

ÉCOLE NATIONALE DES CHARTES

Anne Legrand

*Docteur en histoire de la musique
et musicologie*

Annotation, transcription et structuration automatisées en TEI de sources particulières par le biais de l'apprentissage automatique :

les théories musicales germaniques des premiers « temps modernes » du projet TMG (*Thesaurus Musicarum Germanicum*)

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2021

Résumé

Ce mémoire est réalisé en vue de l'obtention du diplôme de Master 2 « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes. Il a été rédigé à la suite d'un stage de trois mois au sein de l'équipe d'IReMus en lien avec le labex OBVIL, et dont le projet de recherche s'intitule *Thesaurus Musicarum Germanicarum*. Ce projet TMG étudie la théorie musicale allemande entre 1470 et 1750 en exploitant les moyens informatiques actuels. Ces théories n'ont fait l'objet d'aucune transcription ou édition électronique du fait de leur complexité théorique, linguistique et typographique. C'est pourquoi le projet TMG verse à combler ce manque avec une base de données permettant un accès intégral à ces sources. Ce mémoire reconstitue l'ensemble des étapes de traitement des traités musicaux permettant d'aboutir à la création de transcriptions annotées sous la forme de fichiers XML-TEI. Il présente de plus une analyse critique des enjeux, stratégies et résultats envisagés dans le cadre du projet TMG autant que du stage, afin de rendre compte d'un exemple de projet et de développement inscrit dans le cadre des humanités numériques.

Mots-clefs : reconnaissance automatique d'écriture ; transcription collaborative ; annotation sémantique ; OCR-D ; Transkribus ; XML-TEI ; XSLT ; TMG ; théorie musicale allemande.

Informations bibliographiques : Anne Legrand, *Annotation, transcription et structuration automatisées en TEI de sources particulières par le biais de l'apprentissage automatique : les théories musicales germaniques des premiers « temps modernes » du projet TMG (Thesaurus Musicarum Germanicum)*, mémoire de master « Technologies numériques appliquées à l'histoire », dir, Thibault Clérice, École nationale des chartes, 2021.

Remerciements

Je remercie les professeurs du master 2 « Technologies numériques appliquées à l'histoire » à l'École nationale des chartes et les étudiants de la promotion 2019-2020.

Je remercie Christophe Guillotel-Nothmann, chargé de recherche CNRS-IReMus pour le rôle de tuteur qu'il a joué durant ce stage, et l'équipe d'IReMus pour son accueil.

Je remercie les diplômés de l'ADEMEC qui, dans le contexte de crise sanitaire, ont mis leurs mémoires à disposition des étudiants du master 2 « Technologies numériques appliquées à l'histoire » via *Google Drive*.

Je remercie ma famille et mes amis pour leur précieux et indéfectible soutien durant les deux années universitaires.

Liste des sigles et abréviations

- ADEMEC : Association des Diplômés et des Étudiants des Masters de École nationale des chartes
- BnF : Bibliothèque nationale de France
- CACTUS : Corpus en diachronie, textométrie et usages
- CDL : California Digital Library
- CICM : Centre de recherche Informatique et Création Musicale
- CNRS : Centre National de Recherche Scientifique
- DARIAH : Digital Research Infrastructure for the Arts and Humanities
- DNB : Deutsche National Bibliothek
- EMT : Early Music Theory
- ENS : École Normale Supérieure
- EHESS : École des Hautes Études en Sciences Sociales
- IHRIM : Institut d’Histoire des Représentations et des Idées dans les Modernités
- GND : Gemeinsame Normdatei, équivalent à notice d’autorité en français
- GRM : Groupe de Recherches Musicales
- INA : Institut National de l’Audiovisuel
- Ircam : Institut de recherche et coordination acoustique/musique
- IReMus : Institut de Recherche en Musicologie
- Labex : Laboratoire d’excellence
- MUFI : Medieval Unicode Font Initiative
- OBTIC : Observatoire des Textes, des Idées et des Corpus
- OBVIL : Observatoire de la Vie Littéraire qui devient OBTIC en 2020.
- RISM : Répertoire International des Sources Musicales
- SMI : Saggi Musicali Italiani
- TFM : Traités Français sur la Musique
- TGN : Thesaurus of Geographic Names

- TME : Texts on Music in English
- TMG : Thesaurus Musicarum Germanicum
- TMI : Thesaurus Musicarum Italicarum
- TML : Thesaurus Musicarum Latinarum
- TRéMiR : Traité Musicaux de la Renaissance
- UMR : Unité Mixte de Recherche

*

- API : *Application Programming Interface*
- CAO : *Composition Assistée par Ordinateur*
- MEI : *Music Encoding Initiative*
- MIDI : *Musical Instrument Digital Interface*
- OCR : *Optical Character Recognition*
- OCR-D : *Optical Character Recognition Development*
- PAGE : *Page Analysis and Ground-truth Elements*
- PDF : *Portable Document Format*
- TEI : *Text Encoding Initiative*
- TXM : logiciel de *Textométrie*
- WADL : *Web Application Description Language*
- WER : *Word Error Rate*
- XML : *eXtensible Markup Language*
- XSLT : *eXtensible Stylesheet Language Transformations*
- XTF : *eXtensible Text Framework*

Table des figures

1.1	Transcription en TEI du texte de l' <i>Air nouveau</i> sur le site d'OBVIL	11
1.2	Transcription en MEI de l' <i>Air nouveau</i> sur le site NEUMA	12
1.3	Exemple de présentation de l'apparat critique de la base TRéMiR	13
1.4	Interface de consultation de <i>Syntagma Musicum</i> de Michael Praetorius	16
1.5	Le menu « Edition » de l'interface de lecture de <i>Syntagma Musicum</i> de Michael Praetorius	17
1.6	Représentation spatiale et temporelle des noms de personnes par le Geo-Browser	18
1.7	Diagramme des termes recensés par le « Thesaurus » de <i>Syntagma Musicum</i>	19
2.1	L'interface graphique de Transkribus et un essai de transcription de la page 16 du traité de Burmeister	24
3.1	Configration d'un dossier <code>data</code> du <i>repository imageAnnotationGroundTruth</i>	33
3.2	Interface graphique du logiciel TMG_ImageAnnotation et segmentation corrigée manuellement	34
3.3	Exemple des définitions de régions de la méthologie de balisage	35
3.4	Extrait du tableau de balisage de la méthodologie	36
3.5	Extrait du fichier <code>editor.py</code>	36
4.1	Affichage de la transcription dans une nouvelle fenêtre de l'interface graphique	38
4.2	Extrait de la convention de transcription des caractères latins	39
4.3	Transcription des caractères grecs	40
4.4	Extrait de la convention de transcription des caractères musicaux	40
4.5	Contenu du dossier <code>data</code> du traité de Burmeister après la transformation en tei	41
4.6	Exemple de métadonnées du <teiHeader> en tei	42

Introduction

Aujourd’hui, l’outil numérique s’est imposé dans la musicologie. Ce champ disciplinaire s’intéresse aux constructions du savoir musical à partir de documents liés à la musique tels que les partitions, les enregistrements sonores, les instruments de musique et d’autres sources annexes comme les documents d’archives ou iconographiques, les textes relatifs à la réception ou à la création, la correspondance, les écrits de musiciens ou d’interprètes, les études sociologiques, anthropologiques, ethnologiques, historiques, etc.

La musicologie se décline elle-même en sous-disciplines : la musicologie historique, la musicologie systématique, l’ethnomusicologie et l’acoustique musicale qui produisent des formes diversifiées de travaux parmi lesquels on trouve l’édition critique d’œuvres musicales, de traités et d’ouvrages pédagogiques, de la correspondance et d’autres écrits de musiciens ou théoriciens de la musique...

Il est intéressant de noter que cette discipline s’affirme tout au long du XX^e siècle comme en témoigne sa professionnalisation relativement récente avec notamment la création en 1951 de l’Institut de musicologie¹ qui deviendra l’UFR de musicologie de la Sorbonne en 1968. Durant la deuxième moitié du XX^e siècle, la musicologie s’empare de l’outil informatique. Elle l’associe à ses recherches et ses expériences avec l’analyse musicale assistée par ordinateur ou la création musicale assistée par ordinateur². On peut constater que dès la création du Centre Universitaire Expérimental de Vincennes en 1969, le Département de la musique inscrit dans le cadre d’une démarche expérimentale, la Composition assistée par ordinateur (CAO) à son enseignement, et ouvre le centre de recherche spécialisée en informatique et création musicale (CICM). Les institutions françaises vont jouer un rôle centrale dans le développement de la musicologie et des outils informatiques³. On voit apparaître au milieu des années 1970 deux institutions qui vont préserver cette place centrale de l’informatique dans la musicologie jusqu’au début des années 2000 : la création de l’Ircam par le compositeur et chef d’orchestre Pierre Boulez, et le Groupe de recherches musicales (GRM) dans le domaine du son et des musiques électroacoustiques fondé par Pierre Schaeffer en 1958 qui rejoint l’Institut national de l’audiovisuel en 1975.

Les objets de la musicologie concernés par le numérique sont aujourd’hui essentiellement les sources musicales et les textes relatifs à la musique ou aux musiciens, la musique notée (la partition), la musique enregistrée et les objets producteurs de sons employés dans un cadre musical. Pour notre stage, nous nous sommes intéressés au traitement numériques de corpus musicaux particuliers : les sources théoriques allemandes et latines de

1. Danièle Pistone, « Romain Rolland face à la musicologie de son temps », *Cahiers de Brèves*–29 (juin 2012), p. 28, URL : https://www.association-romainrolland.org/image_articles29/Pistone29.pdf (visité le 16/08/2020).

2. Marc Battier, *Musique et informatique : une bibliographie indexée*, dir. Université de Paris VIII, Réédition augmentée, couv. ill. 23 cm. Index., Ivry-sur-Seine, 1978 (Documents).

3. Christophe Guillotel-Nothmann, « Les signes musicaux et leur étude par l’informatique. Le statut épistémologique du numérique dans l’appréhension du sens et de la signification en musique », *Revue musicale OICRM*, 6–2 (2020), p. 47, URL : <https://revuemusicaleoicrm.org/rmo-vol6-n2/signes-musicaux/>

la période 1470-1750. Si l'édition électronique de partitions de musique est maintenant une pratique courante, ce corpus n'en a pourtant pas fait partie, du fait de sa singularité qui consiste à rassembler sur une même page différentes structures de langage comme du texte, des tablatures ou des portées de musique...

Christophe Guillotel-Nothmann comble ce manque avec son projet de recherche *Thesaurus Musicarum Germanicarum* (TMG) qu'il lance en 2015. Il commence à mettre à disposition sous forme de base de données une dizaine de traités musicaux et révèle à travers ces savoirs musicaux « l'imprégnation croissante de la pensée humaniste qui entraîne au 17^e siècle un bouleversement de premier ordre dans la musique occidentale »⁴. Chercheur au CNRS et à l'Institut de Recherche en Musicologie (IReMus), Christophe Guillotel-Nothmann poursuit son projet de recherche et développe un programme en python nommé « TMG_ImageAnnotation » dont le but est de générer automatiquement un balisage structuré en TEI pour une édition électronique de ces traités musicaux. Ce logiciel représente un outil informatique aux enjeux considérables pour les musicologues car ils pourront accéder depuis la plate-forme TMG à un corpus unique et inédit afin d'en prolonger son exploitation scientifique. Soutenu financièrement par le Labex Observatoire de la vie littéraire (OBVIL) en 2019, Christophe Guillotel Notthman proposait un stage de trois mois devant se dérouler en présentiel à l'IReMus durant le premier semestre 2020. La crise sanitaire nous a obligés à réaliser ce stage en télétravail. Les missions présentées dans le cahier des charges, consistaient à :

- Expérimenter des logiciels OCR et de techniques de transcription ;
- Mettre en œuvre l'édition électronique TEI/MEI ;
- Identifier et modéliser des concepts musicaux ;
- Identifier des sources musicales, textuelles et iconographiques en lien avec la source principale.

Christophe Guillotel-Nothmann étant le concepteur du logiciel TMG_ImageAnnotation, nous avons travaillé sous sa direction. L'exécution de ces missions nous a permis d'améliorer l'identification et la transcription automatisées par le logiciel des zones particulières aux traités musicaux. Nous essaierons donc dans ce présent mémoire de justifier le recours à ce logiciel conçu pour ces sources musicales spécifiques plutôt que l'utilisation d'une plate-forme de reconnaissance et de transcription automatique comme Transkribus.

Pour cela, nous contextualiserons dans une première partie le projet TMG et celui du stage. Puis, nous présenterons le corpus et le logiciel avec leurs spécificités. Dans une deuxième partie, nous exposerons les différentes étapes menant à l'identification et la transcription automatisée en TEI du corpus.

4. *Thesaurus Musicarum Germanicarum / TMG*, URL : <http://tmg.huma-num.fr/> (visité le 17/07/2020)

Première partie

Le projet TMG

Ce mémoire fait suite au stage réalisé dans le cadre du projet de base de données *Thesaurus Musicarum Germanicarum* dirigé par Christophe Guillotel-Nothmann. Chargé de recherche au CNRS, affecté à l'IReMus, il souhaite combler un manque dans l'édition électronique et commence à travailler sur un corpus d'une dizaine de traités musicaux allemands de la période qui s'étend du XV^e au XVIII^e siècle. La mise en ligne de ce premier corpus est un enjeu essentiel pour la compréhension de l'imprégnation de la pensée humaniste dans la musique occidentale par la communauté des musicologues. Nous allons retracer le projet qui conduira Christophe Guillotel-Nothmann à concevoir son logiciel TMG_ImageAnnotation. Nous le présenterons d'abord dans son contexte institutionnel.

Chapitre 1

Le projet TMG dans son contexte

1.1 IReMus

1.1.1 Présentation d'IReMus

L'IReMus¹, l'Institut de recherche en musicologue est une Unité Mixte de Recherche (UMR 8223) de Sorbonne Université qui naît de la fusion en 2004 de trois centres : l'Observatoire Musical Français (OMF), le Patrimoine et Langages Musicaux (PLM) et l'Institut de recherche sur le patrimoine musical en France (IRPMF). Placé sous les tutelles du CNRS, de la Bibliothèque nationale de France (BnF) et du ministère de la Culture, IReMus regroupe des chercheurs, des enseignants-chercheurs, des conservateurs, des ingénieurs et des doctorants. Cet établissement d'enseignement supérieur est membre du Collégium Musicæ et se situe actuellement à Paris, rue de Louvois, au premier étage du Département de la Musique de la BnF. Avec une soixantaine de membres permanents répartis en cinq équipes de recherche, cette unité est considérée comme la plus importante quantitativement au niveau national dans le domaine de la musicologie. Son champs d'études s'étend du Moyen Âge à la musique électroacoustique, en passant par le jazz et les musiques actuelles réparti en cinq axes recherches : « Éditer, restituer, valoriser les patrimoines musicaux », « Écrire sur la musique et la musicologie », « Analyser la musique », « Étudier les contextes historiques, culturels et sociaux », « Représentation et réception de la musique ».

Une partie des travaux d'IReMus participe à la valorisation du patrimoine musical conservé en France et comprenne ainsi des éditions critiques d'œuvres musicales, de traités et d'ouvrages pédagogiques, de correspondances et autres écrits de musiciens mais aussi des études de collections, de catalogues de fonds musicaux ou de catalogues thématiques, des analyses d'œuvres et de systèmes théoriques ainsi que le développement d'outils numériques parmi lesquels le logiciel TMG_ImageAnnotation. La gouvernance de l'IReMus a défini cinq axes stratégiques prioritaires pour la période 2019-2023 :

1. *IReMus*, URL : <https://www.iremus.cnrs.fr/> (visité le 31/07/2020)

- Jouer un rôle actif et utile au sein de la communauté musicologique ;
- Constituer une plate-forme pour l’interdisciplinarité ;
- Renforcer les liens avec le monde professionnel ;
- Développer une vulgarisation exigeante et de qualité, incluant notamment le renforcement du lien avec la formation ;
- Favoriser l’accompagnement et l’insertion professionnelle des jeunes chercheurs.

C’est dans ce souci de vulgarisation que l’IReMus a recours aux plates-formes d’humanités numériques qui constituent une manière de partager plus largement le savoir produit au sein de cette unité de recherche². Nous nous intéresserons d’abord à trois projets d’éditions numériques de partitions ou de documents comme la revue *Mercure galant* dont les pages comprennent articles, poèmes et airs chantés, ou à des traités musicaux avant de présenter plus particulièrement le projet TMG.

1.1.2 Les projets d’édition électronique d’IReMus

L’édition critique est l’un des domaines de spécialisation des chercheurs de l’IReMus à travers des travaux comme les éditions monumentales de l’oeuvre de Jean-Philippe Rameau (Société Jean-Philippe Rameau, distribution Bärenreiter), de Gabriel Fauré (Éditions Bärenreiter) et de Claude Debussy (Éditions Durand). Ce travail éditorial est poursuivi dans le domaine de l’édition numérique avec notamment les bases de données *Mercure galant*, NEUMA, et TRéMiR qui retiennent l’attention par leurs enjeux d’éditions électroniques similaires à ceux de TMG.

1.1.2.1 *Mercure galant*

Le *Mercure galant* est un périodique de l’époque de Louis XIV. Il est créé par Jean Donneau de Visé et paraît de 1672 à 1710 à raison d’un ou deux volumes par mois. Tous les volumes sont aujourd’hui consultables en ligne dans Gallica, la bibliothèque numérique de la BnF. Le contenu de cette revue est varié. On y trouve des comptes rendus de fêtes, de cérémonies, de spectacles, des œuvres littéraires comme des poèmes, des fables, des livrets ou lettres mais aussi de la musique et des gravures. Le projet coordonné par Anne Piéjus et Nathalie Berton-Blivet (CNRS-IReMus), propose une édition numérique des articles relatifs à la musique, à la vie musicale, au théâtre et à la critique littéraire. Initiée en 2014, cette édition est hébergée par la plate-forme du Labex OBVIL et propose une indexation fine par noms de personnes, noms de lieux, noms d’institutions, de corporations et de sociétés, ainsi que par mots-clés, dates et titres d’œuvres. Les index de noms seront intégrés à un vaste thésaurus de la France de l’Ancien Régime élaboré en collaboration avec

2. Gilles Demonet, « Partager le savoir en musicologie : un axe stratégique pour l’Institut de Recherche en Musicologie (IReMus) », *Lettre de l’InSHS*–58 (mars 2019), p. 4, URL : https://www.inshs.cnrs.fr/sites/institut_inshs/files/download-file/lettre_infoinshs_58.pdf

le Centre de recherche du château de Versailles³. Un module de cartographie historique permettra de visualiser les lieux sur un fond de carte historique ou moderne. Quant aux 717 airs de musique, ils sont référencés dans un *Catalogue des airs publiés dans le Mercure galant (1678-1700)* au sein de PHILIDOR, le portail de ressources numériques du Centre de musique baroque de Versailles⁴. Une transcription des airs en MEI est disponible sur le site NEUMA où ils sont rendus audibles par une synthèse MIDI.

FIGURE 1.1 – Transcription en TEI du texte de l'*Air nouveau* sur le site d'OBVIL.

[Air noté]
<p><i>Mercure galant, janvier 1678 [tome 1], p. 224-226.</i></p> <p>Enfin, Madame, j'ay trouvé moyen de vous satisfaire, & je vous envoie deux Airs notez que vous ne regardez, s'il vous plaist, que comme un essay de ceux que j'auray soin de vous envoyer tous les Mois. Voicy les Paroles du premier que je mets ici sans les noter, afin que vous les puissiez lire d'abord sans embarras.</p> <p style="text-align: center;">AIR NOUVEAU.</p> <p><i>TOn Troupeau, Sylvie, Peut seul t'engager. Tu passes la vie Sans prendre un Berger. Soûpire, Cruelle, Pour des soins plus beaux, Un Berger fidèle Vaut mille Troupeaux.</i></p> <p>Cet Air est d'un Maistre estimé des Personnes de la plus haute qualité, & comme elles ont le goust bon, je ne doute point que ses Ouvrages ne meritent les éloges qu'elles leur donnent ; mais vous pouvez vous en éclaircir par vous-même, jetter les yeux sur la Note. Vous sçavez parfaitement la Musique, & il ne vous faut qu'un moment pour connoistre la beauté de celle-cy.</p> <p>images/1678-01_224.JPG</p>
<p>cf. <i>Ton troupeau Sylvie, janvier 1678, in Airs du Mercure galant (1678-1710)</i></p>

On voit sur cette transcription un lien qui ouvre l'image de la partition⁵. La référence est située en bas de l'image : images/1678-01_224.JPG. L'encart à droite de la transcription de l'*Air nouveau* propose un lien vers le fichier MEI sur le site de NEUMA⁶. Un visuel permet de suivre la mélodie sur la portée : les notes s'affichant en rouge indiquent les notes entendues. Cet exemple d'édition numérique montre la difficulté d'éditer deux langages différents existants sur une même page : ici le texte et les portées musicales.

3. Centre de recherche du château de Versailles, *Thésaurus historique sur la France de l'Ancien Régime (2016-...)* Centre de recherche du château de Versailles, URL : <https://chateauversailles-recherche.fr/francais/recherche/projets-scientifiques-et-recherche-appliquee/thesaurus-historique-sur-la-france-de-l-ancien-regime-2016> (visité le 01/08/2020)

4. Centre de musique baroque de Versailles PHILIDOR équipe, *Présentation*, Base de données PHILIDOR - CMBV, URL : <http://philidor.cmbv.fr/ark:/13681> (visité le 01/08/2020)

5. Anne Piéjus, Nathalie Berton-Blivet, Alexandre De Craim, Vincent Jolivet et Frédéric Glorieux, *Mercure galant, janvier 1678 — Mercure Galant, OBVIL*, URL : https://obvil.sorbonne-universite.fr/corpus/mercure-galant/MG-1678-01#MG-1678-01_224 (visité le 31/07/2020)

6. *Ton troupeau Sylvie, janvier 1678*, Neuma V2, URL : http://neuma.huma-num.fr/home/opus/timbres:airsmercure:1678_01_224/# (visité le 31/07/2020)

FIGURE 1.2 – Transcription en MEI de l'*Air nouveau* sur le site NEUMA.

1.1.2.2 NEUMA

NEUMA est une bibliothèque numérique de partitions musicales issues des collections patrimoniales basées en France. Les transcriptions sont aux formats MusicXML et MEI. Soutenue dès 2009 par l’Agence Nationale de la Recherche (ANR), la bibliothèque a continuellement été enrichie et propose de télécharger les partitions sous le format PDF. Achille Davy-Rigaux (CNRS-IReMus) assure la responsabilité scientifique de ce projet qui permet grâce à une interface collaborative d’enrichir la collection après une demande d’inscription à la plate-forme par courrier électronique et l’obtention d’un login.

Comme pour les airs de la revue *Mercure galant*, les partitions sont transcrives dans les règles du solfège moderne et sont rendues audibles sur l’interface de lecture, permettant ainsi de suivre la partition tout en l’entendant⁷. Avec les touches du clavier virtuel placé en haut à droite de l’interface de consultation, l’utilisateur peut écrire un motif musical sur une portée puis, rechercher le motif dans une collection en cliquant sur le bouton « recherche ». Malgré cette fonctionnalité d’exploration poussée des corpus, NEUMA ne propose pas de documents superposants des langages différents dans une même page. Cette base de données demeure ainsi un outil essentiel dans le référencement en ligne des incipit musicaux et poursuit ainsi le travail initié par le département de la Musique de la BnF dès 1942 en valorisant leur catalogue d’incipit. NEUMA répond également aux espoirs formulés par Nanie Bridgman en 1959 dans le *Bulletin des bibliothèques de France* : « Il se pourrait enfin que les progrès des machines électroniques puissent offrir à notre catalogue, dans un proche avenir, d’intéressantes possibilités »⁸.

7. *Ibid.*

8. Nanie Bridgman, *Le classement par incipit musicaux*, 1^{er} janv. 1959, URL : <https://bbf.enssib.fr/consulter/bbf-1959-06-0303-002> (visité le 01/08/2020)

1.1.2.3 TRéMiR

La base de données TRéMiR coordonnée par Christophe Dupraz (École normale supérieure - IReMus) entre 2013 et 2016, est un projet d'édition en ligne et d'indexation de Traités Musicaux Romans⁹. Une vingtaine de traités musicaux italiens et espagnols de la Renaissance sont mis à disposition des chercheurs, musicologues ou autres passionnés de l'aspect musical de la pensée humaniste du XVI^e siècle. Il nous a semblé intéressant de rapprocher ce projet d'édition avec celui de TMG car les documents traités présentent des similitudes de contenu et de période.

Pour TRéMiR, les illustrations et les exemples musicaux sont insérés dans le corps du texte sous forme de fichiers images en format JPEG. Afin de montrer l'érudition des auteurs des traités et plus précisément leurs recours aux sources théoriques et aux textes philosophiques, théologiques, rhétoriques, historiques sous forme de citation, d'emprunt, d'adaptation, d'amplification et d'imitation de textes préexistants, une indexation approfondie permet d'afficher dans la zone de l'apparat critique aussi bien les noms de personnes et de lieux pour l'« Index generalis » que les titres d'œuvres qui sont mentionnés dans le traité en cours de lecture pour l'« Exempla ». Christophe Dupraz souhaite faire évoluer son index général en cumulant pour chaque entrée, les données de tous les traités.

FIGURE 1.3 – Exemple de présentation de l'apparat critique de la base TRéMiR

The screenshot shows a page from the TRéMiR database. At the top, there's a header with the project name "TRéMiR" and the author's name "Pedro Cerone". Below the header, there's a menu bar with links like "ante", "Parte 1", "Parte 2", "post", "El Melopeo y Maestro (1613)", "Libro", "Scholia", "Index generalis", "Apparatus criticus", and "Aide ?".

The main content area displays a musical score with two staves. The first staff is in F major (F#), and the second is in C major (C). Below the score, there's a caption: "Chiave di F fa ut. | Chiave di C sol fa ut.".

On the right side, there's a detailed critical apparatus (Apparatus criticus) for the score. It includes a section titled "Prima parte" with numbered notes explaining musical terms. There are also several boxes containing Latin quotations and their French translations, such as "Nature modum, per C cantare solemus : F b mollem notat ; sed G b quadrum ostendit." and "Primus, re la : Secundus, re fa : Tertius, mi fa : Quartus, mi la : Quintus, fa fa : Sextus, fa la : Septimus, ut sol : Octavus, ut fa :". Red arrows point to specific parts of the apparatus, highlighting the link between the musical notation and the critical notes.

Remarquons l'image de l'annotation musicale incluse dans la page sans aucun renvoi ou transcription MEI. Sous le bandeau supérieur de présentation du traité, le distique souligné en pointillé renvoi à l'apparat critique du traité situé dans la zone inférieure de droite. De plus un lien hypertexte entre crochets [f.3] donne accès, ici sur Gallica, au

9. *TRéMiR*, URL : <http://www.ums3323.paris-sorbonne.fr/TREMIR/index.htm> (visité le 02/08/2020)

texte original de la référence indiquée dans le traité par l'auteur, Pedro Cerone.

Tout comme Christophe Dupraz, Christophe Guillotel-Nothmann souhaite mettre en lumière l'humanisme musical présent dans les traités musicaux de la Renaissance et du début de la période moderne, grâce aux nouvelles technologies offertes de l'édition électronique permettant de valoriser le contenu et d'exploiter les données du web. Ces deux membres permanents de l'IReMus proposent ainsi en ligne un important corpus musicologique et le portent à la connaissance de la communauté des musicologues et autres spécialistes mais également d'un large public.

1.2 Le projet de recherche TMG

En 2015, Christophe Guillotel-Nothmann constate que l'édition électronique des écrits musicaux théoriques allemands du Moyen Âge à la période moderne avait été exclue du processus d'édition en ligne. En effet, ce processus commence en 1990 avec le projet Thesaurus Musicarum Latinarum¹⁰ (TML) développé par le Center for the History of Music Theory and Literature de l'Indiana University¹¹. TML donne l'impulsion à des initiatives similaires avec des sources italiennes, anglaises et françaises : Saggi Musicali Italiani¹² (SMI), Texts on Music in English¹³ (TME) et les Traités Français sur la Musique¹⁴ (TFM). Ces éditions sont dépourvues d'interventions éditoriales et d'apparat critique et le premier projet d'édition électronique à utiliser la TEI pour encoder son corpus se trouve en Europe avec le Thesaurus Musicarum Italicarum (TMI). Développée à partir de la fin des années 1990 par Frans Wiering de l'Utrecht University, cette plate-forme multimédia devient « une référence pour de nombreuses années à venir »¹⁵ dans l'édition électronique de traités.

S'inspirant de TMI et du projet TRéMiR de son collègue Christophe Dupraz¹⁶, Christophe Guillotel-Nothmann se lance alors dans la réalisation de la base de données *Thesaurus Musicarum Germanicarum* comprenant des écrits de la sphère germanique, situés en marge de ce processus d'édition électronique en raison de leur complexité théo-

10. *Thesaurus Musicarum Latinarum*, URL : <https://chm1.indiana.edu/tml/> (visité le 06/08/2020)

11. En 2015, TML est victime d'une cyberattaque et doit fermer sa base de données. Deux ans plus tard, une nouvelle édition est mise en ligne, exploitant les normes d'encodage de la TEI et MEI.

12. *Saggi musicali italiani*, URL : <https://chm1.indiana.edu/smi/> (visité le 06/08/2020)

13. *Texts on Music in English*, URL : <https://chm1.indiana.edu/tme/> (visité le 06/08/2020)

14. *Introduction to Traités français sur la musique*, URL : <https://chm1.indiana.edu/tfm/tfmintro.html> (visité le 06/08/2020)

15. Eleanor Selfridge-Field, *Computers and music*, Grove Music Online, URL : <https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040583> (visité le 17/07/2020)

16. Christophe Guillotel-Nothmann, *Thesaurus Musicarum Germanicarum et la "Law of the Stimulative Arrears" ? / TMG*, URL : <http://tmg.huma-num.fr/fr/content/christophe-guillotel-nothmann-thesaurus-musicarum-germanicarum-et-la-law-stimulative-arrears> (visité le 07/08/2020)

rique, linguistique, mais aussi typographique et n'ayant fait l'objet d'aucune exploitation musicologique systématique par le biais des technologies numériques.

1.2.1 L'édition électronique de *Syntagma Musicum*, volume 3, de Michael Praetorius : une première étape du projet

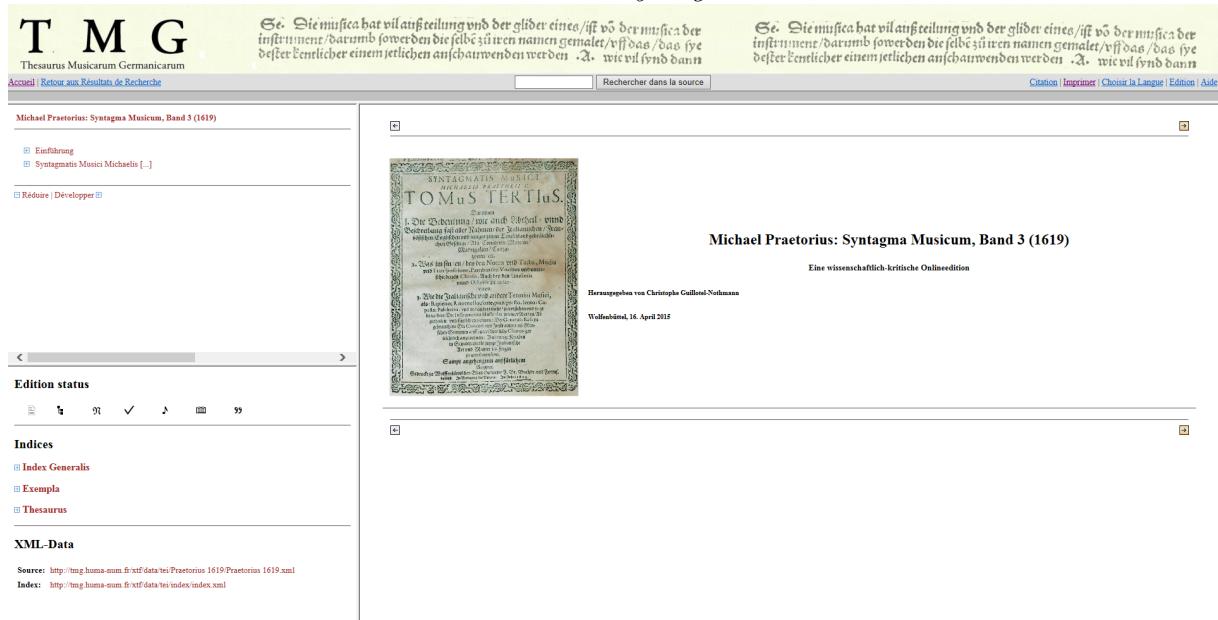
Financé en 2015 par la Herzog August Bibliothek de Wolfenbüttel en Allemagne, ce projet répond à deux défis posés par Christophe Guillotel-Notthmann : premièrement, les défis techniques d'une édition électronique poussée en raison du contenu spécifique des pages comprenant des tableaux, des exemples musicaux ou autres caractères spéciaux et deuxièmement, les défis scientifiques de mise en valeur de la pensée musicale théorique de la fin de la Renaissance et du début du Baroque. En effet, l'œuvre choisie est centrale dans l'histoire de la musicologie en ce qu'elle montre les influences et la réception de la théorie et de la pratique musicales italiennes de cette époque dans les pays germaniques. D'ailleurs, on peut constater que Michael Praetorius, compositeur, organiste et maître de chapelle à la cour de Wolfenbüttel à la fin du XVI^e siècle, écrit et publie en 1919 ce troisième volume de *Syntagma Musicum* en allemand et non en latin.

1.2.1.1 Méthodologie éditoriale

L'interface de consultation du traité est composée de quatre éléments :

- L'en-tête contient dans le menu différents onglets : « Citation » offre la possibilité d'une recherche plein texte, un lien pour une citation de l'ouvrage, « Imprimer » ouvre une version imprimée de tout le traité, « Choose Language » permet le choix de la langue de l'interface entre l'anglais, allemand et le français et « Edition » propose de choisir les interventions éditoriales ;
- En haut à gauche, la table des matière pouvant se dérouler ;
- En bas à gauche, l'apparat critique reposant comme pour TRéMiR sur trois index ainsi que les liens vers le code source de l'édition ;
- Au centre, affichage du texte de la source avec accès au fac-similé électronique, aux exemples musicaux et hyperliens dans le texte.

L'interface de lecture propose donc dans la fenêtre principale le contenu de la source avec une pagination indiquée en haut de chaque page et complétée entre crochets si nécessaire. L'icône en forme de livre placé en haut et à gauche de chaque page, ouvre une nouvelle fenêtre qui donne l'accès au facsimile mis en ligne sur le site de la bibliothèque de Wolfenbüttel. L'encodage au format XML-TEI permet d'afficher les différentes strates de modifications de l'édition électronique. La typographie, les abréviations et la ponctuation ont été normalisées. Pour les afficher, il suffit de sélectionner les options correspondantes dans le menu « Edition ». Par défaut, seuls les changements de ligne des passages en

FIGURE 1.4 – Interface de consultation de *Syntagma Musicum* de Michael Praetorius

vers sont respectés mais l’option « Keep line breaks » affiche les retours à la ligne de l’édition originale. La fenêtre « Edition » offre aussi la possibilité d’afficher les corrections résultant de l’erratum de l’auteur ou celles de l’éditeur. Les caractères surlignés en rouge indiquent la modification apportée par l’édition électronique et les caractères surlignés en gris indiquent ceux de la source.

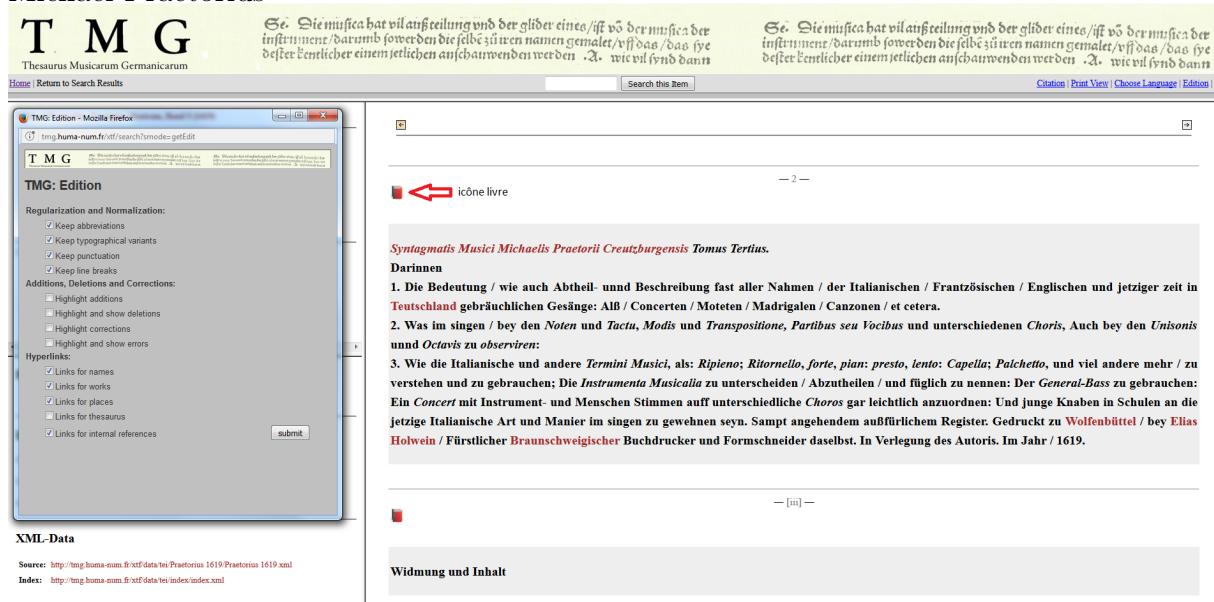
1.2.1.2 Développement de l’apparat critique du projet TMG

Christophe Guillotel-Nothman s’inspire de la méthodologie de l’apparat critique développé par Christophe Dupraz pour sa base de données TRéMiR¹⁷. Christophe Guillotel-Nothman en reprend l’« Index generalis » et l’« Exempla », ajoute le « Thesaurus » et propose un balisage différencié selon les règles de la TEI pour toutes les entrées des index. Il est possible d’empêcher la fonctionnalité des liens générés en les décochant dans le menu « Edition ».

1. L’« **Index generalis** » regroupe les noms de personnes, d’oeuvres et de lieux et permet d’ouvrir les liens de chaque occurrence vers la source. Chaque entrée située dans l’index ou dans la source, grâce à un lien hypertexte s’identifiant par la couleur rouge foncée des caractères, ouvre dans une nouvelle fenêtre, des notices contenant des informations prosopographiques, bibliographiques et géographique. Ces informations proposent des liens vers diverses ressources en ligne comme GND (Gemein-

17. C. Guillotel-Nothmann, « Ressources numériques pour l’étude de la théorie musicale de l’époque moderne », *Revue de musicologie*, 106–2 (2020), p. 467-481

FIGURE 1.5 – Le menu « Edition » de l’interface de lecture de *Syntagma Musicum* de Michael Praetorius

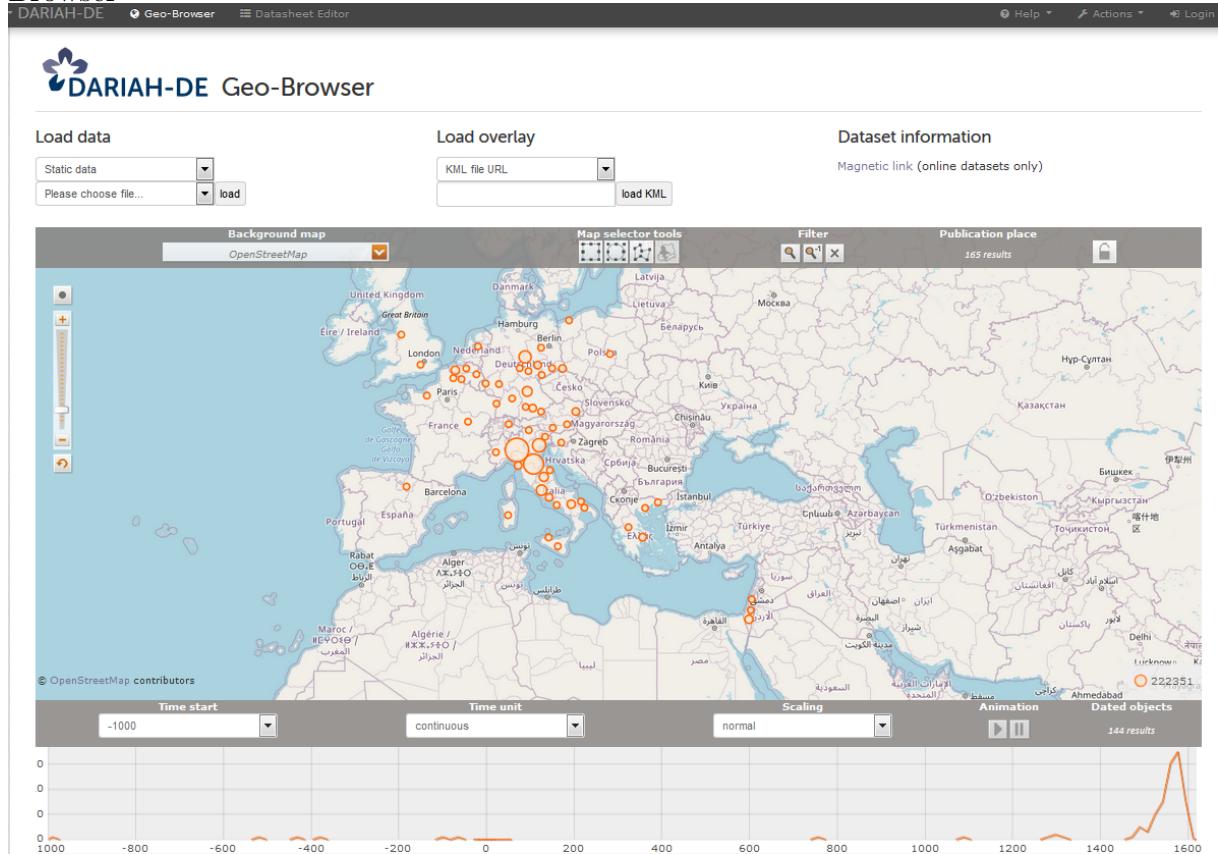


same Normdatei¹⁸), Wikipédia¹⁹, Grove²⁰, le catalogue de la DNB (Deutsche NationalBibliothek²¹), GeoNames²², OldMapsOnline²³, TGN (Thesaurus Geographic Names²⁴)²⁵. Afin de présenter le contexte historique et géographique de manière plus vivante, en cliquant sur l’« Index generalis », une carte générée par le Geo-Browser du DARIAH-DE (Digital Research Infrastructure for the Arts and Humanities²⁶) s’ouvre dans une nouvelle fenêtre. Elle représente de façon spatiale et temporelle les noms de personnes recensés dans le *Syntagma Musicum* et montre que Praetorius a consulté des ouvrages de musique provenant essentiellement d’Italie bien que lui-même n’y ait jamais voyagé.

2. L’« **Exempla** » répertorie les citations qu’elles soient textuelles, musicales ou iconographiques, et en donne les références bibliographiques tout en proposant de les

- 18. *Gemeinsame Normdatei (GND)*, Deutsche Nationalbibliothek, URL : https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html (visité le 19/08/2020)
- 19. *Wikipedia – Die freie Enzyklopädie*, URL : <https://de.wikipedia.org/wiki/Wikipedia:Hauptseite> (visité le 19/08/2020)
- 20. *Oxford Music*, Oxford Music Online, URL : <https://www.oxfordmusiconline.com/> (visité le 19/08/2020)
- 21. *Startseite*, Deutsche Nationalbibliothek, URL : https://www.dnb.de/DE/Home/home_node.html (visité le 19/08/2020)
- 22. *GeoNames*, URL : <https://www.geonames.org/> (visité le 19/08/2020)
- 23. Klokan Technologies GmbH (<https://www.klokantech.com/>), *Old Maps Online*, URL : <https://www.oldmapsonline.org> (visité le 19/08/2020)
- 24. *Getty Thesaurus of Geographic Names (Getty Research Institute)*, URL : <http://www.getty.edu/research/tools/vocabularies/tgn/> (visité le 19/08/2020)
- 25. *Michael Praetorius : Syntagma Musicum, Band 3.(1619)*, URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Praetorius%201619/Praetorius%201619.xml&chunk.id=div_1&toc.id=div_1&brand=default (visité le 07/08/2020)
- 26. *Startseite - DARIAH-DE*, URL : <https://de.dariah.eu/> (visité le 19/08/2020)

FIGURE 1.6 – Représentation spatiale et temporelle des noms de personnes par le Geo-Browser



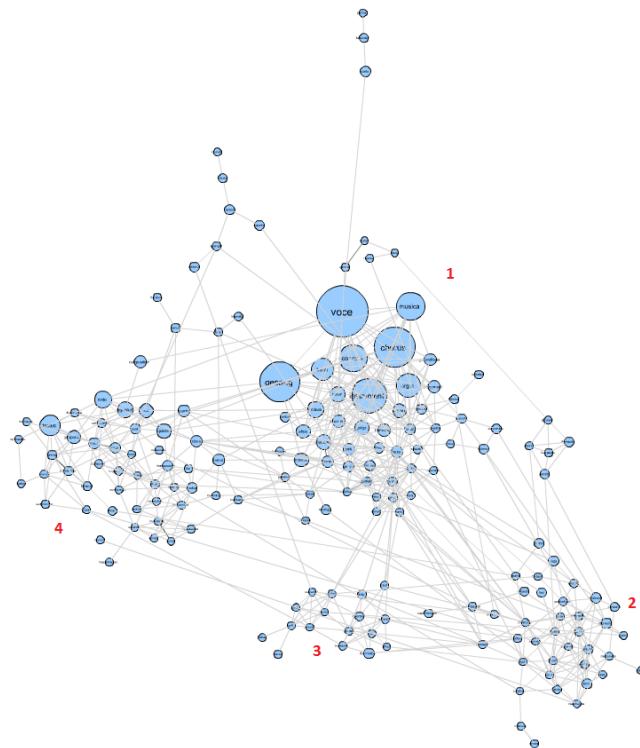
consulter grâce à un lien permettant d'ouvrir le document source disponible en ligne. Trois types de citations sont distingués : les citations explicites et implicites ou les paraphrases. L'« Exempla » propose également une carte générée par le Geo-Browser de Dariah-DE représentant de façon spatiale et temporelle les auteurs des citations.

3. Le « **Thesaurus** » regroupe tous les termes musicaux en cochant au préalable l'option dans le menu « Edition ». Réalisé par l'application yEd²⁷, un diagramme s'affiche en cliquant sur l'onglet « Thesaurus ». Il recense les terminologies musicales dont il identifie quatre pôles : 1) les noms d'instruments et ce qui se rapporte à la voix, 2) les genres, 3) les ornements et 4) la théorie.

1.2.2 Enrichissement de la base de données TMG

Christophe Guillotel-Nothmann fédère une quarantaine de musicologues, historiens, informaticiens, linguistes et spécialistes du livre rattachés essentiellement à des institutions allemandes ou françaises mais aussi belges et néerlandaises et qui s'intéressent comme

^{27.} *yEd Graph Editor*, yWorks, the diagramming experts, URL : <https://www.yworks.com/products/yed> (visité le 19/08/2020)

FIGURE 1.7 – Diagramme des termes recensés par le « Thesaurus » de *Syntagma Musicum*

lui à la mise à disposition de sources comme les théories musicales des premiers temps modernes.

1.2.2.1 Une nouvelle équipe et des logiciels adaptés à l'enrichissement de la base TMG

De septembre 2015 à juin 2016, Christophe Guillotel-Nothmann bénéficie d'un financement du programme Idex (Initiatives d'excellence) de PERSU (Sorbonne Universités Pour l'Enseignement et la Recherche) qui lui permet de recruter trois assistants-ingénieurs et d'enrichir sous sa direction scientifique la base de données TMG. Trois éditions électroniques pilotes viennent augmenter la base TMG et reprennent les champs d'investigation développés dans *Syntagma Musicum* : Anaëlle Le Royer réalise l'édition de *Musica getutscht* (1511)²⁸; Pauline Spychala celle de *Musica* (1507)²⁹ et Adrien Tannhof celle d'*Excerpta musicæ* (c. 1496)³⁰.

L'accès à la base de données se fait par l'interface de recherche où quatre onglets proposent en haut de page quatre options : une recherche avancée qui s'affiche par défaut, une recherche par mots-clés ou une recherche libre ainsi qu'une découverte des contenus

28. *Musica getutscht* 1511, URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Virdung_1511/Virdung_1511.xml (visité le 22/08/2020)

29. *Musica* 1507, URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Cochlaeus_1507/Cochlaeus_1507.xml (visité le 22/08/2020)

30. *Excerpta musicæ* (1496), URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Anonymus_1496/Anonymus_1496.xml;brand=default; (visité le 22/08/2020)

par navigation³¹.

La gestion et l'interrogation de la base de données TMG reposent sur l'infrastructure XTF³², développée par la Californian Digital Library (CDL). TXM³³ complète les moyens informatiques utilisés dans TMG et permet une analyse textométrique des termes musicaux rassemblés dans le « Thesaurus ». Si le texte est encodé au format XML-TEI, les exemples musicaux le sont au format XML-MEI. Le logiciel Verovio³⁴ permet la visualisation des parties musicales et l'infrastructure NEUMA en permet l'interrogation.

Sous l'apparat critique de l'interface de consultation, les fichiers textes au format XML-TEI sont mis à disposition dans la rubrique « XML Data »³⁵. Le code source au format XML-MEI est accessible sous les exemples musicaux et un fichier MIDI le rend audible.

1.2.2.2 Vers une transcription automatisée : TMG_tagger

À terme, le projet TMG souhaite présenter un corpus de 340 sources³⁶ qui contiennent du texte, des images, des tablatures, des tables ou des exemples musicaux. Le volume et la complexité du corpus ne sont pas compatibles avec un encodage détaillé de toutes les sources dans un délai raisonnable, c'est-à-dire inférieur à cinq années.

TMG_tagger a été conçu pour réduire le temps nécessaire à l'encodage des textes et pour aider les chercheurs en sciences humaines à créer avec un minimum de connaissances techniques un document TEI en fonction de leurs besoins. Le logiciel est écrit en Java 8. Il permet un balisage semi-automatique des documents par le biais d'algorithmes de recherche, de normalisation et de dictionnaires de personnes, d'œuvres, de lieux et de termes musicaux (10.000 entrées au mois de mai 2016).

TMG_tagger porte sur la microstructure des documents (mots, signes de ponctuation, sauts de page, passages à la ligne). La nécessité d'améliorations se fait rapidement sentir et il devient nécessaire de développer les fonctionnalités du logiciel afin de permettre la création assistée de l'élément <teiHeader> et la prise en compte par le balisage puisse prendre en compte des niveaux hiérarchiques supérieurs à la phrase.

Le logiciel est disponible sous la forme d'une licence CC BY-NC-SA 3.0. La demande se fait à Christophe Guillotel Nothmann par courrier électronique³⁷.

31. *TMG : Search Form*, URL : <http://tmg.huma-num.fr/xtf/search> (visité le 20/08/2020)

32. *XTF*, URL : <https://xtf.cdlib.org/> (visité le 19/08/2020)

33. *Projet Textométrie*, URL : <http://textometrie.ens-lyon.fr/> (visité le 20/08/2020)

34. *Verovio*, URL : <https://www.verovio.org/> (visité le 19/08/2020)

35. Cf. figure 1.4.

36. *Imprimés / TMG*, URL : <http://tmg.huma-num.fr/fr/corpus> (visité le 10/08/2020)

37. *TMG_tagger / TMG*, URL : <http://tmg.huma-num.fr/fr/content/tmgtagger> (visité le 10/08/2020)

Chapitre 2

Choix des sources et des logiciels

Pour enrichir la base de données TMG, un réel besoin s'exprime autour de la création d'un nouvel outil informatique facilitant l'encodage et l'édition de régions non textuelles comme les tables, les tablatures, les schémas et les exemples musicaux. Nous présenterons d'abord les sources avec lesquelles nous avons expérimenté le nouveau logiciel TMG_ImageAnnotation et entraîné un réseau de neurones pour l'identification automatisée de ces régions.

2.1 Les sources

Pour le stage, nous avons travaillé sur deux traités conservés par la Staatsbibliothek de Berlin¹ qui en propose une version numérisée libre de droits depuis sa base de « Collections numérisées »².

2.1.1 *Cursus Philosophici Encyclopaedia* de Johann Heinrich Alsted

Johann Heinrich Alsted est un théologien et philosophe protestant calviniste allemand. Son oeuvre est importante car il est l'un des premiers à publier des encyclopédies au début du XVII^e siècle dont *Cursus Philosophici Encyclopaedia* en 1620. Nous nous sommes intéressés au livre 14 consacré à la musique, intitulé *Encyclopaedia Liber Decimusquartibus, In quo Musica* et situé entre les pages 812 et 830 de la numérisation³. Les éléments particuliers aux théories de la musique comme les portées ou les tables sont présents au sein du texte de ce livre écrit en latin.

1. Startseite / Staatsbibliothek zu Berlin, URL : <https://staatsbibliothek-berlin.de/> (visité le 20/08/2020)

2. S. B. B. Developers, Digitalisierte Sammlungen der Staatsbibliothek zu Berlin, URL : <https://digital-beta.staatsbibliothek-berlin.de> (visité le 20/08/2020)

3. Johann Heinrich Alsted, *Cursus Philosophici Encyclopaedia : Libris XXVII; Complectens Universae Philosophiae methodum, serie praeceptorum, regularum & commentariorum perpetua ...* 1620, URL : <http://resolver.staatsbibliothek-berlin.de/SBB0001D6E800010000> (visité le 15/07/2020)

2.1.2 *Hypomnematum musicae poeticae... synopsis* de Joachim Burmeister

Le deuxième traité avec lequel nous avons travaillé, est consacré uniquement à la musique. Il est publié à l'aube du XVII^e siècle en 1599 et s'intitule *Hypomnematum musicae poeticae... synopsis*⁴. L'auteur Joachim Burmeister est un humaniste, compositeur et théoricien de la musique allemand. Il est né en 1564 à Lunebourg où il étudie avec le Cantor Christophe Praetorius, oncle de Michael Praetorius⁵. Burmeister entre à l'université de Rostock en 1589 et va y enseigner avant de devenir Cantor. Il publie trois ouvrages sur la musique dont le premier est *Hypomnematum musicae poeticae... synopsis*. Les suivants en seront le développement et les musicologues les étudient et les citent aujourd'hui comme une importante contribution à l'étude de la rhétorique musicale. *Hypomnematum musicae poeticae... synopsis* est écrit en latin et afin de frapper et imprégner les esprits des lecteurs, le traité contient de nombreux tableaux, schémas, tables, tablatures ou synopsis auxquels s'ajoutent des portées musicales, des signes musicaux et des caractères grecs ou allemands gothiques.

2.2 Les logiciels

Pour la continuation du projet TMG, Christophe Guillotel-Nothmann s'est posé des questions de méthodologie et a dû opérer des choix de logiciels. Fallait-il poursuivre l'investigation de la conception d'un logiciel comme TMG_tagger adapté aux sources particulières que représentent les traités de musique de la période des temps modernes ou avoir recours à un logiciel préexistant comme Transkribus ?

2.2.1 Transkribus

Transkribus est une plate-forme de reconnaissance de texte, d'analyse d'images et de reconnaissance de structure de documents historiques qui prend en charge la transcription et l'annotation des documents⁶. Elle est la poursuite de deux projets européens :

4. Joachim Burmeister, *HYPOMNE-//MATVM MVSICAE // POETICAЕ.// A // M. IOACHIMO BVRMEISTERO,// ex Isagoge, cuius et idem ipse auctor est,// Ad Chorum gubernandum, cantumque // componendum conscriptâ,// SYNOPSIS.//*, 1599, URL : <http://resolver.staatsbibliothek-berlin.de/SBB0001DA0D00000000> (visité le 15/07/2020)

5. J. Burmeister, Agathe Sueur et Pascal Dubreuil, *Musica poetica*, Google-Books-ID : zyHhc0Y-seukC, 2007 (Amicus, Renaissance et période préclassique. Domaine Germanique : 1), p. 8

6. Pour cette partie sur la plate-forme Transkribus, je m'inspire et reprend le travail d'Alix qu'elle présente dans son mémoire et dans un atelier de l'ADEMEC le 21 janvier 2020 à l'École nationale des chartes. Alix Chagué, *Constituer un corpus pour la fouille de texte - de la transcription des documents d'archives à l'annotation : exploration d'une méthodologie par l'ANR Time Us*, mémoire de master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2008, URL : https://github.com/alix-tz/M2TNAH_memoire-de-stage.git

tranScriptorium (2013-2015)⁷ et de READ (*Recognition and Enrichment of Archival Documents*)⁸, une infrastructure financée par la Commission Européenne dans le cadre du programme Horizon 2020. Dirigée par l'université d'Innsbruck, l'objectif de la plate-forme Transkribus est de mettre en place des outils pour améliorer l'accès au contenu des objets archivistiques. Le 1^{er} juillet 2019, READ devient une société coopérative européenne (SCE) et READ-COOP devient la seule infrastructure maintenant et développant la plate-forme Transkribus.

Transkribus est disponible en anglais, en allemand et en italien. Ses principales fonctionnalités de traitement automatique sont la reconnaissance optique de caractères (OCR), la reconnaissance d'écriture manuscrite (HTR), l'analyse de la structure (*Layout Analysis*) et le repérage de mots (*Word Spotting*). Le logiciel HTR n'étant pas *open source*, un paiement est demandé au-delà de la transcription de 400 pages manuscrites ou de 2500 pages imprimées.

Pour utiliser la plate-forme, l'utilisateur doit créer un compte et ses données seront synchronisées avec le serveur de Transkribus. Le travail peut être privé et collaboratif sur un ensemble de corpus de textes. L'utilisateur propriétaire (*owner*) d'une collection peut inviter d'autres utilisateurs, en leur attribuant un rôle de propriétaire ou de lecteur (*reader*). Seul l'utilisateur propriétaire peut modifier la collection en ajoutant, modifiant ou supprimant des documents.

L'interface graphique de Transkribus se compose de cinq zones.

1. Une **barre de menu** (A) en haut de la fenêtre permet d'accéder aux fonctionnalités principales, notamment le menu principal, la gestion du profil d'utilisation, l'import et l'export de fichiers ou encore l'actualisation des pages.
2. Le **canevas** (B), où s'affiche l'image à transcrire. Le canevas permet de réaliser manuellement la segmentation du texte ou de l'éditer. La barre de menu sur le côté gauche est complétée d'outils de navigation au sein de la sous-collection et de boutons de zoom.
3. L'**éditeur de texte** (C), où s'affiche la transcription du texte. Chaque ligne d'une zone de texte est numérotée et correspond à un segment de l'image. Une barre de menu en bas permet de modifier le style du texte (texte en gras, italique, souligné ou barré, en exposant ou en indice, etc).
4. Les **onglets** (D) dans le panneau latéral gauche permettent de réaliser un très grand nombre de tâches : 1) gérer l'accès à la collection et à son contenu (*Server*), 2) avoir un aperçu du statut de chaque élément dans la sous-collection en cours de consultation (*Overview*), 3) gérer la structure et la segmentation de la page en cours de consultation (*Layout*), 4) gérer les métadonnées de la page en cours (*Metadata*),

7. *TranScriptorium*, URL : <http://transcriptorium.eu/> (visité le 31/08/2021)

8. *About us*, READ-COOP, URL : <https://readcoop.eu/about/> (visité le 31/08/2021)

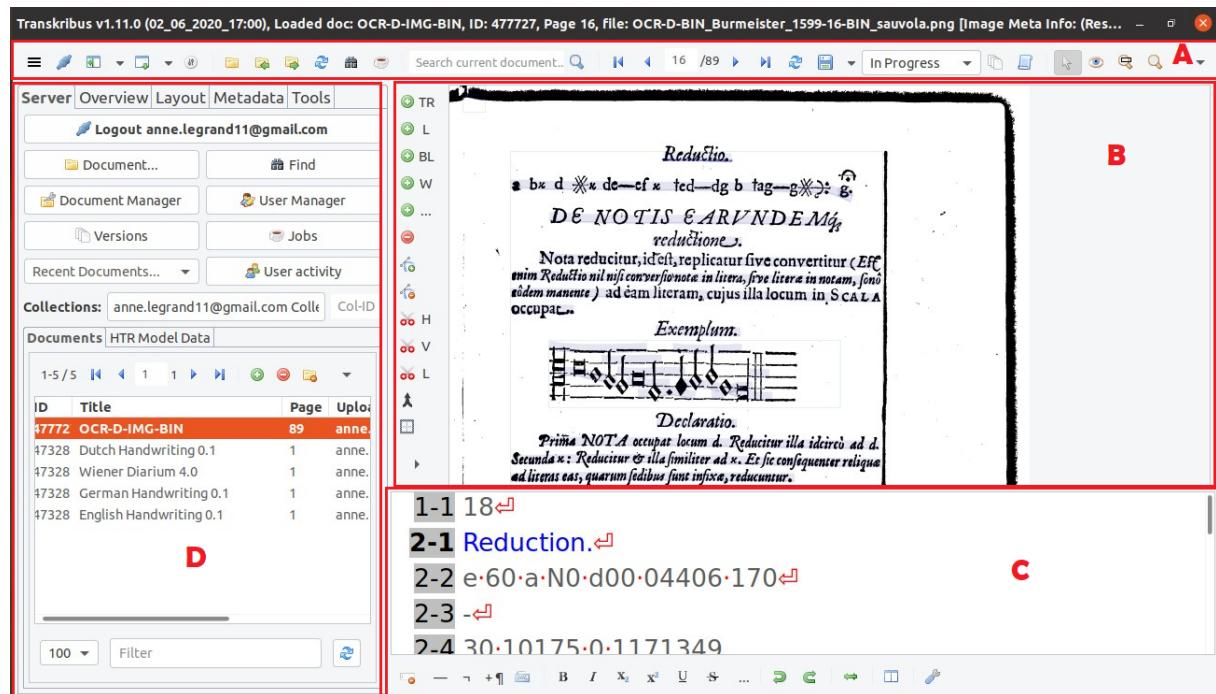


FIGURE 2.1 – L’interface graphique de Transkribus et un essai de transcription de la page 16 du traité de Burmeister.

ou encore, 5) accéder aux différents outils de traitement automatique disponibles (*Tools*). Parmi ces outils, on trouve « *Layout Analysis* », pour le repérage automatique des zones et lignes de texte, « *Text Recognition* », pour la reconnaissance automatique du texte (HTR), ou encore « *Compute Accuracy* », pour l’analyse automatique des taux d’erreur d’une transcription automatique. Chacun de ces cinq onglets principaux dispose d’onglets ou d’options avancées.

La plate-forme propose également d’exporter la totalité ou une portion d’une sous-collection sous plusieurs formes :

- Sous forme de paquets liant fichiers de métadonnées (standard METS⁹), fichiers de texte au format XML (standard ALTO¹⁰ ou PAGE¹¹) et images (JPG).
- Sous forme de PDF, avec ou sans l’alignement de la transcription avec l’image, avec ou sans coloration des annotations.
- Sous forme de fichiers Word (DOCX), avec la possibilité de développer ou non les

9. *Metadata Encoding Transmission Schema*, un standard créé pour conserver les métadonnées et la structure hiérarchique d’objets faisant partie d’une collection numérique, ainsi que les liens vers ces objets.

10. *Analyzed Layout and Text Object*, un standard pour le stockage des données techniques de description de la structure d’un document ayant fait l’objet d’un OCR. Il est généralement articulé avec un fichier METS.

11. *Page Analysis and Ground Truth Elements*, un format pour le stockage des données de description de la structure d’un document OCRisé, ainsi que la transcription (idéale) associée à chacune des zones de texte ; *PRIImA/tools/PAGELibraries*, URL : <https://www.primaresearch.org/tools/PAGELibraries> (visité le 30/08/2020).

abrégations, de conserver ou non les coupures de lignes, et d'inclure ou non les éléments d'annotation comme les passages incertains.

- Sous forme de fichiers XML selon le standard de la TEI¹² avec la possibilité de paramétrier certains aspects du fichier de sortie, parmi lesquels la mise en forme des lignes (`<1b/>` ou `<1>...</1>`) ou encore le choix d'intégrer ou non les coordonnées des zones et lignes de texte.

Quel que soit le format d'export, Transkribus permet de paramétrier les fichiers de sortie en précisant les pages du document courant à exporter.

Mais la transcription proposée par Transkribus pour le traité de Burmeister n'est pas suffisante. On voit dans l'exemple proposé¹³ que le logiciel ne reconnaît que peu de caractères. Le document comprenant trop de caractères musicaux insérés dans le texte ou sur des portées, des tables, des schémas ou des tablatures, nécessite un long travail préparatoire de création de données d'apprentissage pour un résultat satisfaisant d'annotation et de transcription. De plus, l'accès à Transkribus est réservé à des utilisateurs initiés même si la plate-forme propose des guides¹⁴ fonctionnant comme des tutoriels à suivre étape par étape et devant faciliter l'initiation.

Christophe Guillotel-Nothmann s'oriente donc vers une solution « maison » pouvant annoter des zones particulières mais également entraîner un réseau de neurones permettant leur identification automatisée afin de faciliter l'utilisation du logiciel par les musicologues.

2.2.2 TMG _ ImageAnnotation

TMG _ ImageAnnotation est le nouveau logiciel conçu par Christophe Guillotel-Nothmann en langage de programmation Python. Le logiciel prend les données du fichier METS et stocke les annotations de l'image au format PAGE XML. TMG _ ImageAnnotation intègre les huit modules développés par l'OCR-D¹⁵ et améliore notamment l'étape de balisage des pages qui demeure critique dans les processus d'OCR pour les sources spécifiques comme les traités de musique. Le logiciel affiche les images de la source, puis délimite des zones de texte comme les paragraphes, les titres, les légendes, les lettrines, les graphiques, les exemples musicaux... Ces zones sont prédéfinies par le logiciel mais peuvent être modifiées par l'utilisateur. TMG _ ImageAnnotation est un logiciel *open source* et peut être téléchargé depuis le compte Github de Christophe Guillotel-Nothmann¹⁶.

12. *Text Encoding Initiative*, un standard pour la représentation numérique de textes.

13. Cf. figure 2.1.

14. *How-to Guides*, READ-COOP, URL : <https://readcoop.eu/transkribus/resources/how-to-guides/> (visité le 31/08/2021)

15. *Module Projects - OCR-D*, URL : <https://ocr-d.de/en/module-projects> (visité le 07/07/2020)

16. C. Guillotel-Nothmann, *TMG _ ImageAnnotation*, URL : <https://github.com/guillotel-nothmann/imageAnnotation.git> (visité le 30/06/2020)

Deuxième partie

Des images aux fichiers XML

Après avoir présenté le corpus et le logiciel utilisés pendant le stage, nous nous concentrerons sur les principaux moyens permettant l'amélioration de l'identification et la transcription automatisées des zones particulières aux traités musicaux par le biais de l'entraînement du réseau de neurones du logiciel TMG_ImageAnnotation. Nous exposerons successivement les différentes étapes sur lesquelles nous nous sommes concentré durant le stage. Nous commencerons par l'annotation des fac-similés électroniques en dégageant les problèmes rencontrés et les améliorations trouvées. Nous aborderons ensuite les phases de transcription et de transformation en fichiers XML-TEI permettant une édition électronique de ces traités et les rendre accessibles sur la base de données TMG.

Chapitre 3

L'annotation des facs-similés électroniques

TMG_ImageAnnotation nous apporte un environnement d'annotation et de consultation spécifique et adapté à nos sources, permettant de lancer le réseau de neurones mais également de visionner aisément les résultats proposés, de les corriger et de les sauvegarder rapidement. La première partie du stage a consisté à installer et vérifier le bon fonctionnement du logiciel avec un système d'exploitation Ubuntu. Nous nous attarderons sur la présentation de TMG_ImageAnnotation et ses fonctionnalités avant de présenter le travail d'annotation réalisé.

3.1 L'interface graphique et les commandes du logiciel TMG_ImageAnnotation

3.1.1 L'interface graphique

Le logiciel *open source* est donc disponible sur le compte Github de Christophe Guillotel-Nothmann. Pour exécuter les *scripts*, il est nécessaire d'installer un environnement virtuel basé sur **Python 3** et dans lequel les *packages* suivants doivent être téléchargés au préalable :

- **numpy** (diminutif de *Numerical Python*) : interface pour stocker et effectuer des opérations sur les données.¹
- **lxml** : parseur pour les fichiers XML et HTML en Python.²

1. Plongez en détail dans la librairie NumPy, OpenClassrooms, URL : <https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/4740941-plongez-en-detail-dans-la-librairie-numpy> (visité le 17/06/2020)

2. lxml - Processing XML and HTML with Python, URL : <https://lxml.de/> (visité le 17/06/2020)

- **matplotlib** : permet de tracer et visualiser des données sous forme de graphes.³
- **scikit-image** : collection d'algorithmes pour le traitement d'images.⁴ item **tensorflow** (version 1.6.0) : outil d'apprentissage automatique ou *machine learning*.⁵ item **keras** (version 2.1.3) : permet d'interagir avec les algorithmes de réseaux de neurones profonds et d'apprentissage automatique, notamment Tensorflow.⁶ item **h5py** (version 2.9) *Hierarchical Data Format, V5* : format de type conteneur de fichier.⁷

Tous ces *packages* sont listés dans le *repository* intitulé « imageAnnotation » sur le compte Github de Christophe Guillotel-Nothmann où se trouvent également les instructions d'installation. Ensuite dans le terminal, on ouvre le dossier `imageAnnotation` puis ceux nommés `ImageAnnotation/src` et on lance le logiciel avec `python3 main.py`. Il est possible de s'exercer avec aux commandes du logiciel en ouvrant le fichier `mets.xlm` dans le dossier `annotationExample/data`. La deuxième étape consiste à télécharger le *repository* intitulé `imageAnnotationGroundTruth`⁸. Les facs-similés électroniques se situent dans ce *repository* dans des dossiers au nom des auteurs des traités musicaux avec la date de la première édition de la source. Les noms des dossiers s'affichent de la façon suivante : `Burmeister_1599`. Ces dossiers sont composés d'un dossier `data` qui comprend le fichier `mets.xlm` et sera ouvert avec le logiciel, un dossier `OCR-D-IMG-BIN` comprenant les fichiers `.png` et un dossier `OCR-D-SEG-REGION` pour les fichiers `.xml`.

Le fichier `mets.xlm` établit le lien entre le fichier `.xml` au standard PAGE qui contient les coordonnées des zones identifiées, et le fichier `.png` correspondant à une image de la source à partir de laquelle sera réalisée la transcriptions.

L'interface graphique est divisée en deux parties :

- **La barre des tâches** : Elle est située en haut de l'interface et contient les commandes pour ouvrir ou quitter le fichier `mets.xlm`, sauvegarder le travail effectué, afficher l'image suivante ou précédente. Elle contient également des boutons permettant de sélectionner une zone encore non identifiée tout en renseignant en même temps son nom (« Paragraph », « Heading », « Caption »...).
- **La fenêtre de l'image** : L'image s'affiche dans la fenêtre centrale, en dessous de la barre des tâches.

3. *Matplotlib : Python plotting — Matplotlib 3.4.3 documentation*, URL : <https://matplotlib.org/> (visité le 17/06/2020)

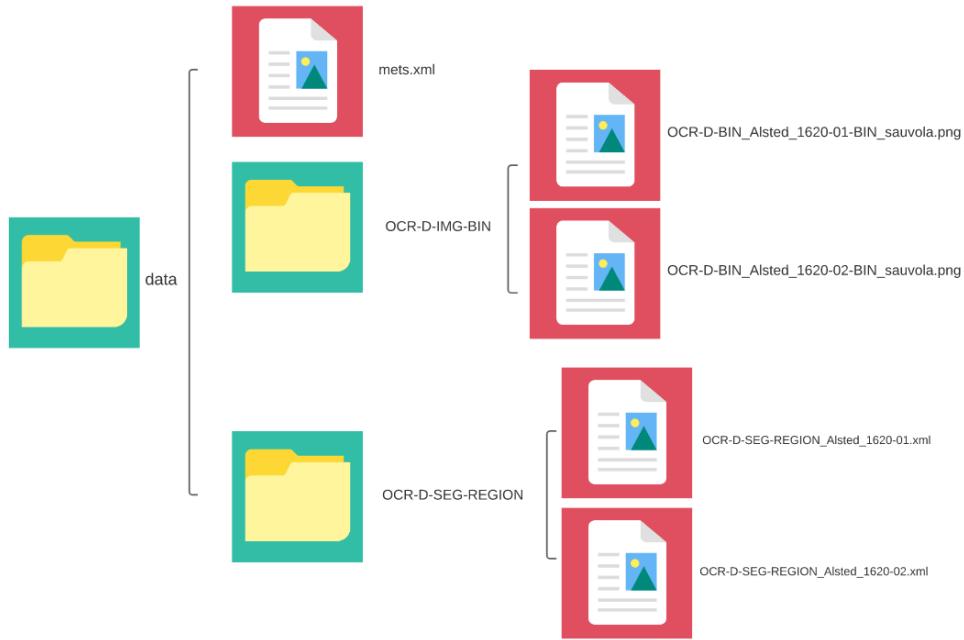
4. Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart et Tony Yu, « scikit-image : image processing in Python », *PeerJ*, 2 (19 juin 2014), e453, DOI : [10.7717/peerj.453](https://doi.org/10.7717/peerj.453)

5. *TensorFlow*, TensorFlow, URL : <https://www.tensorflow.org/?hl=fr> (visité le 17/06/2020)

6. *Keras : the Python deep learning API*, URL : <https://keras.io/> (visité le 17/06/2020)

7. *HDF5 for Python*, URL : <https://www.h5py.org/> (visité le 17/06/2020)

8. C. Guillotel-Nothmann, *imageAnnotationGroundTruth*, juin 2020, URL : <https://github.com/guillotel-nothmann/imageAnnotationGroundTruth.git>

FIGURE 3.1 – Configuration d'un dossier data du *repository imageAnnotationGroundTruth*

3.1.2 Les commandes du logiciel TMG _ ImageAnnotation

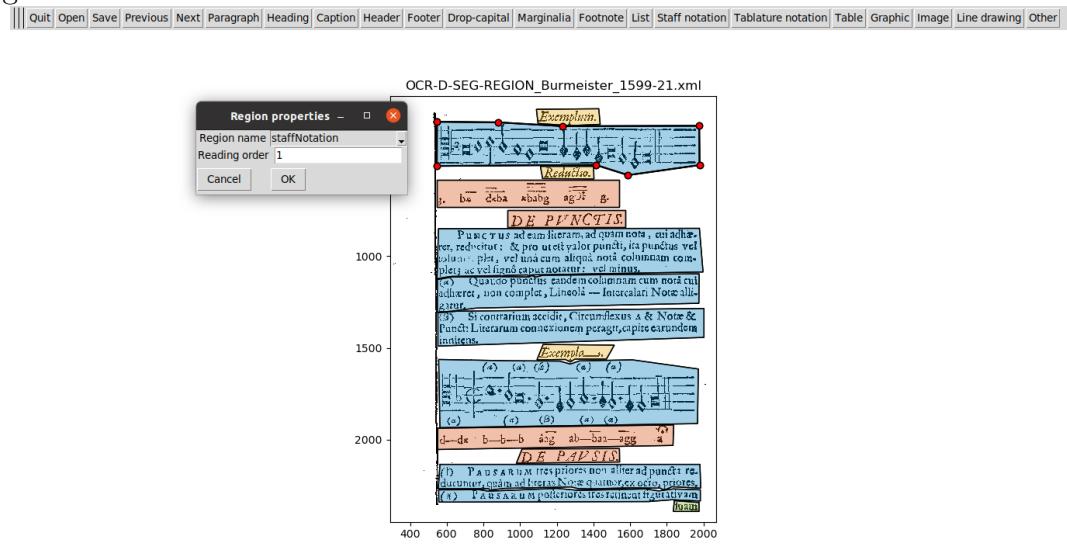
Les commandes pour **Open**, **Quit**, **Save**, **Previous** et **Next** peuvent être exécutées à partir du clavier :

- **Open** : alt+o
- **Quit** : alt+q
- **Save** : alt+s
- **Previous** : <
- **Next** : >

Quand on ouvre un fichier **mets.xml**, une première image s'affiche avec une segmentation des différentes zones réalisée par le logiciel. Cette première étape de pré-traitement des fac-similés de la source fait partie du processus de la reconnaissance optique de caractères ou *optical character recognition* (OCR). Pour améliorer le balisage de ces régions, mon travail consistait à les supprimer ou les rectifier à l'aide des commandes ci-dessous pouvant être réalisées également à partir du clavier :

- **Open** : alt+o
- **Quit** : alt+q
- **Save** : alt+s
- **Previous** : <
- **Next** : >
- **Ajouter un point cardinal** : clic sur la ligne et i

FIGURE 3.2 – Interface graphique du logiciel TMG_ImageAnnotation et segmentation corrigée manuellement



— Supprimer un point cardinal : clic sur le point et d

Le clic droit de la souris est utilisé pour afficher en rouge les points cardinaux de la sélection. Puis en tapant + , une petite fenêtre s'ouvre et propose les informations de la zone comprenant un numéro indiquant sa position par rapport aux autres régions ainsi que son nom⁹. Afin d'éviter que les zones se superposent, une commande permet d'ajouter des points cardinaux sur l'encadrement de la zone. On déplace ensuite un point cardinal afin d'ajuster l'encadrement de la région. Toutes les commandes sont listées dans le fichier READ.me du *repository imageAnnotation*. Le choix des régions peut s'effectuer par les boutons de la barre des tâches ou par une combinaison de touches du clavier. Ces options sont également listées dans le fichier READ.me.

Le travail d'ajustement du balisage réalisé, Christophe Guillotel-Nothmann a pu exercer le réseau de neurones et améliorer la segmentation et l'identification automatique des régions par le logiciel.

3.2 Vers une amélioration du logiciel et un ajustement des zones

3.2.1 Amélioration de certaines commandes

L'utilisation du logiciel a permis d'identifier des difficultés d'utilisation de certaines commandes. Pour supprimer une région, il fallait supprimer successivement les points cardinaux un par un. Les commandes répétitives étaient difficiles et laborieuses. Lorsqu'

9. Cf. figure 3.2. La numérotation des régions commence à 0.

un point cardinal restait en dessous d'une autre région, il devenait invisible et empêchait l'ouverture des fenêtres informatives des autres régions. Christophe Guillotel-Nothmann a donc ajouté une commande au logiciel permettant de supprimer une région dans son intégralité avec une seule manipulation, ce qui a permis de ne plus rencontrer ce problème. Une autre amélioration a été réalisée : la possibilité de se déplacer d'une région à une autre avec les flèches déplaçant le curseur vers le haut ou le bas, ce qui a facilité la vérification et la numérotation des zones. J'ai également constaté une difficulté avec la commande de zoom qui se bloquait à la sélection de la zone à agrandir et provoquait une interruption anormale de l'ordinateur. Une fois identifié, le problème a été résolu.

3.2.2 Ajustement des zones

L'ajustement des zones comprend la correction de la segmentation des régions mais également l'assimilation des termes et de leur usage correct pour chaque zone. Seize régions

FIGURE 3.3 – Exemple des définitions de régions de la méthologie de balisage

Drop-capital

Les lettres capitales sont identifiées dans une région spécifique quand elles s'étendent sur plusieurs lignes et se démarquent graphiquement.

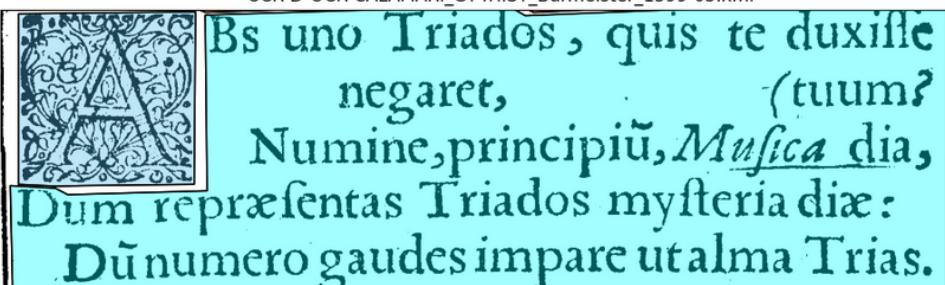
- Exemple 10: Lettre capitale sur plusieurs lignes.



Scala, est quinque linearum musicalium, & inter illos
comprehensorum spaciorum congeries.

- Exemple 11: Lettre capitale ornementée.

OCR-D-OCR-CALAMARI_GT4HIST_Burmeister_1599-05.xml



Bs uno Triados, quis te duxisse
negaret, (tuum?
Numine, principiū, *Musica* dia,
Dum repræsentas Triados mysteria diæ:
Dū numero gaudes impare ut alma Trias.

ont été déterminées et la région *List* a été ajoutée pour définir des listes numérotées. Chaque région a une couleur différente sauf les régions *Paragraph* et *staffNotation* qui ont la même couleur bleue.

La complexité de la structure des textes nous a conduit à réfléchir sur la bonne dénomination de chaque région et leur correcte localisation. En effet, certains caractères musicaux sont insérés dans le texte. Nous avons décidé qu'ils ne représentaient pas une région particulière et donc de les intégrer à la région *Paragraph* sélectionnant le texte. L'indentation des paragraphes a été respectée sauf pour les paragraphes continus sur deux pages qui ont été identifiés dans ce cas comme deux régions *Paragraph* différentes. La région *staffNotation* détermine les exemples musicaux sur portées mais également

la notation par lettres. Des réflexions se sont portées sur la dénomination des légendes des exemples musicaux sur portées, des tables, des tablatures et des images. Elles ont été désignées par la région *Caption*. Il a fallu assimiler la différence entre les régions *tablatureNotation* et les regions *Table*.

Ce travail m'a conduit à rédiger la méthodologie du balisage, disponible dans le fichier `READ.me` du repository `imageAnnotationGroundTruth`¹⁰. Cette méthodologie comprend un tableau récapitulant le nom des balises utilisées et propose leur équivalent au format XML-PAGE ainsi que des exemples pour chaque région et leurs particularités. L'équivalent au format XML-PAGE a été obtenu en faisant correspondre le nom des

FIGURE 3.4 – Extrait du tableau de balisage de la méthodologie

Classe	Page XML	exemples
Caption	<code><pc:TextRegion type="caption"></code>	Exemple 3 Exemple 4 Exemple 5 Exemple 6
Diagram	<code><pc:ChartRegion ></code>	Exemple 6
Drop-capital	<code><pc:TextRegion type="drop-capital"></code>	Exemple 10 Exemple 11
Footer	<code><pc:TextRegion type="footer"></code>	Exemple 9
Footnote	<code><pc:TextRegion type="footnote"></code>	

`regionClass` avec le nom de chaque balise. Le nom des `regionClass` se situe dans le code paramétrant ces `regionClass` dans le fichier `editor.py` du dossier `src` du logiciel. Par exemple, la région `Drop-capital` sera encodée en XML avec la balise `TextRegion`

FIGURE 3.5 – Extrait du fichier `editor.py`

```

if self.regionClass == "TextRegion":
    if self.type in ["paragraph", "caption", "header", "heading",
"footer", "drop-capital", "marginalia", "footnote", "page-number"]:
        self.regionName = self.type

    elif self.type == "other":
        if self.custom == "list":
            self.regionName = self.custom

    elif self.custom == "linegroup":
        self.regionName = self.custom

```

de type `drop-capital` : `<pc:TextRegion type="drop-capital">`. En dessous du tableau, toutes les régions sont définies en proposant de nouveau un exemple pour chaque définition¹¹.

10. *Ibid.*

11. Cf. figure 3.3.

Chapitre 4

Transcription et transformation en XML-TEI

Le travail de relecture et de correction du balisage de la phase de pré-traitement étant réalisé pour le traité de Burmeister, nous pouvions passer à la reconnaissance des caractères et à l'extraction du texte par le logiciel TMG_ImageAnnotation pour commencer la phase de post-traitement consistant à corriger les erreurs d'interprétation. Nous nous intéresserons d'abord à l'évolution de l'interface graphique du logiciel permettant une relecture d'une première transcription avant l'étape de la transformation du fichier au format XML-TEI.

4.1 Transcription et relecture

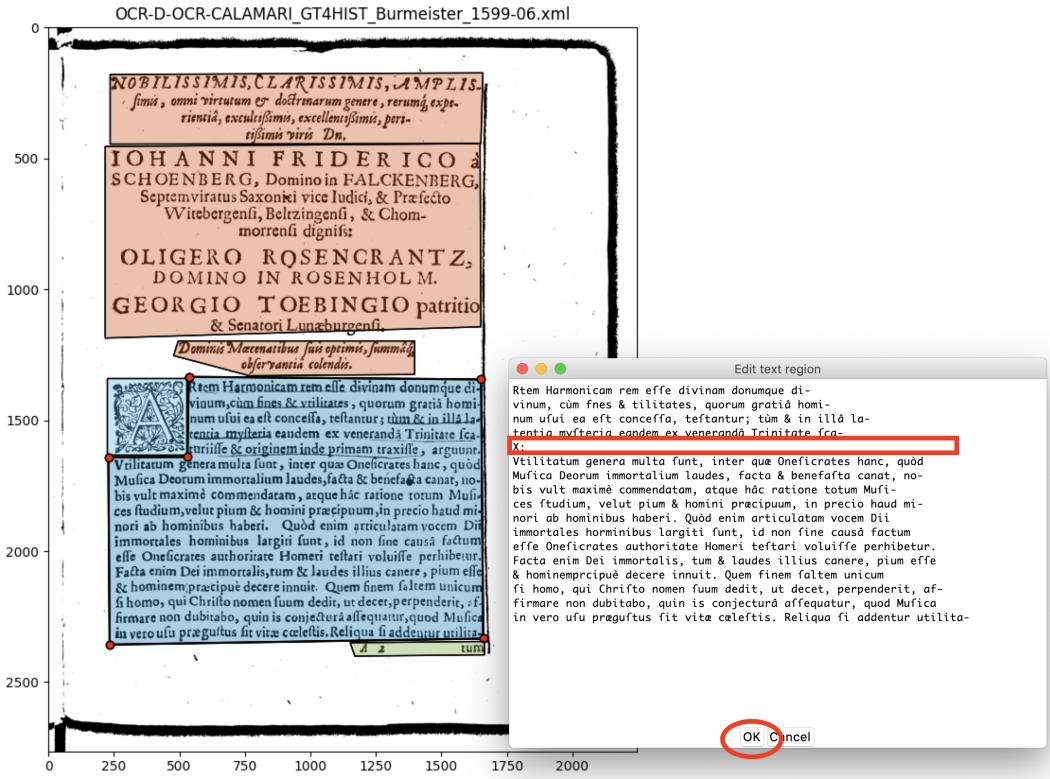
Christophe Guillotel-Nothmann a modifié le logiciel pour afficher dans une nouvelle fenêtre sa transcription. J'ai pu ainsi la relire et la corriger pour arriver à une version la plus fidèle possible à la source.

4.1.1 Une nouvelle fenêtre à l'interface graphique

Pour afficher la transcription, après avoir ouvert le fichier `mets.xlm` du traité, il suffit de sélectionner une région comportant du texte avec le clic gauche de la souris puis de taper sur la touche `#` du clavier. Une nouvelle fenêtre s'ouvre alors avec le texte transcrit par le logiciel. Dans l'exemple suivant, la transcription est relativement propre malgré quelques coquilles et une ligne qui n'a pas été transcrise. La transcription sera donc réalisée manuellement. Quand la relecture et les corrections d'une région sont terminées, il suffit de valider le travail réalisé en cliquant sur le bouton `OK` de la fenêtre et passer à une région suivante. Un nouveau dossier nommé `OCR-D-OCR-CALAMARI_GT4HIST` s'est créé au moment de l'OCR contenant les nouveaux fichiers nommés sur ce modèle `OCR-D-OCR-CALAMARI_GT4HIST_Burmeister_1599-02.xml` pour chacune des pages

du traité. Les données de nos corrections sont donc sauvegardées sur ces fichiers qui comprennent maintenant, non seulement les coordonnées des zones identifiées, mais aussi les transcriptions post-traitées.

FIGURE 4.1 – Affichage de la transcription dans une nouvelle fenêtre de l’interface graphique



4.1.2 Vers une convention de la transcription

À la relecture de la transcription, j’ai relevé des erreurs récurrentes du logiciel sur certains caractères :

- Les lettres « ra » sont souvent transcris par le logiciel comme un « m ».
- La lettre « h » est souvent transcrit par le logiciel comme un « b ».
- Les lettres « is » est souvent transcris par le logiciel comme un « s ».
- Le point virgule est souvent transcrit par le logiciel comme un « 5 ».
- La lettre « q » est souvent transcrit par le logiciel comme un « g ».

J’ai également remarqué que la transcription du logiciel perdait information de l’italique. Ce problème a été relevé mais pas corrigé durant le stage. D’autres problèmes sont apparus avec les caractères latins, grecs, allemands accompagnés ou non de diacritiques, ainsi qu’avec les caractères musicaux et ceux des tablatures intégrés dans le texte d’une région Paragraph. Malgré l’installation de polices de caractères comme *Andron Scriptor Web v.*

3 et *Symbola*, certains caractères grecs et latins ne s'affichaient pas correctement avec le système d'exploitation Ubuntu. Christophe Guillotel-Nothmann a proposé que je réalise une liste des caractères basée sur le travail réalisé par OCR-D¹ afin de pouvoir par la suite l'intégrer au logiciel. J'ai réalisé cette liste en m'a aidant également de la documentation disponible sur la plate-forme MUFU (Medieval Unicode Font Initiative)². La liste propose une colonne pour le caractère, une colonne pour le caractère codé en hexadécimal³, une colonne pour la transcription choisie quand le caractère ne s'affichait pas et une colonne pour l'image du caractère.

FIGURE 4.2 – Extrait de la convention de transcription des caractères latins

Character	Unicode hexadecimal	Transcription	image
á	á		
â	â		
æ	æ		
Æ	Æ		
č	č		
ê	ê		
&	&		
ì	ì		
í	í		
ñ	í̃		
ò	ò		
ó	ó		
ô	ô		
œ	œ		
ꝑ	́	[q+aigu]	ꝑ
ꝑ	q;		ꝑ
ꝑ	q́;	[q:+aigu]	ꝑ

1. Zentrum Sprache Matthias Boenig OCR-D, *Alphabets, Abbreviations and Special Characters*, URL : <https://ocr-d.de/en/gt-guidelines/trans/trFremdsprache.html> (visité le 05/08/2020)

2. Tarrin Wills, « The Medieval Unicode Font Initiative », *Medieval Unicode Font Initiative* (, 19 févr. 2016), URL : <https://skaldic.abdn.ac.uk//m.php> (visité le 05/08/2020)

3. système de numérotation permettant la conversion avec le système binaire des ordinateurs

Cette convention de transcription est disponible dans le fichier `READ.me` du *repository imageAnnotationGroundTruth*⁴, en dessous de la méthodologie de balisage. Elle est suivie par un tableau des caractères grecs comprenant une colonne pour le caractère, et une colonne pour le caractère codé en hexadécimal. Leur affichage n'a pas posé de problème et ils ont été transcrits en les insérant après le préfixe « GR » mis entre crochets :

FIGURE 4.3 – Transcription des caractères grecs

: [GR γάμψα].

Le même modèle a été reproduit pour les caractères musicaux, en insérant le préfixe « MUS » avant la convention de transcription choisie. Comme pour les caractères latins, le tableau contient quatre colonnes pour les caractères, l'Unicode, la transcription et l'image. Pour l'Unicode hexadécimal, nous nous sommes reportés aux plates-formes *babelmap*⁵ et *alt-codes*⁶.

FIGURE 4.4 – Extrait de la convention de transcription des caractères musicaux

Character	Unicode hexadecimal	Transcription	image
†	†		
#	#		
♩	𝆶	[Mus maxima]	
♪	𝆷	[Mus longa]	
♫	𝆸	[Mus brevis]	
♩	𝆹	[Mus sebrevis]	

4. C. Guillotel-Nothmann, *imageAnnotationGroundTruth...*

5. *BabelMap Online (Unicode 13.0)*, URL : <https://www.babelstone.co.uk/Unicode/babelmap.html> (visité le 03/08/2020)

6. *Music Note Symbols*, URL : https://www.alt-codes.net/music_note_alt_codes.php (visité le 05/08/2020)

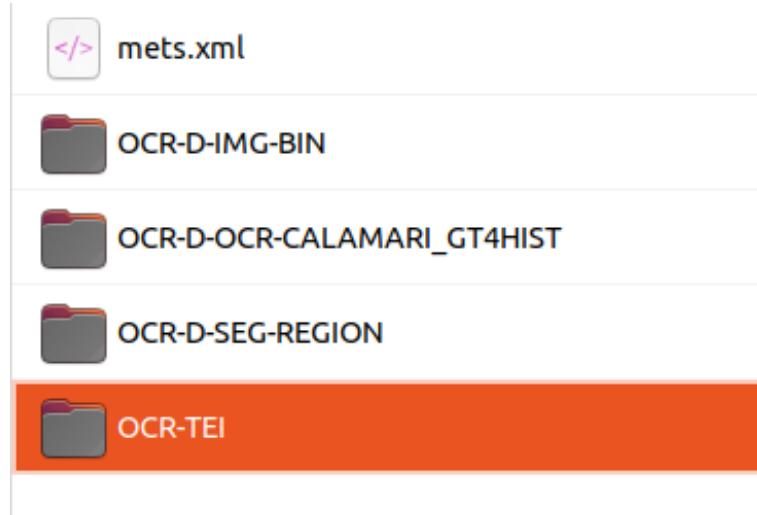
4.2 Transformation au format XML-TEI

Après cette convention de transcription des caractères rencontrant des problèmes d'affichage, Christophe Guillotel-Nothman a souhaité que j'associe le scénario de transformation du fichier nommé `page2tei` au fichier `mets.xml` afin de commencer l'extraction des zones créées en TEI.

4.2.1 Le scénario de transformation `page2tei`

Ce scénario a été conçu par Dario Kampkaspar. Ce chercheur formé à l'Université d'Heidelberg a travaillé à la Herzog August Bibliothek de Wolfenbüttel où il a dirigé différents projets d'humanité numérique dont le DARIAH-DE⁷. Il est aujourd'hui développeur à l'Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) de Vienne. Son scénario est un fichier *open source* et il est téléchargeable depuis le compte Github de Dario Kampkaspar dans un *repository* intitulé `page2tei`⁸. Il est intéressant de noter que Matthias Boenig de la Berlin-Brandenburg Academy of Sciences and Humanities, également membre du projet OCR-D, est l'un des collaborateurs de `page2tei`⁹. J'ai réalisé cette transformation avec le logiciel Oxygen en ouvrant le fichier `mets.xml` puis j'ai configuré le fichier `page2tei-0.xsl` comme scénario de transformation. Un fichier XML-TEI se crée alors, enregistré sous le nom `Burmeister_1599` dans un nouveau dossier `OCR-TEI` du dossier `data`. Ce dernier comprend maintenant le fichier `mets.xml` et quatre dossiers :

FIGURE 4.5 – Contenu du dossier `data` du traité de Burmeister après la transformation en tei.



7. Dario Kampkaspar, URL : <https://www.oeaw.ac.at/acdh/team/current-team/dario-kampkaspar-1> (visité le 30/08/2020)

8. Dario Kampkaspar, `dariok/page2tei`, 30 août 2020, URL : <https://github.com/dariok/page2tei>

9. D. Kampkaspar et Matthias Boenig, `page2tei-0`, 30 août 2020, URL : <https://github.com/dariok/page2tei/blob/master/page2tei-0.xsl>

L'élément `<facsimile>` du nouveau fichier `Burmeister_1599` au format XML-TEI regroupe et reprend les informations de tous les fichiers au format XML-PAGE. À l'intérieur de `<facsimile>`, l'élément `<surface>` représente chaque page du traité dans lequel se décline l'élément `<graphic>` reprenant la référence du fichier `.png` et l'élément `<zone>` comprenant le texte et le référencement de chaque point cardinal par région. À la fin de l'élément `<facsimile>` commence celui du `<text>` que nous n'avons pas pu traiter pendant le stage.

4.3 L'aborescence du `<teiHeader>`

Nous avons commencé à travailler sur la description des métadonnées contenues dans le `<teiHeader>`. Tous les champs ne peuvent être remplis automatiquement et n'obéissent pas à un modèle standard de métadonnées. Par exemple, Christophe Guillotel-Nothmann souhaite que l'élément `<edition>` de l'élément `<editionStmt>` du `<fileDesc>` reporte les informations de l'éditeur scientifique et non celles de la maison d'édition de la source. L'automatisation ne peut pas produire ces informations. La transcription structurée produite par l'automatisation servira de point de départ à une édition scientifique. La construction du `<teiHeader>` a commencé d'être revue et complétée : une intervention manuelle destinée a corrigé le contenu ou a ajouté des informations.

FIGURE 4.6 – Exemple de métadonnées du `<teiHeader>` en tei.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">Joachim Burmeister: Hypomnematum musicae poeticae
          ... synopsis. (1599)</title>
        <author><forename full="yes">Joachim</forename>
          <surname full="yes">Burmeister</surname></author>
        <editor><forename full="yes">[namePart]</forename><surname
          full="yes">[namePart]</surname></editor>
      </titleStmt>
      <editionStmt>
        <edition>Semiautomatic identification of text zones, diagrams and
          musical examples from the TMG_ImageAnnotation
          project.</edition>
      <respStmt>
        <resp>Transcription based on OCR-D optical recognition
          modules. Page-TEI conversion based on page2Tei
          transformation scenario.</resp>
      <name xml:id="zoneIdentification" full="yes" instant="false"
        ><ref type="ext"
          target="https://github.com/guillotel-nothmann/imageAnnotation"
          >TMG_ImageAnnotation</ref></name>
      <name xml:id="ocr" full="yes" instant="false"><ref
        type="ext" target="https://ocr-d.de/"
        >OCR-D</ref></name>
      <name xml:id="page2Tei" full="yes" instant="false"><ref
        type="ext"
        target="https://github.com/guillotel-nothmann/page2tei"
        >page2Tei</ref></name>
    </fileDesc>
  </teiHeader>
</TEI>
```

Conclusion

Le présent mémoire s'est attaché à retracer les différentes missions qui m'ont été confiées durant le stage. Elles m'ont permis de découvrir l'IREMUS et les différents projets d'édition électronique développés par les chercheurs. Elles m'ont également permis d'expérimenter un nouveau logiciel créé en adéquation avec le projet TMG tout en participant à son évolution et à la découverte des problèmes rencontrés. J'ai ainsi pu enrichir mes connaissances au niveau des recherches liées à l'OCR et à l'automatisation de la transcription.

Mes missions associées à la phase d'annotation et de transcription ont produit la méthodologie de balisage et les conventions de transcriptions permettant d'améliorer les données d'entraînement pour une annotation automatique. Ces missions ont constitué une étape cruciale de ce stage. Un aspect reste à améliorer pour la progression de TMG : le recours à une plate-forme collaborative facilitant la transcription participative. Reposant sur la contribution de personnes extérieures au projet, elle permettrait de fédérer des chercheurs d'horizons différents avec des connaissances dans des domaines variés dont les humanités numériques.

Bibliographie

Les sources théoriques allemandes et latines

ALSTED (Johann Heinrich), *Cursus Philosophici Encyclopaedia : Libris XXVII ; Complectens Universae Philosophiae methodum, serie praceptorum, regularum & commentariorum perpetua ...* 1620, URL : <http://resolver.staatsbibliothek-berlin.de/SBB0001D6E800010000> (visité le 15/07/2020).

BURMEISTER (Joachim), *HYPOMNE-//MATVM MVSICAE // POETICAE.// A // M. IOACHIMO BVRMEISTERO,// ex Isagoge, cuius et idem ipse auctor est,// Ad Chorum gubernandum, cantumque // componendum conscriptâ,// SYNOPSIS.//*, 1599, URL : <http://resolver.staatsbibliothek-berlin.de/SBB0001DA0D00000000> (visité le 15/07/2020).

Musicologie et outil informatique

- BATTIER (Marc), *Musique et informatique : une bibliographie indexée*, dir. Université de Paris VIII, Réédition augmentée, couv. ill. 23 cm. Index., Ivry-sur-Seine, 1978 (Documents).
- BRIDGMAN (Nanie), *Le classement par incipit musicaux*, 1^{er} janv. 1959, URL : <https://bbf.enssib.fr/consulter/bbf-1959-06-0303-002> (visité le 01/08/2020).
- BURMEISTER (Joachim), SUEUR (Agathe) et DUBREUIL (Pascal), *Musica poetica*, Google-Books-ID : zyHhcoYseukC, 2007 (Amicus, Renaissance et période préclassique. Domaine Germanique : 1), p. 8.
- DEMONET (Gilles), « Partager le savoir en musicologie : un axe stratégique pour l’Institut de Recherche en Musicologie (IReMus) », *Lettre de l’InSHS*–58 (mars 2019), p. 4, URL : https://www.inshs.cnrs.fr/sites/institut_inshs/files/download-file/lettre_infoinshs_58.pdf.
- Grm*, URL : <https://inagrm.com/fr> (visité le 14/07/2020).
- GUILLOTEL-NOTHMANN (Christophe), « Les signes musicaux et leur étude par l’informatique. Le statut épistémologique du numérique dans l’appréhension du sens et de la signification en musique », *Revue musicale OICRM*, 6–2 (2020), p. 47, URL : <https://revuemusicaleoicrm.org/rmo-vol6-n2/signes-musicaux/>.
- IRCAM*, URL : <https://www.ircam.fr/> (visité le 14/07/2020).
- IReMus*, URL : <https://www.iremus.cnrs.fr/> (visité le 31/07/2020).
- Musique - UFR Arts, philosophie, esthétique*, URL : <https://www-artweb.univ-paris8.fr/?-Musique-> (visité le 14/07/2020).
- PISTONE (Danièle), « Romain Rolland face à la musicologie de son temps », *Cahiers de Brèves*–29 (juin 2012), p. 28, URL : https://www.association-romainrolland.org/image_articles29/Pistone29.pdf (visité le 16/08/2020).
- SELFridge-FIELD (Eleanor), *Computers and music*, Grove Music Online, URL : <https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040583> (visité le 17/07/2020).

Exemples de projets d'édition électronique d'IReMus

PHILIDOR (Centre de musique baroque de Versailles équipe), *Présentation*, Base de données PHILIDOR - CMBV, URL : <http://philidor.cmbv.fr/ark:/13681> (visité le 01/08/2020).

PIÉJUS (Anne), BERTON-BLIVET (Nathalie), DE CRAIM (Alexandre), JOLIVET (Vincent) et GLORIEUX (Frédéric), *Mercure galant, janvier 1678 — Mercure Galant, OBVIL*, URL : https://obvil.sorbonne-universite.fr/corpus/mercure-galant/MG-1678-01#MG-1678-01_224 (visité le 31/07/2020).

Ton troupeau Sylvie, janvier 1678, Neuma V2, URL : http://neuma.huma-num.fr/home/opus/timbres:airsmercure:1678_01_224/# (visité le 31/07/2020).

TReMiR, URL : <http://www.ums3323.paris-sorbonne.fr/TREMIR/index.htm> (visité le 02/08/2020).

VERSAILLES (Centre de recherche du château de), *Thésaurus historique sur la France de l'Ancien Régime (2016-...)* Centre de recherche du château de Versailles, URL : <https://chateauversailles-recherche.fr/francais/recherche/projets-scientifiques-et-recherche-appliquee/thesaurus-historique-sur-la-france-de-l-ancien-regime-2016> (visité le 01/08/2020).

Éditions électroniques de traités musicaux

EMT : Early Music Theory, URL : <https://earlymusictheory.org/> (visité le 06/08/2020).

GUILLOTEL-NOTHMANN (Christophe), « Ressources numériques pour l'étude de la théorie musicale de l'époque moderne », *Revue de musicologie*, 106–2 (2020), p. 467-481.

Introduction to Traités français sur la musique, URL : <https://chmtl.indiana.edu/tfm/tfmintro.html> (visité le 06/08/2020).

Saggi musicali italiani, URL : <https://chmtl.indiana.edu/smi/> (visité le 06/08/2020).

Texts on Music in English, URL : <https://chmtl.indiana.edu/tme/> (visité le 06/08/2020).

Thesaurus musicarum italicarum, URL : <https://euromusicology.cs.uu.nl/> (visité le 06/08/2020).

Thesaurus musicarum italicarum Web, URL : <https://tmiweb.science.uu.nl/> (visité le 06/08/2020).

Thesaurus Musicarum Latinarum, URL : <https://chmtl.indiana.edu/tml/> (visité le 06/08/2020).

Thesaurus Musicarum Germanicarum

Christophe Guillotel-Nothmann, Thesaurus Musicarum Germanicarum et la "Law of the Stimulative Arrears" ? / TMG, URL : <http://tmg.huma-num.fr/fr/content/christophe-guillotel-nothmann-thesaurus-musicarum-germanicarum-et-la-law-stimulative-arrears> (visité le 07/08/2020).

Excerpta musice (1496), URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Anonymus_1496/Anonymus_1496.xml;brand=default ; (visité le 22/08/2020).

Imprimés / TMG, URL : <http://tmg.huma-num.fr/fr/corpus> (visité le 10/08/2020).

Michael Praetorius : Syntagma Musicum, Band 3.(1619), URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Praetorius%201619/Praetorius%201619.xml&chunk.id=div_1&toc.id=div_1&brand=default (visité le 07/08/2020).

Musica 1507, URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Cochlaeus_1507/Cochlaeus_1507.xml (visité le 22/08/2020).

Musica getutscht 1511, URL : http://tmg.huma-num.fr/xtf/view?docId=tei/Virdung_1511/Virdung_1511.xml (visité le 22/08/2020).

Thesaurus Musicarum Germanicarum / TMG, URL : <http://tmg.huma-num.fr/> (visité le 17/07/2020).

TMG - Thesaurus Musicarum Germanicarum / IReMus, URL : <https://www.iremus.cnrs.fr/fr/projets-de-recherche/tmg-thesaurus-musicarum-germanicarum> (visité le 07/08/2020).

TMG : Search Form, URL : <http://tmg.huma-num.fr/xtf/search> (visité le 20/08/2020).

Utilitaires pour la rédaction du mémoire

CAMPS (Jean-Baptiste), *Jean-Baptiste-Camps/biblatex-enc*, original-date : 2014-07-23T13:09:45Z,
6 avr. 2020, URL : <https://github.com/Jean-Baptiste-Camps/biblatex-enc>
(visité le 30/08/2020).

Inclure sa bibliographie : de Zotero à LaTex. - Renoult Jonathan (*muchos*), URL : <http://renoult-jonathan.tilde3.eu/docs/inclure-bibliographie-zotero-latex>
(visité le 30/08/2020).

Logiciels et services utilisés pour la base de données TMG

Projet Textométrie, URL : <http://textometrie.ens-lyon.fr/> (visité le 20/08/2020).
TMG_tagger / TMG, URL : <http://tmg.huma-num.fr/fr/content/tmgtagger> (visité le 10/08/2020).

Verovio, URL : <https://www.verovio.org/> (visité le 19/08/2020).

XTF, URL : <https://xtf.cdlib.org/> (visité le 19/08/2020).

Logiciels et services utilisés pour l'apparat critique de *Syntagma Musicum*

Gemeinsame Normdatei (GND), Deutsche Nationalbibliothek, URL : https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html (visité le 19/08/2020).

GeoNames, URL : <https://www.geonames.org/> (visité le 19/08/2020).

Getty Thesaurus of Geographic Names (Getty Research Institute), URL : <http://www.getty.edu/research/tools/vocabularies/tgn/> (visité le 19/08/2020).

GMBH (HTTPS://WWW.KLOKANTECH.COM/) (Klokan Technologies), *Old Maps Online*, URL : <https://www.oldmapsonline.org> (visité le 19/08/2020).

Oxford Music, Oxford Music Online, URL : <https://www.oxfordmusiconline.com/> (visité le 19/08/2020).

Startseite, Deutsche Nationalbibliothek, URL : https://www.dnb.de/DE/Home/home_node.html (visité le 19/08/2020).

Startseite - DARIAH-DE, URL : <https://de.dariah.eu/> (visité le 19/08/2020).

Wikipedia – Die freie Enzyklopädie, URL : <https://de.wikipedia.org/wiki/Wikipedia:Hauptseite> (visité le 19/08/2020).

yEd Graph Editor, yWorks, the diagramming experts, URL : <https://www.yworks.com/products/yed> (visité le 19/08/2020).

Logiciels et services utilisés pour le stage

- About us*, READ-COOP, URL : <https://readcoop.eu/about/> (visité le 31/08/2021).
- CHAGUÉ (Alix), *Constituer un corpus pour la fouille de texte - de la transcription des documents d'archives à l'annotation : exploration d'une méthodologie par l'ANR Time Us*, mémoire de master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2008, URL : https://github.com/alix-tz/M2TNAH_memoire-de-stage.git.
- Dario Kampkaspar*, URL : <https://www.oeaw.ac.at/acdh/team/current-team/dario-kampkaspar-1> (visité le 30/08/2020).
- DEVELOPERS (S. B. B.), *Digitalisierte Sammlungen der Staatsbibliothek zu Berlin*, URL : <https://digital-beta.staatsbibliothek-berlin.de> (visité le 20/08/2020).
- GUILLOTEL-NOTHMANN (Christophe), *imageAnnotationGroundTruth*, juin 2020, URL : <https://github.com/guillotel-nothmann/imageAnnotationGroundTruth.git>.
- *TMG_ImageAnnotation*, URL : <https://github.com/guillotel-nothmann/imageAnnotation.git> (visité le 30/06/2020).
- HDF5 for Python*, URL : <https://www.h5py.org/> (visité le 17/06/2020).
- How-to Guides*, READ-COOP, URL : <https://readcoop.eu/transkribus/resources/how-to-guides/> (visité le 31/08/2021).
- KAMPKASPAR (Dario), *dariok/page2tei*, 30 août 2020, URL : <https://github.com/dariok/page2tei>.
- KAMPKASPAR (Dario) et BOENIG (Matthas), *page2tei-0*, 30 août 2020, URL : <https://github.com/dariok/page2tei/blob/master/page2tei-0.xsl>.
- Keras : the Python deep learning API*, URL : <https://keras.io/> (visité le 17/06/2020).
- lxml - Processing XML and HTML with Python*, URL : <https://lxml.de/> (visité le 17/06/2020).
- Matplotlib : Python plotting — Matplotlib 3.4.3 documentation*, URL : <https://matplotlib.org/> (visité le 17/06/2020).
- Module Projects - OCR-D*, URL : <https://ocr-d.de/en/module-projects> (visité le 07/07/2020).

- Plongez en détail dans la librairie NumPy*, OpenClassrooms, URL : <https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/4740941-plongez-en-detail-dans-la-librairie-numpy> (visité le 17/06/2020).
- PRImA/tools/PAGE Libraries*, URL : <https://www.primaresearch.org/tools/PAGE Libraries> (visité le 30/08/2020).
- Startseite / Staatsbibliothek zu Berlin*, URL : <https://staatsbibliothek-berlin.de/> (visité le 20/08/2020).
- TensorFlow*, TensorFlow, URL : <https://www.tensorflow.org/?hl=fr> (visité le 17/06/2020).
- The TEI Guidelines*, URL : <https://tei-c.org/release/doc/tei-p5-doc/en/html/> (visité le 30/08/2020).
- Transcriptorium*, URL : <http://transcriptorium.eu/> (visité le 31/08/2021).
- Transkribus*, Page Version ID : 1015231899, 31 mars 2021, URL : <https://en.wikipedia.org/w/index.php?title=Transkribus&oldid=1015231899> (visité le 30/08/2021).
- Transkribus*, READ-COOP, URL : <https://readcoop.eu/transkribus/> (visité le 30/08/2021).
- WALT (Stéfan van der), SCHÖNBERGER (Johannes L.), NUNEZ-IGLESIAS (Juan), BOULOGNE (François), WARNER (Joshua D.), YAGER (Neil), GOUILLOART (Emmanuelle) et YU (Tony), « scikit-image : image processing in Python », *PeerJ*, 2 (19 juin 2014), e453, DOI : [10.7717/peerj.453](https://doi.org/10.7717/peerj.453).

Convention de transcription

BabelMap Online (Unicode 13.0), URL : <https://www.babelstone.co.uk/Unicode/babelmap.html> (visité le 03/08/2020).

MATTHIAS BOENIG OCR-D (Zentrum Sprache), *Alphabets, Abbreviations and Special Characters*, URL : <https://ocr-d.de/en/gt-guidelines/trans/trFremdsprache.html> (visité le 05/08/2020).

Music Note Symbols, URL : https://www.alt-codes.net/music_note_alt_codes.php (visité le 05/08/2020).

WILLS (Tarrin), « The Medieval Unicode Font Initiative », *Medieval Unicode Font Initiative* (, 19 févr. 2016), URL : <https://skaldic.abdn.ac.uk//m.php> (visité le 05/08/2020).

Glossaire

- **API** : *Application Programming Interface* - Ensemble de requêtes, souvent HTTP, permettant d’interagir avec un serveur et ses données sans passer par une interface graphique.
- **Bibliothèque numérique** : Ensemble organisé de documents nativement numériques ou numérisés accessibles à distance par l’Internet.
- **Format** : Manière normalisée de représenter des données ou des fichiers sous la forme d’informations binaires.
- **Gestionnaire de paquets** : Outil permettant d’automatiser l’installation, la mise à jour ou la désinstallation de logiciels ou de *packages* dans un environnement informatique.
- **HTTP** : *HyperText Transfer Protocol* - Protocole pour le transfert d’informations sur le web.
- **Interface-graphique** : Souvent par opposition aux interfaces en lignes de commande, dispositif visuel et symbolique permettant l’interaction entre l’humain et la machine, souvent accompagné d’un pointeur de type souris.
- **Java** : Langage de programmation et plateforme informatique sous licence Oracle Java.
- **JSON** : *JavaScript Object Notation* - Format de données textuelles structurées dérivé de la notation des objets du langage JavaScript.
- **Logiciel libre** : Par opposition aux logiciels propriétaires, un logiciel dont l’utilisation, la copie et la modification sont permises légalement.
- **MEI** : *Music Encoding Initiative* - Projet *open source* de création d’un standard de description des documents musicaux pour XML.
- **METS** : *Metadata Encoding Transmission Schema* - Standard XML permettant de conserver les métadonnées et la structure hiérarchique d’objets faisant partie d’une collection numérique, ainsi que les liens vers ces objets. Développé par la *Digital Library Federation*.
- **MIDI** : *Musical Instrument Digital Interface* - Protocole de communication et format de fichier dédiés à la musique, et utilisés pour la communication entre instruments électroniques, contrôleurs, séquenceurs, et logiciels de musique.

- **MusicXML** : Format de fichiers ouvert basé sur XML pour la notation musicale.
- **Module** : En Python, fichier pouvant contenir des fonctions, des classes et des données, et pouvant être importé dans un script.
- **Open source** : Mouvement visant à garantir la possibilité de distribuer librement des logiciels, d'accéder à leur code source et de créer des logiciels dérivés de ces codes sources. La mise à disposition des logiciels est régie par divers types de licences.
- **Package, library** : Ensemble de modules contenant des outils tels que des fonctions. Pour être utilisé, il doit être importé entièrement ou partiellement, par module.
- **PAGE** : *Page Analysis and Ground Truth Elements* - Standard XML permettant de stocker la description et la transcription de fichiers (*ground truth*) de fichiers transcrits. Développé par le laboratoire PRIma (*Pattern Recognition & Image Analysis*) de l'université de Salford à Manchester.
- **Pixel** : Unité permettant de mesurer la définition d'une image numérique matricielle.
- **Pseudo-classe** : En CSS, mot-clé ajouté à un sélecteur afin d'indiquer l'état spécifique dans lequel l'élément doit être pour être ciblé par la déclaration.
- **PNG** : *Portable Network Graphics* - Format ouvert de compression sans perte pour les images numériques, développé par le W3C depuis 1996.
- **Python** : Langage de programmation informatique à usage général, multi-plateforme et *open source*
- **Repository** : Aussi appelé « dépôt informatique » ; espace organisé de stockage de fichiers.
- **Script** : Programme ou extrait de programme informatique dont l'exécution conduit à la réalisation d'une ou plusieurs actions définies dans le programme.
- **Standard** : Texte de référence reconnu, documenté et élaboré par un groupe de travail spécialisé, visant à harmoniser l'activité d'un secteur donné. Pour XML, les standards prennent la forme de schémas et de règles de balisage permettant de créer des documents de structures comparables au sein d'un même standard.
- **TEI** : *Text Encoding Initiative* - Standard de description de documents textuels pour XML. Développé par le TEI Consortium.
- **TXM** : Plate-forme *open source* d'affichage et de traitement de données textuelles, développé par Serge Heiden (laboratoire IHRIM, équipe CACTUS, ENS Lyon).
- **Verovio** : Bibliothèque *open source* développée par le RISM permettant la visualisation de partitions encodées au format XML-MEI.
- **XML** : *eXtensible Markup Language* - Langage de balisage générique permettant de décrire des informations de manière organisée et standardisée.

- **XSLT** : *Extensible Stylesheet Language Transformations* - Langage basé sur XML permettant de styliser ou transformer des fichiers XML ou HTML.
- **XTF** : *eXtensible Text Framework* - Plate-forme *open source* développée et maintenue par la California Digital Library (CDL) pouvant fournir un accès à du contenu numérique.

Table des matières

Résumé	iii
Remerciements	v
Liste des sigles et abréviations	vii
Table des figures	ix
Introduction	3
I Le projet TMG	5
1 Le projet TMG dans son contexte	9
1.1 IReMus	9
1.1.1 Présentation d'IReMus	9
1.1.2 Les projets d'édition électronique d'IReMus	10
1.2 Le projet de recherche TMG	14
1.2.1 L'édition électronique de <i>Syntagma Musicum</i> , volume 3, de Michael Praetorius : une première étape du projet	15
1.2.2 Enrichissement de la base de données TMG	18
2 Choix des sources et des logiciels	21
2.1 Les sources	21
2.1.1 <i>Cursus Philosophici Encyclopaedia</i> de Johann Heinrich Alsted . . .	21
2.1.2 <i>Hypomnematum musicae poeticae... synopsis</i> de Joachim Burmeister	22
2.2 Les logiciels	22
2.2.1 Transkribus	22
2.2.2 TMG_ImageAnnotation	25

II Des images aux fichiers XML	27
3 L'annotation des facs-similés électroniques	31
3.1 L'interface graphique et les commandes du logiciel TMG_ImageAnnotation	31
3.1.1 L'interface graphique	31
3.1.2 Les commandes du logiciel TMG_ImageAnnotation	33
3.2 Vers une amélioration du logiciel et un ajustement des zones	34
3.2.1 Amélioration de certaines commandes	34
3.2.2 Ajustement des zones	35
4 Transcription et transformation en XML-TEI	37
4.1 Transcription et relecture	37
4.1.1 Une nouvelle fenêtre à l'interface graphique	37
4.1.2 Vers une convention de la transcription	38
4.2 Transformation au format XML-TEI	41
4.2.1 Le scénario de transformation page2tei	41
4.3 L'aborescence du <teiHeader>	42
Conclusion	45
Bibliographie	49
Glossaire	69
Table des matières	73