Quand la préservation passe par la classification : le cas des documents sonores et musicaux

Bouchra Lamrini (1)
Bouchra.Lamrini@ircam.fr

Francis Rousseaux (2)
Francis.Rousseaux@ircam.fr

Raffaele Ciavarella (3)
raffaele.ciavarella@ircam.fr

Alain Bonardi (4)
Alain.Bonardi@ircam.fr

Jérôme Barthelemy (5)
Jerome.barthelemy@ircam.fr

Institut de Recherche et Coordination Acoustique/Musique, Centre Georges Pompidou, 1, place Igor Stravinsky, 75004 Paris (1, 2, 3, 4, 5)

Mots-clés : Classification, data-mining, document numérique, musique contemporaine, préservation, processus numérique temps réel, Traitement Automatique des Langues.

Résumé : La création artistique contemporaine fait aujourd'hui largement appel aux technologies électroacoustiques et numériques. Dans la création musicale spécialement, les dispositifs et les outils logiciels permettant de manipuler les sons en temps réel sont apparus voici une trentaine d'années, et notamment les « patchs » processus numériques temps réel utilisés lors de performances ou de concerts en live. Soumis aux difficultés de la préservation, ces modules logiciels de traitement sonore sont souvent considérés comme des véritables documents numériques, ils sont à la fois supports de création et supports de constitution de connaissances dans la création artistique contemporaine. Pour soutenir les échanges et la construction d'une interprétation collective autour de ce document, nous proposons dans cet article une approche d'analyse et de classification, par les techniques du data-mining, de ces processus numériques afin de former une ontologie du domaine voire une organologie des traitements musicaux et audio numériques.

Introduction

La musique est traditionnellement considérée comme l'art consistant à arranger et ordonner sons et silences au cours du temps et parfois dans l'espace (le musicien sur scène ou dans le public) selon 4 critères : le rythme (c.-à-d. la durée des sons dans l'espace temporel) qui est le support de cette combinaison dans le temps, la hauteur celle de la combinaison dans les fréquences (son grave ou aigu), le timbre (la nature du son qui dépend du spectre produit par la source sonore, l'intensité du son (ce que l'on appelle les effets dynamiques en analyse musicale : piano, crescendo...). Le patch s'appuie sur un ensemble de catégories relevant principalement des critères et des formalismes mathématiques issus des sciences de traitement du signal. Ils permettent donc la réalisation d'intentions musicales particulières, hors des schémas traditionnels, et traduisent un ensemble de savoirs et de savoir-faire relatifs à ces intentions, et à la création musicale contemporaine dans son ensemble. Le patch est donc le résultat d'une activité créatrice, et un élément indispensable de la performance et de l'œuvre musicale.

Dans la perspective d'une pérennisation à long terme de la création musicale, le problème se pose donc de savoir préserver les processus temps réel. Pour répondre à cette question, le projet ASTREE[1] (*Analyse/Synthèse de Processus Temps Réel*), dans lequel s'inscrit ce travail, s'appuie sur une stratégie de développement d'outils permettant de transcrire, documenter, expliciter les processus existants afin d'améliorer leur pérennité, et les rendre indépendants de l'environnement technique sous-jacent. Dans ce contexte, notre contribution consiste

[1] http://www.ircam.fr/sel.html?tx_ircam_pi1[showUid]=46&ext=1

notamment à développer une méthodologie d'analyse et de classification par des techniques de data-mining, afin de constituer des bases de connaissances ou des ontologies du domaine en jeu. Le passage progressif des documents existants à l'ontologie prônée par cette méthodologie sera réalisé par des traitements qui relèvent successivement de la terminologie, de la modélisation des connaissances et enfin de la représentation des connaissances. Nous étendons ces traitements en prenant en compte la structure des documents et des éléments porteurs d'information. Pour cela, nous nous basons sur des documents où les traitements et objets sont transcrits au moyen du langage fonctionnel FAUST[2] (*Functional Audio STreams*). Le découpage structurel de chaque document correspond à une caractérisation sémantique de son contenu. L'idée est de tirer profit de la structure et de la hiérarchisation du document pour souligner la complémentarité entre identification de concepts (objets recherchés) et extraction de relations. Les différentes réflexions exposées pour l'analyse et la classification de ces documents, font appel à des concepts et des techniques provenant de plusieurs domaines, parmi lesquels figurent la Représentation de Connaissances (*RC)*, le Traitement Automatique du Langage Naturel (*TALN*), et l'Intelligence Artificielle (*IA*). Il s'agit d'une recherche de nature pluridisciplinaire dont l'objectif général est de mettre en relation un certain nombre d'hypothèses théoriques, d'explorer plusieurs concepts et d'appliquer des techniques provenant de différentes disciplines afin de proposer une méthodologie pour construire et ensuite maintenir les ontologies obtenues, et en particulier, la découverte de relations entre objets pertinents à l'ontologie. Ce projet de recherche amène une nouvelle réflexion sur les divers paliers à envisager dans une telle démarche de modélisation de connaissances textuelles pour des objectifs de formation d'ontologie du domaine et préservation de l'œuvre interactive et de son exécution. Certes, pour classifier les éléments matériels et logiciels de la musique électronique interactive, une première possibilité est d'envisager par exemple la famille des classes telles que : instruments électriques, synthétiseurs analogiques, synthétiseurs numériques, modules d'effets, synthétiseurs virtuels, et logiciels temps réel. Cette classification statique est facile à établir. Elle est uniquement fondée sur la nature technique des dispositifs utilisés et ne prend pas en compte leurs fonctions musicales. Le cadre de notre travail répondra donc à cette problématique en proposant d'établir des extractions et des classifications dynamiques à partir de l'extraction de descriptions numériques de ces dispositifs. L'association de ces deux classifications, complétée par le savoir faire de l'expert du domaine, permettra ainsi d'envisager des classifications 'homme-machine' dont le but final est de faire émerger une organologie des traitements sonores temps réel.

L'article est organisé comme suit. Nous faisons tout d'abord une brève introduction sur le patch et les défis relatifs à sa documentation et sa préservation à long terme dans la création artistique contemporaine. Nous exposons ensuite notre démarche envisagée en termes d'analyse et de classification, pour construire l'ontologie du domaine en question. Enfin, nous discutons les travaux en cours et les perspectives de la méthodologie présentée.

1 Le patch

1.1 Création et représentation de connaissances

L'apparition du patch date de la fin des années 1980, après deux décennies qui avaient misé sur des solutions hardware propriétaires. De nombreux chercheurs se sont investis pour trouver des solutions permettant d'enrichir les musiques mixtes, associant musiciens humains et sons électroniques. Dans ce contexte, le patch marque le début des approches interactives dans lesquelles la machine est en attente d'informations, acquises par des capteurs, venant de l'artiste, musicien ou interprète. L'une des premières pièces du répertoire IRCAM (*Institut de Recherche et Coordination Acoustique/Musique*) à exploiter ces possibilités était *Jupiter* de Philippe Manoury (1987) d'après [1], associant une flûte (dotée de capteurs et d'un microphone pour obtenir un certain nombre de paramètres du jeu du flûtiste) à un dispositif d'informatique musicale temps réel. Miller Puckette a crée ensuite en 1988 un premier système *Patcher,* dont sont dérivés les deux logiciels les plus utilisés aujourd'hui : *Max/MSP/Jitter*[3] et *PureData*[4].

Les patchs reprennent certains paradigmes des dispositifs électroacoustiques utilisés sur scène depuis plus d'une cinquantaine d'années comme les réverbérations, les processeurs d'effets, etc. À titre d'exemple [Fig.1], nous présentons ci-dessous un module créé avec le logiciel Max/MSP, produisant deux sinusoïdes, à 100 et 200 Hz, et les additionnant, l'amplitude du résultat étant atténuée (multipliée par 0.25). L'aspect du patch fait penser à un montage électrique, comme ceux élaborés sur la paillasse d'un laboratoire de physique, avec des appareils qui sont les modules rectangulaires (exemple de l'objet cycle~ qui produit une sinusoïde, à l'instar d'un générateur) et des câbles, qui sont ici les connections entre objets. Il est extrêmement facile de réaliser ces patchs, sans

[2] http://faust.grame.fr/
[3] http://www.cycling74.com
[4] http://crca.ucsd.edu/~msp/software.html

programmation, simplement en incorporant en quelques clics de souris des objets issus de menus, et en les reliant par des connections.

Ce type de dispositif est largement utilisé [1, 2, 3, 4], dans les arts de la performance tels que la musique, la danse, le théâtre, la vidéo, et les œuvres associant plusieurs arts [Tab.1], pour deux raisons : 1) il fonctionne en temps réel en fournissant une réponse quasi immédiate aux entrées proposées (le délai de calcul des sorties étant considéré comme négligeable par rapport à l'ordre de grandeur temporel des entrées); 2) il permet une construction simple des traitements en s'appuyant sur une représentation graphique. Du point de vue de l'écriture, aussi bien informatique que musicale, le patch est un dispositif interactif offrant à l'utilisateur un mode de travail et de production facile à gérer et maîtriser sans faire appel à la programmation usuelle et aux catégories de la musicologie. Il n'y a plus de code à écrire, modifier et compiler, l'utilisateur adapte les objets utilisés, les connexions et les paramètres des patchs jusqu'à ce qu'il obtienne un résultat satisfaisant. Le mode de travail et de production se fait donc par retouche du patch. En revanche, la démarche créative du compositeur devient dépendante du système et de l'environnement technique utilisé.

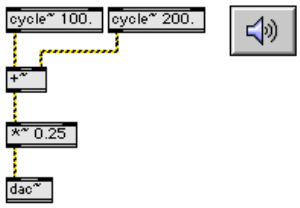

Figure 1. Exemple de patch produisant la somme de deux sinusoïdes.

Tableau 1. Exemples d'utilisation de patchs dans les arts de la performance.

Œuvre	Configuration
Anthèmes II de Pierre Boulez, pour violon et électronique temps réel (1997).	Instrument solo et électronique temps réel
L'écarlate, performance de danse conçue par Myriam Gourfink, chorégraphe, sur une musique de Kasper Toeplitz (2001)	Performance danse et musique
K, musique et texte de Philippe Manoury (2001)	Opéra avec transformations sonores en temps réel

1.2 Problématique et défis de la documentation du patch

Il convient tout d'abord de noter que les actions de préservation concernant les objets numériques mis en œuvre dans une création artistique doivent examiner la nature de l'œuvre originale, et les relations de l'objet numérique avec l'œuvre, avant d'évaluer les actions à effectuer pour le préserver. Il s'agit bien de préserver l'ensemble des connaissances permettant dans le futur d'apprécier le contexte dans lequel l'objet numérique est mis en œuvre : les informations portant sur l'environnement technique numérique, mais aussi l'environnement immédiat (par exemple, les caractéristiques de la salle de concert), les périphériques utilisés (microphone, hauts-parleurs, etc.). On devra aussi tenir compte d'une caractéristique de la démarche artistique, qui est une tendance à repousser les limites d'une technologie donnée, afin de la transcender pour obtenir de nouveaux effets. Il s'agit ici de préserver les intentions qui sous tendent la démarche créatrice. Une approche ou un modèle global et structuré de préservation devra impérativement tenir compte de ces considérations.

Le patch lui-même, document numérique porteur d'informations sur la conception musicale et technique d'une œuvre musicale, est confronté à diverses difficultés (fragilité des supports de stockage, obsolescence des standards techniques, l'instabilité des patchs vis-à-vis de leur complexité, etc.) qui compliquent la tâche de préservation. Le processus de documentation comporte généralement trois activités : la recherche (repérer et identifier les données pertinentes), la préservation (pérenniser les données) et la diffusion (rendre ces données disponibles et les transférer en connaissances utilisables). Une pratique documentaire rigoureuse est d'autant plus essentielle que les sources d'information sur la création contemporaine sont non structurées, et notamment pour les œuvres elles-mêmes. Compte tenu des défis communs aux domaines de la préservation des documents et de l'art numérique, la recherche dans ces documents est bien entendu pertinente. En matière de documents numériques, Margaret Hedstrom [5] fait une recommandation intéressante : « Préserver le contenu, le contexte et la structure tout en préservant la capacité d'afficher, de lier et de manipuler les objets numériques ». La plus grande difficulté réside probablement dans le second aspect, car cela suppose de garder l'accessibilité à une multitude de logiciels et de systèmes d'exploitation. Hedstrom ajoute également : « La préservation de matériel

numérique nécessite souvent la transformation complexe et coûteuse d'objets numériques, pour qu'ils demeurent des représentations authentiques de la version originale ainsi que des sources utiles aux fins d'analyse et de recherche ». Le paradoxe de la préservation de documents numériques réside entre les expressions « transformations » et « représentations authentiques de l'original ». Les spécialistes proposent de plus en plus le concept d'émulation [6] et d'enveloppe contextuelle parmi les stratégies d'archivage numérique. Ce concept est une solution potentielle au problème des documents numériques dépendant des logiciels et, ce qui est plus important, du matériel requis pour y avoir accès. Ce concept permet également de conserver l'entière capacité de traitement des données.

Cette approche suppose de placer le document, conservé dans sa forme d'origine, dans une enveloppe virtuelle contenant toutes les instructions nécessaires pour sa récupération, son affichage et son traitement. Les instructions expliquent comment lier le document à un ensemble d'émulateurs qui servent de passerelle entre le document, qui peut demeurer stable, et le contexte technologique en constante évolution. Ainsi, plutôt que de tenter de modifier une multitude de documents, les gestionnaires d'archives ou de collections numériques n'ont qu'à mettre à jour les émulateurs. Une fiche technique renfermant d'importantes spécifications peut être ajoutée au document sous forme de métadonnées. Cela facilite le repérage de documents qui nécessiteront une certaine intervention en vue de leur conservation future. Grâce aux métadonnées, la description du document peut être incluse dans le document même, faisant de ce dernier sa propre fiche de catalogue. Il existe cependant un inconvénient : les métadonnées peuvent être aussi imposantes sinon plus que le document décrit.

Les patchs, comme document numérique de représentation de connaissances, sont source d'intérêt pour la communauté des chercheurs en musicologie et en informatique musicale. Nous citons dans ce cadre l'exemple du projet de collaboration 'Mustica' entre l'IRCAM et l'INA [7] sur la production d'une base de données ouverte aux organismes souhaitant remonter des œuvres contemporaines. En matière d'œuvre, nous citons par ailleurs, quelques travaux [8, 9, 10] (cités dans [1, 3]) menés pour la maintenance des patchs selon quatre actions : préservation, émulation, migration et virtualisation. L'article de Bullock et Coccioli [11] présente l'exemple de l'œuvre *Madonna of winter and spring* et la tentative d'émulation d'un modèle Yamaha TX816 (pour la synthèse sonore), qui n'existe plus, grâce à des patchs créés sous PureData. A ces exemples s'ajoutent les travaux de [11, 12] dont les auteurs recommandent de sauvegarder les patchs réalisé par MaxMSP au format texte plutôt qu'en codage binaire. Enfin, nous citons le travail d'Andrew Gerzso d'après [2] qui a été réalisé au sien de l'IRCAM sur la recherche des représentations indépendantes d'une implémentation technique, dans le cadre de l'œuvre *Anthèmes II* de Pierre Boulez.

2 Approche pour la préservation à long terme des processus numériques temps réel

2.1 Cadre du travail cadre

Le travail de recherche que nous présentons s'intègre dans une problématique qui considère le document numérique comme porteur de connaissances et doté d'une intentionnalité qui le construit ou le reconstruit pour une préservation à long terme. Dans cette vision du document comme signe, nous portons plus particulièrement notre intérêt sur l'analyse et la classification d'un document, c'est-à-dire sur les objets qui le constituent. Pour illustrer ce besoin de préservation du document, le projet ASTREE dans lequel s'inscrit notre travail, a permis une réflexion nouvelle sur la documentation et la préservation des processus temps réel vis-à-vis de l'histoire d'une création et l'exécution d'une œuvre de l'art de la performance. Cette réflexion consiste à transcrire les processus existants, conçus dans les environnements tels que Max/MSP ou PureData, dans le langage fonctionnel FAUST [13, 14]. Cette réflexion peut être illustrée par les trois points suivants [Fig.2] :

1. Génération automatique de documentation sous une forme indépendante de toute technologie (hardware et software). Cette documentation permettra une réimplémentation manuelle complète du processus originel.
2. Application des techniques de data-mining sur les expressions algébriques issues de FAUST et de la documentation afin de constituer des connaissances sur les processus en question.
3. Génération d'un code optimisé et indépendant des librairies et des systèmes de traitement de signal.

Au moyen d'un concept familier de bloc-diagramme, FAUST permet de décrire facilement des processeurs de signaux qui traitent et transforment des signaux en entrée pour produire des signaux en sortie. L'utilisateur dispose de modules de traitements élémentaires qu'il combine de différentes manières pour obtenir le traitement souhaité. Le compilateur FAUST utilise ensuite cette description pour écrire automatiquement un programme C++ équivalent. Des techniques d'optimisation particulières permettent d'engendrer un code C++ de qualité dont l'efficacité est généralement comparable à celle d'un programme écrit à la main. La figure 3 présente un exemple

de la fonction ‘Envelop’. Cette fonction permet de générer un bruit blanc contrôlé par une enveloppe. Plusieurs paramètres sont utilisés pour ajuster la forme de l'enveloppe, la longueur d'attaque, la durée de décadence, le pourcentage de volume soutenu, et la durée de sortie.

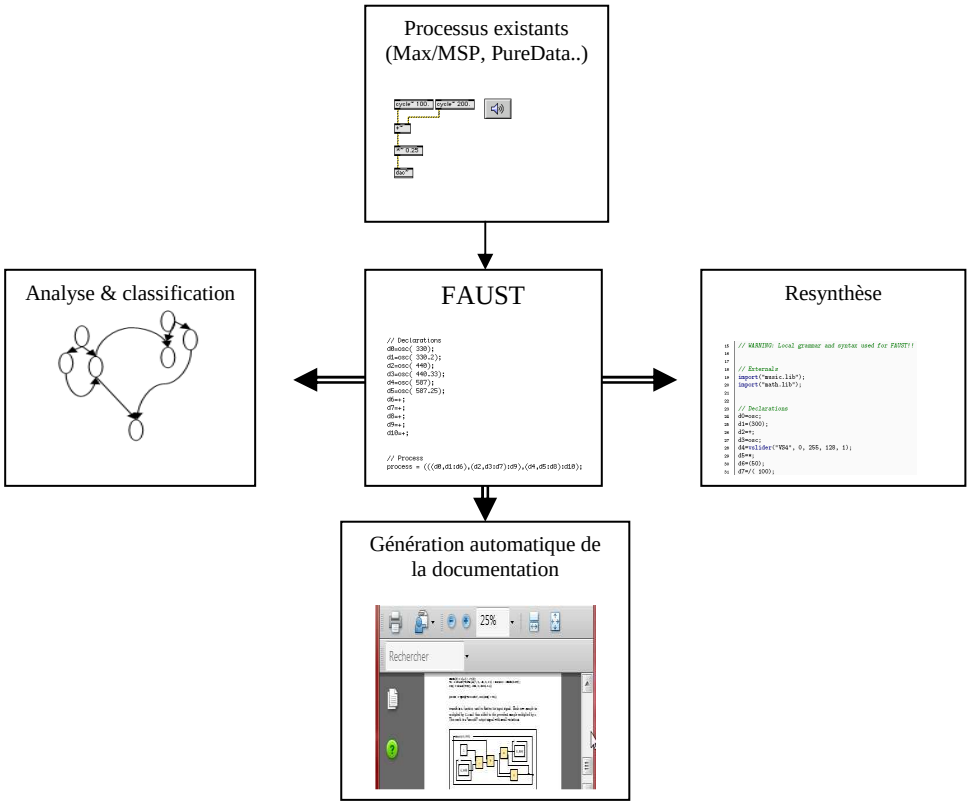

Figure 2. Processus de la transcription des objets existants dans le langage FAUST.

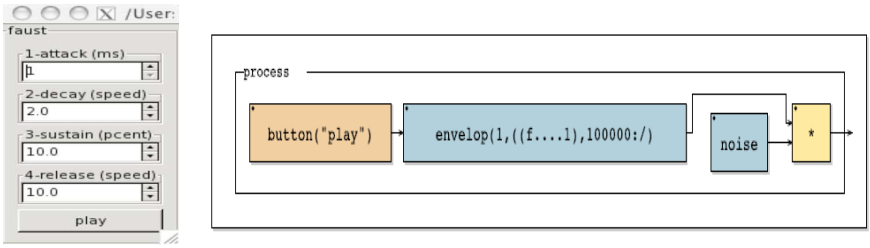

Figure 3. Schéma de représentation du contrôle par une enveloppe d’un bruit blanc au moyen de FAUST.

2.2 Classification des documents numériques et formation d’ontologie

En intelligence artificielle, les ontologies sont apparues comme une réponse aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques. L'importance que revêt aujourd'hui l'usage des ontologies pour le développement de systèmes à base de connaissances n'est plus à démontrer. Les chercheurs du Web ont adopté ce terme ontologie pour référer à un document (ou fichier) définissant d'une façon formelle les relations entre termes [15]. Dans le cadre du Web sémantique, les ontologies sont utilisées comme noyau du système pour accéder à des informations structurées ainsi qu'à des règles d'inférence supportant le raisonnement automatique. Ces ontologies offrent également la possibilité, pour un programme, de retrouver les différents termes désignant un même concept. Il s'agit pour ce cas spécifique, d'ontologies de type domaine. La génération de l’ontologie et l’extraction des données utilisent plusieurs sources de connaissances telles que les ontologies déjà formées par des concepts arrangés hiérarchiquement avec un haut degré de connectivité. Ces ontologies sont fondées sur une conception de nature générale suivant un héritage des propriétés [16] ; des sources de connaissances lexicales, comme la base de données lexicales Wordnet [17] ; des répertoires des expressions régulières conçues pour convenir des items lexicaux structurés (dates, n° tel, mesures...) et qui peuvent fournir par héritage des informations appropriée aux concepts crées [18] ; et des documents d'apprentissage qui contiennent le contenu textuel du domaine spécifique pour l'intérêt de l'utilisateur (langage spécifique du texte).

Comme nous l'avons mentionné auparavant, cet article propose une nouvelle réflexion de mise au point d'une méthode de classification permettant de former de manière automatique une ontologie conceptuelle du domaine musical en termes de traitements musicaux et audio numériques. De manière générale, la définition d'une méthode de classification conceptuelle repose sur deux éléments : la définition d'une distance permettant de comparer les objets à classer, et d'autre part la définition d'un algorithme de classification qui construit la structure arborescente proprement dite. Dans le domaine de la formation des ontologies conceptuelles, la notion de distance entre mots a été étudiée en Traitement Automatique de la Langue (TAL) pour former les classes sémantiques qui forment les nœuds de ces catégories suivant une approche ascendante ou descendante, tandis que les algorithmes de classification ont été plus largement étudiés en analyse de données et en apprentissage automatique. Dans ce contexte, de nombreux outils destinés à l'acquisition automatique/ou semi-automatique de classes sémantiques visant à regrouper de termes proches, sont élaborés. De point de vue sémantique, cette notion de proximité est généralement fondée sur des mesures de distance entre termes en fonction du degré de ressemblance des contextes dans lesquels ils apparaissent. Les descriptions et les régularités recherchées des contextes des termes dépendent de l'approche menée. Ainsi, les contextes peuvent être purement graphiques, i.e. sous forment de co-occurrences dans une fenêtre de mots [19, 20, 21]. Le choix de la distance appropriée pour un corpus est donc un problème posé et peu étudié [22]. La majorité des travaux ne s'intéressent qu'à la seule évaluation de la tâche finale pour laquelle l'apprentissage est effectué. Les critères de cette évaluation restent quantitatifs et donnent rarement lieu à des études comparatives [23]. En revanche, la caractérisation des résultats des méthodes fournit les outils d'aide au choix de la plus pertinente ou la mise au point de nouvelles méthodologies. Il en est de même pour les algorithmes de classification. Elaborer une méthodologie ou un outil permettant de mettre au point un algorithme de classification conceptuelle pour former des ontologies en fonction de la tâche est une tâche ardue. Par ailleurs, les travaux effectués en classification conceptuelle [24, 25] n'ont pas trouvé de propos en apprentissage à partir de corpus. Certes, il faut souligner que la construction d'une ontologie dans le domaine du TAL exige une phase de choix et d'adaptation des algorithmes existants aux problèmes posés par le corpus autour de la distance, voire développer de nouveaux outils et moyens méthodologiques. Soulignons par ceci et à présent l'importance qu'on doit accorder au choix des outils et des techniques pour l'extraction et à la sélection des objets (ici les indicateurs des patchs) contenus dans l'ensemble de nos documents transcrits par Faust.

L'ensemble des étapes de notre démarche proposée est élaboré à partir d'une étude bibliographique sur les travaux en TAL (cités auparavant), notamment la formation de classes sémantiques à partir de corpus et à partir de travaux en apprentissage sur la classification conceptuelle et l'analyse des données. Nous avons donc noté que pour effectuer les diverses tâches de classification, recherche et filtrage de documents, il faut d'abord représenter les textes de manière à la fois économique et significative. On sait que le modèle vectoriel est l'approche la plus courante dans lequel le texte est représenté par un vecteur numérique obtenu en comptant les éléments lexicaux les plus pertinents. Ces vecteurs sont fournis par des prétraitements simples. On commence généralement par éliminer les mots grammaticaux (articles, prépositions, etc.) et par réduire les variantes morphologiques à une forme commune (souvent appelée terme). Puis on compte les occurrences des termes les plus importants de manière à représenter chaque document par un vecteur dans l'espace des termes. Un corpus de documents génère donc une matrice *Document-Terme* [Fig.4] qui permet ensuite d'appliquer les opérations vectorielles usuelles avec des résultats sémantiquement pertinents dans l'ensemble.

$$\begin{array}{c} \\ Doc_1 \\ Doc_2 \\ \vdots \\ Doc_m \end{array} \begin{array}{c} \begin{array}{cccc} Terme_1 & Terme_2 & \dots & Terme_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \end{array}$$

Figure 4. Matrice Document-Terme (x_{ij}= fréquence d'apparition du terme i dans le document j).

Cependant ce type d'approche produit généralement des vecteurs lexicaux de très grande dimension qui sont coûteux à stocker et à traiter. Des vecteurs qui sont partiellement ou entièrement vides (contenant généralement plus de 90% de valeurs nulles), ayant en plus des termes fortement corrélés entre eux. La détection d'une telle synonymie entre termes a des conséquences négatives pour l'indexation et la recherche. Evidement des documents voisins sémantiquement peuvent très bien ne pas contenir les mêmes termes. Détecter les relations entre termes permettra d'améliorer la recherche de documents. Une représentation à partir de ces vecteurs redondants sera donc difficilement lisible par un utilisateur qui voudrait s'en servir pour évaluer rapidement le contenu d'un document et chercher à voir les relations entre divers documents. Devant ce manque, il serait important de trouver la dimension intrinsèque du domaine, c'est-à-dire la dimension minimale permettant de représenter les données sans perte d'information. Pour ce type de traitement, Il y a une gamme de méthodes permettant de calculer un nombre réduit de dimensions pour un ensemble de données, nous citons la méthode Latent Semantic Analysis *(LSA)* qui n'utilise pas la matrice de covariance, mais extrait de nouveaux axes

directement de la matrice document-terme [26], l'Analyse Factorielle, dont la méthode la plus connue est l'Analyse en Composantes Principales (*ACP*), mais elle ne permet pas en pratique de traiter des vecteurs de très grande dimension comme dans le domaine des documents. En revanche, la plus prometteuse est la version neuronale de l'ACP, appelée Algorithme Hebbien Généralisé (*AGH*), permettant de traiter de tels vecteurs avec de bons résultats.

L'objectif de cette étude bibliographique était d'esquisser un panorama des méthodes de data-mining et de text-mining, notamment les méthodes de catégorisation et de classification non supervisée en se plaçant dans un cadre méthodologique pour identifier les indicateurs, liés par exemple à la représentation des documents, à la sélection des termes décrivant le contenu des documents et aux algorithmes de catégorisation et de clustering, à prendre en compte pour employer l'une plutôt que l'autre et savoir comment les évaluer sur notre problème de classification de ce type de documents numériques.

2.3 Présentation du corpus

Les objets pris en compte dans nos travaux sont des objets issus des expressions algébriques contenues dans la documentation transcrite par le langage fonctionnel FAUST. Pour illustrer la nature du problème à résoudre, nous avons choisi un exemple de documentation sur un filtre passe band 'BPF' comme un processus temps réel conçu dans Max/MSP et transcrit dans le langage FAUST. La figure 5 présente un zoom sur les types de structures contenues dans le corpus en question. Les concepts que l'on cherche à repérer dans la documentation se trouvent dans les codes Faust, la documentation générée en latex, les trois librairies importées (maxmsp, music, et math) et les commentaires inclus dans les codes et les librairies.

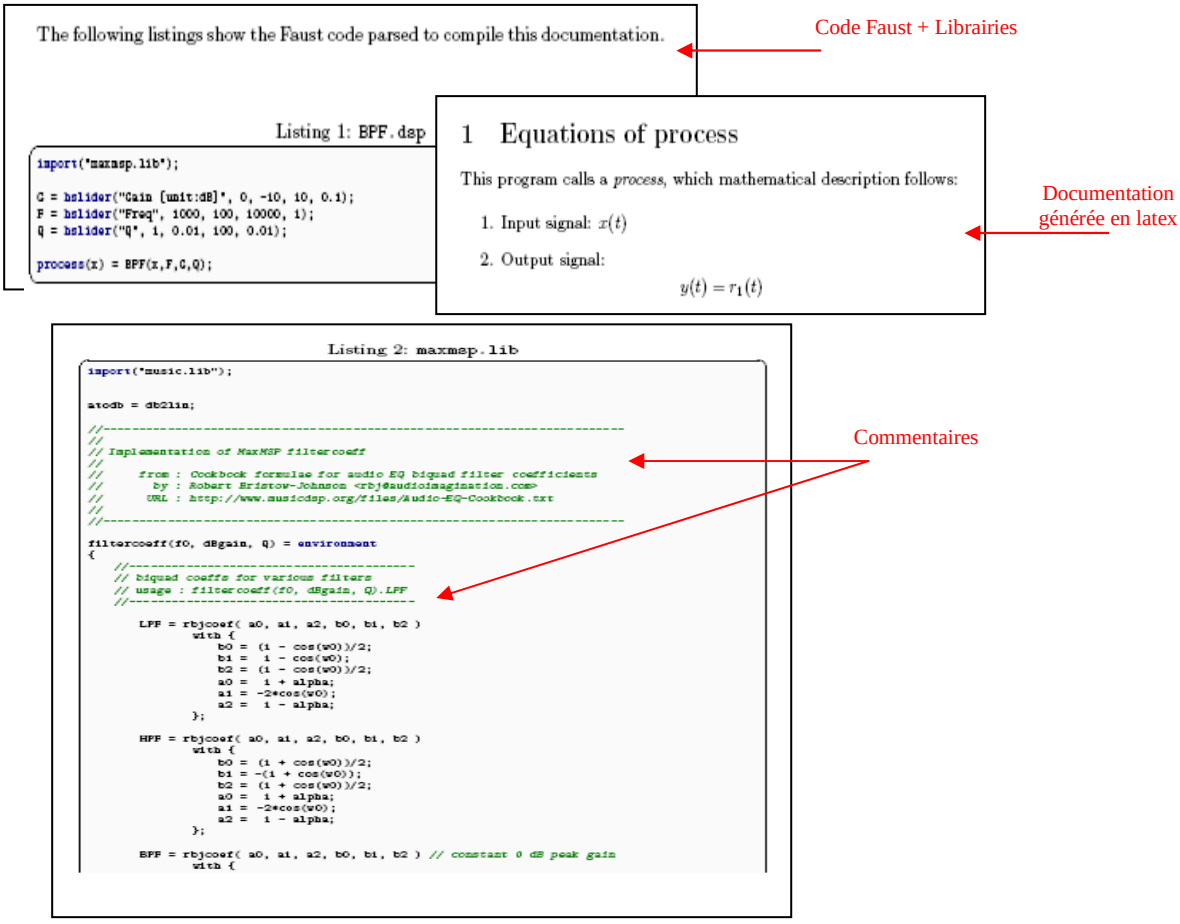

Figure 5. Structures contenues dans le corpus à analyser

Si nous nous intéressons par exemple seulement aux objets provenant de Max/MSP, les trois fragments du document permettent de fournir une information [Fig. 6] sur les objets clés à savoir :

- le filtre passe-bande 'BPF' présent dans la librairie ''maxmsp.lib''. Ceci est donné par l'expression décrivant le processus générale ''process (x) = BPF(x, F, G, Q)''.

- présence d'une entrée x(t), ceci est indiqué dans le texte descriptif et par l'expression algébrique du processus.
- les coefficients du filtre, ici le gain 'G', la fréquence 'F' et le facteur de qualité 'Q'.
- les expressions fonctionnelles qui lient le processus aux coefficients du filtre comme la présence de la fonction de transfert ''biquad''. La présence de cette fonction entraine également la présence d'autres paramètres de conversion et de fonctions de calcul comme les fonctions trigonométriques.
- et également un domaine de variation de chaque coefficient approprié au filtre en question (Exp : *''Freq'', 1000, 100, 10000, 1*).

Les processus dans leur expressions algébriques peuvent être vus comme des objets clés qui génèrent une sorte de connaissances appropriées aux termes recherchés comme 'Filter, BPF, APF, Oscillator, Osc…', et les termes 'Q, G, F,…' comme des attributs de ces termes recherchés. En s'appuyant sur les différents documents transcrits et normalisés à présent, nous avons commencé à constituer une base de connaissances sur les objets clés, les termes associées ainsi que d'autres données permettant de fournir une information sur la proximité, la corrélation des objets, la proximité des classes objets, la distribution des objets autour d'un objet de la même classe (chaque objet a une distance proportionnelle à son degré d'association. En classification conceptuelle ascendante par exemple (car c'est la plus recommandée pour des raisons de complexité algorithmique et 'd'explicabilité' pour la validation par l'utilisateur en cours d'apprentissage) l'algorithme général consiste à réunir successivement par paires, les objets proches, ou les objets et les classes objets proches, afin de former des hiérarchies ou des graphes de classes d'objets.

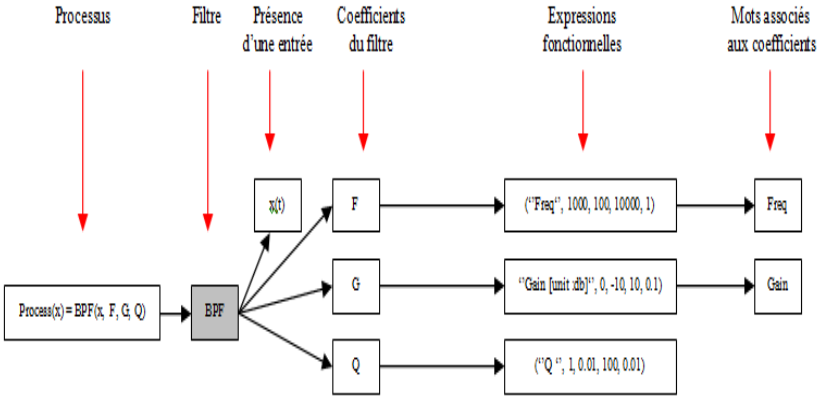

Figure 6. Digramme illustrant les différentes termes clés qui peuvent fournir une connaissance sur le filtre passe-bande 'BPF : Band Pass Filter '

2.4 Méthodologie proposée pour l'analyse et la classification des patchs

Sur notre problème d'analyse et de classification des patchs, un système d'extraction d'information classique seul n'est pas capable d'analyser les différents segments du corpus en question, notamment sa structure qui s'appuient en général sur des expressions algébriques, combinées d'une certaine manière à décrire les entrées et les sorties des processus réalisés d'une part, et les fonctions de transformation effectuée entre ces entrées et ces sorties d'autre part. En effet, pour analyser une phrase, un système d'extraction d'information effectue successivement une analyse lexicale (segmentation de la phrase en chaines de caractères qui représentent des mots), une analyse morpho-syntaxique (étiquetage des mots par leur catégorie syntaxique et association de chaque mot à sa forme canonique..), une analyse syntaxique (analyse de la structure de la phrase), et en fin une analyse sémantique (compréhension du sens des mots et des relations entre les mots). Or, notre document numérique est le résultat d'une transcription à base d'un langage fonctionnel. Sous sa forme observable, ce document peut être considéré comme une suite de mots ordonnés selon des règles qui ne reflètent pas forcément l'ordre dans lequel les mots s'appliquent les un aux autres pour former l'interprétation sémantique fonctionnelle. Plusieurs voies que nous avons citées auparavant donc peuvent être explorées pour résoudre ce type de problème de classification, et par ailleurs celle de la grammaire catégorielle combinatoire applicative [27].

Dans le cadre de ce travail, notre méthodologie envisagée d'analyse et de classification s'appuie sur deux pistes complémentaires, qui s'emboîtent pour donner une classification finale des objets en question, et notamment jugée par l'expert du domaine [Fig.7].

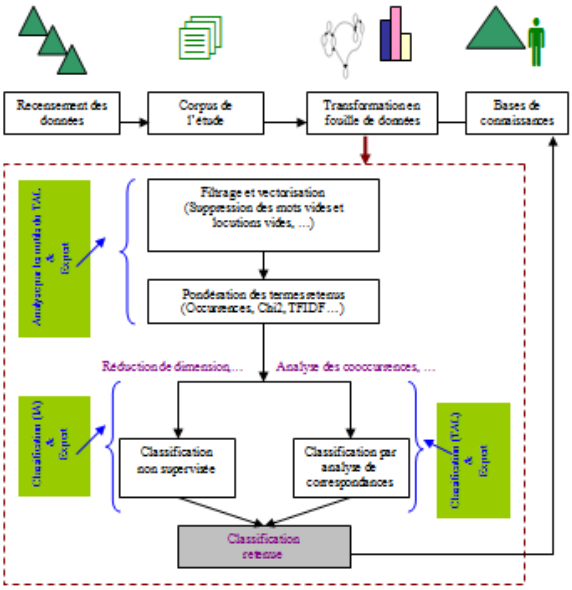

Figure 7. Etapes de la méthodologie proposée pour la classification des patchs.

Notre idée dans un premier temps est de réaliser une série de prétraitements classiques au moyen des outils du TAL afin d'éliminer les objets non-clés pour notre étude et fixer le type de pondération qu'on doit affecter aux objets retenus pour la classification. Cette phase de prétraitement est guidée notamment par l'expert du domaine. L'intervention de l'expert permettra de valider chaque étape avant la validation finale du la classification attendue par notre étude. Ensuite, on a deux voies à parcourir pour avoir une classification satisfaisante selon nos objectifs fixés au début de l'étude. La première voie est de poursuivre les analyses par les mêmes outils du TAL tels que : les analyses des cooccurrences d'objets-clés, par exemple faire des comparaisons entre les paires d'objets-clés de la famille Filter : 'APF-BPF' et 'APF-HPF', et également l'analyse d'autres objets associés à ces objets-clés ; les analyses thématiques des unités de contexte, autrement dit opérer et modeler les objets qui émergent des différents unités du corpus et qui sont décrits à travers leur vocabulaire caractéristique, c'est-à-dire à travers des ensembles de mots-clés (lemmes ou catégories) co-occurrents. Ces objets émergents peuvent être utilisés pour obtenir de nouvelles variables qui peuvent être utilisées dans des analyses ultérieures. Les analyses comparatives des sous-ensembles du corpus qui donnera lieu à une classification par une analyse des correspondances. Comme toutes les techniques factorielles, les analyses comparatives permettent l'extraction de facteurs qui ont la propriété de récapituler d'une façon organisée l'information significative contenue dans les innombrables cellules des tableaux de données. En outre, cette technique d'analyse permet la représentation graphique, dans un ou plusieurs espaces, des points qui détectent les objets en lignes et colonnes, et qui dans notre cas sont les entités linguistiques (mots, lemmes, segments de texte, textes, etc.). Les résultats d'analyses permettront d'évaluer des rapports de proximité/distance - ou de similitude/différence - entre les objets considérés. La figure 8 illustre un schéma approximatif de ce que nous envisageons de réaliser par la démarche proposée dans ce travail après une pré-étude du corpus transcrits par FAUST ; elle présente un exemple de la famille 'Filter'. La figure 9 présente les distances qui peuvent être analysées entre les classes pour relever et mettre en valeur les différentes corrélations traduisant une certaine connaissance cachée ou mal connue en termes de fonctionnement des patchs comme des processus temps réel.

Pour la deuxième voie, nous envisageons de réduire la dimension de corpus par l'application de l'une des techniques d'analyse factorielle citées auparavant, la plus adaptée pour notre cas est la version neuronale de *l'ACP*, appelée *AGH* (citée dans § 2.2). Cela permettra de projeter les documents dans un espace beaucoup plus compact avec des concepts plus significatifs. En effet ces nouvelles dimensions représentent essentiellement des corrélations entre termes dans un corpus donné, révélant de la sorte les thèmes principaux du corpus presque aussi bien qu'une classification des documents. Ensuite, nous utilisons les vecteurs dans le nouvel espace (donnés par les valeurs de sortie du réseau AHG) pour effectuer une classification non supervisée des documents. Les algorithmes les plus connus sont soit l'algorithme des centres mobiles (k-means) qui emploie une distance euclidienne pour ne pas avoir à normaliser les vecteurs, soit les réseaux de neurones compétitifs par la règle du Kohonen comme les cartes de SOM [28]. Nous citons ici les travaux de Nguyen et Zreik [29, 30, 31] sur le développement du système *Hyperling* pour reconnaitre les langues dominantes dans un site Web multilingue et dont les deux algorithmes K-means et SOM ont été implémentés et testés avec succès.

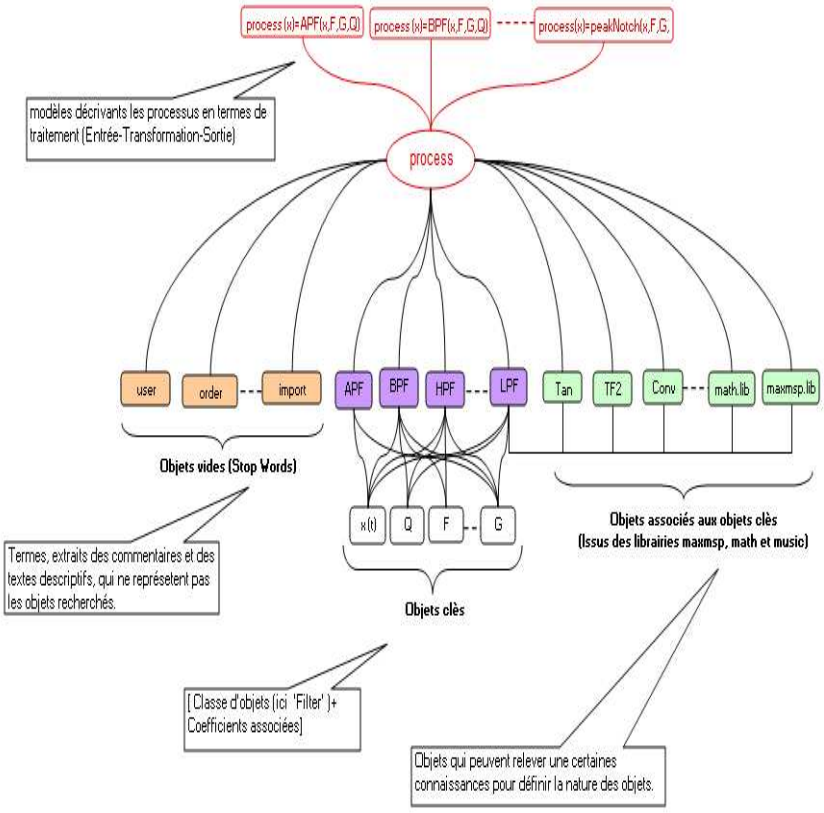

Figure 8. Schéma approximatif illustrant l'un des types d'information constituées dans un corpus décrivant la famille 'Filter'.

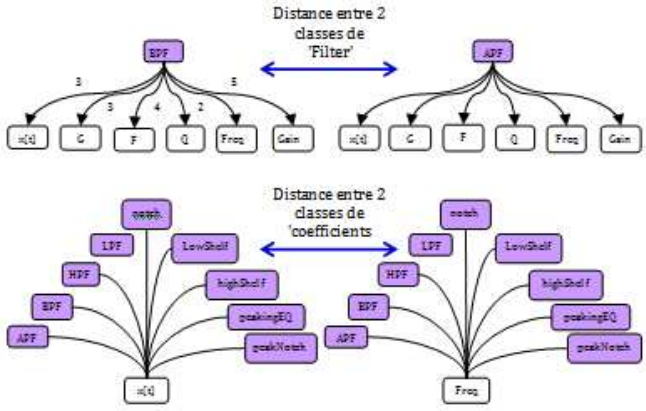

Figure 9. Schéma illustrant les types de distance entre les classes des objets.

3 Conclusion et perspectives

En génie documentaire, les sources documentaires s'accroissent et les applications issues du Traitement Automatique du Langage et deviennent applicables à une large variété de problèmes dans ce contexte. De nombreux travaux ont été proposés pour construire de façon plus ou moins automatique des ontologies de type hiérarchies conceptuelles à partir de corpus analysés. Profitant ainsi des moyens qu'offrent les outils du TAL à savoir : l'extraction d'information, l'indexation de documents, la désambiguïsation syntaxique, etc. A l'opposé, il manque les outils et les méthodologies qui permettent d'évaluer et comparer les points faibles et forts des différentes approches pour un corpus spécifique à un domaine et une tâche donnée. Le travail que nous réalisons dans le cadre du projet ANR ASTREE, permet dans ce sens de proposer une nouvelle réflexion qui permet d'associer les techniques du TAL pour la classification conceptuelle à celles provenant de l'intelligence

artificielle pour la catégorisation et la classification non supervisée afin d'analyser la nature de ces processus temps réel en question et d'en relever les propriétés cachées. La spécificité de notre challenge scientifique est donc de constituer, par extraction et classification dynamique, des connaissances permettant de générer et faire émerger une organologie des traitements sonores temps réel qui contribuera ainsi à des pratiques de préservation de l'œuvre de l'art de la performance. Il est à noter que le corpus à analyser dans ce cadre ne demande pas forcément de disposer de connaissances additionnelles (terminologique ou sémantique) pour guider l'apprentissage ou évaluer les résultats. Les travaux en cours et futurs se concentrent sur les moyens qui permettent d'obtenir et d'évaluer les classes attendues par nos objectifs, à savoir les critères de constitution du corpus d'entrée, les distances, et les critères d'évaluation des résultats, nivaux d'abstraction (lié aux objets qui doivent être isolés de la classification, etc.). L'efficacité et la fiabilité de notre approche seront notamment validées par une phase de simulation numérique pour mettre en évidence les limites de l'approche développée. A travers ce travail, nous avons également pu montrer le rôle du patch comme document numérique dans la création artistique contemporaine et les différentes approches mises au point pour leur maintenance, dont la classification est certes l'une des solutions recommandées pour sa préservation à long terme dans la création de l'art contemporain.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence *2008 CORD 003 01*. Les auteurs tiennent à remercier les partenaires du projet ASTREE, et notamment ceux du Centre National du Création Musicale Grame, à l'origine du langage FAUST, et de l'outil de génération automatique de documentation sur lequel notre travail se base.

Bibliographie

[1] A. Bonardi et J. Barthélemy, Le patch comme document numérique : support de création et de constitution de connaissances pour les arts de la performance, *10ième Colloque International sur le Document Electronique (CIDE.10)*, 2-4 juillet, Nancy, France 2007.

[2] A. Bonardi, M. H. Serra et M. Fingerhut, Documentation musicale et outils hypermédias, *Deuxième Colloque International sur le Document Electronique (CIDE.2)*, 5-7 juillet, 295-309, Damas, Syrie 1999.

[3] A. Bonardi, J. Barthélemy, G. Boutard et R. Ciavarella, Préservation de processus temps réel. Vers un document numérique. *Document Numérique*, 11 (3-4), 59-80, 2008.

[4] P. Bottoni, S. Faralli, A. Labella et M. Pierro, Mapping with planning agents in the Max/MSP environment: the GO/Max language, *In Proceedings of the 2006 International Conference on New Interfaces for Musical Expression*, Paris, France 2006.

[5] M. Hedstrom, Digital Preservation: A Time Bomb for Digital Libraries, *Computers and the Humanities*, 31, 189–202, 1998.

[6] J. Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Published by ECPA (European Commission for Preservation and Access), Edo H. Dooijes, Amsterdam. 1998. http://www.clir.org/pubs/reports/rothenberg/contents.html

[7] B. Bachimont, J. F. Blanchette, A. Gerszo, A. Swetland, O. Lescurieux, P. Mahoudeaux, N. Donin et J. Teasley, Preserving Interactive Digital Music: A report on the Mustica Research Initiative, *In Proceedings of the Third International Conference on WEB Delivering of Music (WEB'03)*, Leeds, England 2003.

[8] H. Bosma, Documentation and Publication of Electroacoustic Compositions at NEAR, *In Proceedings of the Electroacoustic Music Studies Network International Conference* (EMS 05), Montreal, Canada 2005.

[9] D. Teruggi, Preserving and Diffusing, *Journal of New Music Research*, 30 (4), 403-405, 2001.

[10] V. Tiffon, Les musiques mixtes entre pérennité et obsolescence, *Revue Musurgia*, XII/3, Paris. 2005.

[11] J. Bullock et L. Coccioli, Modernising Live Electronics Technology in the Works of Jonathan Harvey, *In Proceedings of the International Computer Music Conference*, Barcelona, Spain 2005.

[12] M. Puckette, New Public-Domain Realizations of Standard Pieces for Instruments and Live Electronics, *In Proceedings of the International Computer Music Conference*, Miami 2004.

[13] Y. Orlarey, S. Letz et D. Fober, Multicolore technologies en Jack and Faust, *In Proceeding of the international Computer Music Conference-ICMA*, 2008.

[14] Y. Orlarey, D. Fober et S. Letz, FAUST : an efficient Functional Approach to DSP Programming, *New Computational Paradigms For Computer Music*, Delatour, France 2009.

[15] T. Berners-Lee, J. Hendler et O. Lassil, The Semantic Web, *Scientific American Magazine, 2001*.

[16] K. Mahech, et S. Nirenburg, A situated Ontology for Pratical NLP, *In Proceeding of Workshop on Basic Ontological Issues in Knowledge Sharing* : *International Joint conference on Artificial Intelligence (IJCAI-95)*, August 19-20, Montreal, Canada 1995.

[17] C. Fellbaum, *Wornet : An Electronic Lexical Database*, Combridge, Ma :MIT Press. 1998.

[18] D.W. Embly, D. M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y. K. Ng et R.D. Smith, Conceptual model based data extraction from multiple record Web, *Data Knowledge and Engineering*, 31(3), 227-251, 1999.

[19] K. Sparck Jones et E. B. Barber, What makes an automatic keywords classification effective?, *Journal of the ASIS*, 18, 166-175, 1971.

[20] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai et R.L. Mercer, Class-based n-gram models of natural language, *Computational Linguistic*, 18(4), 283-298, 1992.

[21] K.W. Church et P. Hanks, Word Association Norms, Mutual Information, and Lexicography, *In proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 76-83, 1989.

[22] B. Habert, A. Nazarenko et A. Salem, *Les linguistiques de corpus*, Ed Armand Collin. 1997.

[23] I. Dagan, L. Lee et F. Pereira, Similarity-Based Methods For Word-Sense Disambiguation, *In proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'96)*, 56-63, 1996.

[24] D. H. Fisher, Knowledge Acquisition via Incremental Conceptual Clustering. *In Machine Learning Journal*, 2, 139-172, 1989.

[25] G. Bisson, Conceptual Clustering in a First Order Logic Representation, *In Proceedings of 10th European Conference on Artificial Intelligence (ECAI'92)*, 458-462, Vienna 1992.

[26] S. Deerwester, S. T. Dumais, G. W. Furnas., T. K. Landauer et R. Hashman, Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

[27] J-P. Desclés et I. Biskri, Logique combinatoire et linguistique: grammaire catégorielle combinatoire applicative, *Mathématiques et sciences humaines*, Tome 132, 39-68, 1995.

[28] T. Kohonen, Self-organization of very large document collections: state of the art, *In Proceeding of ICANN'98*, London 1998.

[29] D. Nguyen, Nouvelle méthode syntagmatique de vectorisation appliquée au Self-organizing map des textes vietnamiens, *JEP-TALN-RECITAL (Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, 19-22 avril 2004, Fès, Maroc 2004.

[30] D. Nguyen et K. Zreik, HYPERLING : Système de reconnaissance et de classification des hyperdocuments multilingues, *In International Conference in Computer Science « Research, Innovation and Vision of the Future» RIVF05, 21-24 February, University of CANTHO, Vietnam 2005.*

[31] K. Zreik, et D. Nguyen, Catégorisation de documents multilingue : le système Hyperling, *In CIDE.8*, 25-28 Mai, Beyrouth, Liban 2005.