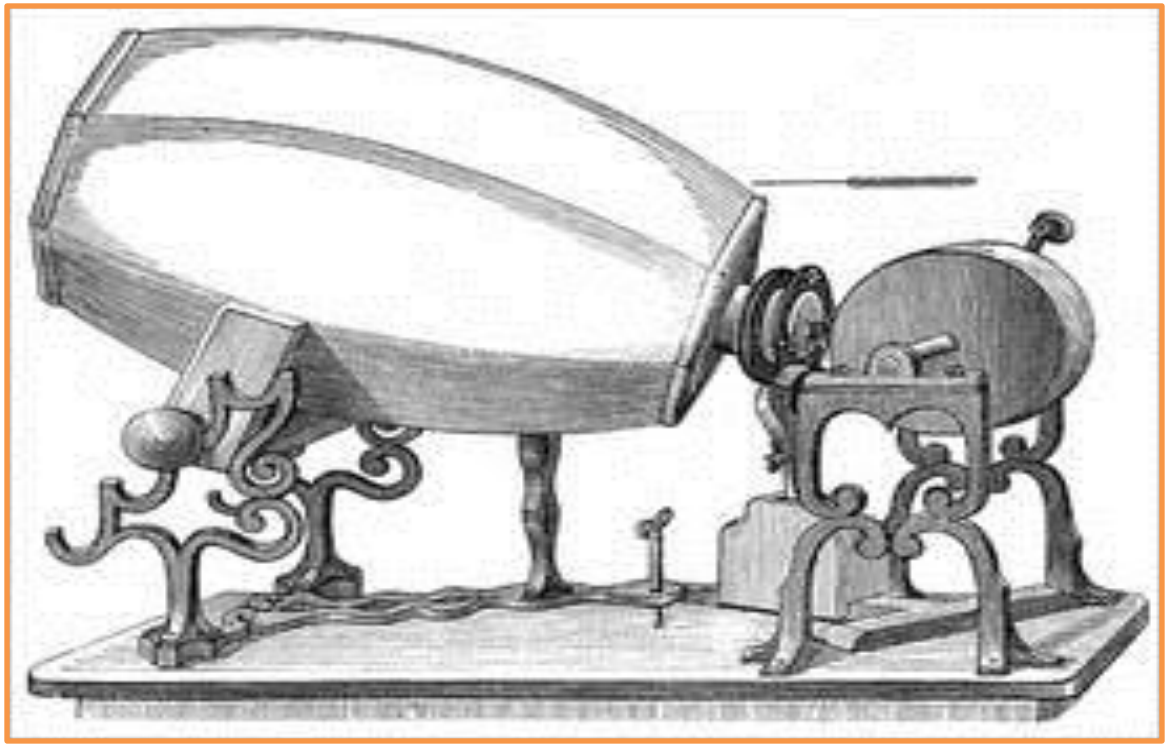
Geler/dégeler les paroles : pourquoi, comment ?

Benoît Habert

ICAR – ENS Lyon & EDF R&D

Que nos voix demeurent



Phonautographe d'Edouard-Léon Scott de Martinville – modèle 1859
In Franz Josef Pisko *Die neuere Appparate der Akustik*, Vienne, 1865

Que nos voix demeurent

- Plus vieil enregistrement sonore au monde (150 ans)
  - 9 avril 1860, dix secondes - probablement une femme : « Au clair de la lune, Pierrot répondit… »
  - Edouard-Léon Scott de Martinville – Phonautographe, inscrivant une ligne blanche ondulant sur une bande papier recouverte de noir de fumée
  - 17 ans avant le phonogramme de Thomas Edison, 28 ans avant le premier enregistrement (rouleau de cire – oratorio d'Haendel)

Que nos voix demeurent

- « Phonautogramme » découvert par David Giovannoni et Patrick Feaster (association First Sounds) – dépôts INPI (Institut national de la propriété intellectuelle) et Académie des sciences
- Phonautographe: dispositif pour enregistrer le son visuellement, pas pour le rejouer
- Reconstitution à l'aide d'un laser de la forme du sillon, analogue au signal sonore initial – technique mise au point et adaptée par Carl Haber et Vitaliy Fadeyev (Lawrence Berkeley National Laboratory)

Que nos voix demeurent

- Paroles dégelées

Alors [Pantagruel] jeta pour nous sur le tillac de pleines poignées de paroles gelées... Après avoir été un peu réchauffés entre nos mains, [les mots] fondaient comme neige. Nous les entendions bien mais nous ne le comprenions pas, car c'était une langue barbare.

Rabelais *Le quart livre*, ch. 56

- Hergé *Les oranges bleues* Tournesol enregistre un message dans des glaçons : on entend le message quand les glaçons fondent

Que nos voix demeurent : enjeux

- Patrimoine historique : le français tel qu'on l'a parlé (ESLO)
- Le français parlé aujourd'hui dans sa diversité géographique, sociale, générationnelle, situationnelle

Linguistique

- « Écrire » l'oral pour
  - le donner à lire (vitesse, handicap)
  - l'indexer (chercher des informations)
- « Parler » l'écrit pour
  - qui ne peut lire
  - automatiser des tâches répétitives (horaires, etc.)

Ingénierie linguistique

Que nos voix demeurent : enjeux

… l'inflexion des voix chères qui se sont tues.

De sa mère, elle se rappelle les yeux, les mains, la silhouette, pas la voix, ou sinon de façon abstraite, sans grain. La vraie voix est perdue, elle n'en possède aucune trace matérielle. Mais des phrases lui viennent souvent spontanément aux lèvres, que sa mère utilisait dans le même contexte, des expressions qu'elle n'a pas le souvenir d'avoir utilisées avant, « le temps est mou », « il m'a tenu le crachoir », « chacun son tour comme à confesse », etc. C'est comme si sa mère parlait par sa bouche et avec elle toute une lignée de gens.

Annie Ernaux *Les années* Gallimard 2008

Que nos voix demeurent : conditions

- Linguistique
  - Besoin de corpus transcrits et annotés très précisément, sur des dimensions spécifiques
  - Transcriptions coûteuses (jusqu'à x60)
- Ingénierie linguistique (reconnaissance/synthèse de la parole)
  - Utilisation des enchaînements mots les plus probables de sons, de
  - Besoin de très vastes corpus transcrits
  - Transcriptions moins fines (industrialisation)

Que nos voix demeurent : étapes

- Enregistrer les « paroles qui volent »
- Transcrire / annoter

- Le dur désir de durer
  - Quel apport du numérique ?
  - Quelles fragilités du numérique ?
  - Quelles mesures prendre face aux fragilités du numérique ?

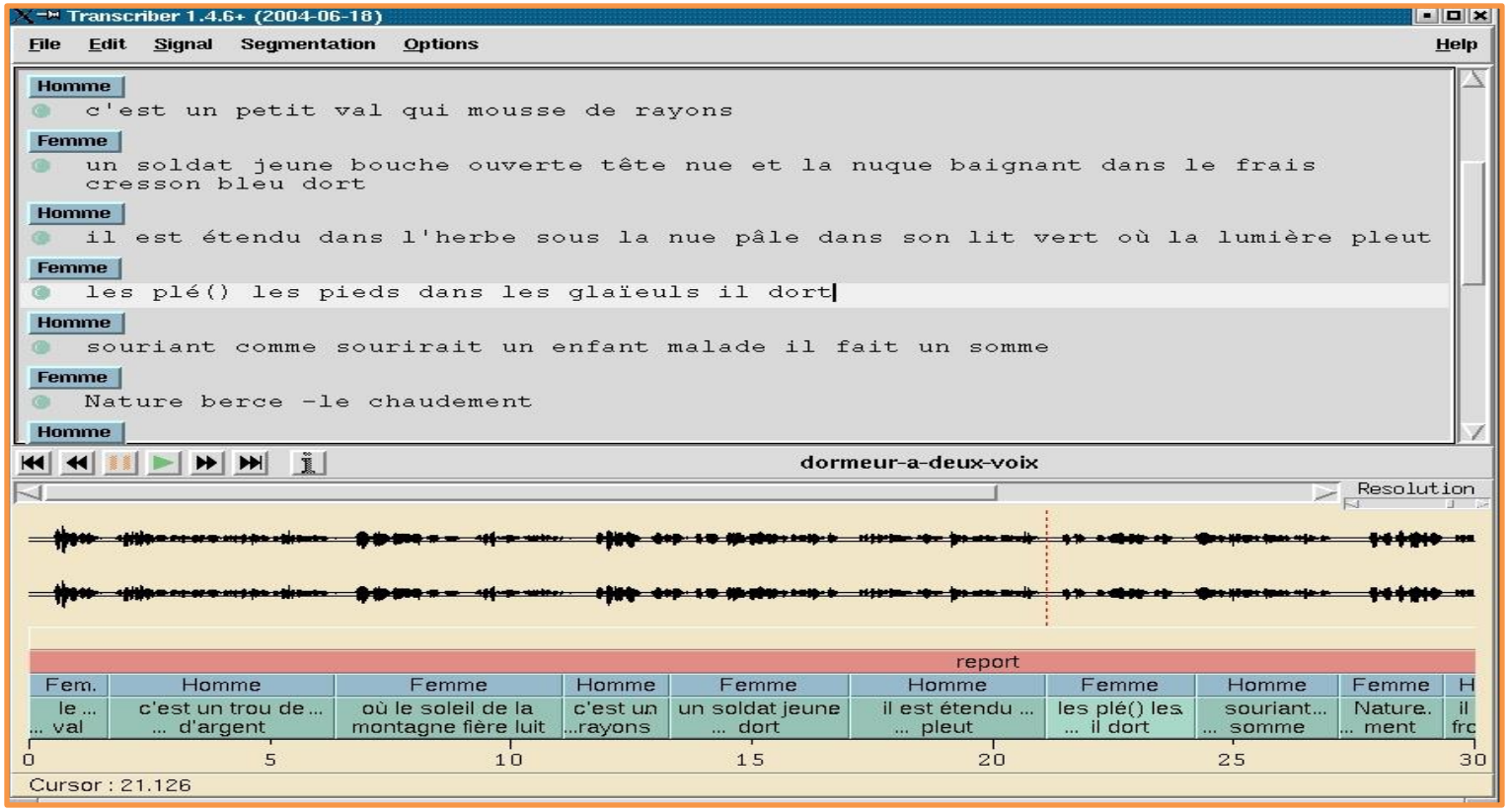
Plan

- Que nos voix demeurent
- **De l'analogique au numérique : fragilisations**
- Archivage numérique de données orales
- Mises en perspective

Enregistrer : analogiquement

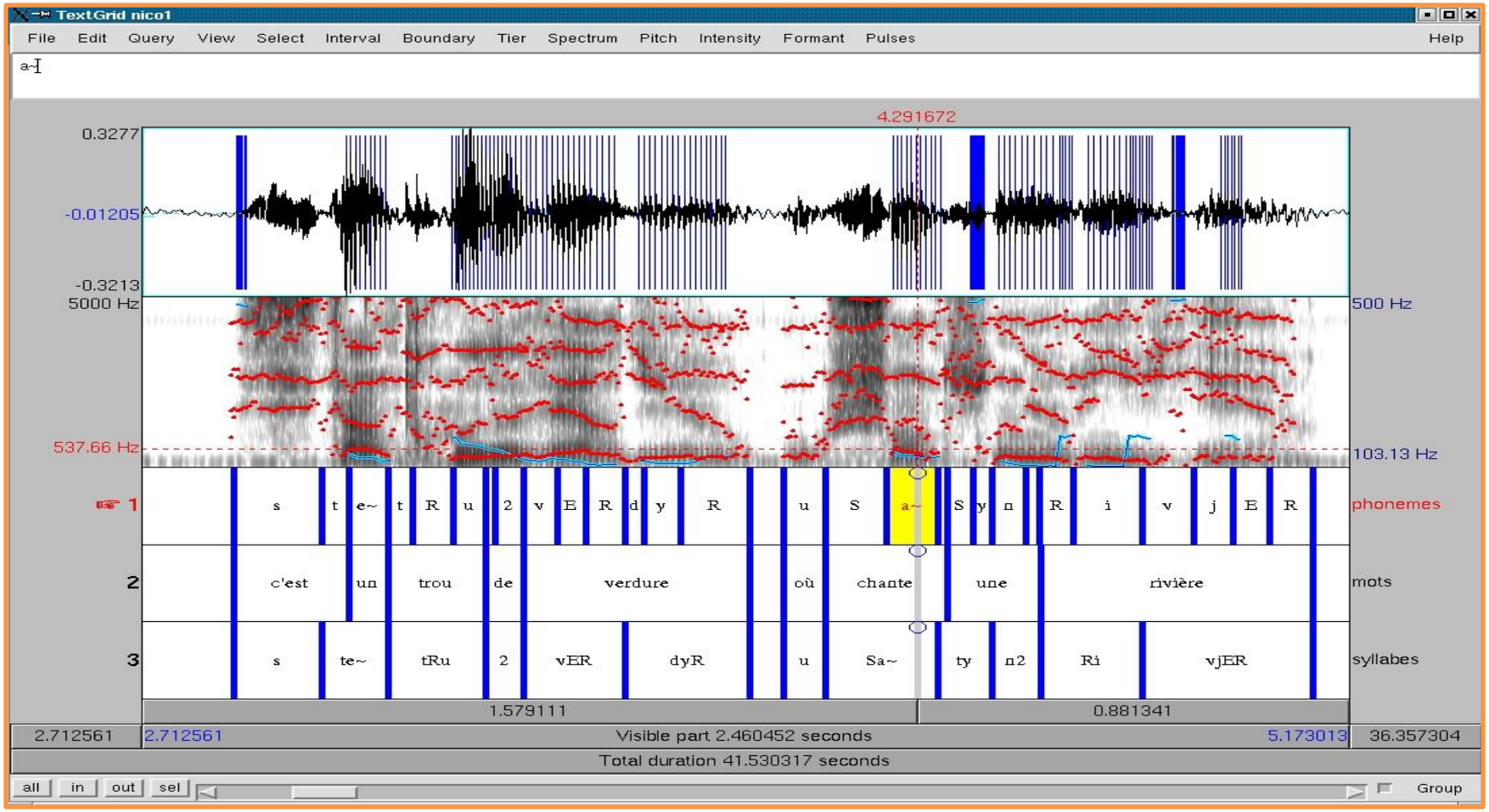
« L'invention du magnétophone portatif devrait être considérée comme une date déterminante pour le développement de la linguistique. Pour la première fois, l'invention technique donnait à chacun la possibilité d'étudier des échantillons de sa propre langue parlée, en les conservant aussi stables qu'on avait pu le faire pour des échantillons de langue écrite. » (M. Halliday *Spoken and Written Language*, Oxford University Press, 1985) - cité par C. Blanche-Benveniste *Approches de la langue parlée en français*, Ophrys, 1997

Enregistrer/annoter : numériquement



Avant : bande magnétique pour son et papier pour transcription
Maintenant : plus de séparation entre type d'information et médium

Enregistrer/annoter : numériquement



Fragilités insignes du numérique

- Les données, primaires et secondaires, ne sont plus indépendantes des truchements techniques au sens large (« machines », logiciels, protocoles) et des savoirs et savoir faire correspondants
- Altérations/disparitions possibles
  - Support (dégradation souvent aléatoire)
  - Dégradation “catastrophique” et non graduelle
  - Format obsolète
  - Absence de dispositif d’accès aux données
  - Contexte : donnée devenue orpheline, flottante, inane

Défaillances matérielles

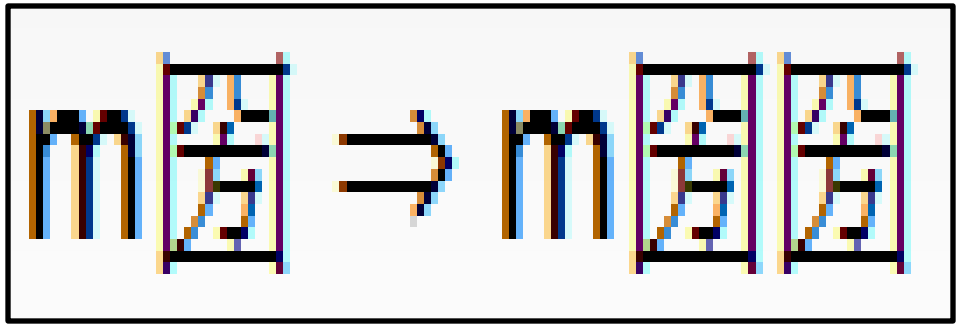
Une étude réalisée par Google sur son propre parc informatique a ainsi révélé que 8% des disques vieux de deux à trois ans devaient être remplacés en raison de leur défaillance.

(Hoog, 2009)

Dégradations minimes

sexe ∅ 1 ou 2

millesime_naissance ∅ 00... 99



Humaines machines

Ce qui réside dans les machines, c'est de la réalité humaine, du geste humain fixé et cristallisé en structures qui fonctionnent.

Gilbert Simondon *Du mode d'existence des objets techniques*, Aubier, 1958

... ou qui cessent de fonctionner...

Disjecta membra

... je passai une journée entière à glaner, comme si de ces disjecta membra de la bibliothèque devait me parvenir un message... Souvent, à partir d'un mot ou d'une image survivante, je reconnus de quel ouvrage il s'agissait... À la fin de ma patiente reconstitution se profila dans mon esprit comme une bibliothèque mineure, signe de la majeure disparue, une bibliothèque composée de morceaux, citations, périodes incomplètes, moignons de livres.

Umberto Eco, *Le nom de la rose*, fin.

Un numérique à courte vue

- L'essentiel du numérique actuel (natif ou non) a une espérance de vie limitée : 5-10 ans
- Les sciences humaines et sociales courent le risque de (re)produire des données et des connaissances sous forme numérique (native ou non) qui se révèleront inutilisables à long terme

Plan

- Que nos voix demeurent
- De l'analogique au numérique : fragilisations
- **Archivage numérique de données orales**
  - Origine et modèle normatif
  - Projet pilote (2008-2010)
  - Strates du numérique
  - Vocabulaire
- Mises en perspective

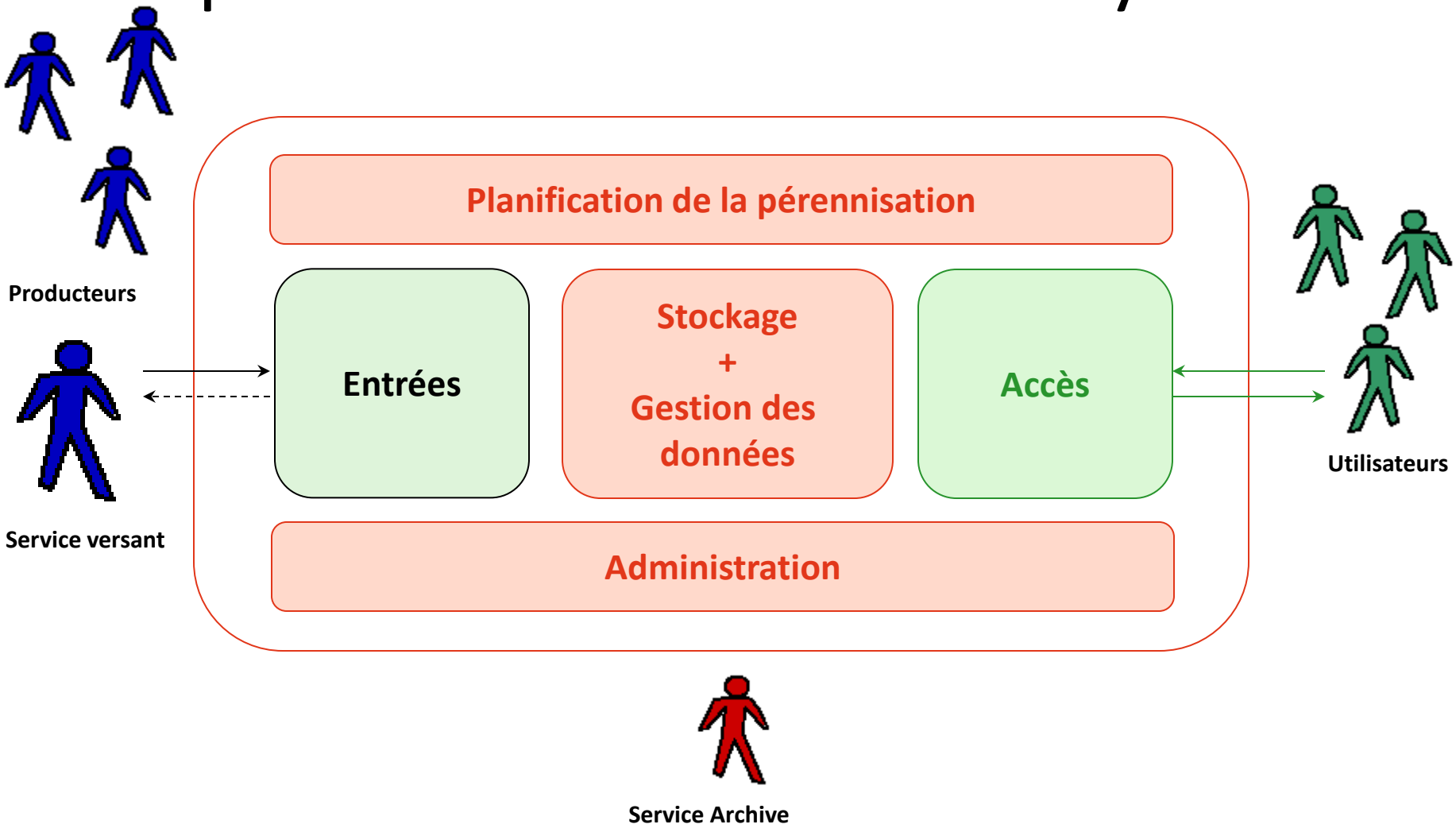
Archivage numérique : origines

- Données spatiales (NASA, CNES)
  - Nativement numériques dès la fin des années 60
  - Coûteuses et non reconstituables (comète, éruption solaire, carte des forêts)
- Mise en place d'un standard à la fin des années 90 puis d'une norme ISO 14 721 OAIS – Open Archival Information System – en 2003
  - Données spatiales : rien de spécifique
  - Norme nécessairement abstraite vu la rapidité des changements dans le numérique

Niveaux de normalisation

- Standard : homologation explicite par une communauté d'utilisateurs
- Norme : définie par une instance de normalisation, nationale, comme l'AFNOR (Association française de normalisation) ou internationale, comme l'ISO (International Organization for Standardization)

Le modèle OAIS
Open Archival Information System



Un projet pilote d'archivage cadre

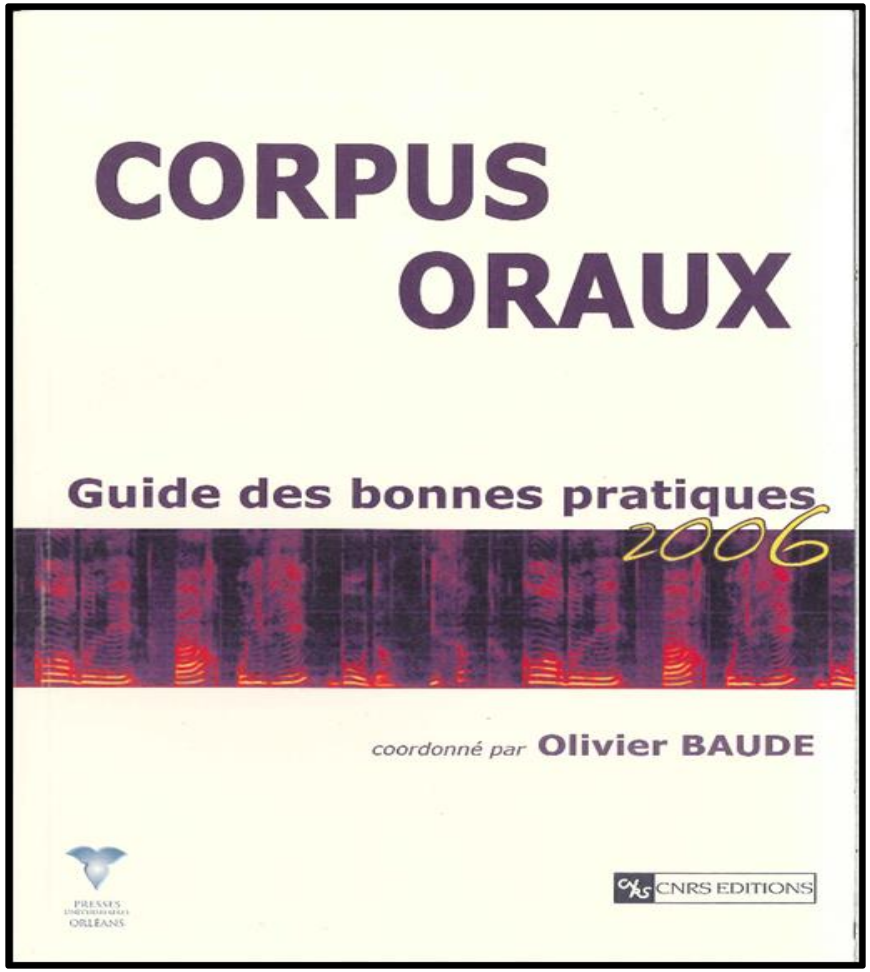
- Contexte : les TGIR (Très grandes infrastructures de recherche)
  - Origine : physique
  - Equipements dépassant les limites en durée et en montant des cadres habituels
  - Parangon : LHC (Large Hadron Collider) : 10 000 scientifiques de 100 pays, 3 milliards d'euros, 15 Po de données attendues par an
- Extension à d'autres secteurs
  - 2007 TGE Adonis
  - 2008 Feuille de route ESFRI : 3 autres TGIR (bibliothèques ; données socio-économiques ; corpus)

Un projet pilote d'archivage démarche

- Diversité des situations selon les secteurs des SHS
  - Pas de solution unique immédiate
  - Choix d'un projet pilote
- Cadrage
  - Etude externe (CERN)
  - Basé sur 2 « gros » centres de calcul
  - Evaluation externe (Université de Montréal) + validation par la Direction des Archives de France

Un projet pilote d'archivage choix des données orales

- Diversité et maturité des communautés scientifiques + grand public (DGLFLF)
- Complexité (audio, vidéo, annotations) et volume (2 To) adéquats
- Structuration (fédérations) et dynamisme des communautés
- Structure métier : CRDO



Entrées et accès

- CINES
  - 40 ingénieurs et techniciens
  - Archivage : mission (ministère) , département (une archiviste et plusieurs ingénieurs) et plate forme (PAC)
    - Thèses (via l'ABES)
    - Revues numérisées (Persée – Lyon 2)
    - HAL
- CC-IN2P3
  - 70 ingénieurs et techniciens
  - Très bonne connectivité RENATER
  - Très grands volumes en ligne (5 Po – 1 Po = 1000 To) et en stockage (30 Po)
  - Expertise calcul et stockage distribués
  - Travail avec les SHS (3D en archéologie) depuis 2005

Entrées et accès

- CINES

- CC-IN2P3

Truchement : le CRDO

- Absence de compétences CINES/CC-IN2P3 sur les données orales
- Limite des compétences techniques et des moyens des producteurs de données orales
- CRDO
  - Mutualisation de ces moyens
  - Connaissance des producteurs et appui transverse

CRDO Aix

Les dépôts les plus récents >> plus — page 1 >>

[testARK] Outil *ProsodyPro* (Yi XU)
UCL division of psychology and language sciences (UCL, London UK)
A Praat script for large-scale systematic prosody analysis [Détails]
(*computational_linguistics*)

Voir

ProsodyPro by Yi Xu is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

2010-09-21
Version 1

[testARK] Outil *PENTATrainer* (Yi XU)
UCL division of psychology and language sciences (UCL, London UK)
A Praat script for extracting pitch targets from vocal signals [Détails]
(*computational_linguistics*)

Voir

PENTATrainer by Yi Xu is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

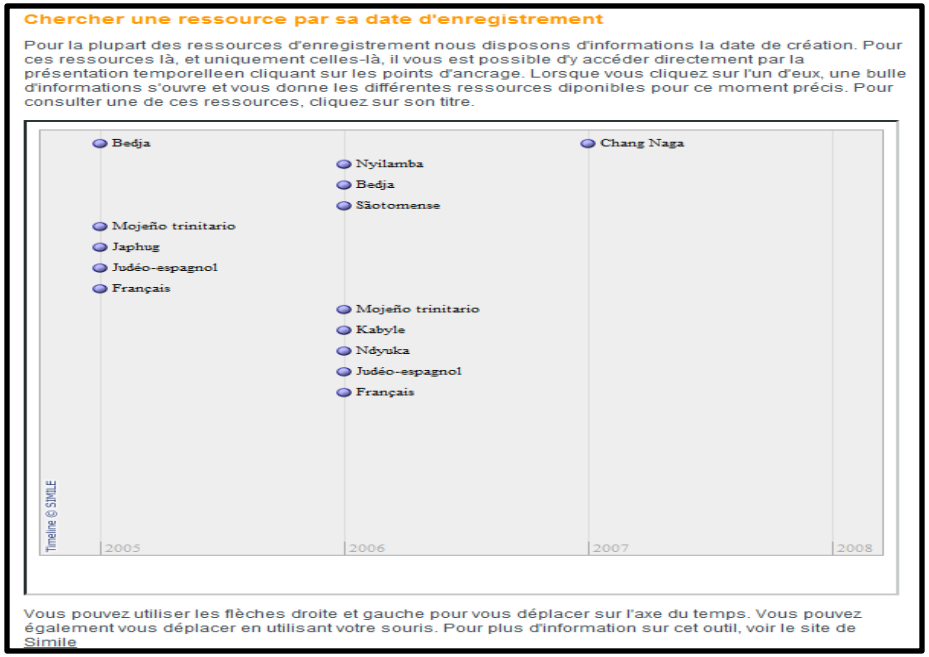
2010-09-18
Version 1

[testARK] Données primaires (corpus) Corpus Double Marquage de Genre (masculin/féminin) - Entretiens (Julie ABBOU)
Laboratoire parole et langage (LPL, Aix-en-Provence FR) -> source
Département de lettres modernes, Université de Provence (Aix-en-Provence FR)
Corpus audio de 4 entretiens semi-dirigés (environ 5h30 au total).
Les entretiens portent sur la féminisation des textes (double-marquage) en contexte politique libertaire, avec des locuteurs issus de ces cultures politiques.
Productions métadiscursives sur les stratégies de féminisation des locuteurs.
[Détails]
(*applied_linguistics, sociolinguistics*)
français

2010-09-12
Version 1

CRDO Paris


Chercher une ressource par la géographie
Nous disposons pour un certain nombre de ressources, d'informations sur le lieu précis où a été effectuée l'enquête. Pour ces ressources là, et uniquement celles-là, il vous est possible d'y accéder directement par la carte géographique en cliquant sur les points rouges. Lorsque vous cliquez sur l'un d'eux, une bulle d'informations s'ouvre et vous donne les différentes ressources diponibles pour ce point précis. Pour consulter une de ces ressources, cliquez sur son titre.
[patientez jusqu'à ce que les points rouges aparaissent sur la carte]
Français
Submit
Plan
Satellite
Mixte
5000 mi
5000 km
Imagerie ©2010 NASA – Conditions d'utilisation


Chercher une ressource par sa date d'enregistrement
Pour la plupart des ressources d'enregistrement nous disposons d'informations la date de création. Pour ces ressources là, et uniquement celles-là, il vous est possible d'y accéder directement par la présentation temporelleen cliquant sur les points d'ancrage. Lorsque vous cliquez sur l'un d'eux, une bulle d'informations s'ouvre et vous donne les différentes ressources diponibles pour ce moment précis. Pour consulter une de ces ressources, cliquez sur son titre.
Bedja
Mojeño trinitario
Japhug
Judéo-espagnol
Français
Nyilamba
Bedja
Sãotomense
Mojeño trinitario
Kabyle
Ndyuka
Judéo-espagnol
Français
Chang Naga
Timeline © SIMILE
2005
2006
2007
2008
Vous pouvez utiliser les flèches droite et gauche pour vous déplacer sur l'axe du temps. Vous pouvez également vous déplacer en utilisant votre souris. Pour plus d'information sur cet outil, voir le site de Simile


Accueil > Chercher une ressource
Rechercher:
*omelette
Catégories
Operateurs
Langues
Résultats affichés
Toutes
near
Toutes
15
Valider
Vous pouvez utiliser dans vos recherches les jokers *, ? ainsi que les rangées de caractères du type [a-zA-Z]
Ressources 1 à 15 (total: 96 trouvées)
Suivants >>
fichier
titre
déposant
langue
fichiers liés
ESLO - Omelette ban...
Université d'Orléan...
Français
consulter
ESLO - Omelette ban...
Université d'Orléan...
Français
consulter

Les strates du numérique

- 00110001001101010011100000110000001110000011011100110101001101100011010100110110001100010011100100110010001100011010100
- 00110001 00110101 00111000 00110000 00111000 00110111 00110101 00110110 00110101 00110110 00110001 00111001 00110010 00110110 00110101
- 49 53 56 48 56 55 53 54 53 54 49 57 50 54 53
- 1 5 8 0 8 7 5 6 5 6 1 9 2 6 5

- flux de 125 bits → découpage en 15 octets (ensemble de 8 bits) → un octet représente un nombre en notation binaire → nombres interprétés comme des numéros d'ordre dans un jeu de caractères → nombre à 15 chiffres

Les strates du numérique

- 158087565619265
- 158 087 565 619 265
- 1 58 08 75 656 192 65

- Le contexte donne sens aux données
- 2 « vues » possibles (au moins)
  - des nombres (manipulables comme tels : addition, moyenne...)
  - des identifiants (relevant d'autres opérations)
- La structure n'est pas présente, c'est une interprétation

Les strates du numérique

- 1 58 08 75 656 192 65
- Structure
  - sexe : 1
  - millésime naissance : 58
  - mois naissance : 08
  - lieu de naissance (ici : département : 75, subdivision : 656)
  - numéro d'ordre : 192
  - clé de contrôle : 65

Les strates du numérique

- http://xml.insee.fr/schema/nir.html
- L'INSEE gère le répertoire national d'identification des personnes physiques depuis 1946 [RNIPP].
- Le numéro d'inscription au répertoire (NIR) est l'identifiant unique des individus inscrits au RNIPP.

Les strates du numériques

Le NIR est un numéro à treize caractères dont la composition est précisée dans l'article 4 du décret n° 82-103 du 22 janvier 1982 : "Le numéro attribué à chaque personne inscrite au répertoire comporte 13 chiffres. Ce numéro indique successivement et exclusivement le sexe (1 chiffre), l'année de naissance (2 chiffres), le mois de naissance (2 chiffres), et le lieu de naissance (5 chiffres ou caractères) de la personne concernée. Les trois chiffres suivants permettent de distinguer les personnes nées au même lieu, à la même période."

Les strates du numérique

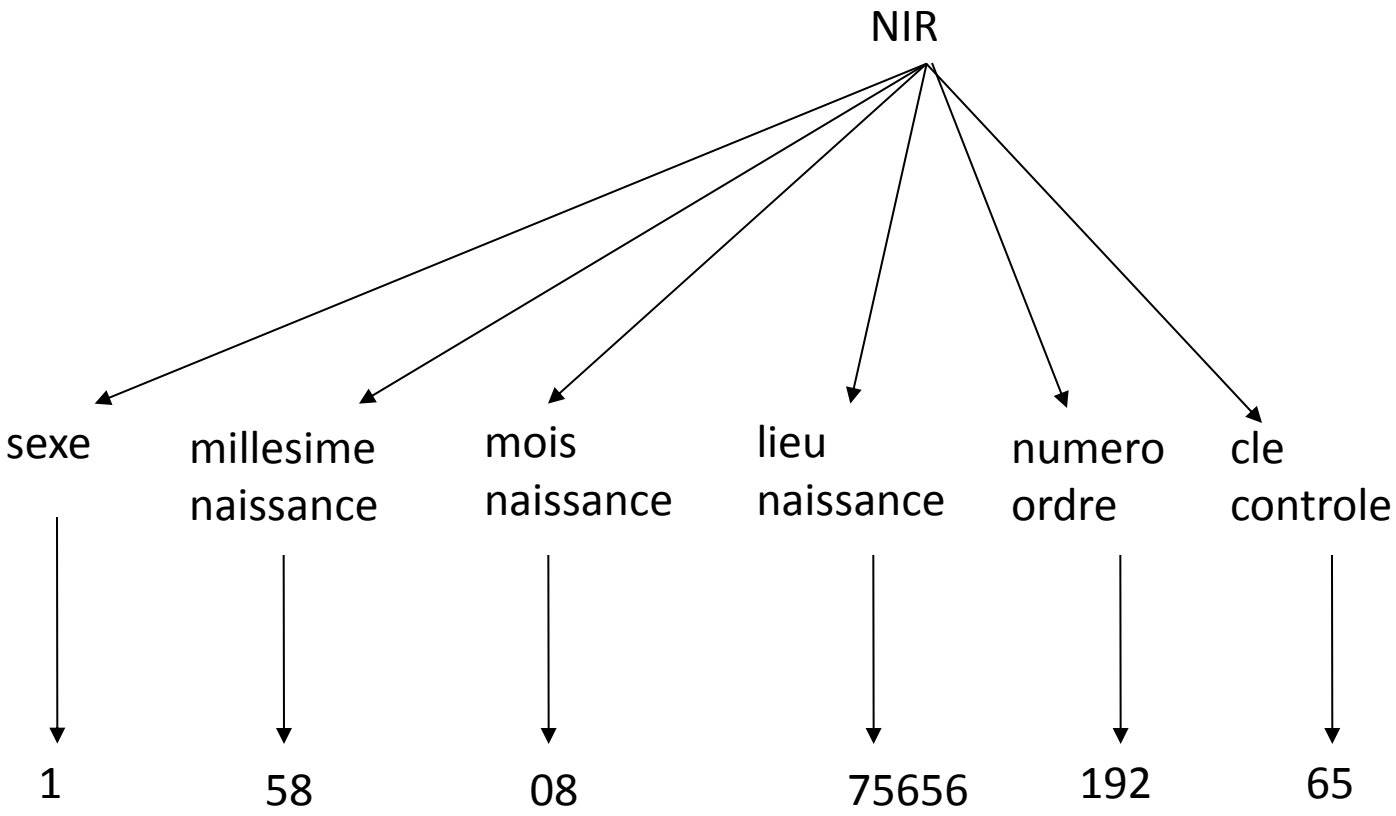
Le numéro NIRAvecCle est un identifiant numérique de 15 chiffres composé du NIR (13 chiffres) et de sa clé (2 chiffres). Il permet de réduire les erreurs de saisie ou de transmission des identifiants NIR et est couramment utilisé dans le domaine médical, notamment dans les déclarations de sécurité sociale.

Les strates du numérique

- Clé de contrôle : moyen de vérifier que tous les chiffres précédents ont bien été transmis
  - Calcul : 97 – (nombre formé par les 13 premiers chiffres modulo 97)
    NB Modulo = reste de la division entière

  - 158087565619265 → 1580875656192 + 65 → 97 – (1580875656192 modulo 97) = 97 – 23 = 65 → *OK*
  - 1580875**4**619265 → 1580875**4**6192 + 65 → 97 – (1580875**4**6192 modulo 97) = 97 – 14 = 83 → *Pas OK*

Structure implicite



Structure explicite

<NIR>
    <sexe>**1**</sexe>
    <millesime_naissance>**58**</millesime_naissance>
    <mois_naissance>**08**</mois_naissance>
    <lieu_naissance>**75656**</lieu_naissance>
    <numero_ordre>**192**</numero_ordre>
    <code_controle>**65**</code_controle>
</NIR>

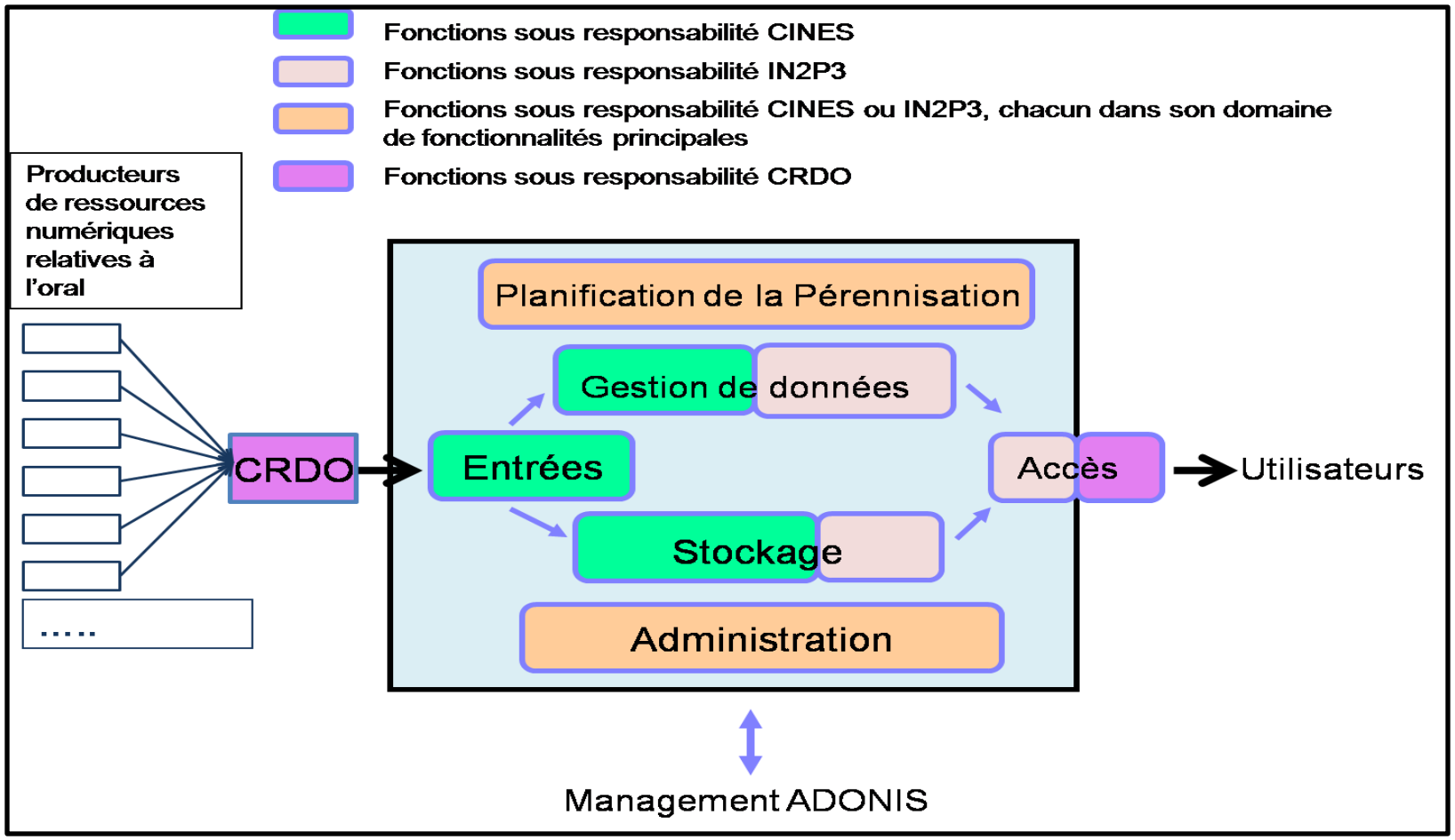
« Envelopper »

- Couches informationnelles
  - Donnée numérique de départ
  - « Mode d'emploi » (jeu de caractères, structure signifiante)
  - Contexte (provenance, droits, empreinte, histoire)
- Ex :
  - 00110001001101010011100000110000001110000011011 10011010100110110001101010011011000110001001110 01001100100011011000110101
  - ISO-Latin1 + NIR
  - N° SS de X, utilisable par…

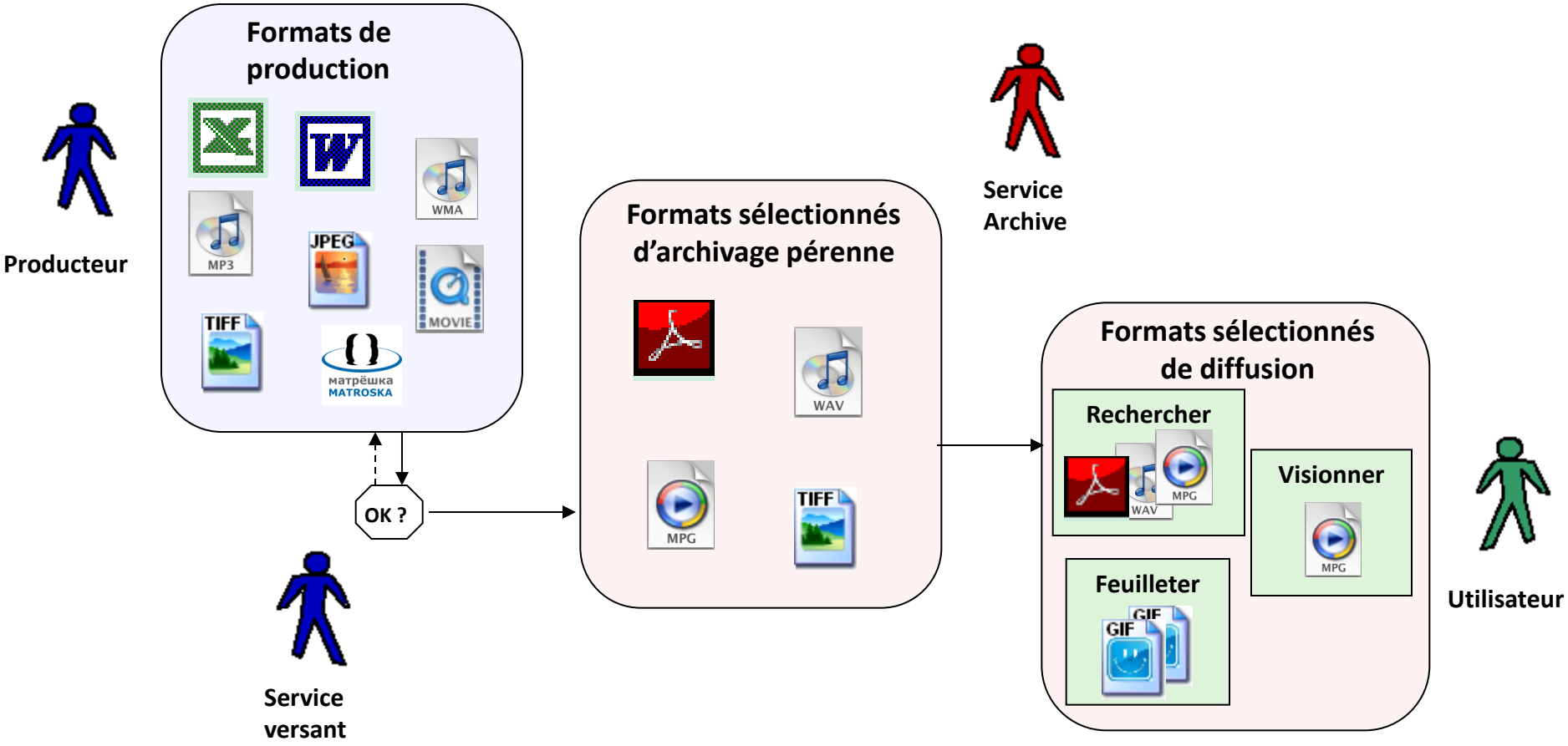
S’accorder pour geler/dégeler

- 2 contraintes
  - Avoir des « formes » compatibles (manière de fournir l’information)
  - Correspondre à ce qui est attendu (type d’information)

OAIS pour le projet pilote



Du producteur au consommateur



Les formats acceptés au CINES

Ce sont des formats identifiés et vérifiables
- publiés
- largement utilisés
- si possible normalisés

Type	Format
Texte	HTML, PDF, TXT, XML
Image	GIF, JPEG, TIFF, PNG, SVG
Audio	WAV, AIFF, AAC, VORBIS
Vidéo	MJPEG2000, MPEG4, THEORA

- Le système PAC est interfacé avec différents outils pour contrôler le format des fichiers transférés

Janus archivistique

S20	JK: décrivez voir un peu
S21	FD 237: comment
S22	JK: comment vous faites
S23	décrivez un peu ce que vous faites
S24	FD 237: pour faire l'omelette
S25	JK: pour faire l'omelette
S26	FD 237: ah vous tombez bien c'est moi qui les fait toutes ici
S27	JK: ah bon alors j'écoute
S28	FD 237: je casse les oeufs
S29	[mic:noise:instantaneous]
S30	JK: oui
S31	FD 237: dans un bol
S32	ou dans un saladier
S33	sel
S34	poivre
S35	des petits morceaux de beurre
S36	que je mets
S37	avec les
S38	que je bats avec les oeufs
S39	je mets des tranches de baccon dedans
S40	je tourne tout et dans la poêle très chaude
S41	voilà l'omelette pour moi
S42	[rire en fond:noise:instantaneous]
S43	JK: eh bien
S44	FD 237: c'est moi qui les fait toutes
S45	femme: ah oui

```
- <S who="JK" id="elso006s20">
    <AUDIO start="29.325" end="30.876"/>
    <FORM kindOf="ortho">décrivez voir un peu</FORM>
  </S>
- <S who="FD 237" id="elso006s21">
    <AUDIO start="31.379" end="32.301"/>
    <FORM kindOf="ortho">comment</FORM>
  </S>
- <S who="JK" id="elso006s22">
    <AUDIO start="31.379" end="32.301"/>
    <FORM kindOf="ortho">comment vous faites</FORM>
  </S>
- <S who="JK" id="elso006s23">
    <AUDIO start="32.78" end="34.511"/>
    <FORM kindOf="ortho">décrivez un peu ce que vous faites</FORM>
  </S>
- <S who="FD 237" id="elso006s24">
    <AUDIO start="34.511" end="35.583"/>
    <FORM kindOf="ortho">pour faire l'omelette</FORM>
  </S>
- <S who="JK" id="elso006s25">
    <AUDIO start="35.583" end="37.852"/>
    <FORM kindOf="ortho">pour faire l'omelette</FORM>
  </S>
- <S who="FD 237" id="elso006s26">
    <AUDIO start="35.583" end="37.852"/>
  - <FORM kindOf="ortho">
      ah vous tombez bien c'est moi qui les fait toutes ici
    </FORM>
  </S>
- <S who="JK" id="elso006s27">
    <AUDIO start="37.852" end="39.732"/>
    <FORM kindOf="ortho">ah bon alors j'écoute</FORM>
  </S>
```

Janus archivistique

Corpus Double Marquage de Genre (masculin/féminin) - Entretiens

Laboratoire parole et langage (LPL, Aix-en-Provence FR) -> source
Département de lettres modernes, Université de Provence (Aix-en-Provence FR)
http://crdo.fr/crdo000714/fr
oai:crdo.fr:crdo000714 (voir)
ark:/87895/2.7-17282 (?)
http://crdo.fr/wiki/crdo000714

[retour]

[Télécharger]

Type d'objet	Données primaires (corpus)
Identifiant	crdo000714
Vitrine (public)	
Langue du corpus	français
Type discursif selon OLAC	*dialogue*
Type linguistique de données selon OLAC	*primary_text*
Domaine(s) linguistique(s)	*applied_linguistics* *sociolinguistics*
Responsable	Contact Julie ABBOU
Lien vers la page wiki	wiki/crdo000714
Description	Corpus audio de 4 entretiens semi-dirigés (environ 5h30 au total). Les entretiens portent sur la féminisation des textes (double-marquage) en contexte politique libertaire, avec des locuteurs issus de ces cultures politiques. Productions métadiscursives sur les stratégies de féminisation des locuteurs.


```
<DocDC>
<title>fr: Corpus Double Marquage de Genre (masculin/féminin) - Entretiens</title>
<creator>Julie ABBOU</creator>
<subject>fr: métadiscours; genre; féminisation; libertaire</subject>
<subject>fra</subject>
<subject>applied_linguistics</subject>
<subject>sociolinguistics</subject>
<description>fr: Corpus audio de 4 entretiens semi-dirigés (environ 5h30 au total).<br/>Les entretiens
portent sur la féminisation des textes (double-marquage) en contexte politique libertaire, avec des locuteurs
issus de ces cultures politiques.<br/>Productions métadiscursives sur les stratégies de féminisation des
locuteurs.<br/></description>
<publisher>Laboratoire parole et langage (LPL, Aix-en-Provence FR)</publisher>
<publisher>Département de lettres modernes, Université de Provence (Aix-en-Provence FR)</publisher>
<contributor>Julie ABBOU (creator)</contributor>
<contributor>Claire (speaker)</contributor>
<contributor>Éva (speaker)</contributor>
<contributor>Gaël (speaker)</contributor>
<contributor>Thomas (speaker)</contributor>
<contributor>Julie Abbou (recorder)</contributor>
<date>2010-11-01</date>
<type>corpus</type>
<format>MP3</format>
<source>NR</source>
<language>fra</language>
<relation>NR</relation>
<coverage>ISO3166: FR</coverage>
<rights>CRDO licence; rightsHolder = Julie ABBOU</rights>
<rights>Autorisations d'accés en attente des signatures des contributeurs</rights>
<rights>license:URI http://crdo.fr/licenceCRDO.php?version=1</rights>
</DocDC>
<DocMeta>
<authenticite>oui</authenticite>
<dureeConservation>P10000Y</dureeConservation>
<identifiantDocProducteur>crdo000714</identifiantDocProducteur>
<docRelation>
<typeRelation>maj</typeRelation>
<sourceRelation>PAC</sourceRelation>
<identifiantSourceRelation>17282</identifiantSourceRelation>
</docRelation>
<noteDocument>Version 1 : fichiers sources MP3 et leur décompactage au format WAV </noteDocument>
<serviceVersant>CRDO-Aix</serviceVersant>
<structureDocument>index.xhtml</structureDocument>
</DocMeta>
```

Un vocabulaire « flottant »

Je me souviens que Georges Perec voulait faire adopter l'appellation « un matiouze » pour l'expresso très serré que je commandais invariablement, usant de la formule de Maxine G. : « Un double express dans une toute petite tasse ». Au café ou au restaurant, il commandait « un matiouze », puis expliquait le mot, dans l'espoir - impossible - qu'il s'imposerait.

Harry Matthews *Le verger* POL 1986

Un vocabulaire « flottant »

- Stocker
- Sauvegarder
  - Copie « telle quelle », éventuellement multiple
  - Pas de métadonnées obligatoirement associées : contexte limité (nom de fichier, type)
- Répliquer
  - Copie telle quelle « ailleurs » (de préférence site distant)

Un vocabulaire flottant

- Migrer
  - Changer l'archive de format
    - Nouvelle version d'un format d'archivage
    - Nouveau format d'archivage
  - Paradoxe
    - Nécessaire pour que l'archive ne meure pas
    - Ne garantit pas la préservation du contenu informationnel – cf. « saut » entre Microsoft Office 2003 et 2007

Un vocabulaire « flottant »

- Archiver
  - S'accorder avec qui a versé sur ce qui a été versé (résumé ou empreinte numérique) et pouvoir le lui « rendre » tel quel (réversibilité)
  - « Certifier »
    - Données
    - Métadonnées : contexte nécessaire à la pérennisation
  - Préserver des dégradations ou destructions de supports
    - Réplication sur site distant
    - Vérification de la « stabilité » de l'archive (empreinte numérique)
  - Migrer données et métadonnées en fonction des évolutions des formats
  - Donner accès aux données en fonction des métadonnées

Un vocabulaire « flottant »

- Archive
  - En français, connotation « rat de bibliothèque », documents inertes pendant un long délai
  - En anglais, accent mis sur l'accès immédiat et aisé (cf. arXiv, l'ancêtre de HAL)
- Des archives vives, pas des « natures mortes »
  - Pouvoir amender une annotation (changement de conventions)
  - Liens entre données et (nouvelles) annotations

Plan

- Que nos voix demeurent
- De l'analogique au numérique : fragilisations
- Archivage numérique de données orales

- **Mises en perspective**
  - Bilan sur le projet pilote
  - Désir et mal de mémoire en SHS
  - Apprendre à garder et à oublier

Eléments de bilan

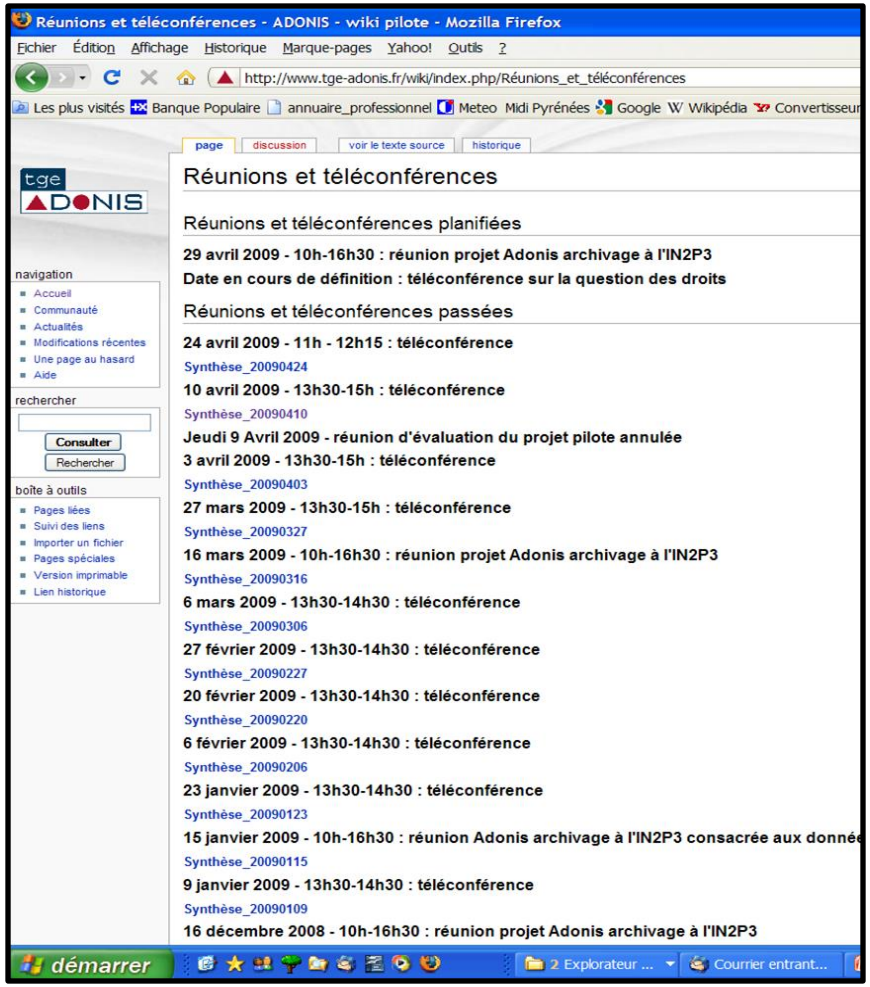
- Fiabilité et stabilité des piliers du dispositif : CINES et CC-IN2P3
- Généricité et extensibilité
  - Prochaine étape : données 3D en archéologie
- Souplesse pour l'accès : local (Aix) / au CC-IN2P3 (Paris)
- Un sain partage des tâches
  - Allège la tâche des chercheurs sur l'oral
  - Profite des compétences CINES et CC-IN2P3 et les accroît
  - Truchement « métier » du CRDO

S’accorder

- Archiver
  - Pas essentiellement un problème technique
  - Suppose surtout des représentations partagées
- Se mettre d’accord sur
  - L’organisation des données et des métadonnées
  - La division précise des responsabilités et des tâches
- Un travail lent, patient, encadré (consultant) : plus d’1 année et demi entre lancement (septembre 2008) et passage en production (juillet 2010)

Le temps, les moyens pour s'accorder

- Moyens de travail
  - Wiki
  - Liste de diffusion
- Réunions
  - 1 téléconférence / 15 jours
  - 1 réunion / 1 mois ½
- Coordination technique assurée par un consultant : Claude Huc (ex. CNES, groupe PIN)



S’accorder

- Cf. travaux d’Alain Desrosières sur la conventionnalisation des notions statistiques et des indicateurs
  - Conventions stabilisées comme « boîtes noires » (cf. Latour) permettant l’action collective
  - Ex. monographies / indice de pauvreté (et Europe) ; diversité ethnique
- Archivage numérique pérenne : début d’une telle conventionnalisation

S’accorder

- Les conventions des transcriptions et des annotations « sédimentent » les débats et accords provisoires des communautés de recherche
- La forme (structuration) de ces conventions peut faciliter leur transmission et leur pérennisation
  - Place des standards et normes
  - Partage de cette forme dans les communautés

Présence du passé en sciences humaines et sociales

- Claude Lévi-Strauss *Tristes tropiques* – 1955
- Etude interdisciplinaire de Plozévet (entre Quimper et la pointe du Raz) autour d'Edgar Morin au milieu des années 60 : souhait – de la part d'historiens, de sociologues, d'ethnologues – de disposer aujourd'hui des données de terrain, des journaux de recherche pour une étude des évolutions « au long cours »

Besoin et mal d'archive en SHS

- Départs en retraite massifs
- Un numérique envahissant
  - Campagnes de numérisation
  - Données nativement numériques (projets ANR)
- Fragilités
  - Equipes « petites »
  - Culture technique et moyens humains limités
  - Recherche de plus en plus par projets limités
  - Mutualisations variables selon les disciplines

Construire des communautés d'utilisateurs

- Les équipements mutualisés dans les sciences dures viennent servir des communautés structurées
- Humanités numériques en France : encore dans les limbes (cf. séparation des universités)
- Lyon : probablement un lieu privilégié pour des projets en humanités numériques (ENS, ENSSIB, Persée, CC-IN2P3…)

Chercher : autrement

- Des liens nouveaux entre données primaires et données secondaires
- Donner une place à la production de données de recherche « durables » ?
- Standardisation, normalisation, industrialisation et division des tâches
  - D'où des infléchissements des formations

Pérenniser la pérennisation

- Incertitude sur l'engagement sur la mutualisation en sciences humaines et sociales (prévision TGIR SHS en 2008 : 1.5% du budget des TGIR français)
- Nouvel équilibre Etat / régions / universités
- Modèle économique à rendre plausible
  - Les économies d'échelle ne sont pas immédiates
  - Elles dépendent surtout de coûts cachés

EDF R&D

- Contexte : déménagement campus Saclay (2014)
- Autre dimension : données administratives, valeur probante (cf. Direction Commerce)
- Circuit de production/utilisation des grands types de documents (*record management*)
- Objets complexes : mail, simulations 3D, codes de calcul et jeux d'essai, wikis et blogs

Désir de mémoire pour peuples « recrus d'histoire »

Aujourd'hui, la mémoire et son culte font office d'« agents de liaison » entre un passé fantasmé, un présent inquiétant et un futur indéfinissable.

E. Hoog, *Mémoire année zéro*, Seuil 2009

Musée Andy Warhol, Pittsburgh 608 « time capsules », cartons dans lesquels AW vidait périodiquement son bureau, qu'il scellait et datait

Régine Robin *La mémoire saturée* Stock 2003

Désir de mémoire

- Selon les estimations du cabinet IDC, 420 milliards de photos ont été prises dans le monde en 2007, soit près de 50 millions par heure (Hoog, 2009)
- France: 6 commémorations nationales fixées entre 1880 et 1999, 6 entre 2000 et 2006
- France : vaste succès des *Lieux de mémoire* de Pierre Nora (5 000 pages)

Désir de mémoire

… aujourd'hui, la technique incite à tout garder, quelle que soit la nature de l'objet concerné. […] on ne conserve plus parce que c'est important, mais c'est parce que l'on conserve que c'est important. Ou plutôt qu'on lui permet de le devenir

E. Hoog, *Mémoire année zéro*, Seuil 2009

Désir de mémoire

La recherche du temps perdu passait par le web. [...] On était dans un présent infini.
On n'arrêtait pas de vouloir le « sauvegarder » en une frénésie de photos et de films visibles sur le champ. Des centaines d'images dispersées aux quatre coins des amitiés, dans un nouvel usage social, transférées et archivées dans des dossiers - qu'on ouvrait rarement - sur l'ordinateur. Ce qui comptait, c'était la prise, l'existence captée et doublée, enregistrée à mesure qu'on la vivait, des cerisiers en fleur, une chambre d'hôtel à Strasbourg, un bébé juste né. Lieux, rencontres, scènes, objets, c'était la conservation totale de la vie. Avec le numérique, on épuisait la réalité.

Annie Ernaux *Les années* Gallimard 2008

Rester sensible

Au fond, une vie de recherche en manuscrits est un parcours qui va de boîte [d'archives] en boîte. Un jour, je les regretterai, ces boîtes à " malices ", parce que de plus en plus les documents sont numérisés, microfilmés, interdisant le rapport au toucher, au feuilletage, à l'originalité de l'écriture, à la sensation manuelle et aux odeurs. Une partie de ce qui fait le bonheur de la recherche - émotion, visualisation concrète sur le papier de ces vies ramassées en quelques mots ou feuillets - finira par disparaître.

Arlette Farge *Quel bruit ferons-nous ?* Les prairies ordinaires 2005

Rester sensible

Mais quel bruit ferons-nous de ces corps sans voix où l'écrit, faiblement, est venu apporter quelque lumière ?

Arlette Farge *Le bracelet de parchemin* Bayard 2003

Ce que je lis dans les archives, curieusement, je le vois en couleur et je l'entends ; de plus, je le sens. Ce n'est peut-être pas une bonne façon de faire de l'histoire, mais cela se passe ainsi, c'est violent et coloré ; tout y est décrit, avec beaucoup d'indications sur les corps, les blessures, les odeurs et les sons. Les archives sont d'abord visuelles ; il faut se cramponner pour prendre de la distance et pouvoir réfléchir sur elles car on est happé par des sensations physiques.

Arlette Farge *Quel bruit ferons-nous ?* Les prairies ordinaires 2005

A. Farge *Essai pour une histoire des voix au XVIIIe siècle* Bayard 2009

Formes de mémoire

- Travail du deuil par opposition à la mélancolie (autodépréciation) – Freud *Deuil et mélancolie*
- Compulsion de répétition : le passé fait retour en acte / remémoration et perlaboration : passé repris, travaillé et devenu acceptable par le sujet – Freud *Remémoration, répétition, perlaboration*

- Repris par Ricoeur *La mémoire, l'histoire, l'oubli* 2000

Litanie d'œuvres disparues

H. Lefebvre *Les unités perdues* Virgile 2004

- 10 500 films réalisés en pellicule nitrate avant 1950 aux Etats-Unis se sont autodétruits
- Morcellement de la toile Exécution de Maximilien (Manet)
- Mallarmé avant de mourir demande à sa femme et à sa fille de brûler ses archives
- Il ne reste que 11 des lettres de Lewis Carroll à Alice Liddel (la mère d'Alice a détruit les autres)
- …

Formes de l'oubli

L'oubli est nécessaire à la société comme à l'individu.

L'oubli … est la force vive de la mémoire et le souvenir en est le produit.

L'opérateur principal de la mise en « fiction » de la vie individuelle et collective, c'est l'oubli.

M. Augé *Les formes de l'oubli* Payot et Rivages 1998

Qu'est-ce qui est précieux pour nous, aujourd'hui ?

- Nous entrons dans un monde numérique où chacun(e) trouve difficile de dire :
  - Ce qui est précieux et doit être pérennisé
  - Ce qui ne l'est pas et peut/doit être jeté, délaissé
- La difficulté à déterminer ce qui du passé peut nous aider pour aujourd'hui ou pour demain pousse à
  - Accumuler des masses énormes de données non structurées
  - Nous reposer sur des outils de recherche imparfaits pour nous y retrouver (la Toile comme horizon de la mémoire individuelle et collective)

Faire mémoire

La mémoire est l'organisation collective d'un oubli sélectif.

Rony Brauman in R. Brauman & A. Finkielkraut, *La discorde: Israël--Palestine, les Juifs, la France --- Conversations avec Élizabeth Levy*. Fayard 2006

Merci

Des questions ?

(Cyber)bibliographie

- [ADONIS] Site du TGE Adonis http://www.tge-adonis.fr/
- [Au clair de la lune] http://www.futura-sciences.com/fr/news/t/technologie-1/d/au-clair-de-la-lune-ecoutez-le-plus-vieil-enregistrement-sonore-du-monde_15096/
- Augé, M. (1998) *Les formes de l'oubli*. Paris : Payot.
- Banat-Berger, F. ; Duplouy, L. ; Huc, C. (2009) *L'archivage numérique à long terme – les débuts de la maturité ?* Paris : La Documentation Française.
- [BARRING] Présentation des conclusions d'O. Bärring sur la mutualisation de l'hébergement et de l'archivage http://www.tge-adonis.fr/?Le-point-de-vue-d-Olof-Barring-du
- Baude, O. (ed) (2006) *Corpus oraux – Guide des bonnes pratiques 2006*. Paris : Presses universitaires d'Orléans & CNRS Éditions. Egalement en ligne : http://www.dglflf.culture.gouv.fr/recherche/corpus_parole/Corpus_Oraux_GBP%202006_version_imprimee.pdf
- [CC-IN2P3] Centre de Calcul de l'Institut national de physique nucléaire et de physique des particules http://cc.in2p3.fr/

- [CINES] Centre Informatique National de l'Enseignement Supérieur http://www.cines.fr/ et particulièrement http://www.cines.fr/-D-I-S-T-.html pour l'archivage pérenne
- [CRDO] Centre de ressources pour la description de l'oral, dans son versant parisien : http://crdo.risc.cnrs.fr/exist/crdo et aixois : http://crdo.up.univ-aix.fr/
- Desrosières, A. (2000) *La politique des grands nombres - Histoire de la raison statistique*. Paris : La Découverte.
- Desrosières, A. (2008) *L'argument statistique I - Pour une sociologie historique de la quantification*. Paris : Presses de l'École des Mines.
- [DUBLIN CORE] Jeu de métadonnées faisant l'objet d'un large consensus http://dublincore.org/
- [ESFRI] Coordination européenne des TGIR http://cordis.europa.eu/esfri/
- Freud "Deuil et mélancolie", in *Métapsychologie*, Gallimard, 1968
- Freud « Remémoration, répétition, perlaboration », in *La technique psychanalytique*, PUF, 1970

- Hoog, E. (2009) *Mémoire année zéro*. Paris : Seuil.
- [OAIS] CCSDS, 650.0-B-1, *Reference Model for an Open Archival Information System (OAIS)* ,ISO 14721, janvier 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf
- [PAC] Plate-forme d'Archivage au CINES http://www.cines.fr/-l-application-PAC-.html
- [PIN] Groupe de travail sur la préservation de l'information numérique http://www-pin.aristote.asso.fr/
- Ricoeur, P. (2000) *La mémoire, l'histoire, l'oubli*. Paris : Seuil.
- Robin, R. (2003) *La mémoire saturée*. Paris : Stock.
- [SPAR] Système de préservation et d'archivage réparti de la Bibliothèque nationale de France http://www.bnf.fr/pages/infopro/numerisation/num_spar.htm

- [SYNERGIES] TGIR canadienne en SHS http://www.synergiescanada.org/
- [TGIR-CNRS] Site du CNRS sur les TGIR http://www.cnrs.fr/fr/recherche/ups3019/feuilles-route-infrastructures.htm
- [TGIR-EU] Principales TGIR européennes en SHS : DARIAH http://www.dariah.eu/ ; CESSDA http://www.cessda.org/ ; CLARIN http://www.clarin.eu/
- [TGIR-MESR] Site du Ministère de l'Enseignement Supérieur et de la Recherche (MESR) sur les TGIR http://www.roadmaptgi.fr/
- [WIKI-ARCHIV] Wiki du projet pilote du TGE Adonis sur l'archivage pérenne de données orales http://www.tge-adonis.fr/wiki/index.php/Accueil_Projet_pilote