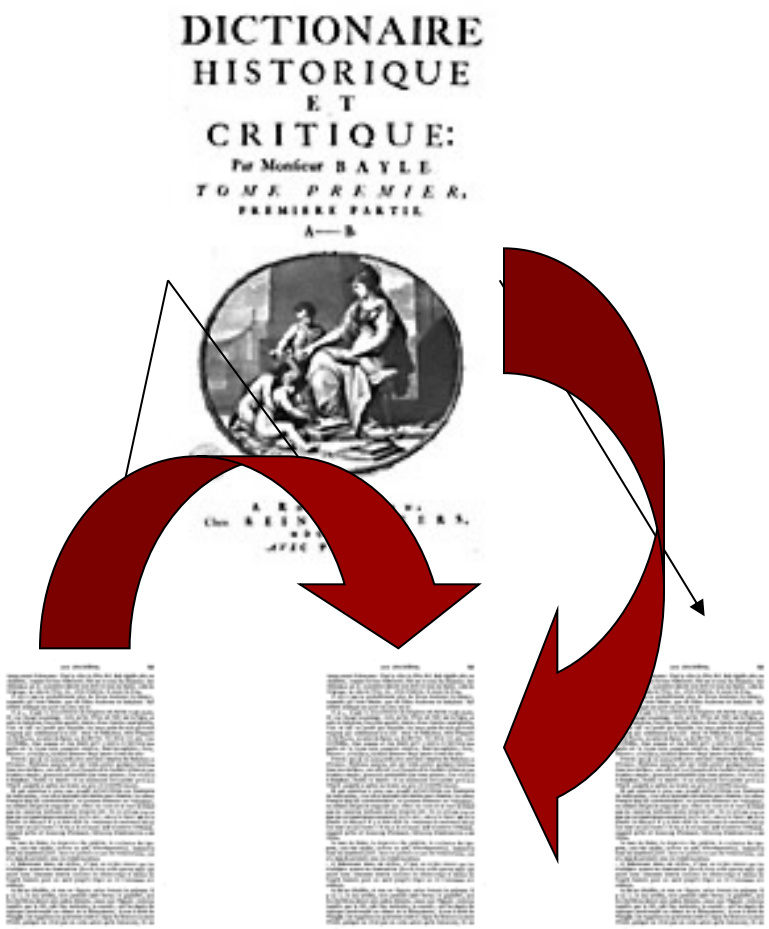
Ce qui est vu est ce qui est publié, i.e. transformé…

Quelle authenticité et fidélité pour le lecteur ?



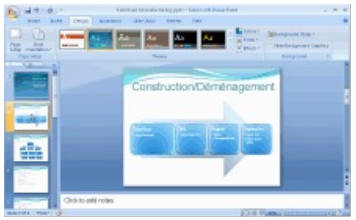
La tradition imprimée

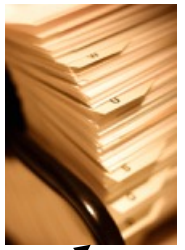
- On a une présentation canonique du contenu
- Chaque lecture peut être confrontée à cette présentation canonique.
- Les lecture deviennent comparables et confrontables car leurs différences apparaissent du fait qu'elles renvoient à la même présentation canonique.
- L'objectivité que prend la version canonique permet l'interprétation.

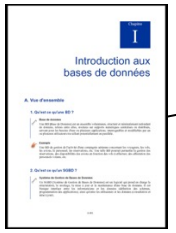


Etablir la référence !

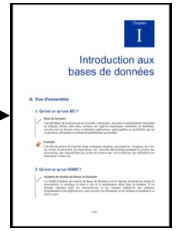
Philologie:
Etablir des versions de référence







Reziropeizop epozir fsdkf mlsdkfml skdfml ksmdkf mlsdkf mksdmf ksmkf mlkffrlmakk mlk amk amlk m km lkm lkml kmlk mlk mlk mlk mk mlk mk mlk mkl mlk mlk mlk mlk mlk mlk mlk mlk mlk mlk ml mlk mlk mkm kml kmlk mlk ml lk lmk mlk lmk lmk lmk lm kjdlkds jkqsjdlkqj lksqjd kjqslkdjj pzeerslkdsqmlkmlqskd kmlqskd lq Repzoirpoezi zi poezir qkdlmsqk mlsqkdlkdslmq rezr ezrez erzr zer zr zre zer zer ezr zer zer ze ze ze z z z z er ezrz e ze zer zer zer ze zer ez zr zer zer



Herméneutique : construire des interprétations

Mais comment établir une telle référence ?

- Tradition de l'écrit:
  - Symétrie entre le l'établissement du texte et son interprétation.
  - Lire c'est déconstruire, déconstruction rendue possible par cette symétrie.
  - Quand la déconstruction est impossible, le document devient propagande ;
- Qu'en est-il pour les nouveaux médias ? Audiovisuel, télévision, multimédia ?
  - Dissymétrie dans l'audiovisuel entre le réalisateur et le lecteur/spectateur.
  - Réception passive interdisant une analyse du contenu.

Le cas de la production audiovisuelle

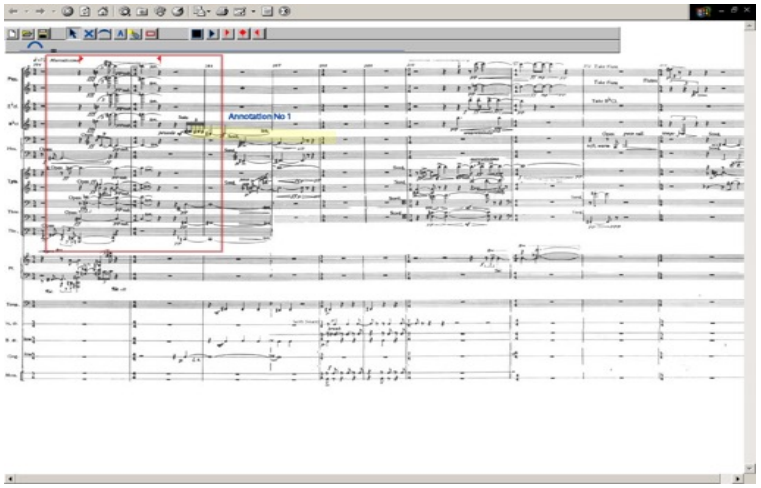
auteur

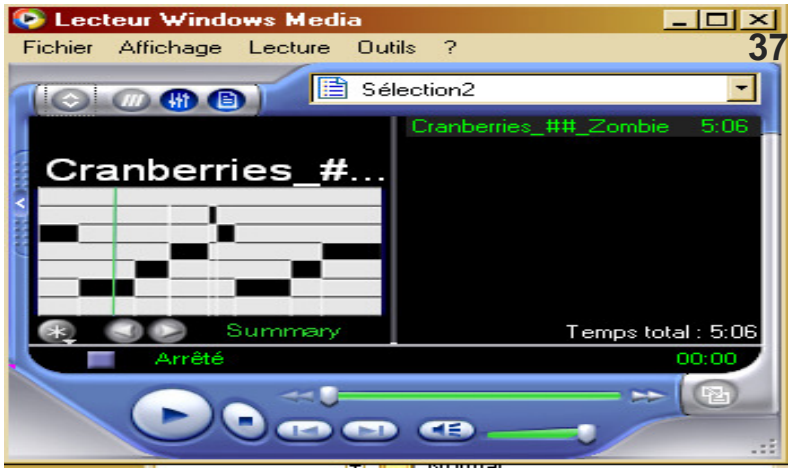
lecteur

?

?

- Production élaborée de contenus divers
- Réception passive d'un contenu apparemment homogène et cohérent.





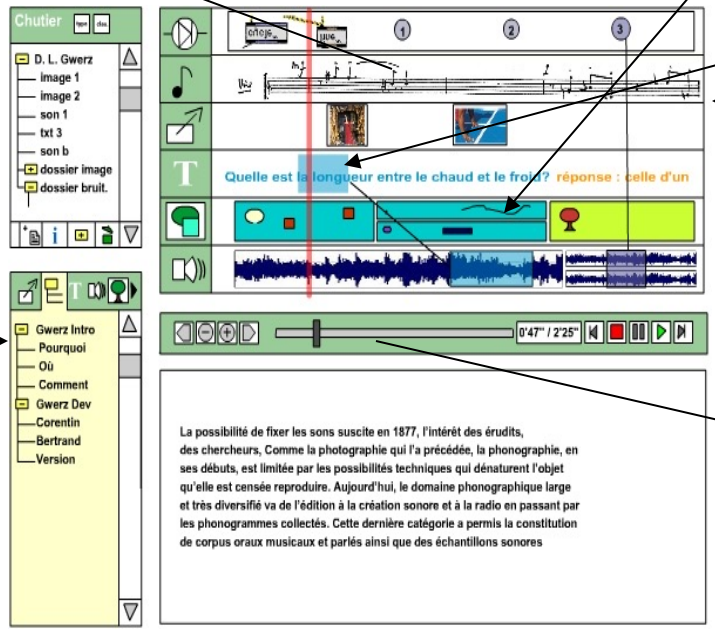
Annotating content

intra-document browsing

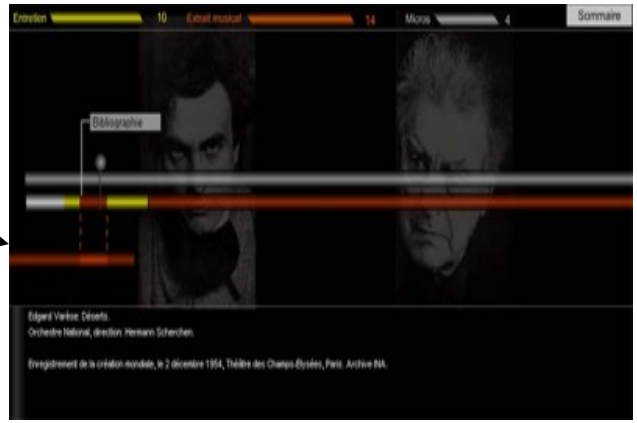
editing, synchronising, linking

Sense units

Selected and indexed units of content



Extracting modules



Multiple delivery device.

Entre philologie et forencics

Imaginez ça une seconde: un homme avec le contrôle total des données volées de milliards de personnes, leurs secrets, leurs vies, leur avenir.

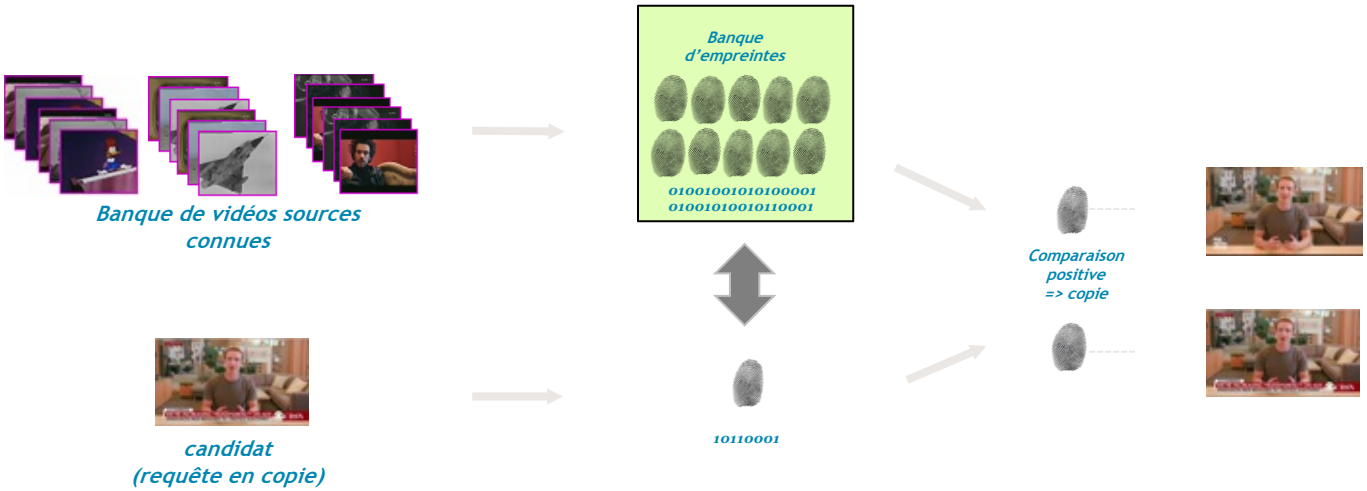
Je dois tout cela à *Spectre*.

Spectre m'a montré que quiconque contrôle les données contrôle l'avenir.

INA Signature | Copie de vidéo surprenante

ina INA Signature

L'outil « signature » : un outil diplomatique et philologique



INA Signature | Détection avec Signature

ina INA Signature

Mise en œuvre

Candidate video	Start
Video_by_bill_posters_uk-ByaVigGFP2U	00:00:03.033

Vidéo transformée - 18 secondes

Reference video	Start
WATCH_Facebook_s_Mark_Zuckerberg_to_cooperate_on_Russian_investigatior	00:04:20.567

Source - 8 minutes 30 secondes

INA Signature | Résultats détection

ina INA Signature

Revenir à la version authentique

INA Signature |
Extrait correspondant dans l'original

(…)
Même dans une situation où nos employés ne sont pas impliqués directement dans la vente [de chaque publicité], nous pouvons faire mieux.

Bon, je ne vais pas prétendre d'ici que nous allons attraper tous les contenus mauvais dans notre système.

Nous ne vérifions pas tout ce que disent les gens avant qu'ils le disent, et franchement, je ne pense pas que la société devrait souhaiter qu'on le fasse.

La liberté signifie que l'on n'a pas à demander la permission avant de faire les choses.
(…)

ina INA Signature

L'origine du délit

BILL POSTERS

BIOGRAPHY

"Interesting fella"

– GCHQ

Working under the pseudonym Bill Posters, Barnaby Francis is an artist-researcher, author and activist who is interested in art as research and critical practice. Poster's works often interrogate persuasion architectures and power relations that exist in public space and online. He works collaboratively across the arts, sciences and advocacy fields on conceptual, intervention, synthetic, net art, photography and installation-based projects. Recently he has established the field of synthetic art with works created using emerging synthetic media (deep fake) technologies.

ina INA Signature

Des photos qui n'en sont pas…

VAN RIPER Frank, Manipulating Truth, Losing Credibility. Washigton Post, 2003

Des photos qui n'en sont plus…

Di Tzeitung : Situation Room de la Maison-Blanche lors de l'assassinat de Ben Laden le 1er mai 2011

Video forensic

Une photo d'archive

Une photo qui a circulé

Des transformations pour mettre en évidence… la manipulation

Paramètres TSL inversés

Teinte 200°, saturation 200%, luminosité 60%

Des similitudes troublantes

Recadrage de la première photo centré sur les nuages provoqués par le 3ème missile (le faux) et le quatrième.

Et pour les archives

- Colorisation
- Segmentation spatiale
- Segmentation temporelle
- Modification / apport d'une bande son sur le flux d'image
- …
- On profite du support, format pour actualiser les archives dans le présent intemporel du moment de leur consultation :
  - Elles sont toujours contemporaines
  - **Présentisme des archives numériques**

Tension

- Renforcer l'immersion :
  - Permettre l'empathie historique
  - Mais abolir la distance critique et tomber dans l'anachronisme psychologique ;
- Mettre à distance le contenu :
  - Permettre la distance critique
  - Mais abolir l'empathie historique par manque d'appétence pour le contenu.
- Comment concilier l'appétence pour le contenu et leur compréhension comme archive ?

Enjeu : temporaliser les données

- L'intégrité n'est pas un acquis à conserver, mais une identité perdue à reconquérir ;
  - Transformer les images et les sons n'est pas à proscrire en soi
- Le gain d'appétence pour les contenus doit permettre de montrer la distance et l'étrangeté des images :
  - Éviter l'immersion ou la sidération pour des contenus où seul le ressenti devient la modalité réceptive
- Paradoxe :
  - Le travail sur les archives doit en faire une promesse qui échoue :
    - Promesse d'une appropriation de contenus comme si on les comprenait comme un soi ;
    - Échec car ces contenus montrent leur étrangeté et leur résistance à l'anachronisme.
  - La promesse d'une appropriation échouant en une objectivation donnant lieu à l'empathie (la promesse) historique (l'échec ouvrant sur la distance critique).

Enjeux de la critique

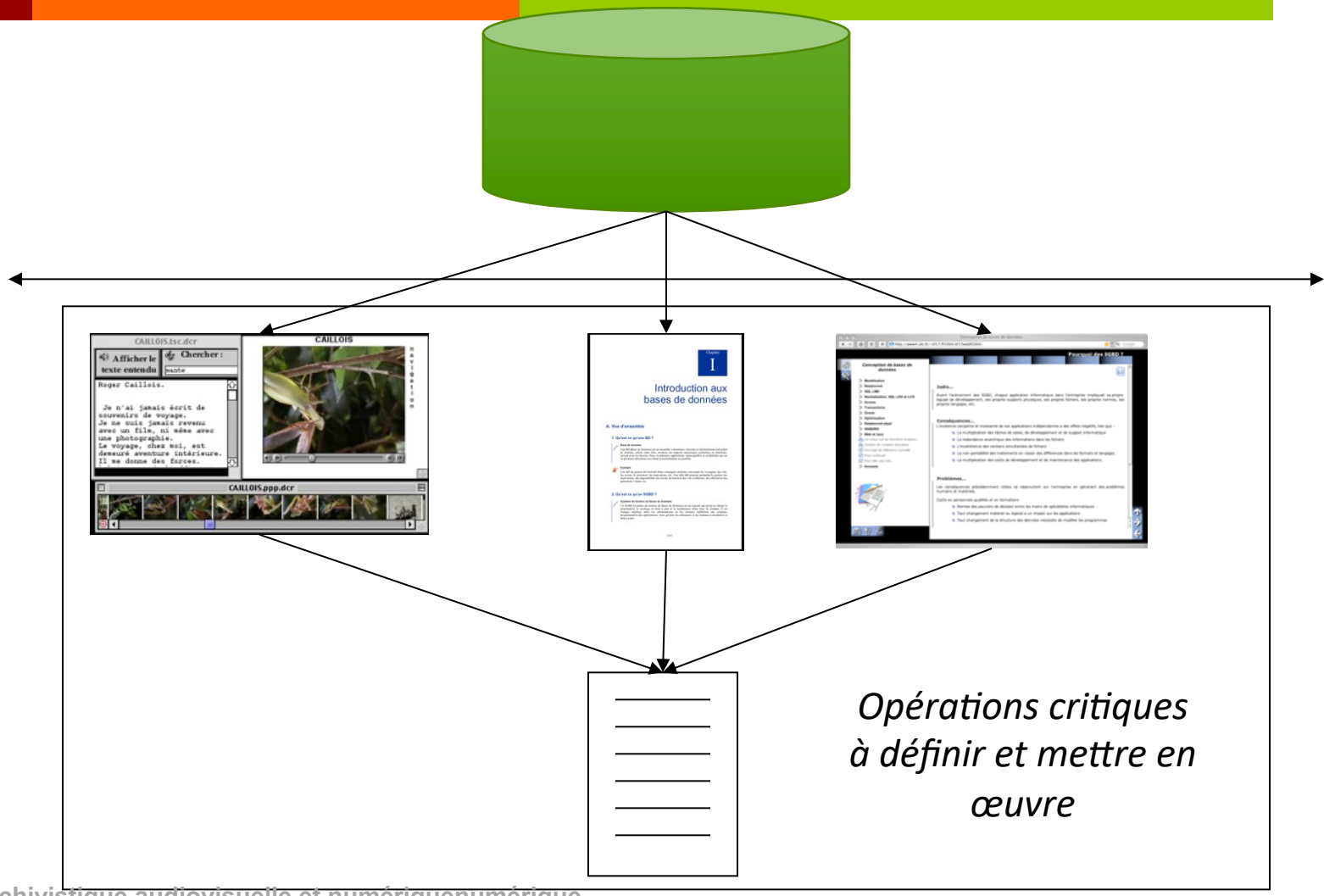
- Enjeu:
  - Rendre intelligible les données et interroger leur authenticité (ce sont les bonnes) et leur intelligibilité (on les comprend).
- Approche :
  - Aborder un récit des données : raconter leur histoire pour comprendre l'histoire qu'elles sont à nous raconter.
  - S'assurer qu'on ne se raconte pas histoires, ou qu'on nous ne raconte pas des histoires / salades…

En résumé

Ressources

Vues

Vue canonique



Un vieux problème !

L’intégration des IA

Des contenus aux données

Données : 4 étapes

1. Le calcul scientifique (1940 – 1950)
   - Ordinateur : *number cruncher*
2. Le transaction de gestion (1950 – 1980)
   - Ordinateur = *symbolic processor*
   - Symbole = nombres ou lettres
3. L'ingénierie des contenus (1980 – 2000)
   - Tout contenu peut être codé et calculé
   - On passe de l'informatique au numérique :
     - E.g. Audiovisuel numérique
4. Le traitement des data (2000 -)
   - Tout enregistrement devient annoté (linked data) ou intégré (big data) pour être porteur d'information.

Calcul : 4 étapes

- 1957 : Le Perceptron
  - Une architecture de neurones formels permet d'apprendre à classer / reconnaître des formes
- 1969 : Minsky montre les limitations de cette approche
  - La recherche se porte sur l'intelligence artificielle symbolique
- 1985 : Rétropropagation du gradient
  - Une architecture de neurones multicouches permet d'apprendre toutes sortes de population
- Années 2000 :
  - Intelligence artificielle statistique : massification des données, puissances de machines.

Plusieurs Webs

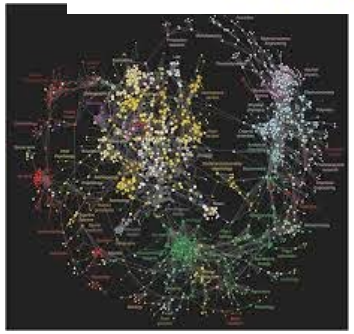
1.0
- Web des documents
- Annuaires, moteurs de recherche

2.0
- Web social
- Folksonomies

3.0
- Web des données (et des objets)
- Métadonnées et ontologies
- Linked Data

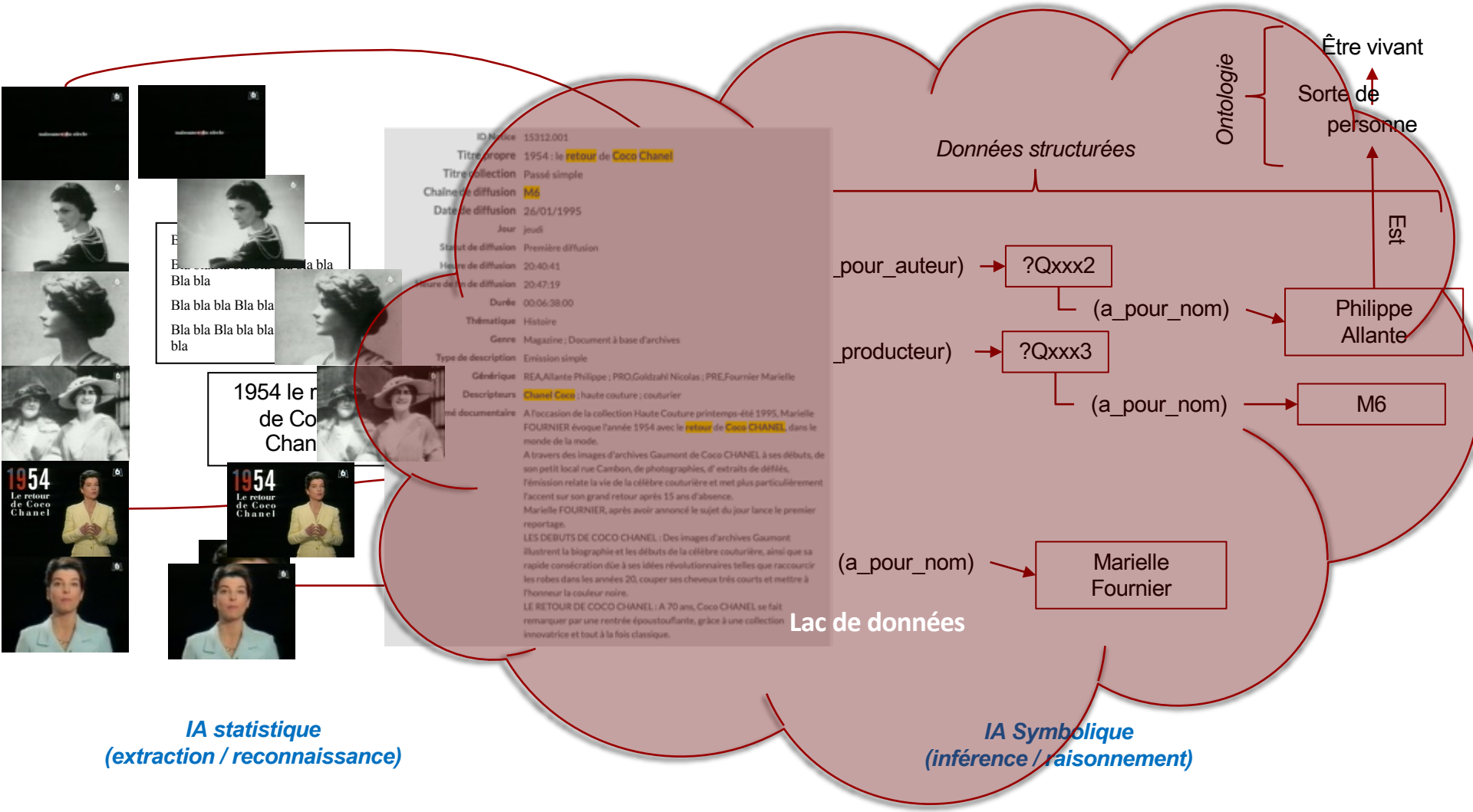
L'exemple des documents

- Du document indexé…
  - Recherche documentaire
    - Paradigme: bibliothèque
- À la ressource annotée…
  - Recherche d'information, agrégation et publication
    - Paradigme : web sémantique
- À la donnée manipulée
  - Données collectées et traduites de manière globale en visualisation.
    - Paradigme : Big Data

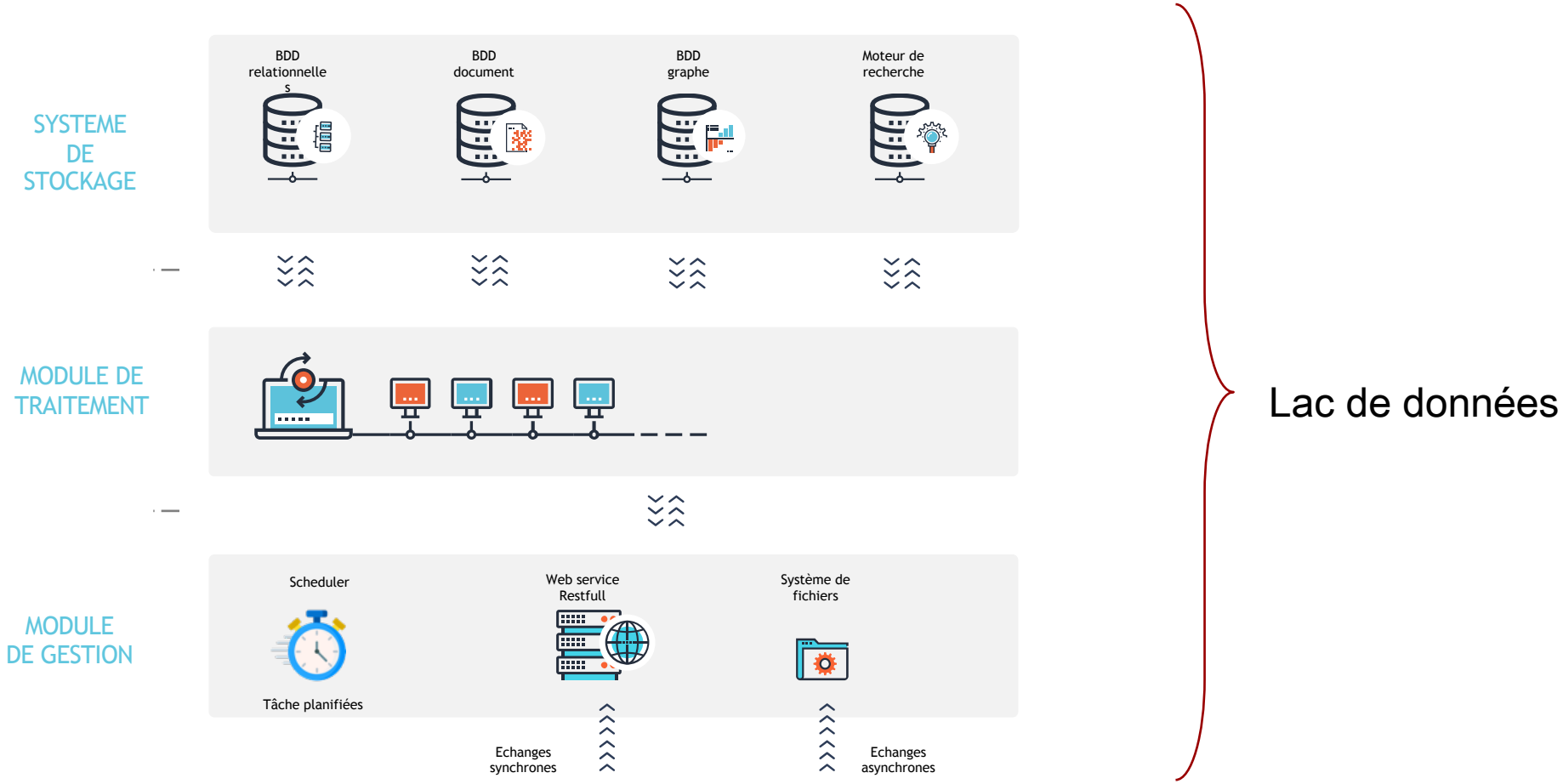


IA symbolique : des notices aux données

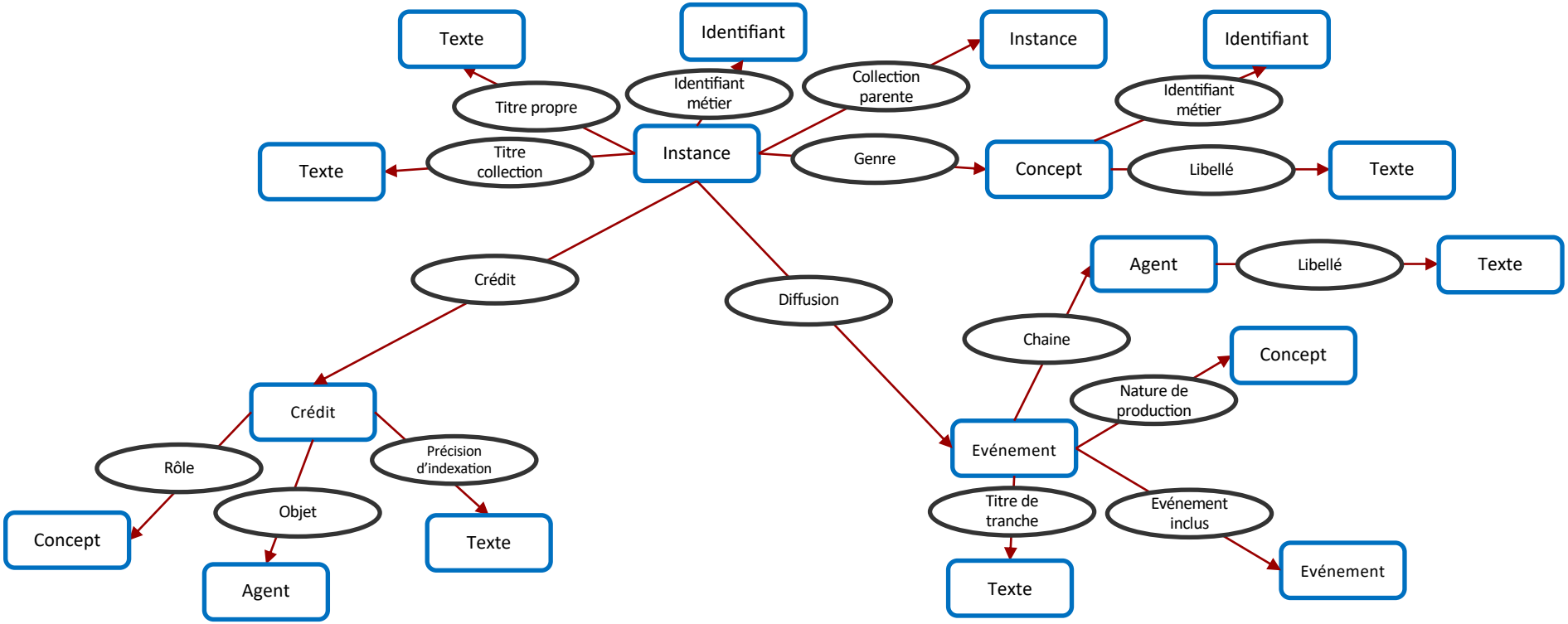
Organiser les données


Bla bla
Bla bla bla Bla bla
Bla bla Bla bla bla
bla
1954 le r
de Co
Chan
1954
Le retour de Coco Chanel
1954
Le retour de Coco Chanel
Titre propre 1954 : le retour de Coco Chanel
Titre collection Passé simple
Chaîne de diffusion M6
Date de diffusion 26/01/1995
Jour jeudi
Statut de diffusion Première diffusion
Heure de diffusion 20:40:41
Durée 00:06:38:00
Thématique Histoire
Genre Magazine ; Document à base d'archives
Type de description Emission simple
Générique REA,Allante Philippe ; PRO,Goldzahl Nicolas ; PRE,Fournier Marielle
Descripteurs Chanel Coco ; haute couture ; couturier
A l'occasion de la collection Haute Couture printemps-été 1995, Marielle FOURNIER évoque l'année 1954 avec le retour de Coco CHANEL dans le monde de la mode.
A travers des images d'archives Gaumont de Coco CHANEL à ses débuts, de son petit local rue Cambon, de photographies, d' extraits de défilés, l'émission relate la vie de la célèbre couturière et met plus particulièrement l'accent sur son grand retour après 15 ans d'absence.
Marielle FOURNIER, après avoir annoncé le sujet du jour lance le premier reportage.
LES DEBUTS DE COCO CHANEL : Des images d'archives Gaumont illustrent la biographie et les débuts de la célèbre couturière, ainsi que sa rapide consécration dûe à ses idées révolutionnaires telles que raccourcir les robes dans les années 20, couper ses cheveux très courts et mettre à l'honneur la couleur noire.
LE RETOUR DE COCO CHANEL : A 70 ans, Coco CHANEL se fait remarquer par une rentrée époustouflante, grâce à une collection innovatrice et tout à la fois classique.
Lac de données
Données structurées
Ontologie
Être vivant
Sorte de personne
Est
_pour_auteur)
?Qxxx2
(a_pour_nom)
Philippe Allante
_producteur)
?Qxxx3
(a_pour_nom)
M6
(a_pour_nom)
Marielle Fournier
IA statistique
(extraction / reconnaissance)
IA Symbolique
(inférence / raisonnement)

Exemple à l'Ina



Modèle de données cible



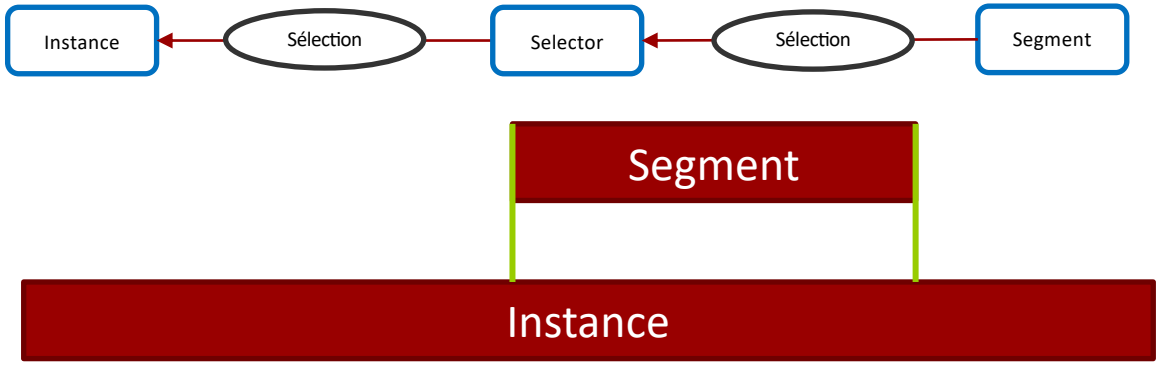
De la notice au nouveau modèle

Identifiant de la notice extrait :	I05007967
Titre de l'extrait :	Jean Rouch sur le racisme
Identifiant de la notice integrale :	CPF86614340
Titre propre de l'intégrale :	Le racisme : 1ère partie
Titre collection de l'intégrale :	Faire Face
Société de progr. de l'intégrale :	RTF (Radio Télévision Française)
Canal de diffusion de l'intégrale :	1ere chaîne
Date de diffusion de l'intégrale :	11/09/1961 00:00:00
Thèque de l'intégrale :	CP (Vidéothèque production)
Durée de l'extrait :	00:13:44
Producteurs (Aff.) :	Société de production - Radiodiffusion Télévision Française (RTF) - Paris - 1961
Genre :	Documentaire ; Interview entretien ;
Thematique :	Sciences ; Société ;
Générique (Aff. Lig.) :	PAR Rouch, Jean ;
Descripteurs (Aff. Lig.) :	DET: racisme ; DET: société ; DET: interview ; DET: cinéma ; DET: anthropologie ; DET: Schweitzer, Albert ; DET: sexualité ;
Description de l'extrait :	L'ethnologue cinéaste Jean Rouch explique en quoi le racisme n'est pas fondé scientifiquement. Les caractères raciaux utilisés en anthropologie physique ; le métissage généralisé de l'humanité ; exemple d'un étudiant africain génie des mathématiques. Il donne son avis sur l'évolution du racisme, causé par le maintien de la suprématie d'une race sur une autre. Un extrait de son film "La pyramide humaine" ponctue ses propos. La survivance de stéréotypes racistes, dans tous les groupes humains ; le contre-exemple du docteur Albert Schweitzer, incapable selon Jean Rouch de s'intéresser à la culture où il vit (les Bantous) ; l'amitié de Jean Rouch avec ses amis noirs, fondés sur l'honnêteté. A partir d'extraits de "Chroniques d'un été", Jean Rouch appelle au métissage généralisé, citant Léopold Sedar Senghor. Il regrette le "racisme sexuel".
Corpus (Aff.) :	Corpus: Afrique et Europe : entre deux cultures - (Corpus INA > PERSONNALITE > ROUCH JEAN > Propos > Afrique et Europe : entre deux cultures)
Document dévolu INA :	Dévolu INA

Date de création :	07/01/2005
Date de modification :	30/09/2022
Documentaliste :	GER
Dernier intervenant :	EDC
Type de l'extrait :	Extrait Thématique
Type de fonds de l'intégrale :	Production
Couleur / NB :	NOIR ET BLANC
Muet/Sonore :	Document sonore
Indexation (info.) :	Atteint: Oui Validé: O Date: 21/08/2006

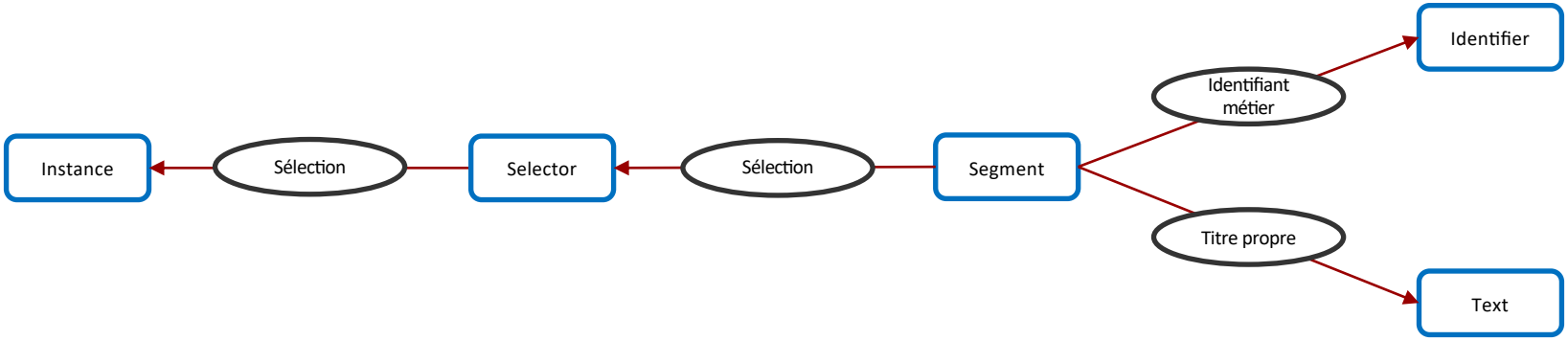
Segment et instance

La notice I05007967 désigne un **extrait (segment)** relié à une **intégrale (instance)**



Text et identifiant du segment

L'identifiant « source » I05007967 est porté par un **identifier du segment**
Le titre de l'extrait est porté par un **text du segment**



IA statistique (DL et ML) : des contenus aux données

reconnaître les données

IA et traitement des archives

- L'enjeu est de reconnaître et de rendre manipulable l'information contenue dans les archives :
  - Outils de reconnaissances de formes fondées sur les IA neuronales (deep learning) ou statistique (machine learning)
  - Par exemple :
    - Reconnaissances des visages, des voix, scènes, des écritures manuscrites, etc.

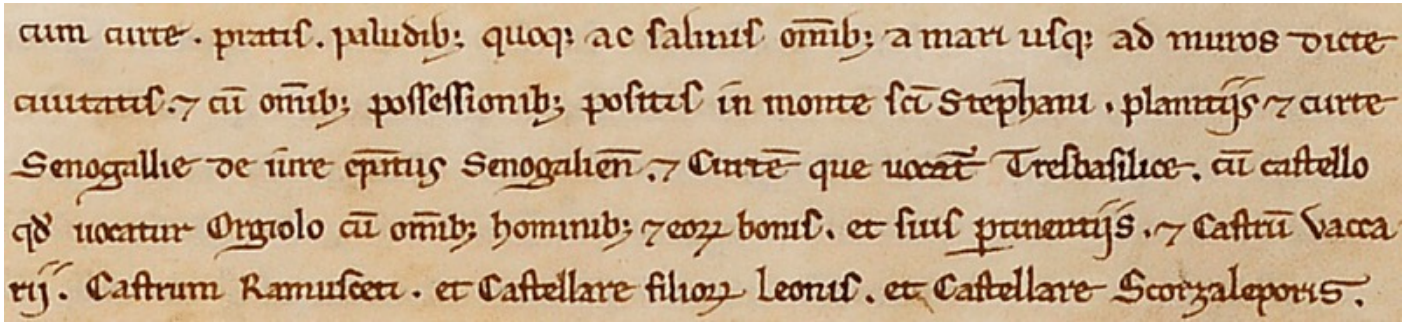
Un exemple : Projet In codice Ratio

Etude de Mathilde Agaisse Camille Garcia, InaSup 2021

Des archives du Vatican

- **85 km de rayonnages**, plus de 600 collections d'archives
- Des textes allant **du VIII au XX è siècle**
- Fonds composé de manuscrits écrits en **minuscule caroline**

Écriture caroline



Ecriture caroline

Claude Médiavilla
Calligraphie
Editions de
l'Imprimerie
Nationale 1993

Une difficulté : la segmentation

- Le problème à traiter :
  - Une segmentation imprécise des caractères dans la minuscule caroline
- Une solution :
  - considérer les mots comme une suite de segments et non comme une suite de caractères

Exemple du mot latin "anno" donnant lieu à de nombreuses interprétations par la ROC : “aiiiio”, “aimo”, “amio”, “aniio’, “aiino”, and “ainio”. (aucun mot latin).

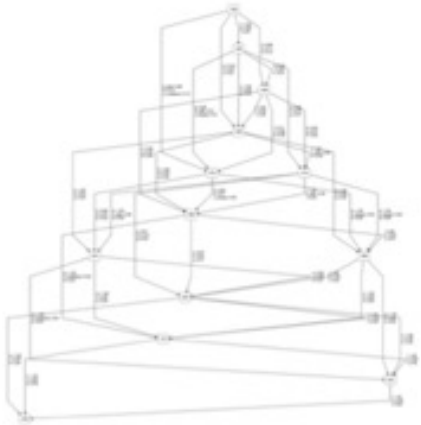
Segmentation en tranches

Segmentation puzzle

Approche

- Entrainer leur programme à **reconnaître l'épaisseur** des traits de plume.
- Les traits les plus fins sont généralement ceux **liant les caractères** les uns aux autres.
- **Utiliser les probabilités :**
  - Un segment "**iii**" est probablement un "**m**" mal déchiffré

(a)



(b)

Fig. 3: Cut-points for the word "culpam" and corresponding lattice. We use green for actual character boundaries, and red otherwise.

path	prob
culpam	
cullum	$1 \cdot 10^{-4}$
culpam	$8 \cdot 10^{-7}$
culluni	$3 \cdot 10^{-7}$
criminis	
criminis	$8 \cdot 10^{-7}$
crinunis	$6 \cdot 10^{-8}$
crinuius	$4 \cdot 10^{-8}$
uiuscemod(i)	
uiufemod	$2 \cdot 10^{-2}$
uuifemod	$2 \cdot 10^{-3}$
uiiifemod	$5 \cdot 10^{-4}$

Table 1: Path probabilities for the words in Figure 4.

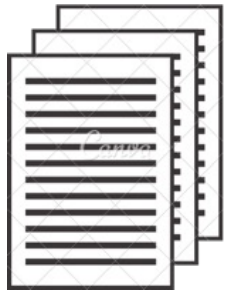
Un préalable: les données d'apprentissage

- Avant d'utiliser l'outil : **le nécessaire collecte de datas.**
- Construction d'une **application de crowdsourcing.**
- **Participation de lycéens** pour entraîner la machine.

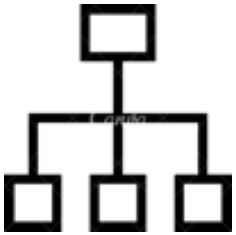
Processus global

Image en haute résolution d'un manuscrit

Classification entre combinaison de traits reconnues et inconnues



La page est segmentée par **crowdsourcing** en mots, les mots sont segmentés en traits



Classement des probabilités de résultats

Conclusion générale

- Le numérique contribue à :
  - Déconstruire l'origine
  - Démultiplier les versions et les variantes
  - Le contenu numérique est apatride et gyrovague.
- Conséquence
  - Déconstruire les variantes pour reconstruire l'origine, supposée et effectuer la critique des contenus.
  - Outre la nécessité du récit critique de l'archive (cf. philosophie de l'archive), on doit avoir la déconstruction critique de ses variantes.
  - Le numérique ne fait que renforcer l'exigence critique.