Numériser l'oral

Michel Jacobson

Numériser l'oral

Jacobson, Michel

LACITO (CNRS, Universités Paris III et Paris IV)
7 rue Guy Môquet
94801 Villejuif cedex

jacobson@idf.ext.jussieu.fr

RÉSUMÉ. *L'histoire de la conservation de l'oral est fortement liée à l'histoire des techniques utilisées pour sa capture. Ces dernières ont fortement évoluées au cours du siècle dernier jusqu'à un bouleversement récent apporté par les techniques numériques et par l'outil informatique.*
Après un bref exposé de la situation actuelle, nous présenterons dans ce qui suit un ensemble de préoccupations nouvelles et anciennes en matière de conservation et de diffusion de données orales. Nous illustrerons notre propos par l'exemple apporté par le Centre de Ressources pour la Description de l'Oral (CRDO), centre nouvellement crée par le CNRS pour assurer le partage de corpus oraux au sein de la communauté scientifique.

ABSTRACT. *The history of preservation of oral corpora is strongly linked to the history of the technical issues for his capture. In the last century, these issues have advanced greatly until an upheaval brought by digital technology, computer tools and networks.*
After a brief presentation of the current state of the art, we present a set of new and old concerns in the preservation and dissemination of oral corpora. We illustrate our point with the example provided by the Centre de Ressources pour la Description de l'Oral (CRDO1), centre newly created by the CNRS to ensure the sharing of oral corpora within the scientific community

MOTS-CLÉS : *corpus oraux, archives ouvertes, OAI, standards, normes, XML, Dublin-Core, OLAC.*

KEYWORDS: *oral corpora, open archives , OAI, standards, XML, Dublin-Core, OLAC.*

1. Une nouvelle forme d'écriture pour l'oral

Depuis très longtemps, l'homme cherche à garder la trace de ses productions orales. Une des premières tentatives a certainement été l'invention des écritures (alphabétiques, syllabiques, idéographiques). C'est dans la continuité de cette approche, qu'à la fin du 19ème siècle, la mise au point de l'alphabet de phonétique international par l'association du même nom[1], a permis aux linguistes de noter la parole dans ses aspects fonctionnels eux-mêmes basés sur des critères articulatoires.

Ce n'est que beaucoup plus récemment que les aspects acoustiques eux-mêmes ont pu être conservés dans des formes et sur des supports adéquats. Les premières techniques d'enregistrements analogiques (sur rouleaux, fils, disques, bandes magnétiques) du début du 20ème on fait place aujourd'hui à des enregistrements numériques stockés sur des supports magnétiques (disques durs, bandes DAT, disquettes, etc.), optiques (cédéroms, DVD), ou hybrides (MiniDisc) pour ne citer que les plus connus.

Bien que les observations de l'oral utilisent aussi bien les supports audio, que vidéo ou parfois même englobent d'autres mesures de l'activité physiologique, nous nous limiterons dans ce présent papier aux aspects liés aux supports audio.

1.1. *Les propriétés d'une écriture acoustique*

Le principe d'une écriture acoustique consiste simplement à indiquer les variations dans le temps de la pression de l'air causées par la source de bruit. Pour cela nous avons donc besoin de connaître pour chaque moment *m* de l'enregistrement la valeur mesurée de la pression de l'air et de savoir l'écrire sur un support afin de pouvoir relire celui-ci à loisir en dehors de la situation de production.

Nous n'aborderons pas ici les techniques de mesure par des microphones, afin de nous concentrer uniquement sur les aspects de représentation de ces mesures sur les supports de stockage actuels.

Pour caractériser un enregistrement audio on utilise en général les trois informations suivantes:

- La fréquence d'échantillonnage qui correspond au nombre de mesures de pression effectuées en une seconde (exprimé en Hz).
- La taille des échantillons qui correspond au nombre de valeurs différentes autorisées pour coder une mesure de pression (exprimé en bits).

[1]L'Association de Phonétique International a été créée en 1885 par les linguistes Paul Passy et Daniel Jones, et diffuse ses travaux à travers sa publication « le maitre phonétique »

- Le nombre de canaux (mono, stéréo ou plus) qui doit correspondre en principe au nombre de sources d'enregistrements distinctes utilisées lors de l'enregistrement.

À ces caractéristiques générales peuvent s'ajouter un certain nombre d'autres telles que:

- La technique de codage utilisée c'est-à-dire le système utilisé pour passer d'une représentation de l'information à une autre. Par exemple dans un codage *étendu*, les mesures sont représentées par leurs valeurs, alors que dans un codage de type *différentiel*, c'est la différence avec la mesure précédente qui est représentée.
- Le type de compression utilisé (technique dont l'objectif est de diminuer le volume occupé par les données sur le support de stockage)

Enfin pour les supports informatiques, une dernière caractéristique importante est le format de fichier utilisé. Un format détermine l'organisation logique de l'information codée dans un fichier. Il existe un grand nombre de formats plus ou moins liés aux codages employés, aux plateformes, aux constructeurs, etc.

1.2. *Un changement de paradigme pour la conservation*

Alors que la conservation des traces écrites de l'oral est une préoccupation assez ancienne, celle des traces acoustiques de l'oral est évidemment elle bien plus récente. Jusqu'à l'arrivée des supports numériques et des traitements informatiques, la conservation des enregistrements oraux passait par la conservation du support sur lequel il était stocké (rouleau, disque, bandes, etc.). Cette conservation était et reste toujours problématique parce que ces supports vieillissent, se dégradent avec le temps et que leur copie ne peut se faire qu'avec une perte d'information.

Avec l'arrivée des supports numériques, des traitements informatiques et des réseaux, la conservation s'est déplacée de la conservation des supports à la conservation des données. En effet, la duplication de données binaires pouvant se faire sans perte et à l'infini, la pérennisation des ressources (données) peut être assurée facilement avec une bonne organisation des duplications. Ce qui pose le plus de problèmes est plutôt la conservation des connaissances que l'on a des codages et formats utilisés ou la gestion des migrations d'un codage ou d'un format vers un autre. En effet le monde informatique évoluant à une vitesse bien plus grande que le monde de l'édition papier, ces codages et formats sont parfois d'un usage très ponctuel avec parfois des effets de mode. Quelques principes de choix doivent donc impérativement être suivi afin d'être en mesure d'assurer la conservation à long terme de ces ressources.

Afin de faire face à l'afflux de ces nouvelles ressources numériques, qu'elles soient issues de la numérisation de sources analogiques anciennes ou qu'elles soient directement créées dans un format numérique, le CNRS a mis en place une organisation: les *Centres de Ressources Numériques*. Il en existe actuellement cinq:

- CN2SV - Centre National pour la Numérisation de Sources Visuelles
- CNRTL - Centre National des Ressources Textuelles et Lexicales
- CRDO - Centre de Ressources pour la Description de l'Oral
- M2ISA - Méthodologies pour la Modélisation de l'Information Spatiale Appliquée aux SHS
- TELMA - Traitement électronique des manuscrits et des archives

Le CRDO[2] est le centre spécialisé sur les ressources orales. Entre autre activité[3], il maintient une archive de ressources sur les langues et le langage qui contient à ce jour un peu plus de mille heures d'enregistrements audio ou vidéo dans environ 90 langues, accompagnées parfois de documents associés tels que des transcriptions, des traductions et autres annotations.

***Figure 1**: points d'enquête des ressources présentes dans le CRDO concernant la métropole.*

[2] http://crdo.risc.cnrs.fr/

[3] Il existe deux branches du CRDO: La branche Parisienne centrée sur les aspects de gestion documentaire et de constitution d'un réservoir de données que nous présentons ici, et la branche Aixoise, centrée, elle, sur les aspects outils pour la recherche.

Cette archive a fait des choix de formats, d'outils pour la création, la gestion et la diffusion de ce type de ressource. Il s'agit d'une *archive ouverte* au sens de l'OAI[4]. Les ressources qu'elle met à disposition sont celles issues du travail du monde de la recherche. Il s'agit d'objets à la fois culturels et scientifiques. Certaines de ces ressources ont acquis un caractère patrimonial indiscutable du fait de leur apport à la recherche, de leur ancienneté et parfois parce qu'il s'agit tout simplement des rares traces, ou même des dernières traces d'une langue ou d'un état de langue disparu. Conserver et donner accès à ces données est donc non seulement un outil de mutualisation pour la communauté académique mais aussi une responsabilité vis à vis des sociétés qui ont participé à leur élaboration. Nous illustrerons nos propos, tout au long de ce papier, par des exemples pris dans les ressources et l'activité de ce centre.

2. La numérisation de l'oral et sa conservation

Nous n'aborderons pas ici les problèmes de conservation des sources anciennes sur des supports analogiques, qui demandent des équipements, des locaux et une organisation assez lourde. Ces problèmes sont assez bien connus puisqu'ils sont partagés au moins en partie avec ceux que posent d'autres supports matériels comme le papier, les peaux ou les tissus.

Nous ne présenterons ici que deux aspects: 1) la numérisation qui est l'opération qui permet de passer du monde analogique au monde numérique et 2) l'accès aux ressources et leur diffusion, une fois qu'elles sont numériques, qu'elles le soient par une opération de numérisation ou parce qu'elle sont nées directement numériques.

2.1. *La numérisation*

La conservation d'enregistrements analogiques anciens posant des problèmes que l'on ne sait pas traiter de manière satisfaisante et qui, de plus, sont très couteux, la solution passe donc la plupart du temps par une opération de numérisation. Mais alors qu'une numérisation avec comme seul but de rendre des ressources immédiatement consultables est assez simple techniquement, la numérisation dans un objectif de pérennisation de l'information est plus complexe et doit se faire avec des préoccupations de fidélité, de traçabilité et de normalisation.

2.1.1. La fidélité

La fidélité de la version numérique à la version analogique d'origine est bien sûr importante d'un point de vue scientifique mais des contraintes économiques, techniques et même conceptuelles empêchent qu'elle soit parfaite. La numérisation est toujours le résultat d'un compromis où l'on accepte de perdre une partie de

[4]Open Archive Initiative

l'information. Afin de guider les conservateurs dans le choix des caractéristiques à utiliser lors des opérations de numérisation, l'association internationale des archives sonores et audiovisuelles (IASA), par exemple, définit dans ses recommandations[5] pour les données audio une qualité plancher avec une fréquence d'échantillonnage de 48 kHz, une taille des échantillons à 24 bits, un codage sans compression et un format RIFF/WAV ou BWF, tout en conseillant des caractéristiques de qualité plus haute (192 kHz, 24 bits).

La perte due à l'échantillonnage se juge facilement à l'aune du théorème de Nyquist. Ce dernier nous dit que la fréquence la plus haute correctement restituée sera inférieure à la moitié de la fréquence d'échantillonnage choisie. En effet, il faut pour qu'une fréquence soit représentée avoir au moins deux points de mesure par période., sachant par ailleurs par des études en physiologie que la bande passante de l'audition humaine se situe entre 30 Hz et 20.000 Hz.

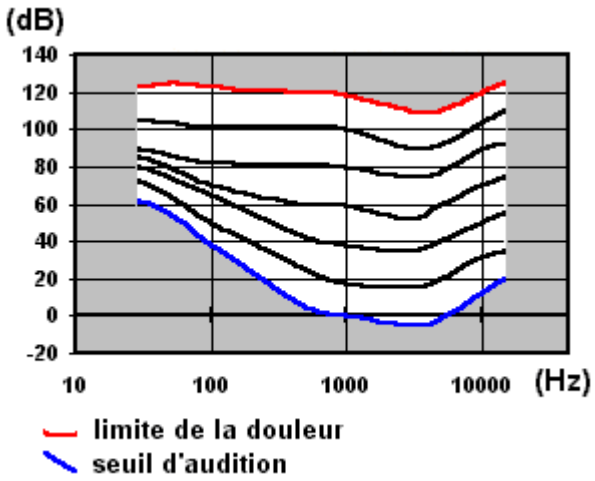
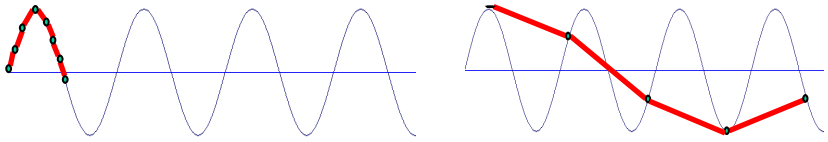


***Figure 2**: Diagramme de Fletcher et Munson[6].*

[5] http://www.iasa-web.org/IASA_TC03/TC03_French.pdf
[6] Ces courbes montrent la sensibilité moyenne de l'oreille en fonction de la fréquence.



Plusieurs points de mesure par période

Moins de deux points de mesure par période

***Figure 3**: Fréquence d'échantillonnage*

La fréquence d'échantillonnage pour la norme CD-AUdio, par exemple, est de 44,1 kHz. Ce format ne restituera donc correctement les fréquences que jusqu'à hauteur de 20 kHz, ce qui couvre bien la plage audible par l'humain. Toutefois a) l'homme peut produire des fréquences en dehors de cette plage, b) d'autres sources de bruits peuvent avoir été enregistrées de manière concomitante avec la parole au moment de la prise de son, c) l'équipement utilisé à l'origine pour capter le son puis l'écrire sur un support peut avoir ajouté des bruits (par exemple des bruits de moteurs électrique autour de 50Hz), enfin, le vieillissement du support peut lui aussi avoir déjà eu de l'influence sur les fréquences. La numérisation dans un but de préservation doit se faire avec la meilleure qualité possible afin d'éviter des pertes et d'avoir à nouveau recours à l'original analogique.

Afin de garantir la meilleure fidélité possible à l'original, la numérisation devra aussi se faire sans aucun traitement sur les données (copie droite). On évitera donc tout filtrage des fréquences, amplification, normalisation et autres traitements surtout s'ils sont non documentés et non réversibles. Ces traitement pourront toujours être faits plus tard sur des versions dérivées des ressources, dans un but esthétique, de confort d'écoute, d'homogénéisation, etc.

A titre d'exemple le travail de numérisation qu'effectue le CRDO se fait actuellement en 96KHz et 24 bits, sans aucun traitement autre qu'une éventuelle préparation physique de la bande (ajout d'amorces, collage en cas de cassures, déroulage, dessication en cas de bandes collantes, etc.). L'original numérique ainsi obtenu est alors déposé sur un système de stockage alors qu'une copie dégradée à 22KHz, 16 bits, mono alimente systématiquement un serveur dédié à la diffusion sur Internet. La distinction entre ces deux environnements permet aussi d'effectuer des traitements liés à la conjoncture et aux aspects de diffusion comme la création d'extraits, l'anonymisation, la compression, la mise dans des formats propriétaires pour la diffusion en *streaming,* opérations que l'on ne souhaiterait sans doute pas pratiquer sur le document d'origine.

2.1.1. La traçabilité

La traçabilité des ressources répond à un certain nombre de besoins techniques (comme celui de retrouver les supports d'origine) et de besoins d'usages (tel que retrouver tout le contexte social et culturel de l'époque de production). C'est aussi ce qui permet à ceux qui consultent les ressources de s'adresser à ceux qui ont contribué à leur création (auteurs, éditeurs, chercheurs, etc.) afin de leur communiquer des informations ou de leur en demander. C'est enfin une exigence de tout travail scientifique de s'exposer à la critique et aux contradictions.

L'organisation de cette traçabilité passe en grande partie par la documentation, la standardisation de celle-ci et la manière de communiquer cette information. La documentation recouvre tout un ensemble de préoccupations qui vont du technique au scientifique en passant par le juridique. Pour clarifier l'ensemble des préoccupations à couvrir pour la documentation d'une ressource, on peut distinguer quelques grands acteurs qui sont ses créateurs, les gestionnaires et les utilisateurs.

Les créateurs d'une ressources peuvent être nombreux, on peut citer à titre d'exemple, le chercheur à l'initiative d'une l'enquête, les enquêteurs qui ont pratiqués les enregistrements, les locuteurs enregistrés, l'organisme qui a financé ce travail, les transcripteurs, traducteurs et autres annotateurs qui ont enrichis cette ressource dans le cadre de l'exploitation prévue par les objectifs l'enquête. Pour tous ces participants il est souhaitable de les identifier, de qualifier le statut de leur participation et de rattacher ces derniers aux droits qu'ils peuvent avoir dessus (droits d'auteur, droit au respect de la vie privée[7]).

Les gestionnaires des ressources, eux, ont besoin d'informations techniques sur la ressource, afin de pouvoir avoir des actions sur celle-ci. Par exemple son format pour éventuellement effectuer des migrations, la source et le procédé qui ont donnés lieu à la forme électronique dans le cas de documents non nativement numériques. Ils ont aussi besoin de connaître des informations juridiques tels que la liste des ayant-droit et les conditions dans lesquelles des ressources peuvent être communiquées à autrui, modifiées, dupliquées, etc.

L'utilisateur, finalement, a besoin lui de renseignements sur la forme de la ressource (son format, sa taille ou sa durée, son emplacement) afin d'adapter au mieux les outils qu'il va utiliser pour l'exploiter, de renseignements sur son contenu intellectuel afin de déterminer si cette ressource est pertinente pour lui et d'informations juridiques afin de savoir ce qu'il a le droit de faire avec et à qui il peut s'adresser pour en savoir plus, qui il doit citer pour faire référence à la ressource dans une publication.

[7]Baude Olivier (Ed.) -- *Corpus oraux : Guide des bonnes pratiques 2006*

Toutes ces informations à propos d'une ressource s'appellent des métadonées. Plus celles-ci sont riches et explicites plus la découverte et l'exploitation d'une ressource en sera facilitée.

Au CRDO, chaque ressource, c'est-à-dire en fait chaque fichier déposé par un chercheur fait l'objet d'une description comportant tous les renseignements cités plus haut. L'ensemble des fiches de renseignements constitue le catalogue des ressources du centre. Le catalogue public correspondant à toutes les ressources librement accessibles est communiqué de manière standardisé à tous ceux qui le demandent, c'est-à-dire en pratique à des fournisseurs de services comme OAIster[8], OLAC[9], Google, etc. qui offrent après des moteurs de recherche sur des ensembles plus vastes de données.

2.1.1. La normalisation

Finalement la fidélité à l'original et la possibilité de tracer l'historique des objets sont des précautions méthodologiques guidée par des préoccupations scientifiques. Pour faciliter l'usage d'une ressource, la forme de celle-ci ainsi que sa description et son procédé de communication doivent être compréhensibles pour l'utilisateur final. Enfin, pour sa conservation à long terme, il faut que sa forme soit correctement documentée, ce qui nous conduit naturellement à l'utilisation de normes et de standards autant que possible. Ce point est particulièrement critique pour les formats de fichiers et les codages utilisés. Une norme peut être vue pour certains de ces aspects comme un langage partagé et explicite qui permet de décrire un objet, un procédé. Une description normalisée permet une interprétation non ambiguë, dont la stabilité est garantie par le concept même de norme (résultat de la réflexion d'un grand nombre que l'on stabilise et que l'on fige dans une version, puis qui est maintenu par un organisme dont c'est la mission et qui n'est pas directement influencé par des intérêts privés, commerciaux ou industriels).

Les fichiers audio du CRDO sont soumis dans des formats WAV ou BWF avec un codage PCM[10]. Les fichiers d'annotations sont soumis dans un format XML bien-formé, accompagné éventuellement d'un schémas d'annotation permettant la validation ou bien en fichier texte seul (ASCII ou Unicode). Si les annotations sont manuscrites il est également possible de les déposer sous forme de scans encapsulés dans des fichiers PDF.

Les schémas utilisés pour exprimer les métadonnées au CRDO sont ceux du Dublin-Core et ceux de OLAC. La norme Dublin-Core[11] définit quinze étiquettes de

[8] http://www.oaister.org/
[9] Open Language Archives community (http://www.language-archives.org/)
[10] Pulse Coded Modulation
[11] Dublin-Core normalisé ISO...

signification très génériques (title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights). Chacune de ces rubriques peut être précisée plus finement. Par exemple, à la place de la rubrique date il est possible d'utiliser l'une de ces étiquettes (created, valid, available, issued, modified, dateAccepted, dateCopyrighted, dateSubmitted). Dublin-Core définit enfin un certain nombre de codages possibles pour ces étiquettes, par exemple W3CDTF pour les dates ou URI pour les identifiers, etc. OLAC reprend l'ensemble de ces spécifications de codage et vient ajouter ou préciser un certain nombre de rubriques pour l'adapter aux besoins de la communauté sur l'oral (discourse-type, language, linguistic-field, linguistic-type, role) et entretient pour chacun de ces ajouts des vocabulaires contrôlés. Par exemple, le rôle d'un contributor peut prendre l'une de ces valeurs (annotator, author, compiler, consultant, data_inputter, depositor, developer, editor, illustrator, interpreter, interviewer, participant, performer, photographer, recorder, researcher, research_participant, responder, signer, singer, speaker, sponsor, transcriber, translator)

2.2. *L'accès aux ressources du CRDO et leur diffusion*

Le CRDO permet aux laboratoires, aux chercheurs, aux projets et à des communautés de conserver, d'avoir accès, de partager des ressources. C'est aussi un outil qui permet d'augmenter la visibilité de celles-ci en facilitant leur découverte et leur exploitation. Afin d'atteindre cet objectif, le CRDO à mis en place une architecture basée sur le modèle de diffusion de l'Open Archive Initiative. Les grandes caractéristiques de cette architecture tournent autour des deux types d'acteurs principaux que sont les fournisseurs de ressources et les fournisseurs de services. Les fournisseurs de ressources peuvent être repartis sur le territoire, ce qui est le cas dans la communauté à laquelle on s'adresse puisqu'il existe de nombreux corpus dans de nombreux laboratoires, universités etc. et peu de tentatives de centralisation. Tous ces fournisseurs de ressources doivent savoir communiquer des informations sur le contenu de leurs entrepôts de données à l'aide d'un protocole défini par l'OAI: OAI-PMH[12]. Les fournisseurs de services peuvent alors interroger autant de fournisseurs de ressources qu'ils le souhaitent en utilisant cet unique protocole. Ils centralisent alors, pour les besoins de leurs services l'ensemble des métadonnées des catalogues interrogés. Cette action s'appelle le *moissonnage* et seules les métadonnées sont concernées, les ressources elles-mêmes restant dans les entrepôts d'origine. Chaque entrepôt peut choisir le ou les schémas avec lesquels il va encoder ses métadonnées à partir du moment où il s'engage à les délivrer aussi en utilisant le Dublin-Core. Le CRDO peut les délivrer soit en Dublin-Core soit en OLAC.

2.2.1. L'accès

Le CRDO mis en place un système de droit d'accès distinguant trois niveaux:

[12]Open Archive Initiative – Protocol for Metadata Harvesting

- Public: Les ressources sont librement accessibles sans mot de passe et les métadonnées qui les concernent sont publiques et diffusées à travers le protocole OAI-PMH.
- Privé: Les ressources et leurs métadonnées ne sont accessibles qu'avec un mot de passe.
- Semi-public: Les ressources ne sont accessibles qu'avec un mot de passe mais les métadonnées qui les concernent sont publiques et diffusées à travers le protocole OAI-PMH. Un champ de ces métadonnées précise cette politique (accessRights= « Access restricted (password protected) » ou accessRights = « Freely accessible for non-commercial use »)

Bien évidement on ne facilite que la découverte des ressources qui ne sont pas privées. Tous les fournisseurs de services qui le souhaitent peuvent moissonner l'archive du centre. Par exemple des organismes comme OLAC, OAIster, et même Google moissonnent régulièrement cette archive et proposent des moteurs de recherche qui donnent accès aux ressources du CRDO ainsi qu'à des collections beaucoup plus vastes. A titre d'illustration, la figure 4 montre le résultat d'une recherche sur le mot-clef *nelemwa* sur le portail OAIster. Les réponses comportent un lot en provenance du CRDO (des enregistrements et leurs transcriptions), un autre en provenance de HAL (préprints ou postprints d'articles sur cette langue) et un dernier lot en provenance de Persée (articles sur cette langue dans des périodiques numérisés). Toutes ces ressources ont été déposées par un même chercheur, ce qui n'aurait certainement pas été le cas si notre recherche avait porté sur un langue plus étudiées comme le français.

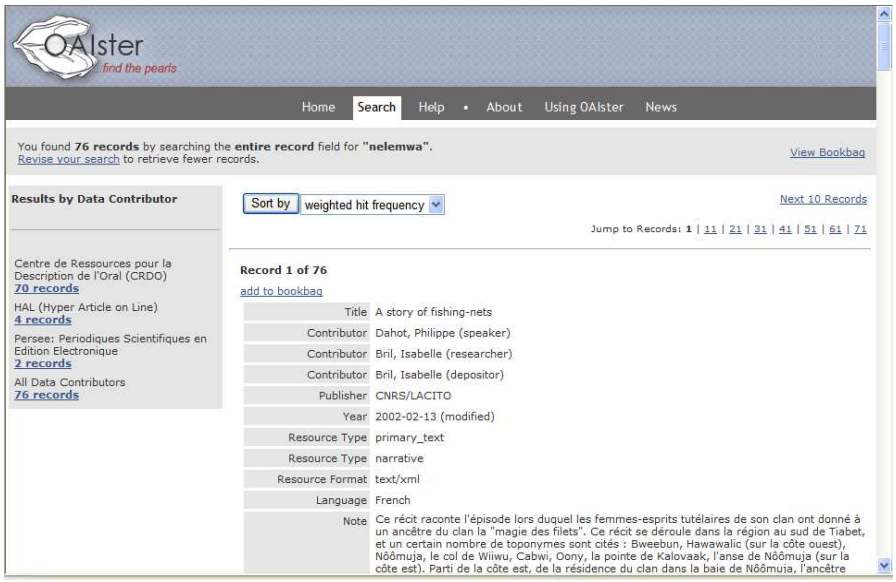


***Figure 4**: recherche du mot-cléf "nelemwa" sur le portail OAIster*

Un accès est aussi possible sur le portail du CRDO, qui offre un moteur de recherche qui exploite complètement l'ensemble du jeu d'étiquette d'OLAC/Dublin-Core, qui permet la recherche dans les environnements privés, à partir du moment où la personne s'est identifiée et qui enfin donne quelques outils d'exploitation des ressources trouvées.

2.2.1. La valorisation

En premier ce qui facilite l'exploitation des ressources est leur aspect normalisé. Cela signifie que tout un chacun peut, s'il s'en donne les moyens, développer ses outils ou les choisir dans ceux qui respectent les normes utilisées. En général, plus une ressource est normalisé plus vaste est le choix d'outil pour l'exploiter. La valorisation de certaines ressources sur le site du CRDO (interfaces de consultation multimédia, recherche d'occurrences, etc.) ou en dehors (portails communautaires sur le français et les langues de France, portails institutionnels pour un laboratoire, ou portails dédiés à un projet) permettent de faciliter d'exploitation en adaptant l'exploitation au public visé.

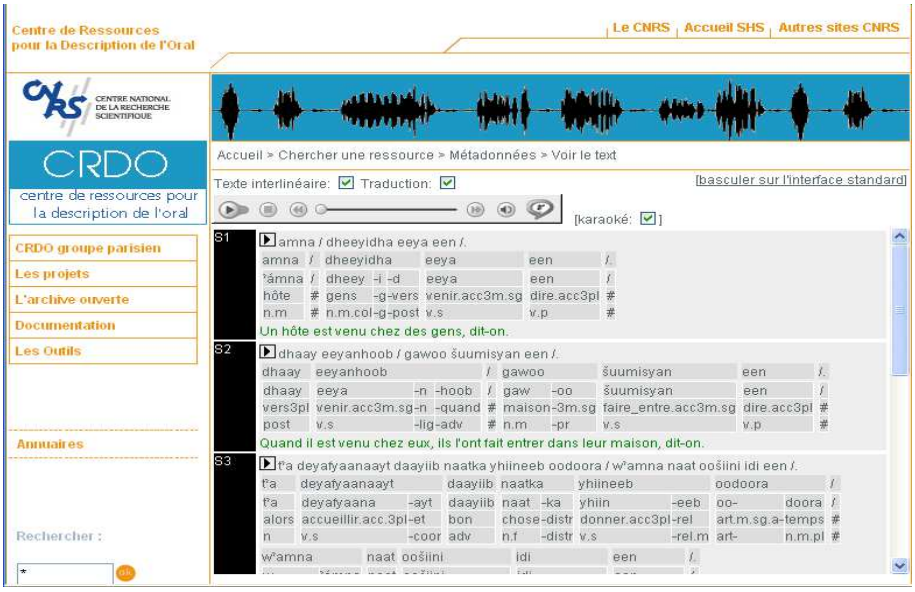


***Figure 5**: présentation multimédia (transcription + enregistrement audio)*

Pour le moment deux DTD (Transcriber, Archivage) font l'objet d'une valorisation particulière sur le site du CRDO. Ces deux DTD permettent une visualisation multimédia directement sur le site alors qu'une ressource qui utiliserait une DTD

propriétaire ne pourrait être consultée qu'en mode *download*. Parallèlement des expérimentations ont lieu sur les schémas de TEI et de CHILDES.

3. Perspectives

La dématérialisation puis la conservation de l'oral numérique passe donc par une série d'opérations délicates qui demandent des compétences dans différentes disciplines allant de l'ingénierie du son à la gestion documentaire en passant par le développement informatique et des métiers plus nouveaux qui sont à cheval sur ces disciplines.
Le prototype d'architecture mis en place au CRDO est principalement tourné autour des fonctions de stockage, d'accès et de diffusion. Des efforts restent à faire sur les aspects organisationnels, la mise en place de procédures et la conservation pérenne, qui pourraient être aidés en particulier par la mise en place d'un modèle d'architecture inspiré de l'OAIS[13].

4. Bibliographie

Baude Olivier (Ed.), Blanche-Benveniste Claire, Calas Marie-France, Cappeau Paul, Cordereix Pascal, Goury Laurence, Jacobson Michel, de Lamberterie Isabelle, Marchello-Nizia Christiane, Mondada Lorenza -- *Corpus oraux : Guide des bonnes pratiques 2006* -- Paris : CNRS éditions, 2006.- 203 p.

Bray, T., Paoli, J. et Sperberg-McQueen, C. M. (Eds), « Extensible Markup Language (XML) Version 1.0 », recommandation du Word Wide Web Consortium, 10 février 1998.

Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) : standard CCSDS, CCSDS 650.0-B-1 (F).

Sperberg-McQueen, C. M., et Burnard, L., « TEI Guidelines for Electronic Text Encoding and Interchange (P3) », Chicago and Oxford: ACH/ACL/ALLC Text Encoding Initiative, 1994.

The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 du 14 juin 2002 (http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm)

[13]Open Archival Information System