ANNEXES

BILAN DES OBJECTIFS ET LIVRABLES

I.ENCODAGE ET DONNEES MUSICALES

I.1. VALORISATION DES FORMATS OUVERTS POUR L'ENCODAGE DES DONNEES MUSICALES

CONTEXTE ET OBJECTIFS

Comparée à la plupart des disciplines constituant le champ des humanités numériques, la musicologie a pour particularité de disposer de nombreux logiciels, formats et fichiers en vue d'encoder les données musicales. Ce constat concerne entre autres les multiples manières d'éditer et de manipuler les partitions[3]. À titre d'exemple, MUSICA2 recense à ce jour plus d'une vingtaine de formats et fichiers destinés à l'édition de partitions. Parmi eux, certains standards, comme MusicXML, côtoient des formats ouverts moins populaires, mais encore en usage (RDG ou encore ABC). Cette pluralité est la conséquence d'attentions diverses en matière d'encodage – accent mis sur le graphisme ou sur la sémantique – et de répertoire : les musiques folkloriques, les tablatures anciennes ou même le chant grégorien.

Cependant, force est de constater que le recours à des formats ouverts n'est pas systématique. Un exemple éloquent, qui déborde largement sur la musicologie publique, est la base de données Choral Public Domain Library[4], de plus en plus exploitée par les musicologues. L'utilisation d'éditions préexistantes et numériques s'avère un gain de temps non négligeable pour de nombreux projets de recherche. Si la qualité des partitions doit néanmoins être vérifiée, le choix des utilisateurs de CPDL est également tributaire des formats présents, dépendants uniquement des contributeurs. Il est ainsi fréquent que des fichiers de logiciels propriétaires soient proposés (Finale ou Sibelius). Dans d'autres cas, il n'est pas rare qu'un fichier PDF[5] apparaisse à défaut d'un fichier MusicXML, pouvant être manipulé par l'ensemble des utilisateurs.

---

[3] Dans le cas des fichiers audio et des images (iconographie, numérisation de source, etc.), l'usage de standards ouverts est d'ores et déjà bien implanté (Wave, MP3, JPG, etc.).

[4] https://www.cpdl.org/wiki/index.php/Main_Page.

[5] Bien que le format PDF soit de nos jours considéré comme l'égal d'un format ouvert (proprement "standardisé"), cette qualité ne dépend que de sa popularité. Il est effectivement possible d'ouvrir un fichier PDF indépendamment du système utilisé, ce qui en fait alors un « standard » de description de page. Cependant, en matière de manipulation, le PDF se comporte comme un fichier propriétaire. En effet, il est dans ce cas nécessaire d'obtenir une licence Adobe (licence privée) afin d'en modifier le contenu ou la structure. Qui plus est, dans le cadre d'une partition, les fonctionnalités proposées par Adobe ne permettent pas de manipuler concrètement le contenu musical. C'est pourquoi le PDF illustre selon Tim Berners-Lee le plus faible niveau du Linked Open Data. On comprend également l'engouement autour des OMR ou des logiciels de traitement des PDF (PDFtoMusicPro) dont les licences sont le plus souvent onéreuses.

Afin d'améliorer l'interopérabilité et la récupération des données musicales, MUSICA2 a pour ambition de valoriser l'usage des formats ouverts et plus particulièrement de certains standards comme MusicXML ou MEI. Le dessein n'est pas de restreindre l'utilisation des formats fermés ou du PDF, mais plutôt de sensibiliser la communauté musicologique au partage de fichiers ouverts et interopérables afin d'agrémenter les bases de données.

LIVRABLES (I.1.A)

MUSICA2 propose ainsi la **réalisation d'un Livre blanc** (I.1.a) – guide de bonnes pratiques – destiné à l'encodage des données musicales. Il recensera l'ensemble des formats ouverts, des fichiers, des logiciels et des systèmes de conversion (également de formats fermés vers des formats ouverts). Chaque format sera répertorié en fonction de son niveau d'interopérabilité afin d'orienter le lecteur vers le meilleur choix ou compromis. Par exemple, un fichier ly (LilyPond) ne peut être recommandé au même titre qu'un fichier au format MusicXML (standard), leur degré d'accessibilité étant fondamentalement différent. Un tel classement suggère dès lors de définir plusieurs paramètres d'évaluations :

- Nombre de logiciels ou d'outils *online* compatibles (format propriétaire ou non-propriétaire) ;
- Nombre de conversions cibles ;
- Nombre de conversions sources.

I.2. VALORISATION DES FORMATS MUSICAUX OUVERTS AUPRES DES CENTRES INFORMATIQUES ET PLATEFORMES CHARGES DE LABELLISER LES FORMATS PERENNES

CONTEXTE ET OBJECTIFS

Si les formats ouverts de fichiers image (GIF, JPG, etc.) ou texte (TEI, XML, etc.) sont acceptés sur la plateforme d'archivage PAC du CINES (Centre Informatique National de l'Enseignement Supérieur), aucun fichier relatif à l'encodage des partitions musicales n'est pour l'heure valable (vérification de la sémantique et de la syntaxe du format) ou archivable (dépôt sur PAC).

La pérennité des formats dépend de plusieurs facteurs. Voici la définition que donne le CINES : un format pérenne est "avant tout un **format publié**. C'est-à-dire un format dont les spécifications internes sont librement accessibles. Seul un format publié peut en effet permettre, dans le plus extrême des cas, à un développeur d'écrire un programme spécifique de relecture du fichier. Un format durable est aussi un **format très utilisé**, ou appelé à le devenir. Un format durable, c'est enfin et si possible **un standard voire une norme**"[6].

Ces divers facteurs expliquent pourquoi tous les formats ouverts validés par le CINES ne sont pas nécessairement archivables sur PAC (APNG ou DAE). Ces différents éléments apportent également

[6] https://www.cines.fr/archivage/un-concept-des-problematiques/le-concept-darchivage-numerique-perenne.

quelques réponses sur le manque de formats destinés à l'encodage musical. Ces derniers sont moins représentés que la plupart des images, textes ou autres documents. Les plateformes d'archivage se conformant aux exigences des bases de données, la question des formats musicaux pérennes ne se pose que si la communauté musicologique en manifeste le besoin. Pourtant, c'est là une dimension essentielle des données musicales sans laquelle leur maintien durable ne peut être atteint. En effet, NAKALA[7] s'appuie fortement sur les recommandations du CINES, bien que les formats autorisés puissent être parfois fermés. C'est pourquoi MUSICA2 souhaite valoriser certains formats ouverts (notamment MusicXML et MEI) auprès du CINES en vue de leur labellisation en tant que formats pérennes.

LIVRABLES (I.2.A ; I.2.B)

Afin de réaliser cet objectif, MUSICA2 propose de promouvoir notamment le MusicXML et la MEI auprès du CINES. Ces deux formats sont de fait les plus prometteurs : tous deux sont non propriétaires (indépendants d'un logiciel payant spécifique) et sont les plus utilisés dans les bases de données musicologiques, dans des proportions toutefois variables. S'il ne fait aucun doute que le MusicXML est le standard en matière d'encodage musical, la MEI est le format le plus apprécié des chercheurs spécialisés dans le catalogage, l'édition critique et l'analyse computationnelle.

À travers des **journées d'études et groupes de travails sur l'archivage** (I.2.a), MUSICA2 souhaite donc encourager et sensibiliser la communauté musicologique à archiver leurs données sur les plateformes spécifiquement dédiées (comme PAC ou NAKALA), et ainsi augmenter le nombre de services versants et de détections exercées par le CINES. Cette démarche va également de pair avec la mise en place de discussions entre le CINES et MUSICA2, avec l'appui d'Huma-Num, afin d'échanger sur la pérennité du MusicXML et de la MEI et d'aboutir à une **procédure de labellisation des deux formats** (I.2.b).

I.3. AMELIORATION DE L'INTEROPERABILITE ENTRE LES DIFFERENTS FORMATS SPECIFIQUEMENT DEDIES A L'ENCODAGE MUSICAL

CONTEXTE ET OBJECTIFS

Outre la question des formats ouverts et pérennes, la pluralité des formats musicaux nuit considérablement à l'interopérabilité. Bien que le MusicXML soit largement considéré comme un standard, tous les formats ne peuvent être exportés ainsi. D'ailleurs, ce manque d'interopérabilité ne vise pas nécessairement les formats les moins usités. Certains problèmes rencontrés par le programme de recherche The CRIM Project illustrent ce point. Dans l'intention d'alimenter les partitions MEI de CRIM, une équipe indépendante de musicologues a accepté de partager ses propres éditions. Or, ces dernières étaient réalisées sous LilyPond. Si LilyPond demeure apprécié par de nombreux enseignants, musiciens et musicologues, un aperçu du *workflow* requis pour exporter en MusicXML ou en MEI suffit à décourager toute collaboration scientifique :

[7] https://nakala.fr.

Ly → (LilyPond/Frescobaldi) → MIDI → MuseScore (MXL) → MusicXML→ Sibelius (Sib) → MEI

De fait, pas moins de trois logiciels et six formats différents ont été utilisés. L'accumulation des conversions, dans des formats relativement distincts, entraîne également des altérations du fichier initial. Dans le cas précis des messes de la Renaissance partagées par l'équipe de musicologues, des segments entiers ont dû être réécrits (notamment les métriques 3/2 à l'évidence non supportés durant l'export MIDI). Ce manque d'interopérabilité oblige donc les musicologues à répéter la fastidieuse tâche qu'est l'édition et induit une perte de temps non négligeable.

Un autre objectif sera par conséquent d'améliorer les conversions entre les formats. Les efforts se concentreront initialement sur les formats les plus utilisés ou ceux dont l'usage est requis, sans solution alternative. Par exemple, sans passer par le code MEI et la visualisation Verovio[8], une édition convenable et non propriétaire du chant grégorien (notes carrées) peut difficilement se faire sans Gregorio (GABC/GregorioTex).

LIVRABLES (I.3.A ; I.3.B ; I.3.C)

Pour ce faire, plusieurs chemins s'offrent au consortium. D'une part, **l'amélioration ou la création ex nihilo de programmes de conversion** semble inévitable (I.3.a). Cette entreprise suppose l'aide de codeurs, si possible spécialisés dans les formats concernés. Elle suppose également de déterminer les conversions sur lesquelles se concentrer, sur la base de l'évaluation de l'interopérabilité ainsi que les bonnes pratiques à adopter. À titre d'exemple, le programme open source lytoxml[9], qui ne permet pour l'instant l'export de fichiers Lilypond vers MusicXML que pour des partitions simples, conduirait, une fois le code amélioré, à réduire considérablement le *workflow* énoncé plus haut.

Afin de s'inscrire dans une démarche parfaitement "FAIR", les programmes en question devront être, d'une part, *open source* et facilement accessibles. Outre le partage du code sur des plateformes comme Github[10], l'idée d'un web service plus ergonomique pourrait être envisagée.

D'autre part, une autre approche concerne le **recensement des conversions actuelles** (I.3.b), travail qui s'appuie grandement sur l'évaluation des formats citée précédemment, en vue de la création d'un annuaire en ligne. En effet, bien que diverses initiatives existent déjà, leur utilisation est contrecarrée par un manque d'accessibilité : diversité des sites, manque de popularité, maintenance arrêtée ou processus implicite et complexe.

Dans tous les cas, MUSICA2 proposera les **différents *workflows*, créés ou améliorés, sous la forme d'un Standardization Survival Kit**[11] (I.3.c). Introduit dans le cadre H2020 PARTHENOS avant d'être intégré par Huma-Num ainsi que Social Sciences & Humanities Open Marketplace, le SSK permettra de décrire

[8] https://www.verovio.org/index.xhtml.
[9] https://github.com/uliska/ly2xml.
[10] https://github.com.
[11] http://ssk.huma-num.fr/#.

de manière simple et didactique divers scénarios de conversion, adaptés à des objectifs spécifiques. De cette manière, les SSK s'avèreront particulièrement utiles en vue d'illustrer les axes du Livre blanc et de représenter précisément les bonnes pratiques à adopter.

I.4 STANDARDISATION DES METADONNEES DES FICHIERS

CONTEXTE ET OBJECTIFS

L'addition des métadonnées est un sujet qui excède le cadre de la donnée numérique pure. Chaque plateforme d'hébergement de données met aujourd'hui un point d'honneur à garantir leur signalement dans un format harmonisé. Généralement, en l'absence d'une ontologie spécifique, le standard le plus utilisé sur le web est le Dublin Core (HTML, XML). Tandis que ce dernier est, depuis 2000, exprimé en RDFa, il fait également l'objet de la norme ISO 15836 depuis 2003. Enfin, il est au cœur du protocole OAI-PMH afin de faciliter les échanges des métadonnées entre les archives ouvertes et les bibliothèques numériques. Au Dublin Core s'ajoutent d'autres standards, comme la famille des formats MARC ou le MODS. Tous sont préconisés par la Music Library Association.

En marge des catalogues spécifiquement destinés à la publication ou à l'archivage web, les métadonnées des fichiers s'avèrent bien moins standardisées. Cette disparité résulte de la pluralité des formats dont le schéma et le contenu des métadonnées, embarquées ou externes, sont distincts. La numérisation d'une partition musicale au format JPG ne disposera pas de mêmes champs de métadonnées internes (Exif) qu'une partition au format MusicXML représentant potentiellement la "même" œuvre ou la "même" édition. Dans le premier cas, les métadonnées sont, par commodité, souvent extérieures au fichier lui-même et ajoutées dans un autre fichier (CSV, TXT ou encore XML), parfois par l'intermédiaire d'un logiciel de gestion bibliographique – l'Exif ne permet pas, en effet, d'intégrer et de manipuler aisément des métadonnées scientifiques. À l'inverse, dans le second cas, le *header* d'un fichier en MusicXML permet d'ajouter simplement diverses métadonnées, au sein même du fichier, dont la sémantique est précisée. Dès lors, les manières se multiplient à mesure que s'accumulent les formats (MEI, PDF, XML, etc.).

C'est pourquoi MUSICA2 propose de déterminer les bonnes pratiques à adopter quant au signalement des métadonnées selon les fichiers utilisés et la nature des données. L'idéal serait, dans un premier temps, d'identifier un ensemble commun et minimal de métadonnées, sous une forme équivalente ou du moins interopérable. Un tel objectif suppose également, dans un second temps, le recours à des schémas de métadonnées d'ores et déjà standardisés afin d'insérer le consortium dans une dynamique préexistante. De plus, la manière de renseigner les métadonnées devra être adaptée en fonction des fichiers. Pour ne citer qu'un exemple, un fichier MEI a la capacité d'intégrer un nombre important de métadonnées, au point de fournir pas moins de trois principales sous parties modulables au sein du *header* (*file description*, *encoding description* et *work description*)[12]. La façon d'encoder les métadonnées (sémantique et schémas utilisés) sera en outre pensée en

[12] https://music-encoding.org/guidelines/v4/content/metadata.html.

fonction de l'usage et de la spécificité des données (publications HTML, XML ou RDF ; catalogage généraliste ou circulation de données techniques).

LIVRABLES (1.4.A ; 1.4.B ; 1.4.C ; 1.4.D)

MUSICA2 propose donc d'élaborer, dans un premier temps, un **ensemble de recommandations de métadonnées minimales proprement dédiées aux fichiers de données musicales** (I.4.a). Ce schéma commun pourra être agrémenté, dans un second temps, d'autres éléments en fonction de la richesse ou de la nature des données en question. Dans la mesure où le Dublin Core (DC) et ses affinements demeurent des standards, il semble judicieux de se baser initialement dessus : les métadonnées élémentaires accompagnant chaque fichier pourraient être facilement manipulées, extraites et utilisées par l'ensemble des gestionnaires de bibliothèques numériques et de plateformes d'archivage. De plus, l'expression en RDF des éléments du Dublin Core peut permettre d'ouvrir facilement la voie vers le web sémantique. Pour finir, l'aspect généraliste anciennement critiqué du Dublin Core laisse aujourd'hui place à davantage de précision (Dublin Core étendu + MARCRelators)[13] et à une articulation qui, suivant le DCAM (Dublin Core Abstract Model), invite à une syntaxe beaucoup plus complète, de type RDF, des métadonnées.

Par ailleurs, le Dublin Core est déjà au cœur de certains formats de fichiers musicaux ou connexes. D'une part, le format MusicXML dispose de balises de métadonnées largement empruntées au Dublin Core. Pour ne citer qu'un exemple, l'élément « *composer* » est un affinement de « *creator* » . Si la MEI s'inspire moins du Dublin Core, la précision qu'offre la sous-partie du *header* <respStmt>, recommandé pour l'échange des métadonnées par MEI.org, permet un alignement comparable : <persName role="composer">Exemple</persName>. Le lien avec le Dublin Core est dans ce cas indirect, mais bel et bien présent : MEI.org préconise en effet d'employer les MARCrelators qui ne sont autres que des affinements du Dublin Core.

Ces recommandations de métadonnées s'appuieront donc sur la détermination de sémantiques communes, applicables et pertinentes à l'égard des données musicologiques. **L'organisation de journées d'étude autour des métadonnées musicales** permettra de recenser l'ensemble des champs de métadonnées descriptives utilisés par les partenaires **afin d'aboutir à un « set » musicologique** (I.4.b). Ces moments seront l'occasion de réfléchir à de potentiels *mappings* et autres alignements afin d'œuvrer, par la suite, à leur création ou à leur optimisation. Des améliorations sont encore à faire, notamment sur la conversion des métadonnées entre le MusicXML et la MEI.

Sur ce point, des collaborations avec des ingénieurs en TEI seront d'une aide précieuse. Soucieuse de préserver la richesse des métadonnées naturellement exprimées suivant l'ISBD (International Standard Bibliographic Description), tout en assurant leur interopérabilité, la communauté TEI a développé différents *data mappings* et conversions du *header* vers d'autres formats bibliographiques : MARC, MODS (TEI2MODS), Dublin Core (XSLT), etc. TEI.org préconise également d'utiliser le container <element XenoData> pour y ajouter des métadonnées dans des schémas ou

[13] https://www.loc.gov/marc/relators/relaterm.html.

formats autres que TEI. Ces différentes perspectives ouvrent ainsi un champ des possibles particulièrement vaste pour la MEI que le consortium, fort de l'expérience de la communauté TEI, s'efforcera d'affiner.

Le consortium proposera également des **modèles plus spécifiques en fonction de la nature des données (codicologique, analytique, etc.) et de leur usage** (I.4.c). Dès lors, il semblera inévitable de se tourner vers des standards plus complets que le Dublin Core. Outre l'élaboration potentielle d'une « extension » DC (*Application profile*) dans l'esprit du Darwin Code, il sera notamment intéressant de voir comment le modèle FRBR, encouragé par la Music Library Association et bénéficiant d'un module au sein de la MEI, permet d'élargir l'horizon des métadonnées musicales en vue du catalogage bibliographique. Le renseignement des métadonnées s'accompagnera d'un ***workflow* précis publié en ligne** (I.4.d), plus particulièrement pour des fichiers qui, comme la MEI, disposent de champs et conteneurs hétérogènes pour les métadonnées.

I.5. VALORISATION DE LA MEI

CONTEXTE ET OBJECTIFS

La Music Encoding Initiative joue un rôle désormais central au sein des FAIR data en musicologie, plus particulièrement en matière d'édition et d'analyse computationnelle. Toutefois, la difficulté de visualiser et de manipuler « musicalement » le code MEI provoque un désintérêt ou une certaine méfiance de la part des musicologues – rappelons que la MEI a été avant tout créée dans le but d'être lisible par une machine. Si d'importants efforts ont été faits, surtout par l'intermédiaire de Verovio[14] (visionneuse), l'interaction reste élémentaire. Face à la richesse sémantique de la MEI et à ses objectifs initiaux, il apparaît cependant vain d'attendre plus d'une interface graphique. La communauté musicologique doit en ce sens être sensibilisée aux fonctions de la MEI, aussi bien en matière de préservation des œuvres et de leurs métadonnées, que d'analyse computationnelle.

Néanmoins, l'interaction entre les musicologues et la MEI peut être simplifiée grâce à des initiatives répondant à des besoins spécifiques. L'exemple le plus connu est MEI Friend[15], dont l'interface (sur la version *online*) permet d'intégrer facilement les symboles musicaux usuels de la notation occidentale. Plus orienté vers les musiques anciennes, Ricercar a récemment développé un programme en python, à l'aide de formulaires, destiné à ajouter des symboles éditoriaux modernes relatifs aux musiques anciennes (coloration, *ficta* et ligature). En plus de permettre une interaction simple avec la machine, ces programmes assurent une modification convenable du code MEI.

LIVRABLES (1.5.A ; 1.5.B)

Afin de valoriser l'usage de la MEI, MUSICA2 organise des **journées d'initiation dédiées aux laboratoires, en France et à l'étranger** (I.5.a). L'objectif est essentiellement de familiariser les

[14] https://www.verovio.org/index.xhtml.
[15] https://mei-friend.mdw.ac.at.

musicologues et étudiants avec les bases de la MEI à travers la présentation d'outils spécifiquement conçus pour faciliter l'encodage MEI (plug-in Sibelius, conversion avec MEIGarage[16], Verovio, modification avec MEIFriend). Plus important encore, ces journées se concentrent sur les bonnes pratiques à adopter pour obtenir un fichier MEI correct sans nécessairement intervenir dans le code. Afin de s'inscrire dans une approche FAIR autant que possible, les différents *workflows* sont à la fois présentés sur Sibelius et MuseScore.

La valorisation de la MEI s'effectuera aussi à travers la **création et l'amélioration d'outils destinés à simplifier la manipulation des fichiers** (1.5.b). À l'instar de MEI Friend, d'autres programmes connexes peuvent également être mis en œuvre en vue de projets ou répertoires spécifiques. Cet objectif requiert en amont de définir les attentes des musicologues, notamment à l'occasion des journées d'initiation. Une fois les besoins identifiés, le consortium proposera des solutions numériques, sous la forme d'outils *open source*. Suivant les ambitions précédentes concernant les métadonnées, MUSICA2 offrira la possibilité d'intégrer rapidement des métadonnées contrôlées au sein des *headers*, conformément aux schémas prédéterminés.

II. GESTION ET STOCKAGE DES DONNEES

II.1. MODELES, RELATIONS ET NORMES DES BASES DE DONNEES

CONTEXTE ET OBJECTIFS

L'interopérabilité entre les bases de données repose essentiellement sur des modèles structurels et sémantiques communs. Concernant la structure, le modèle relationnel (organisation des données à l'aide de tables interconnectées) est de nos jours le plus utilisé en raison de sa gestion intuitive et de sa souplesse en matière de requêtage. Une telle forme permet de croiser facilement de nombreux attributs de données et entités. Cependant, tandis que la pluralité des systèmes de gestion des bases de données relationnelles (BDR), comme PostgreSQL, Oracle, MariaDB ou MySQL, peut nuire à l'interopérabilité, des relations disparates entre les tables peuvent également être un frein à l'intégration ou la comparaison des données. Dans l'intention d'améliorer l'interopérabilité des BDR, MUSICA2 souhaite définir des modèles relationnels génériques. En effet, si la plupart se concentrent sur des entités comme les « œuvres », les « personnes », les « sources » ou encore les « événements », les relations qui les régissent, ainsi que leurs attributs, peuvent être articulées de différentes façons (recours à des tables intermédiaires ou cardinalités diverses).

LIVRABLES

Au-delà des problématiques spécifiquement musicologiques, le choix du système de gestion est donc la première étape d'une interopérabilité optimale. Sur ce point, il semble judicieux de suivre

[16] https://meigarage.edirom.de.

les directives gouvernementales (circulaire Ayrault du 19 septembre 2012)[17] à propos de l'usage de logiciels libres. La circulaire encourage ainsi le recours à PostgreSQL (logiciel sous licence BSD), aux dépens des logiciels propriétaires comme Oracle, DB2 ou encore MySQL. C'est entre autres à l'aide de PostgreSQL que Ricercar (CESR) élabore sa nouvelle BDR agrégeant l'ensemble de ses anciennes bases de données, ou que le projet Dezède (Universités de Rouen et de Montpellier) a été développé dès septembre 2011.

Cette **promotion des logiciels libres de gestion des BDR** (II.1.a) s'accompagne également de la **détermination d'un modèle générique commun entre les BDR** (II.1.b), de manière à assurer une interopérabilité minimum entre les BDR. Outre la déduction d'entités majeures et partagées (entités de haut niveau), le modèle aura pour objectif de représenter les aspects techniques améliorant la structure globale des bases de données à l'aide de relations, cardinalités et tables intermédiaires appropriées. Par ailleurs, le modèle devra intégrer certaines normes générales concernant l'implémentation des données, notamment numériques (dates, numéros de catalogage, etc.), en s'appuyant sur les *Directives sur la gestion et valorisation des données* définies par OTELo et l'INIST[18]. Le modèle structurel sera ainsi décrit au sein de la deuxième partie du guide de bonnes pratiques dédié aux BDR.

II.2. DETERMINATION DE REFERENTIELS COMMUNS ET USAGE D'IDENTIFIANTS PERENNES

CONTEXTE ET OBJECTIFS

Le choix des référentiels demeure une étape particulièrement importante durant l'élaboration d'une base de données. C'est en effet à travers l'usage de référentiels partagés que les données pourront être facilement manipulées et croisées avec des données extérieures. Le paysage des humanités numériques comprend aujourd'hui différents thésaurus et vocabulaires visant à standardiser la sémantique de disciplines spécifiques. Loterre[19] (CNRS/INIST), plateforme dédiée au partage de terminologies scientifiques, recense par exemple plusieurs vocabulaires, tels que ceux développés par le Getty pour l'histoire de l'art et l'archéologie.

Le recours à des référentiels communs est intimement lié à leur identification et pérennité dans le contexte du web. L'exemple le plus fréquent concerne l'identification des personnes. Par le biais d'identifiants numériques jugés pérennes (PIDs), de multiples référentiels d'autorités (VIAF, IdRef, ISNI ou encore Rameau) permettent de citer précisément des personnalités et, ainsi, de résoudre les ambiguïtés relevant de l'homonymie. L'appariement des référentiels (VIAF et BiblissimaData) permet dès lors "aux chercheurs d'identifier des noms, des lieux, des œuvres et des expressions tout en préservant les préférences régionales en matière de langue, d'orthographe et d'écriture"[20]. Par

[17] https://www.april.org/files/201209-circulaire-ayrault-disic-35837-pdf-normal.pdf.
[18] https://hal.archives-ouvertes.fr/hal-01275841/document.
[19] https://www.loterre.fr.
[20] https://www.oclc.org/en/viaf.html.

ailleurs, certains identifiants objets, entre autres DOI et ARK, s'accompagnent de métadonnées dans des schémas prédéfinis (ex. Dublin Core pour Ark).

Le domaine musical comprend peu de vocabulaires normalisés, qui plus est accompagnés d'identifiants pérennes. Certes, de nombreux compositeurs, interprètes, ensembles ou encore éditeurs sont représentés dans les répertoires listés précédemment. Les identifiants objets demeurent cependant plus rares. Citons à ce propos le thesaurus MIMO, standard en matière d'organologie, MusicBrainz, qui recouvre les œuvres, les enregistrements ou encore les lieux de performance, ainsi que l'ISCR, pour les œuvres enregistrées. Les efforts de l'ANR Doremus sont, à ce propos, révélateurs de l'absence de vocabulaires contrôlés pérennes : plusieurs vocabulaires, parfois empruntés à d'anciens labels, ont à la fois été créés et pérennisés à l'aide d'URIs (Uniform Ressource Identifiers) spécialement développés par Doremus (Diabolo, Itema3, Carrier Types, etc.)[21]

Livrables (II.2.a ; II.2.b ; II.2.c)

En plus d'encourager le recours à des logiciels libres de gestion de BDR, MUSICA2 a également pour objectif de déterminer des référentiels contrôlés. Certaines initiatives peuvent déjà être constatées, avec, par exemple, le recours aux identifiants ISNI pour la base Dezède. Dans ce cas, chaque individu ou chaque ensemble de la base est référencé par son identifiant ISNI – quand il existe –, avec pour objectif à terme de communiquer avec d'autres BDR. À ce stade, plusieurs directions peuvent être envisagées. D'une part, le consortium peut se diriger vers des référentiels actuels (MIMO, référentiels Biblissima, VIAF, etc.) qui, le cas échéant, permettraient **d'intégrer les BDR au sein d'un répertoire plus vaste de données** (II.2.a). L'usage d'un identifiant unique et préexistant (URI dans le cadre du web sémantique), en marge des identifiants internes aux BDR (généré par les logiciels de gestion), permettrait ainsi d'identifier sans ambiguïté les données de chaque propriété d'une table.

D'autre part, en l'absence de standards convenables, il est également possible de **recourir à de nouveaux référentiels**, spécifiquement dédiés aux besoins des BDR, et **alignés avec des référentiels existants par l'intermédiaire de Loterre ou Opentheso**[22] (II.2.b). Cet aspect rejoint l'entreprise de Doremus et permettrait de se projeter dans la perspective du web sémantique et des ontologies[23].

Dans ce but, MUSICA2 formera des **groupes de travail se concentrant sur la réalisation ou l'amélioration (pérennisation des données) de vocabulaires** (II.2.c) destinés à des répertoires – musique ancienne, musique contemporaine, etc. – ou des contextes spécifiques. Le consortium devra ainsi définir en amont les domaines jusqu'alors ignorés des ensembles de référentiels et identifiants pérennes. En effet, si les référentiels sont très développés pour les individus, ils restent encore à améliorer largement pour les œuvres (messe, opéra, symphonie, quatuor, romance, etc.), les extraits d'œuvres (mouvements) et les versions (traductions, transcriptions, arrangements, etc.)

[21] https://data.doremus.org/vocabularies.

[22] https://opentheso.huma-num.fr/opentheso.

[23] L'axe III.1, dépendant de celui-ci, reviendra plus particulièrement sur l'enrichissement d'ontologies à l'aide de nouveaux vocabulaires contrôlés exprimés en SKOS.

II.3. VALORISATION DE L'ARCHIVAGE PERENNE

CONTEXTE ET OBJECTIFS

Outre l'identification de référentiels d'autorités, la pérennité des bases de données se rapporte aux problématiques générales concernant l'archivage à long terme des autres contenus scientifiques numériques (articles, notices, graphiques, etc.). Ces problématiques se concentrent à la fois sur la gestion des métadonnées et la création d'identifiants pérennes. De la même manière que pour les identifiants contributeurs, désormais reconnus par les communautés scientifiques (ORCID, idHal ou IdRef), il est aujourd'hui nécessaire que chaque publication, peu importe sa forme, dispose d'un PID (*persistent identifier*). Tout comme pour l'intégration des métadonnées, plusieurs approches permettent aux chercheurs d'attribuer un identifiant pérenne à leurs productions : soit par l'intermédiaire d'un entrepôt adressant automatiquement un DOI ou un identifiant local, soit en sollicitant une agence spécialisée (CrossRef) ou partenaire du consortium DataCite.

Afin de faciliter l'archivage pérenne des données, Huma-Num a développé le service NAKALA, "permettant à des chercheurs, enseignants-chercheurs et équipes de recherche de partager, publier et valoriser tous types de données numériques documentées (fichiers textes, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé afin de les publier en accord avec les principes du FAIR data"[24]. Étroitement lié à la plateforme d'archivage PAC (CINES), NAKALA encourage le dépôt de documents ouverts et incite au signalement de métadonnées (Dublin Core qualifié). Cependant, force est de constater que les deux plateformes d'archivage susmentionnées ne disposent pas de collections musicologiques significatives. Bien que le type « partition » soit présent, la grande majorité des éditions accessibles sur ISIDORE[25] (moteur de recherche exploitant en partie NAKALA) sont issues de Gallica, mais également des carnets de recherche Hypothèses, des articles de revues hébergées par OpenEdition Journals ou de projets plus isolés (collection Dezède). De même, le contenu associé à la discipline « Musique, musicologie et arts de la scène » est faible.

La valorisation de l'archivage pérenne est donc naturellement l'un des objectifs de MUSICA2. La diffusion des données musicologiques au sein de l'environnement Huma-Num (NAKALA et ISIDORE) permettrait non seulement de participer à la dynamique FAIR de la TGIR, mais également de valoriser les recherches et contributions des partenaires, notamment dans le web sémantique (conversion RDF des métadonnées, création d'URIs, etc.).

LIVRABLES (II.3.A ; II.3.B)

Le consortium aura ainsi pour rôle de **définir les jeux de données susceptibles d'être archivés dans NAKALA et référencés sur ISIDORE** (II.3.a). Si besoin, il sera possible de recourir à certaines phases de test, comme le permet par exemple test.nakala. Une fois les données quantifiées et classées sous forme de collections (structure recommandée par Huma-Num), les **collections** seront

[24] https://documentation.huma-num.fr/nakala.
[25] https://isidore.science.

déposées dans l'entrepôt OAI-PMH de NAKALA et seront **indexées par le moteur de recherche ISIDORE** (II.3.b). Suivant le partenariat entre Huma-Num et le CINES, l'apport de fichiers musicaux aboutira en outre à la valorisation de leurs formats, dans l'optique de l'axe I.2.

III. WEB SEMANTIQUE

III 1. DEFINITION D'UN STANDARD ONTOLOGIQUE POUR LA MUSICOLOGIE ET MISE EN ŒUVRE DE BASES DE DONNEES RDF

CONTEXTE ET OBJECTIFS

Depuis la création du CIDOC Conceptual Reference Model[26] en 1996 (normalisé en 2006) et de FRBR (Functional Requirements for Bibliographic Records) en 1997, le recours à des ontologies dans les humanités numériques est devenu plus fréquent – bien que d'une ampleur restreinte par rapport aux sciences fondamentales, au milieu biomédical ou aux secteurs privés. Les deux modèles précités ont en effet offert une base sur laquelle élaborer divers ensembles spécifiques de termes et de concepts. Outre FRBRoo, né en 2006 de la fusion entre le CIDOC CRM et FRBR, le premier modèle comprend désormais de nombreuses extensions comme CRMtext, CRMgeo ou encore CRMba. Au côté de FRBR, RDA est pour sa part utilisé par les institutions anglo-américaines depuis 2010 en matière de références bibliographiques.

Dans le cadre de la musique, il est possible de mentionner quelques ontologies. Certaines concernent des aspects particulièrement précis du domaine acoustique ou sonore, à l'exemple de The Audio Effect Ontology (spécificité des effets studio) et Audio Features Ontology Specification (spécificité du signal audio, reposant principalement sur l'ontologie Event). D'autres ontologies relèvent davantage des universaux musicaux. The Music Ontology[27] permet pour sa part d'organiser aisément des données musicales au sein du Web sémantique. La démarche intègre parfaitement le processus des FAIR data, dans la mesure où The Music Ontology repose sur diverses ontologies préexistantes : FOAF (vocabulaire des personnes ou des groupes), The Event Ontology, The Timeline Ontology ainsi que FRBR ontology. De fait, The Music Ontology peut être appréhendé comme une extension de FRBR pour la musique. Cependant, l'ensemble des concepts permet avant tout de renseigner la formation d'un groupe ou de décrire une performance. Ainsi, l'ontologie se révèle assez pauvre en précision musicologique.

Face aux lacunes de The Music Ontology, l'ANR Doremus a élaboré une nouvelle ontologie à partir de FRBRoo, née d'un projet visant à convertir des répertoires musicaux de la BnF, de Radio France et de la Philharmonie de Paris. Si certaines classes apparaissent redondantes (notamment les niveaux les plus élevés), Doremus comprend 84 classes, contre 54 pour The Music Ontology. Le nombre de

[26] https://www.cidoc-crm.org.
[27] musicontology.com.

propriétés est également supérieur. Outre cette densité différente, les deux ontologies s'avèrent structurellement distinctes. Il est à noter que Doremus exploite davantage les classes de FRBRoo.

Doremus est ainsi devenu l'ontologie musicale et musicologique la plus complète développée à ce jour. Dans le but de valoriser et de diffuser les données musicologiques au sein de *triplestores*, il serait judicieux de s'appuyer sur les capacités de Doremus tout en enrichissant ses vocabulaires. Cet objectif rejoint sensiblement la définition de nouveaux vocabulaires (II.2), dans la mesure où il tend à utiliser les référentiels d'autorités des BDR, élaborés ou améliorés, dans un environnement RDF. L'usage d'une ontologie commune entre les partenaires permettrait d'automatiser et de croiser facilement les données (requêtes SPARQL) tout en créant des connexions avec d'autres données issues du web sémantique.

LIVRABLES (III.1.A ; III.1.B ; III.1.C ; III.1.D)

Le *mapping* des BD actuelles vers une ontologie préexistante est un travail progressif. Dans un premier temps, il sera fondamental pour le consortium de **définir les bases de données susceptibles d'être mappées en « RDF »** (III.1.a). De même, il sera nécessaire de s'interroger sur la cohérence des données : un ou plusieurs *triplestores* ? Une base généraliste ou des bases spécifiques ? Dès lors, il sera dans un second temps essentiel de déterminer les vocabulaires (suivant l'axe II.2.) avec lesquels agrémenter l'ontologie. Doremus ayant originellement pour rôle de publier les données de la BnF, de Radio France et de la Philharmonie de Paris, les vocabulaires et thesaurus employés se limitent conséquemment à ces institutions.

Il sera impératif par la suite **d'exprimer les vocabulaires en SKOS (Simple Knowledge Organization System) et de les publier en *open source*** (III.1.b). Ce vocabulaire RDF, recommandation du W3C, permet de représenter de manière structurée des thesaurus et vocabulaires dans le cadre du web sémantique. De même, chaque élément des vocabulaires question devra être convenablement implanté dans Doremus.

Pour finir, il sera non seulement nécessaire de **déterminer les *mappings* appropriés** (III.1.c), en fonction des formats de données (XMLtoRDF, D2RQ, etc.), mais également d'adapter ces derniers à l'ontologie Doremus (alignement des référentiels avec les classes, relations et vocabulaires). La réalisation de tableaux d'équivalences s'avérera dès lors fondamentale. C'est sur ces équivalences que s'appuieront les ingénieurs responsables d'élaborer les *mappings* spécifiques, en vue de finalement **publier les données en RDF** (III.1.d). Le consortium pourra notamment compter sur l'aide des ingénieurs en web sémantique de la MSH Val de Loire. Une première phase de test, avec un échantillon de quelques bases de données, est d'ailleurs programmée pour la rentrée 2022-2023.

III.2. AMELIORATION DE L'INTERACTION ENTRE LES MUSICOLOGUES ET LE WEB SEMANTIQUE

CONTEXTE ET OBJECTIFS

Si le langage RDF ouvre de nouveaux horizons en matière d'interopérabilité et d'analyse des données, le requêtage des *triplestores* demeure encore l'affaire des informaticiens. Le langage

SPARQL (SPARQL Protocol and RDF Query Language), soit l'équivalent des requêtes SQL pour les BDR, nécessite dans l'absolu une solide connaissance de sa syntaxe.

Quelques efforts ont été réalisés afin de simplifier l'usage du SPARQL. D'une part, plusieurs services de documentation, à l'image de DataBnF, proposent des requêtes préconçues avec leurs points d'accès SPARQL. Toutefois, ces formulaires s'avèrent nécessairement limités et ne permettent pas aux utilisateurs de naviguer librement au sein des *triplestores*. Ce manque d'accessibilité est dès lors un frein à la valorisation des bases de données. D'autre part, SPARNA[28], société spécialisée dans l'organisation des données numériques, a récemment développé SPARNAtural[29], composant Javascript destiné à faciliter l'exploration des graphes de connaissances RDF. SPARNAtural transpose la syntaxe SPARQL en des symboles et relations visuelles adaptés à l'ontologie employée. Ainsi, une approche intuitive et didactique se substitue à la requête SPARQL habituelle.

SPARNAtural est utilisé par DataBnF, mais aussi par OpenArchaeo[30], fruit du consortium Huma-Num MASA, afin d'améliorer l'accessibilité aux données archéologiques. Suivant le même intérêt pour la valorisation des données, MUSICA2 mettra en œuvre des interfaces similaires visant à simplifier l'expérience des utilisateurs désirant naviguer au sein des *triplestores*. Cette étape parachèvera ainsi l'intégration des données musicologiques dans le web sémantique.

LIVRABLES (III.2.A ; III.2.B)

Une fois le ou les *triplestores* finalisés, des **points d'accès SPARQL** seront **mis à disposition** (III.2.a). MUSICA2 proposera un partenariat avec SPARNA en vue d'un **SPARNAtural spécifiquement dédié aux bases de données RDF musicologiques** (III.2.b). Cet objectif impliquera d'adapter le composant aux divers éléments de l'ontologie (classes, relations, objets, etc.). Suivant les orientations du consortium en matière de diffusion des données RDF, le SPARNAtural musicologique sera intégré à l'ensemble des points d'accès SPARQL fournis.

IV. MANAGEMENT DU CONSORTIUM

IV.1. COORDINATION

LIVRABLES (IV.1.A ; IV.1.B ; IV.1.C ; IV.1.D ; IV.1.E ; IV.1.F ; IV.1.G)

La coordination du consortium implique, d'une part, de réaliser différentes réunions de gestion : **comité de pilotage** (IV.1.a), **réunions plénières avec les partenaires** (IV.1.b) et **groupes de travail** (IV.1.c). Programmées régulièrement sur l'année, ces rencontres sont notamment l'occasion pour le consortium de faire un bilan des objectifs remplis et, si nécessaire, de se concerter collectivement

[28] http://www.sparna.fr.
[29] https://sparnatural.eu.
[30] http://openarchaeo.huma-num.fr.

autour de problématiques. D'autre part, le comité de pilotage se doit de produire différents documents de gestion afin de consigner les démarches adoptées par le consortium et, de fait, de garantir son bon fonctionnement : **procès-verbal** (IV.1.d), **bilan des objectifs et livrables** (IV.1.e) et **bilan de fin d'année destiné au Conseil scientifique** (IV.1.f).

Parmi les missions du consortium, le comité de pilotage, à la suite d'une demande motivée à l'aide d'un formulaire disponible sur le carnet Hypothèses MUSICA2, peut également décider de **soutenir logistiquement et financièrement un projet partenaire isolé** (IV.1.g).

IV.2. PUBLICATIONS ET VISIBILITE

LIVRABLES (IV.2.A ; IV.2.B ; IV.2.C)

Outre les diverses manifestations programmées dans le cadre des objectifs susmentionnés, MUSICA2 entend accroître la visibilité de ses travaux à l'aide d'un **carnet Hypothèses régulièrement mis à jour** (IV.2.a). Le carnet a non seulement pour but de présenter les avancées du consortium ainsi que ses missions, mais également de publier divers contenus didactiques destinés à la communauté musicologique. De même, plusieurs **listes de diffusion, newsletters et listes de diffusion internes** (IV.2.b) ont été prévues pour le mois de septembre 2022.

En complément de cette activité sur internet, le consortium entreprend aussi de participer à plusieurs **conférences et colloques – en France ainsi qu'à l'international** (IV.2.c).