Automatic indexation and analysis of ethnomusicological archives : issues and new challenges.

Marie-France Mifune

Automatic Indexation and Analysis of Ethnomusicological Archives: Issues and New Challenges

Marie-France Mifune

UMR 7206 Eco-anthropologie et Ethnobiologie, CNRS – MNHN – Université Paris Diderot – Paris Sorbonne Cité – Sorbonne Universités

As very large scientific databases for social science research become increasingly available, their management raises new fundamental questions and new research challenges. The main challenges with scientific databases in ethnomusicology such as that of the *CNRS – Musée de l'Homme* include the preservation of audio content and scientific data, questions about intellectual property and ethical concerns for data publication, and the development of tools and technologies for cataloging, documenting, and accessing the data. In ethnomusicology, researchers work with sound recordings as well as other multimedia documents such as videos and photographs. The digitization of these resources allows us to preserve and easily access such materials.

Since 2011, these digitized archives have been available through a web platform called Telemeta. Today, 49,300 items from 5800 collections of the *CNRS – Musée de l'Homme* are catalogued on this database (See inside this volume the article by Simonnot for the description of the *CNRS – Musée de l'Homme* Archive and Telemeta platform). The purpose of the Telemeta web platform is to provide researchers and archivists with a system for preserving and accessing sound archives. Telemeta allows sound items to be shared, along with associated metadata that contain information about the context in which the music was produced (such as the title of the musical piece, the instruments played, the population of origin of the musicians, cultural elements related to the musical item, the collector, the year of

the recording and so on).

Indexation in Telemeta is currently based on a semantic search entirely depending on the information provided by the collector and the depositor of a sound item. Telemeta would benefit from the addition of new and complementary metadata extracted directly from the audio signal, which could provide information about the global content or specific temporal zone of each audio item when contextual information is missing. The DIADEMS project aims to provide new automatic tools for indexing and analyzing the audio content, which will be implemented on the Telemeta web platform.

The structure of the article is as follows. The first section presents the DIADEMS project, the different partners involved and the interdisciplinary work on the development of computer tools. The second section describes our work on the automatic indexation of the audio content in the ethnomusicological archives. This represents one possible approach to the issue and is currently a work in progress. The final section discusses new perspectives in ethnomusicology, and beyond, raised by the DIADEMS project.

1. DIADEMS project

The DIADEMS[1] project began in 2013 with the aim of developing computational tools to automatically index the audio content of the ethnomusicological archive of the *CNRS – Musée de l'Homme.* The tools are designed to extract metadata directly from audio signals in order to improve indexing, and to provide efficient access to sound items in the archive through the Telemeta web platform (Fillon et al. 2014).

The DIADEMS project was elaborated during a meeting of researchers from different backgrounds at the "Sciences and Voice" Summer School organized by the French National Center for Scientific Research (CNRS) in 2010. During this meeting, acousticians and ethnomusicologists recognized that automatic tools developed by computer scientists could

facilitate the analysis and storage of their digitized audio archives.

DIADEMS is a multidisciplinary project involving research laboratories from several different disciplines such as the humanities and social sciences, the science and technology of information and communication, and the Parisson company, which is involved in the development of Telemeta. Participants include specialists in speech processing and Music Information Retrieval, IT developers, acousticians, anthropologists, linguists, ethnomusicologists, sound engineers, archivists and experts in multimedia development. To better develop tools to detect analytical categories of interest, we combined the complementary skills of researchers in the four laboratory partners in DIADEMS (IRIT[2], LABRI[3], LIMSI[4], LAM[5]), which specialize in the analysis of audio signals. The development of automatic indexation tools is carried out through collaborations between computer scientists who develop the computer tools and ethnomusicologists, archivists, ethnologists and linguists (CREM-LESC[6], EREA-LESC[7], MNHN[8]) who are both primary users of the platform Telemeta and experts of the recordings contents.

Ethnomusicologists, acousticians, computer engineers, linguists and archivists, analyze sound with different objectives and disciplinary backgrounds with different theories, methodologies, concepts and terminologies. Thus, because interdisciplinary work between specialists from these disciplines requires the sharing of concepts and terminology, we produced a glossary of analytical categories and parameters used in the project to improve communication and comprehension among the DIADEMS partners.

The new technologies developed by computer scientists aim to help ethnomusicologists extract information that is relevant to their research interests from large amounts of archival content. To develop tools that would be useful to acousticians and ethnomusicologists, it was important for the computer scientists to become familiar with key issues regarding archiving and retrieval in ethnomusicological databases, and with acoustic

characteristics of ethnomusicological recordings that may differ from those of other sound data, such as television and radio recordings. Most of the recordings in the *CNRS – Musée de l'Homme* archives are unpublished recordings of music and oral traditions from around the world, collected during field expeditions by researchers. The ethnomusicological archives include 28,000 recordings spanning a wide variety of cultural contexts worldwide starting in the 1900s until today as well as various settings (inside, outside, studio settings) and a wide variety of content (musical practice, speech, dance, and so on).

In addition to heterogeneity in sounds and production contexts, ethnomusicologists and archivists must consider how the automatic indexation tools will be used as well as questions about intellectual property rights and the wishes of depositors. Many items in the collection include very little contextual information. The use of automatic indexation tools will help archivists to index such sound items and add new content information. Automatic indexation will also help researchers to better identify the quality and the content of the recordings. Thus, ethnomusicologists and archivists must precisely define categories that will both capture the content of audio recordings, and facilitate searches and analyses by researchers.

In this highly interdisciplinary context, we organized the development of computer tools in four stages. In the first stage, ethnomusicologists, acousticians and archivists defined the categories (e.g. speech and song) and associated parameters (e.g. fundamental frequency ranges) they wished to detect automatically, bearing in mind the relevance of these categories for other future users of the Telemeta platform. Then, ethnomusicologists and archivists manually annotated the audio contents of a representative sample of sound items accessible on Telemeta to create a learning dataset. In the second stage, computer scientists used the learning dataset to develop computational tools to automatically index these audio recordings according to the categorization and parameterization defined by ethnomusicologists. Third,

the computational tools were applied to another set of recordings serving as the test dataset. Validation was conducted by comparing manual annotations made independently by the ethnomusicologists on the test dataset with the automatic segmentation computed by the software. Categorization parameters provided by ethnomusicologists based on the learning dataset may not always adequately transpose to the test dataset, hence producing automatic annotation errors during the validation procedure. Thus, computer scientists and ethnomusicologists went back and forth to refine the categorization parameters on the learning dataset for improving the efficiency and accuracy of the automatic annotation software. The fourth and final stage will integrate the newly developed computational tools directly into the Telemeta online platform for automatic indexation of any *CNRS – Musée de l'Homme* audio recording. To do so, ethnomusicologists, archivists and computer scientists jointly designed the front end and user interface of these new tools.

2. Definition and detection of analytical categories

Categories of vocal productions

Computer scientists have previously developed tools to automatically detect analytic categories such as speech, music, and song, primarily from French radio and television recordings. While such categories may be efficient to index sound recordings from radio and television, they are largely unable to account for the diversity of the sound recordings in the *CNRS – Musée de l'Homme* ethnomusicological archives. Indeed, the archives contain a large variety of musical performances and vocal practices ranging from speech to song. In fact, liminal utterances between speech and song [9] have been characterized by previous ethnomusicological studies. They showed that various modes of utterance can be characterized by specific acoustic features in the Yezidis from Armenia (Amy de la Bretèque 2010) and the Toraja from Indonesia (Rappoport 2005). Nevertheless it is extremely

challenging for ethnomusicologists to define an efficient categorization of vocal productions based only on acoustic criteria and equally efficient in all cultural practices worldwide. François Picard proposed that differences between recitation, declamation, chanting and singing are not universal, but differ between distinct genres and cultures as exemplified by the comparison of liminal utterances in speech and song in Chinese Buddhist ritual and Beijing Opera (Picard 2008).

The DIADEMS project proposed to formally test this hypothesis by evaluating the efficiency of automatic indexation of a wide variety of vocal productions, ranging from speech to song sampled from different cultures, based only on acoustic parameters. By doing so, we wished to reduce the traditional gap in academic studies between sound and semantics and to develop combined analytical tools for the study of vocal production. Classic ethnomusicological approaches focus on endogenous categorizations of musical practices, thus specific to each culture and never solely based on acoustic criteria. Thus, ethnomusicologists needed to fundamentally change their usual methodological paradigms, concepts, and analysis tools to develop automatic indexation tools based only on acoustic criteria.

We chose to classify vocal productions in two general categories: speech and song. Then, according to the database, we subsequently identified and defined subcategories such as talking, storytelling, recitation, chanting and singing (Figure 1), not based on style or genre, but on acoustics features only.

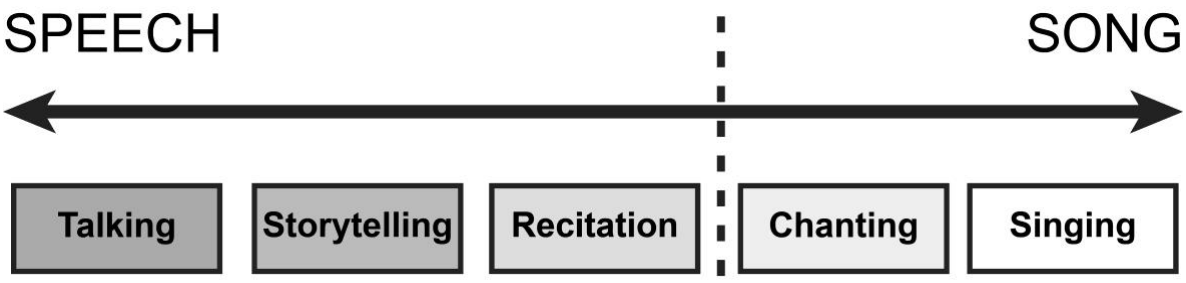

Figure 1 Speech and Song categories

We define *speech* as a vocal production with a significant proportion of unvoiced sounds. This feature is relevant for all the subcategories such as *talking*, *recitation* and *storytelling*. We distinguish *talking* from *storytelling* based only on the mode of realization. *Talking* is characterized by dialogue and *storytelling* by monologue with or without back-channel signal (i.e., an expression or word used by a listener to indicate that he or she is paying attention to the speaker). *Recitation* is characterized by more regular breath rate and rhythmic flow than talking, and a monotonous statement with low frequency range variations.

Alternatively, *song* is defined as a vocal production with a significant proportion of lengthened syllables and voiced sounds. *Chanting* is characterized by a very limited vocal range, close to recto-tono. *Singing* is characterized by ordered pitches and relative stability of fundamental frequencies. Table 1 synthesizes this categorization of vocal production and gives some examples from the *CNRS – Musée de l'Homme* database.

Level 1	Level 2	Characterization	Mode of realization	Example from the database
SPEECH	Talking	* *Significant proportion of unvoiced sounds* * Syllabic flow is faster than singing * Fundamental frequency shows rapid random variation (short intervals)	Dialogue	CNRSMH_I_2007_001_033_02 (Gabon)
	Storytelling	* *Significant proportion of unvoiced sounds* * Syllabic flow is close to talking	Monologue with / without back-channels	CNRSMH_I_1970_012_007_01 (Mali)
	Recitation	* *Significant proportion of unvoiced sounds* * Breath rate is more structured than talking breath rate * Rhythmic flow is more regular than talking rhythmic flow * Low frequency range variations make the statement monotonous	Monologue with / without back-channels	CNRSMH_I_1965_006_001_21 (Vietnam)
SONG	Chanting	* *Significant proportion of lengthened syllables and voiced sounds* * Words are uttered on a single frequency or a very limited vocal range, close to recto-tono * Fundamental frequency shows some dropouts at regular intervals	Monologue or Singing turns	CNRSMH_I_1959_006_001_01 (South India)
	Singing	* *Significant proportion of lengthened syllables and voiced sounds* * Pitches are ordered * Fundamental frequencies show relative stability	Solo / Choir / Singing turns	CNRSMH_I_1977_006_010_02 (North India)

Table 1: Characterization of vocal categories

These different categories characterized by ethnomusicologists have been then evaluated by the acousticians from the DIADEMS project (Feugère et al. 2015). They tested some pitch features for the characterization of intermediate vocal productions. They tested recordings' excerpts totalizing 79 utterances from 25 countries around the world. Among the tested features, the note duration distribution proved to be a relevant measure. The results show that the proportion of 100-ms notes and the duration of the longest note are useful for classifying singing, chanting and speech but not for discriminating speech categories. Talking, storytelling and recitation categories cannot be distinguished based only on the acoustic features here tested. Results support the definitions given by ethnomusicologists: talking and

storytelling only differ by the mode of realization (monologue/dialogue), which is not embedded in pitch features. Furthermore, results corroborate the distinction made by ethnomusicologists between speech and song and their respective subcategories (talking, storytelling, recitation concerning speech and singing, chanting concerning song). These first results show that note duration can be a relevant measure to discriminate between speech, singing and chanting.

The other challenges for computer scientists concern the automatic detection of audio contents with overlapping speech from multiple speakers, speech over music, and instrumental music mixed with singing and/or spoken interventions. Further testing and validation will determine whether these computational tools can be successfully applied to other vocal productions from the archives.

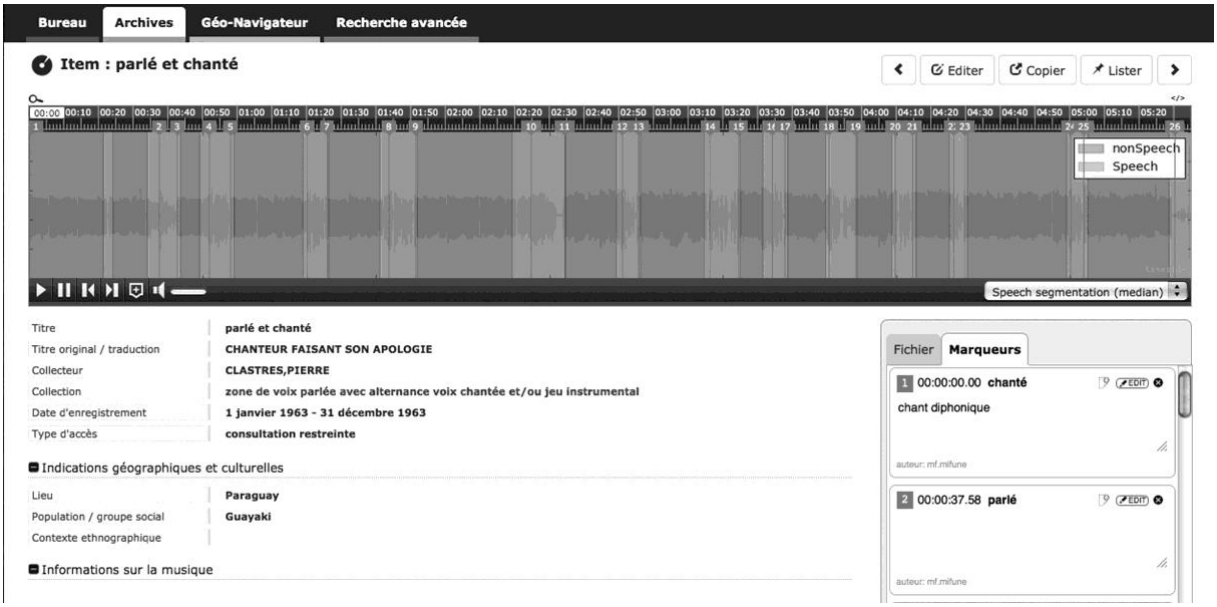

Figure 2 Example of automatic tools to detect speech

The detailed definitions of categories and their criteria are still under development. Nevertheless, these definitions are a first attempt based on the ethnomusicological archives with which researchers[10] in DIADEMS are most familiar. Ultimately, the goal is to refine

these categorical definitions while expanding the corpus of recordings considered. In particular, we are aware that some of the terminology here used can be inappropriate for some specific practices or can be confusing among ethnomusicologists. These major lexicon issues will be overcome by discussing and confronting our newly develop tools with the broader scientific community.

Finally, we also developed sets of tools for the detection of alternation between groups, such as speakers, soloists and choirs, to identify the structure of the recording. This automatic detection can also help researchers to easily access the portions of a recording of specific interest without having to listen to the full recording, hence further improving the efficiency of the analysis.

We have worked extensively on vocal productions since it is the predominant content of the archives and it is also the most challenging task for all the partners. Thus, in this paper we will succinctly present the detection of other types of content such as musical instruments and sounds generated by the recording device that could be useful for indexing and analyzing audio content of sound items from the *CNRS – Musée de l'Homme* archives.

Categories of musical instruments

The DIADEMS project also started to develop useful tools for detecting families of musical instrument. For the detection of instrumental music, we chose the four organological categories defined by Hornbostel and Sachs (1914) and by Dournon (1981): aerophone, chordophone, membranophone, and idiophone. This classification is based on the nature of the material producing sounds: air, string, membrane, or the body of the instrument itself. We choose the playing technique (such as plucked, struck and bowing) as a second criterion to characterize a musical instrument.

The computer engineers started to test several audio features for timbre characterization based on ethnomusicological recordings excerpts previously categorized in different playing techniques within each organological category (Fourer et al. 2014). The computer engineers chose only solo excerpts (where only one monophonic or polyphonic instrument is active) in order to reduce the interference problems, which may occur during audio analysis. This method was found accurate and efficient to automatically classify timbre (Fourer et al. 2014). Thus, within an organological category, the playing technique is relevant for differentiating one instrument from another. The computer engineers are currently developing tools to further test this line of research. The development of timbre characterization could transform the previous classification of musical instruments in four organological categories, and could foster the development of a new categorization of musical instruments in ethnomusicology.

Identification of other types of sound

In addition to speech, song and musical instrument, numerous ethnomusicological recordings contain other types of sounds that are categorized as sounds from the environment such as rain, insect or animal sounds, engine noise, and sounds generated by the recording device itself such as the wind in the microphone or sounds resulting from a damaged recording medium. These types of sounds are usually identified as noise interfering with the object of the recording. While sometimes disturbing, some of these sounds can be used for the analysis and automatic segmentation of the audio content.

A primary concern when analyzing a field recording is to identify the beginning and the end of a recording session. When a collector stops or starts a recording session, there are some specific sounds made by these interruptions (especially on a magnetic-tape recorder) that can be automatically detected and used to segment the recordings. Ethnomusicologists

usually record several sessions in the field one after the other, and most of the time announce or otherwise indicate the change of session. However in some circumstances, recordings of very long duration (such as recordings of rituals) can present several recording interruptions not explicitly announced by the researcher. Automatically detecting these interruptions is of particular interest to identify the changes of context especially when recording cuts are numerous or barely detectable in prolonged recordings. The detection of sounds from recording media is also useful for assessing the quality of the recording prior to listening. The LAM laboratory currently conducts the exhaustive inventory and classification of the various types of sounds generated by the recording processes in the *CNRS – Musée de l'Homme Archive*.

3. New research perspectives in ethnomusicology

The Telemeta web-platform provides ethnomusicologists with greater access to the *CNRS – Musée de l'Homme* recording archive, one of the largest archives of ethnomusicology in Europe, and with computational tools for automatic analyses, annotation, and diffusion of recorded data. In particular, the new technologies developed by DIADEMS will allow efficient indexation and analysis of the recordings, directly from the audio content. Users will have the option to automatically detect and segment sound recordings using a wide variety of parameters through a dynamic audio player embedding audio streaming, visualization of audio signals and its analytical segmentation.

Tools available through Telemeta will also offer new possibilities for diachronic and synchronic comparative studies in the field of musical analysis. Detecting musical similarities based on several parameters (pitch and rhythmic features for instance) will allow users to identify the structure of a piece, as well as rhythmic or melodic patterns that can be used to highlight the possible relationships between musical pieces within and among collections.

The Telemeta platform offers numerous opportunities for collaboration and data sharing within the scientific community. Users can share content including new computer tools, manual annotations of musical pieces and results of automatic analyses. Furthermore, the aim of the platform is to contribute in a collaborative way to the enrichment of the archives both by adding new audio material and metadata. The tools will also be useful for researchers from other fields, such as linguists and anthropologists, interested in audio content such as spoken word recordings related to oral traditions. Most importantly, Telemeta will provide a key online database for musicians and communities worldwide, for listening, enriching and preserving their intangible heritage, as well as communicating with researchers interested in their oral traditions.

The Telemeta open-source framework provides a new platform for researchers to efficiently work with sound archives and to share various kinds of metadata. This web platform provides a new way to interact with sound archives by seeing them not only as repositories for storing the results of previous studies, but as primary sources for new ethnomusicological research (Seeger 2011). The new technologies available through Telemeta promote this transformation of how archives are used.

In the DIADEMS project, interdisciplinary work among ethnomusicologists, archivists, and computer scientists has allowed us to apply cutting-edge computer science research to facilitate studies in ethnomusicology. The development of these computational tools has helped us to enhance our own analytical tools, concepts and research objectives in ethnomusicology. We hope that this project will foster future development of online collaborative tools and that the resource we have developed will evolve according to the growing needs of archivists, researchers, teachers and new users of the Telemeta platform.

Acknowledgements

This work is supported in part by the ANR (French National Research Agency) grant ANR-12-CORD-0022.

References cited

Amy de la Bretèque, Estelle, 2010. “Des affects entre guillemets. Mélodisation de la parole chez les Yézidis d’Arménie”. *Cahiers d’ethnomusicologie* 23: 131-145.

Dournon, Geneviève, 1981. *Guide pour la collecte des musiques et instruments traditionnels*, Paris: UNESCO.

Feugère Lionel, Doval Boris, Mifune Marie-France, “Using Pitch Features for the characterization of intermediate vocal productions”. Proceedings of *The Fifth International Workshop on Folk Music Analysis*, 2014, Paris.

Fillon Thomas, Simonnot Joséphine, Mifune Marie-France, Khoury Stéphanie, Pellerin Guillaume, Lecoz Maxime, Amy de La Bretèque Estelle, Doukhan David, Fourer Dominique, Rouas Jean-Luc, Pinquier Julien, Mauclair Julie and Barras Claude, 2014. “Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis”. Proceedings of the *First International Digital Libraries for Musicology Workshop*.

Fourer Dominique, Rouas Jean-Luc, Hanna Pierre, Robine Matthias, “Automatic timbre classification of ethnomusicological audio recordings”. Proceeedings of the *International Society for Music Information Retrieval Conference* (ISMIR 2014).

Hornbostel Eric von and Sachs Curt, 1914. “Systematik der musikinstrumente”. *Zeischrift für Ethnologie*, 46: 553-590.

Picard, François, 2008. “Parole, déclamation, récitation, cantillation, psalmodie, chant”. *RTMMAM* 2: 1-16.

Rappoport, Dana, 2005. “Les langues frétillantes. Modalités de profération de la parole rituelle chez les Toraja d’Indonésie”. *Second Congress of Asia Network*, EHESS Paris.

Seeger, Anthony 2005. “New technologies Requires New Collaborations: Changing Ourselves to Better Shape the Future”. *Musicology Australia*, 27: 94-111.

---

[1] The Diadems project started in January 2013. It is funded for three years by the French National Agency of Research (ANR). The acronym of the project stands for Description / Indexation / Access to Ethnomusicological Documents and Sounds. http://www.irit.fr/recherches/SAMOVA/DIADEMS/fr/welcome/

[2] The team SAMoVA from the IRIT laboratory (CNRS / Université Paul Sabatier / INPT, UMR 5505, Toulouse, France) is specialized in the primary analysis of the audio signal to identify the language, the speaker, to detect music, speech, songs and key sounds such as applauses.

[3] The LaBRI laboratory (CNRS / Université de Bordeaux, UMR 5800, Bordeaux, France) is specialized in building algorithms to estimate the level of similarity of musical pieces as functions of various criteria, such as harmonics, metrics.

[4] The LIMSI laboratory (CNRS / Université Paris Sud, UPR 3251, Orsay, France) is specialized in modeling speech patterns and in automatically processing speech recordings.

[5] The team LAM from the Institut Jean Le Rond d’Alembert (CNRS / Université Pierre et Marie Curie / Ministère de la culture et de la communication, Paris, France) is specialized in the study of sound from the point of view of engineering sciences (physics, acoustics) and social sciences (cognitive psychology and linguistics).

[6] The team CREM from the LESC laboratory (CNRS / Université Paris Ouest Nanterre La Défense, UMR 7186, Nanterre, France) is specialized in the study of musical practices and knowledge and manages the archive of CNRS-Musée de l'Homme.

[7] The team of linguists EREA from the LESC laboratory (CNRS / Université Paris Ouest Nanterre La Défense, UMR 7186, Nanterre, France) is specialized in ritual speech from the Maya.

[8] The team of ethnomusicologists from the Eco-anthropologie et ethnobiologie laboratory (CNRS / MNHN / Université Paris Diderot / Paris Sorbonne Cité/ Sorbonne Universités, UMR 7206, Paris, France) is specialized in African music studies.

[9] A colloquium on liminal utterances between speech and song has been organized by the International Council for Traditional Music in May 2015 and hosted by the Center of Research in Ethnomusicology (CREM). A round table has been dedicated to the presentation of the main results and findings of the project DIADEMS.

[10] This work on vocal categories has been realized during several meetings by the following researchers of the project: Joséphine Simonnot, Aude Julien-Da-Cruz-Lima, Jean Lambert, Estelle Amy de la Bretèque (CREM-LESC), Susanne Fürniss, Sylvie Le Bomin, Marie-France Mifune (MNHN), Valentina Vapnarsky, Aurore Monod Becquelin, Marie Chosson (EREA-LESC), Boris Doval (LAM) and David Doukhan (LIMSI).