Michel Jacobson

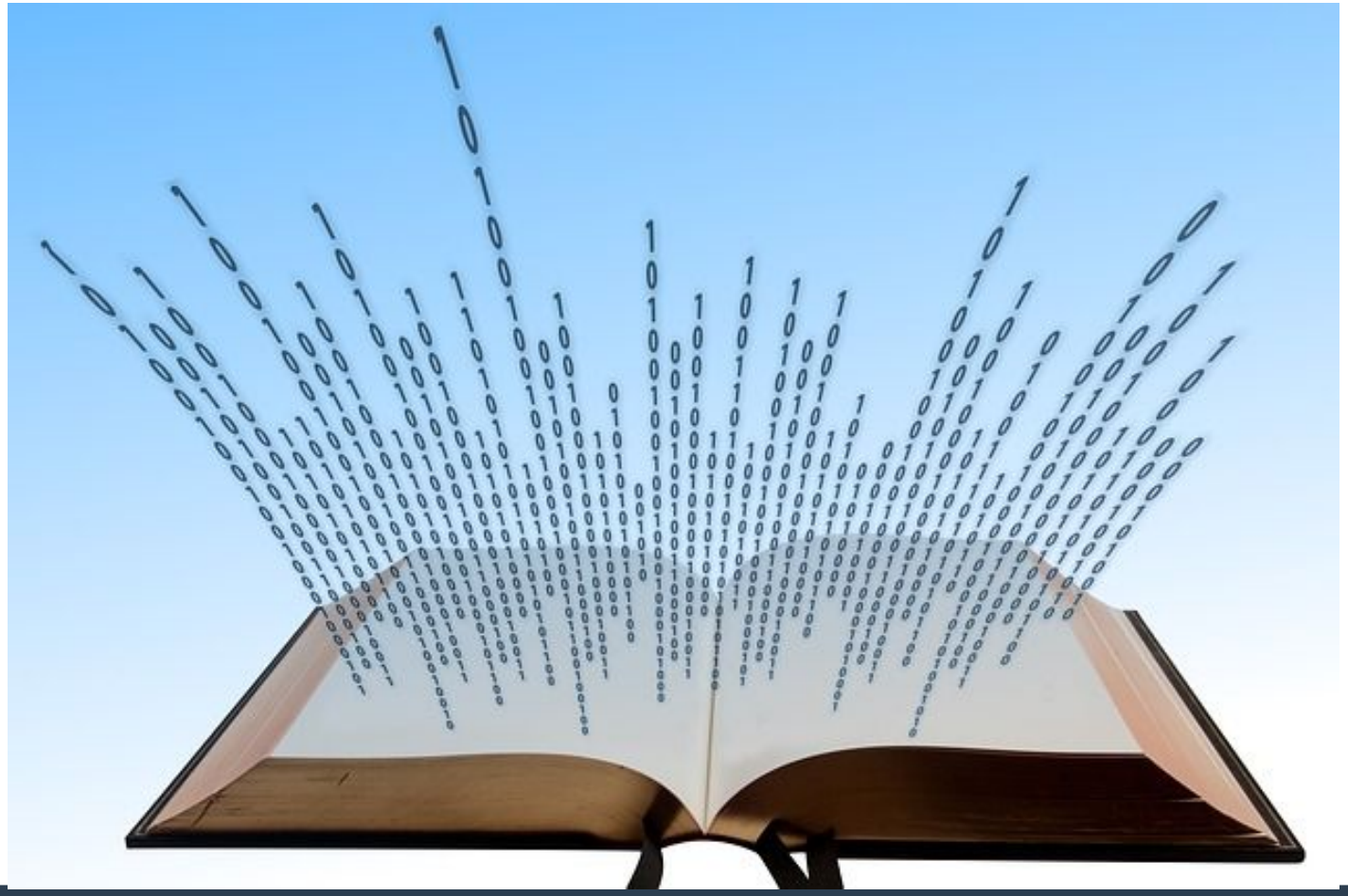
La pérennisation des données

-

retour d’expérience sur les données de l'oral

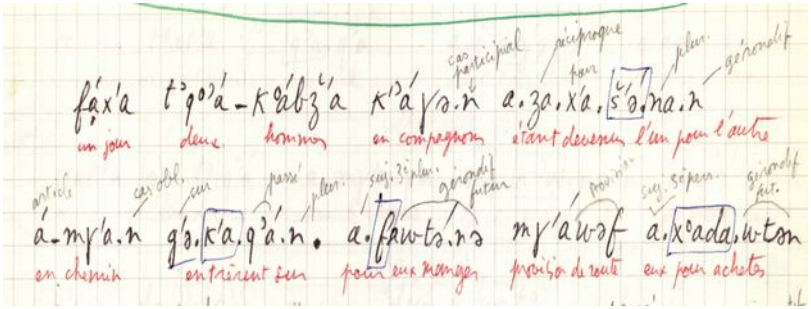
Historique (saison 1)

L'âge de la numérisation



Historique (saison 1)

Historique (saison 1)



1. cíhêdéé ka cíhê bwaaoléé pwö- a péi kucukucu
   légende/et/parler/aigle/dessus-/*n*.le/rocher/Kucukucu/

2. è bwö mu ka è bwö cìnì wíi ìhîm
   il/alors/demeurer/et/il/alors/griller/manger/bancoulier/

\ref T120-C12 003

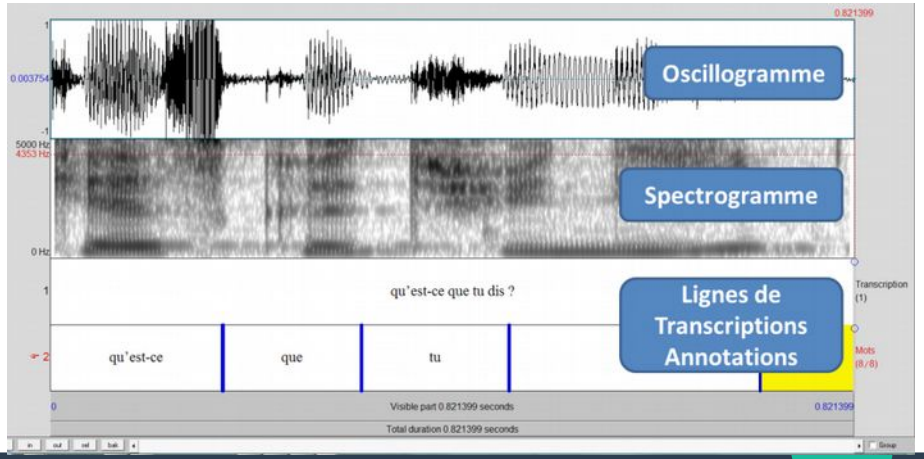
\tx	**mè**	**né**	**hǵgǵ,**	**tɔrìɔŋ**	**ʔą**	
\rm	mɔ̀	né	hǵgǵ	tɔrìɔŋ	ʔą	
\ma	chose	être	comme ça	petit grillon	3S	
\am	N	PRED	ADV	N	PR	

\tx	**gɔ̀nà**			**zǵ**	**kǵą,**		
\rm	BHa-	gɔn	-H	zǵ	kɔ́	ʔą,	
\ma	Acc-	découper	-D	herbes	de	3S	
\am	MV-	V	-FCT	N	FCT	PR	

\tx	**gɔ̀nà**			**zǵ**	**gásá**	**zǵ**	
\rm	BHa-	gɔn	-H	zǵ	gásá	zǵ	
\ma	Acc-	découper	-D	herbes	grand	herbes	
\am	MV-	V	-FCT	N	AV	N	

\tx	**hé**	**mè**	**né**	**gbàmbɔ̀ndɔ́**	**mèí**		**gá.**
\rm	hé	mɔ̀	né	gbàmbɔ̀ndɔ́	mè	-lí	gá.
\ma	comme	chose	être	herbes sp	là-bas	-anaph	comme
\am	SUB	N	PRED	NPR	ADV	-MOD	SUB

\tr C'est ainsi que le Grillon il a délimité son territoire de chasse, il a délimité un territoire, un grand territoire, comme qui dirait Gbambondo là-bas.



Historique (saison 1)

Historique (saison 1)

Actions

- Faciliter la conservation
  - Numérisation des supports analogiques
    - CD-Audio, CD-mixte, Disque-durs
- Faciliter l'accès
  - Mise en place d'un serveur de diffusion
- Faciliter le partage
  - Référencement
    - OAI-PMH
- Faciliter l'édition et la consultation des donnée
  - Définition d'un modèle de données pour les annotations
  - SoundIndex, Karaoke

Historique (saison 2)

L'âge de la mutualisation

Historique (saison 2)

Épisode 1

- Ouverture du service existant dans le laboratoire aux autres laboratoires de la fédération TUL
- Recherche d'une solution d'hébergement
  - Serveur local
  - Serveur hébergé dans le service réseau/système du campus
  - Sollicitations de la délégation et du DSI : échecs
  - Hébergement par une l'unité de services RISC

Historique (saison 2)

Épisode 2 : les centres de ressources numériques

- Suite à un appel conjoint SHS et DIS (2006), mise en place du centre de ressource pour la description de l'oral (CRDO)
  - 2 antennes : Paris (entrepôt de ressources) + Aix-en-Provence (outils)
- Périmètre
  - Ressources numériques
  - Communauté SHS en France
  - Données primaires : enregistrements de parole (audio, vidéo, mesures d'activités physiologiques)

Historique (saison 2)

- Acquisition d'un serveur et d'une baie de stockage
- 1er don des archives du LACITO à la BnF
- Sortie de l'ouvrage collectif « Corpus Oraux : guide des bonnes pratiques » issue d'un groupe de travail constitué de linguistes, juristes, informaticiens et conservateurs.
- 1ère école d'été « Linguistique de Corpus: Constitution, Archivage, Évaluation » juin 2004

Historique (saison 3)

L'âge de l'éveil de la conscience

Historique (saison 3)

- Statut des données
  - Archives publiques ou archives privées ?
- Gestion du cycle de vie
  - Conservation avec ou sans limite de temps ?
    - Le CNRS n'a compétence que pour la conservation d'archives publiques sur les âges courants et intermédiaires au-delà c'est une compétence du réseau des archives.
    - Les éliminations d'archives publiques sont encadrées (visa de l'administration des archives.
- Description des données
  - Quel modèle pour décrire les archives afin de les contextualiser ?

Historique (saison 3)

- 2008
  - Détachement à la Direction des archives de France - DAF (puis Service interministériel des archives de France - SIAF) comme chef de projet sur l'archivage électronique.
    - Mise en œuvre de la norme OAIS dans un pilote d'archivage Pilae pour les archives nationales
    - Suivi des aspects standardisation, normalisation et certification du domaine
    - Suivi des demandes d'agrément
    - Suivi de projets d'archivage et de dématérialisation
  - Au CNRS, le TGE-Adonis se lance dans un programme d'archivage pour les données de SHS

Historique (saison 3)

Programme archivage du TGE-Adonis

- Pré-étude du CERN pour savoir sur quel service adosser ce programme
  - Choix de 2 centres de calcul (CINES CC-IN2P3)
- Équipe projet associant :
  - TGE-Adonis (maîtrise d'ouvrage)
  - CINES (opérateur d'archivage)
  - CC-IN2P3 (site secondaire + fonctionnalités d'accès)
  - DAF (tutelle ministérielle)
  - CRDO (pour tester un type de données)
  - Un consultant du CNES spécialiste de la norme OAIS

Historique (saison 3)

Raisons du choix du CRDO

- Une communauté déjà organisée
  - Un modèle de données : OLAC
  - Une réflexion sur les pratiques : « Corpus Oraux : guide des bonnes pratiques »
  - Organisation des laboratoires de linguistiques en fédérations
- Une typologie étendue de ressources
  - Enregistrements (audio + vidéo)
  - Annotations (XML, PDF, textes, images)

Historique (saison 3)

Les premiers travaux

- Faire monter en compétences l'ensemble des acteurs sur le modèle OAIS
- Étude des formats utilisés par le CRDO
  - Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles.

Historique (saison 3)

Étude sur les formats

- Quelques particularité des formats audio-visuel
  - Encodage vs formatage : MP3
  - Hiérarchies de dépendances : WAV > BWF
  - Formats conteneurs : MP4
- Une méthodologie réutilisables pour les études à venir (par ex. les formats PDF)
- Des critères pour évaluer les formats
  - Ouvert
  - Normalisé
  - Largement utilisé
  - Existence d'outils de contrôle de la conformité du format avec sa spécification

Historique (saison 3)

Les premiers travaux (suite)

- Spécifications des fonctionnalités supplémentaires à ajouter dans la plate-forme d'archivage du CINES (PAC)
  - Nouvelles transactions : mise à jour des métadonnées, versions
  - Relations de parentés entre les paquets
  - Attribution d'identifiants pérennes indépendants de la plate-forme : choix de ARK
- Spécifications de fonctionnalités d'accès sur le CC-IN2P3
  - Choix de Fedora : Piste abandonnée
  - Les Centres de ressources numériques conservent les fonctionnalités d'accès

Historique (saison 3)

Demande d'agrément du CINES

- Opérateur d'archivage pour des tiers (CNRS)
- Occasion d'auditer l'ensemble des fonctionnalités d'archivage électronique couverte
  - Intégrité
  - Authenticité
  - Lisibilité
  - Traçabilité

Historique (saison 3)

Mise en production en 2010

- Urgence 1 : les enregistrements audio librement communicables
- Urgence 2 : les enregistrements vidéo librement communicables
- Urgence 3 : les enregistrements non librement communicables
- Urgence 4 : les annotations

Historique (saison 3)

Développement d'un outil de gestion

- Tableau de bord : statut des ressources
- Gestion des échanges :
  - Fabrication des paquets
    - Versement initiaux
    - Mise à jour des métadonnées
    - Dépôts de nouvelles versions
  - Suivi du workflow
    - Soumission
    - Réception des messages du CINES (accusé réception, certificat d'archivage, avis d'anomalie)
    - Mise à jour de la base de production

Historique (saison 4)

L'heure des bilans

Bilan

Une organisation en place

- Adossée à des structures plus pérennes que des laboratoires
- Une pérennisation prise en charge par des professionnels
  - La perte de donnée coté Cocoon n'est pas catastrophique. On peut récupérer les données.
- Recentrage sur les fonctions en contact avec les producteurs et utilisateurs des données
  - dépôt et accès
- Utile à d'autres projets : plusieurs sont en cours d'aboutir

Limites

Ne concerne que les données numériques

- On complète par une possible prise en charge des supports à la BnF
  - Supports conservés à la BnF
  - Contenus numérisés par la BnF
    - Accessibles en salle chercheur
    - Archivés dans le système d'archive numérique de la BnF (SPAR)
    - Les copies numériques sont déposées dans Cocoon pour permettre un accès recherche en mode web. Ces données ne sont pas re-déposées au CINES

Chiffres

Utilisation

- 2500 heures d'enregistrement déposées au CINES
- 2000 heures d'enregistrement déposées à la BnF

Limites

Cocoon ne prend en charge que les ressource orales (audiovisuelle + annotations + documentation)

- Quid des autres types de documents liés à une collecte (photo, dessins, cartes, objets, etc.) ?
  - Dépôt dans d'autres plate-formes
  - Dépôt aux AN
  - Dépôt à la BnF

Limites

Dans tous les cas, on a besoin de gérer la dispersion des informations

- Entre les institutions
- Entre les entrepôts de données et les référentiels