TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 3.5

Multimodal Music Information Alignment v2

Grant Agreement nr	770376
Project runtime	May 2018 – April 2021
Document Reference	TR-D3.5-Multimodal Music Information Alignment v2
Work Package	WP3 - Automated Music Data Processing and Linking
Deliverable Type	Report
Dissemination Level	PU-Public
Document due date	April 2020
Date of submission	April 2020
Leader	MDW
Contact Person	Werner Goebl (goebl@mdw.ac.at)
Authors	David Weigl (MDW)
Reviewers	Álvaro Sarasúa (VL)

Executive Summary

The alignment of multimodal musical resources, for unified exploration, discovery, and retrieval, represents an important priority in enriching digital music archives. In this document, we provide a description of various types of information resource subject to potential alignment within TROMPA; detail a number of candidate algorithms to perform alignment activities; and present a data model that provides a framework for the application of alignment outcomes in TROMPA use cases. We also present a set of initial resources subject to proposed alignment in future development of this deliverable, comprising Beethoven piano compositions, and early Spanish choral works.

Information resources subject to alignment include:

- ❖ Resources exposing events on a timeline (e.g., audio/video recordings; performance metadata streams; derived audio features)
- ❖ Resources exposing spatial coordinates (e.g., composer, performer, or stage images; video recordings; digitised score images)
- ❖ Resources exposing named (identified) structural elements (e.g., symbolic music encodings; encodings of textual resources such as librettos, concert programmes)

This deliverable discusses these resources in greater detail, and introduces a set of MEI encodings of Beethoven's piano compositions (Sonatas and other works) acting as initial musical repertory resources (alignment targets) in the technical development of this task.

Notable *Candidate alignment algorithms* (available for inclusion within TROMPA workflows) include:

- ❖ **SMAT** (Symbolic Music Alignment Tool), a MIDI-to-MIDI HMM matcher
- ❖ **MAPS** (Matcher for Alignment of Performance and Score), a MIDI-to-Score HMM matcher with native MEI support developed with in-house involvement from mdw
- ❖ **MATCH** (Music Alignment Tool Chest), a tool based on Dynamic Time Warping
- ❖ **Peachnote** performance alignment tool, applying LCS and HMM approaches

The deliverable discusses these algorithms, alongside a number of other alternatives with less straightforward applicability within the TROMPA project, in greater detail.

A discussion of alignment formats and representations is provided, first outlining the typical tabular formats used by tools such as MATCH; then introducing the more flexible, reusable, and web-addressable graph-based, semantic representations implemented by the Music Alignment and Linked Data (MELD) framework, building on the Segment Ontology, a convenient means of bridging music-generic structural representation with music-domain-specific multimodal resources. MEI is described as a particularly suitable encoding schema for musical scores in this context, in terms of the comprehensive, addressable MEI hierarchy which interacts synergistically with the multi-level structural representations ("segment lines") offered by the Segment Ontology.

Based on this discussion, the TROMPA alignment data model proposed is presented, which both: builds on the Music Encoding and Linked Data semantic framework, combining and extending widely used Semantic Web models including the Music, Segment, PROV, and Audio Feature ontologies, and the Web Annotation Data Model; and, which serves as a specialisation of the Schema.org based data model of the TROMPA Contributor Environment (see deliverable D5.1, Data Infrastructure). Using the

proposed alignment model, alignments can be performed at different layers of abstraction - e.g. sections, measures, beats, or individual notes - as the individual use case necessitates.

Two implemented alignment workflows for MIDI-to-Score (MEI) alignment are then detailed: one for real-time alignment (e.g., to highlight notes or enable automatic page turns in coordination with a performance) using MAPS, and one for offline alignment (to enable more thorough analyses of renditions after a performance) using SMAT.

Version Log		
#	Date	Description
v0.1	16 January 2019	Initial version submitted for internal review
v0.2	26 February 2019	Revised version after internal review
v1.0	28 February 2019	Final version submitted to EC
v2-draft	7 April 2020	Version 2 of deliverable, submission for internal review
v2-RC	27 April 2020	Version 2 revision after internal review
v2.0	30 April 2020	Version 2 final submission to EC

Table of Contents

1 Introduction

From the proposal document: “The goal of this task is to adapt existing alignment technology to connect score information of selected standard repertoire to performance properties (timing, dynamics) of multiple available performances into unified data representations. Score information is integrated from T3.1 (Data resource preparation) and T3.4 (Visual analysis of scores) and WP4 (Crowd annotation and incentivization) and sources of performance metadata include audio recordings of renowned artists, high-quality symbolic data resources available online or through data collected within TROMPA (user pilots, WP6 and WP7). Methods established with this task are able to deal with incomplete or partial resources such as partially transcribed scores or excerpts of performances.”

In this document we consider the variety of types of musical resources subject to alignment within TROMPA (section 2); catalogue available technologies to accomplish the alignment process (section 3); briefly discuss representation issues relating to the outcomes of these processes (section 4) before introducing a data model building on a graph representation consisting of a specialisation of the TROMPA Contributor Environment knowledge graph (section 5). Finally, we detail two implemented workflows for performance-to-score alignment in section 6: one operating in real-time to provide initial feedback as a performance is unfolding, the other providing more comprehensive feedback immediately after a performance ends.

This document represents the second and final version of this deliverable. It presents a refinement and extension of the initial deliverable document, updated to describe the state of development supporting the first release of working WP6 prototypes in M24. Further iterative refinements to the described workflows and their components are anticipated over the coming months.

2 Musical resources subject to alignment

2.1 Resource types

We aim to support alignments between various types of multimodal musical resources (Table 1). These include resources adhering to a timeline (e.g. audiovisual recordings; performance- and audio-derived feature data); resources adhering to a symbolic structure (e.g., musical scores encoded as MEI); and resources adhering to a spatial representation (e.g., digitised score images).

Table 1. Resources subject to alignment via the TROMPA data infrastructure.

Media type	Example	Typical source	Anchor types
Image-based	Digitised score images Photographs Diagrams	TROMPA-contributed External	Page Fragment coords IIIF
Structured music encodings	Sparse scores Part scores	TROMPA-contributed External	Named elements XPath selector

	Orchestra scores Piano reductions		CSS selector
Structured textual encodings	Libretto Concert programme	TROMPA-contributed External	Named elements XPath selector CSS selector
Audio	Rehearsal recording Performance recording Studio recording	TROMPA-contributed External	Timed offset
Video	Performance recording	TROMPA-contributed External	Timed offset Fragment coords
Performance metrics stream (e.g. CEUS key positions)	Rehearsal recording Performance recording	TROMPA-contributed	Timed offset
MIDI/OSC stream	Rehearsal recording Performance recording	TROMPA-contributed	Timed offset
Content-derived feature data	Audio descriptors Performance descriptors	TROMPA-contributed	Timed offset Feature entity URI

2.2 Corpora targeted for initial alignment

While we aim to establish and implement a set of workflows and data models to enable alignment of arbitrary classical music resources in line with long-term project aims, current development is driven by an initial target corpus of Beethoven piano works. We have assembled a comprehensive collection of high quality MEI encodings of Beethoven's works for solo piano, incorporating the 32 Sonatas (converted from Humdrum encodings by Craig Sapp [c]) and all variations, bagatelles, rondos, and other pieces (among them Phantasie, Polonaise, "Für Elise", Andante favori), generated by an OMR process from digitised public-domain score images obtained from IMSLP, hand-corrected by Werner Goebl at mdw [d]. For many of these works, we have access to permissively licensed audiovisual recordings and performance metadata (multiple recordings per work) obtained from an international piano competition. This corpus is chosen to bootstrap development of the piano performance companion use case.

3 Alignment technologies

Given the varied kinds of multimedia resources subject to alignment, it is convenient to identify a particular reference structure into which these resources can be interwoven according to their relevant anchor types. The structure offered by musical score encodings is well suited for this purpose, providing a performance-independent canonical reference (unlike individual performance recordings) with fine-grained musical semantics (unlike digitised score images).

The challenge, then, is to anchor the various multimodal resources into a musical score structure. This requires algorithmic processes to map between anchors expressed in the source modalities, and identified score elements. Largely, this involves mapping from

I. timed information streams (audio/visual media, MIDI, performance- and audio-derived metadata), and
II. spatial information (page coordinates).

These tasks are accomplished through (I.) audio-score (or MIDI-score) alignment, where the audio (or MIDI) timeline corresponds to a performance recording, and serves as a clock-provider for performance- and audio-feature metadata; and (II.) Optical Music Recognition (OMR), where salient image regions are identified within images of musical score and mapped to matching granular representations within the symbolic score. As OMR techniques are subject to the “visual analysis of scores” task, with their own deliverable document D3.4, we focus on technologies concerned with aligning timed information streams and musical scores in this section.

While approaches to audio alignment, score following, and related tasks have received considerable attention in the MIR literature, the availability of “off-the-shelf” solutions is limited. Here we briefly detail several algorithms performing audio- or MIDI-score alignment that are (or could potentially be made) available. These algorithms were taken into consideration to arrive at the implemented workflows detailed in section 5.

3.1 Symbolic Music Alignment Tool (SMAT) [1]

- Authors: Eita Nakamura, Kazuyoshi Yoshii, Haruhiro Katayose
- Modality: MIDI-MIDI
- Supports offline alignment of partial performances and partial scores
- Accommodates note insertions / deletions, arbitrary repetition, large skips.
- Open source (MIT) license.

3.2 MAPS (Matcher for Alignment of Performance and Score)

- Authors: Martin Bonev, Carlos Cancino-Chacón
- Modality: MIDI-score (Audio-score under development at JKU Linz University and likely to become available within the course of the TROMPA project)
- Module of the *ACCompanion* piano accompaniment system [2], implementing score following with skips using a specialised variation of a Hidden Markov Model (HMM) inspired by Nakamura [1].
- Supports alignment of partial performances and partial scores; real-time and offline (batch-mode) alignment.
- Natively supports direct alignment from MEI, MusicXML, and MATCH format encodings.
- Accommodates note insertions / deletions, arbitrary repetition, large skips.
- Under active developed with in-house involvement from mdw (in collaboration with OFAI, the Austrian Research Institute for Artificial Intelligence)
- License status tbc - software likely open-licensed - definitely can be incorporated into TROMPA workflows.

3.3 MATCH (Music Alignment Tool Chest) [3]

- Authors: Simon Dixon, Gerhard Widmer.
- Modality: Audio-Audio
- Employs Dynamic Time Warping (DTW) to calculate minimal-cost alignments between two audio signals. Thus, this is an audio-audio alignment; audio-score is performed by first synthesising the score to audio (e.g. via MIDI). Similarly, MIDI-MIDI or MIDI-score alignment is accomplished through synthesis.
- Batch alignment, i.e. cannot align in real-time.
- Accommodates note insertions / deletions, but may get confused by larger structural deviations (i.e. arbitrary repetitions, large skips) that are likely in rehearsal practice.
- Available as open-source code [a], although no license specified.
- Also available as a VAMP plugin [b] for Sonic Visualiser [4] - potentially be used for crowd-sourced / crowd-corrected alignments?

3.4 (PHENICX) Piano Music Companion [5]

- Authors: Andreas Arzt et al.
- Modality: Audio-Audio
- Employs DTW, building on and improving the MATCH approach by Dixon et al.
- Capable of real-time alignment, supports partial performance
- Accommodates structural deviations (repetitions, skips)
- Code appears to be not publicly available, although we could approach the authors. Unclear whether we could incorporate into TROMPA workflows.

3.5 Niedermayer/Widmer tool [6]

- Author: Bernhard Niedermayer and Gerhard Widmer
- Modality: Audio-Audio
- Another DTW-based approach employing a multiple-pass method for improved alignment accuracy.
- Batch alignment, i.e. cannot align in real-time.
- Code appears to be not publicly available, although we could approach the authors. Unclear whether we could incorporate into TROMPA workflows.

3.6 Peachnote performance alignment tool

- Authors: Vladimir Viro and Julian Schmidt
- Modality: MIDI-MIDI
- Two approaches, based on LCS alignment (faster) and on HMMs (more flexible)
- Accommodates structural deviations (repetitions, skips)
- Can do batch and real-time
- Code not publicly available, but can be incorporated in TROMPA workflows

4 Alignment formats and representations

Fundamentally, the alignment technologies discussed in the previous section produce a collection of tuples that each express the connection of anchor points (e.g., timed offsets, MIDI ticks, note identifiers) on either of the two information resources subject to alignment. These connections are typically expressed in a tabular form that is simple to parse but becomes unwieldy when alignments between newly contributed resources are expected to be accommodated over time, in a reusable, web-addressable way.

Graph-based representations offer an alternative that overcomes these limitations. Here, a structural representation of a canonical resource - ideally, a score encoding, or perhaps an audio recording deemed as representative of a work - offers a semantic spine through which all information resources subject to alignment can be interwoven. Newly contributed resources are simply incorporated into this structural representation as they arrive, becoming immediately available for retrieval. The Segment Ontology [7] (SO) provides a convenient means of expressing such a music-generic structural representation in order to bridge domain-specific multimodal resources. MEI is particularly well suited as an encoding schema for musical scores in this context, as the comprehensive, addressable MEI hierarchy interacts synergistically with the ability to specify structural representations (“segment lines”) at multiple levels of abstraction within SO.

Music Encoding and Linked Data (MELD) [8] provides a semantic framework and JavaScript client library that implements these ideas. It has been applied to a number of use cases, including music performance, composition, real-time score annotation, and musicological scholarly communication in previous work (e.g. [9-11]). It has been developed with involvement of TROMPA consortium members at mdw and GOLD, and forms the basis for our proposed alignment data model here.

5 Alignment data model

The alignment data model (Figure 1) serves as an extension and specialisation of the TROMPA Contributor Environment (CE) data model detailed in D5.1 (Data Infrastructure).

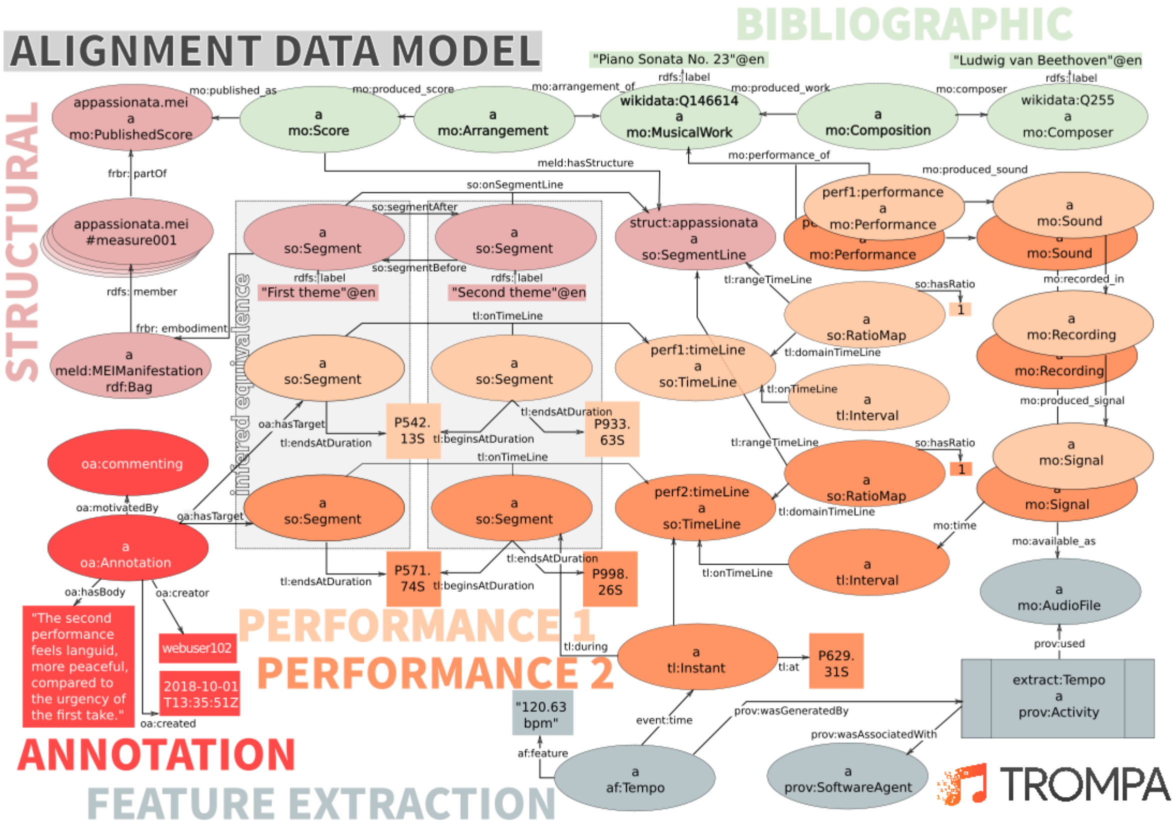

Figure 1. TROMPA's alignment data model, an implementation of the Music Encoding and Linked Data (MELD) framework, supports the capture and interlinking of multimodal information around musical works and performances.

The model builds on the Music Encoding and Linked Data framework, combining several pre-existing ontologies in order to express multimodal alignments between performance recordings (Timeline and Event ontologies) and musical score (Segment ontology with MEI anchoring), while interlinking with bibliographic information (Music ontology) and performance feature data (Audio Feature ontology). The Linked Data representation affords the creation of Web Annotations addressing (fragments of) individual media representations, or indeed their combination in the form of structural segments.

Figure 1 illustrates a simple instantiation of the model in which a music encoding ("appasionata.mei"), described with bibliographic metadata by reference to an external authority (here, Wikidata) is associated with a section-wise structural segmentation, which is mapped to distinct timed intervals along two timelines (corresponding to two distinct performance renditions). Each timeline is associated with a recording of the corresponding performance, available as an audio file, which is further described with extracted feature data and associated provenance information. Every entity within the data model has its own URI, and thus can be targeted by Web Annotations; in the simple illustrated example, an annotation expresses a subjective description of two corresponding timeline intervals (performances of the same structural segment across two performance recordings).

Alignments can be performed at different layers of abstraction - e.g. sections, measures, beats, or individual notes - as the individual use case necessitates. The section-level alignment illustrated above is sufficient to serve as a navigational model e.g. for the piano rehearsal (performance companion) use case by associating performance audio, score information, and aggregated feature data (e.g. average tempo) for salient large-scale musical sections (e.g. allowing users to jump to a particular theme across multimodal representations). Finer-grained alignments can be expressed by specifying additional segment lines, which can coexist hierarchically with other segment lines (via the use of segmentLineMaps, and the "contains" relationship between segments, in order to capture information in fine detail (e.g. for the scholarly analysis of onset timings of individual notes).

6 Implementation of alignment workflows

At the present stage of development, two complementary workflows have been implemented, comprising real-time and an offline MIDI-to-MEI alignment. These are achieved using the MAPS (real-time) and SMAT (offline) tools introduced in section 3. The output of these tools is converted into Linked Data (RDF) adhering to the alignment data model detailed in section 5, which may be made available to end-users through specialised MELD applications or other Linked Data clients.

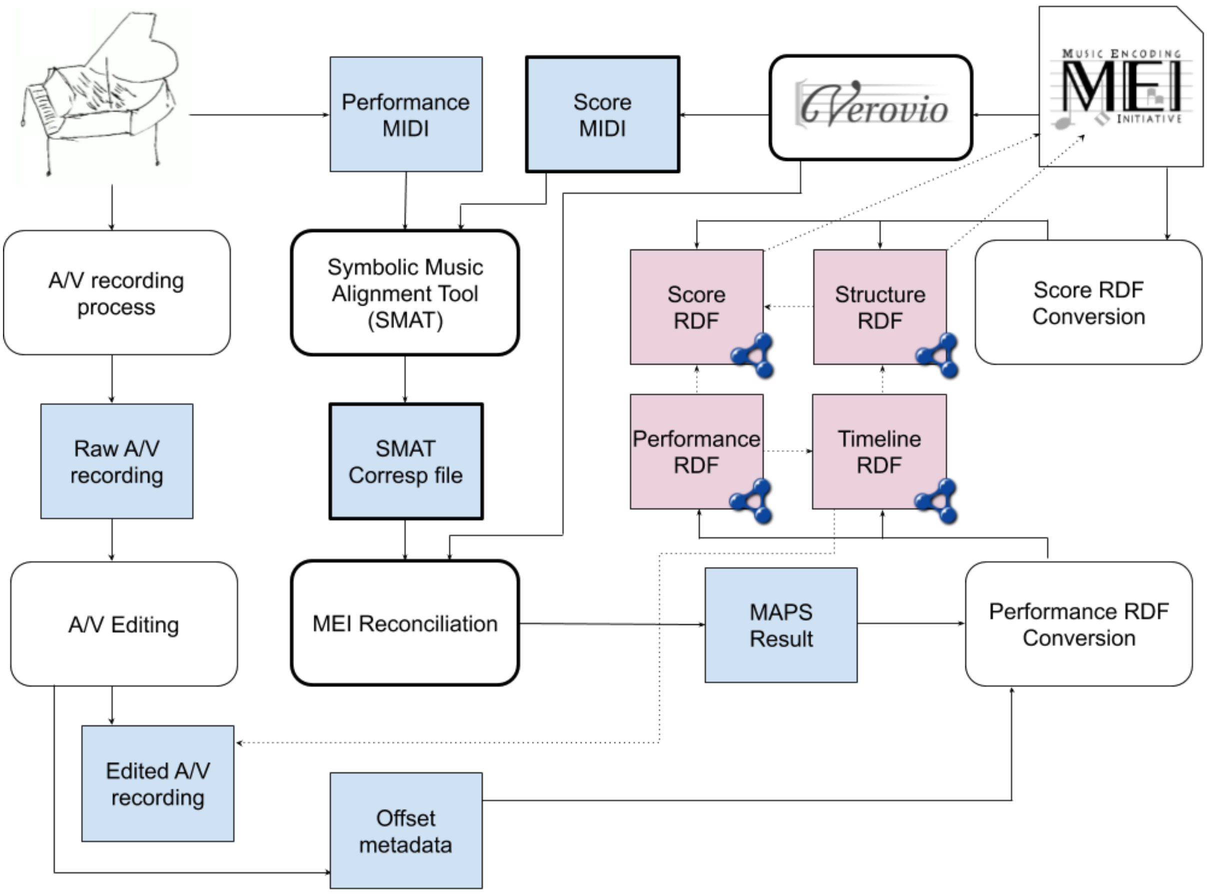

Figure 2. Offline alignment workflow. Solid arrows represent process flow; dotted arrows represent reference-by-URI. Blue-shaded rectangles represent information resources; white rounded rectangles are processes; pink-shaded rectangles are RDF files expressing graphs adhering to the MELD framework, available for consumption by TROMPA user-facing applications and other Linked Data

clients. Bolded outlines indicate steps for offline alignment (using SMAT); these are simplified into one MAPS alignment step for real-time alignment.

6.1 Real-time alignment

The MAPS tool natively handles MEI files (supporting MIDI-to-MEI alignment). As such, MAPS-generated alignment data structures provide a convenient base framework for conversion to RDF adhering to the alignment data model. We have developed a Python module and command-line tool [e] to perform such translations.

Further, we have extended MAPS to support real-time transmission of alignment data to a MELD client through an HTTP connection. This is achieved using Server Sent Events (SSE) [f]: the client application opens a persistent SSE connection to a MAPS server using HTTP GET; the server then continuously yields new alignment events through this connection as they become available, closing the connection only after the MIDI stream remains quiet for a configurable time period (i.e., the performance is finished). The yielded events are transmitted in the JSON-LD format, an RDF serialisation convenient for processing by MELD applications. This enables real-time display and interaction with alignment events, e.g. by colouring in the notes of a score rendering as the notes are played. The workflow is a simplified version of the one illustrated in Figure 2, in which the MAPS alignment process (with native MEI support) replaces several steps (bolded in the figure) required to synthesise and reconcile MIDI from MEI for SMAT (see Section 6.2).

6.2 Offline alignment

The offline alignment workflow is illustrated in Figure 2. At present, SMAT is used to generate offline alignments, as it operates considerably faster and with greater alignment accuracy than the corresponding MAPS alignment mode. This implementational choice may change depending on future progress in MAPS development.

As SMAT does not natively support MEI encodings, pre-alignment steps are required in order to synthesise a MIDI representation of the MEI score, and to align the events within this MIDI representation with note identifiers in the MEI. These are performed using a script [g] which uses Verovio to perform the MIDI synthesis, and to generate a timemap locating each note (by its identifier) at a millisecond offset within the generated MIDI.

From there, SMAT is run between the synthesised reference-MIDI and a given performed MIDI rendition to arrive at the alignment data. Unfortunately, due to a mismatch in the rounding of sub-millisecond time durations between Verovio and SMAT, the correspondence of SMAT alignments (anchored within the synthesised reference-MIDI) and MEI identifiers necessitates an additional fuzzy alignment step, where reference-MIDI events occurring within a 1-millisecond threshold of the MEI note timestamps (obtained from the Verovio timemap) are deemed to match, assuming they share a corresponding pitch height. Even with these pre-processing steps, the time required to fully align a MIDI performance to an MEI score is on the order of a few seconds (for short pieces) to around 40 seconds (for ~15-minute long piano renditions of Beethoven’s 32 Variations in c minor).

The alignment outcomes produced by SMAT are translated into MAPS alignment data structures, from where they are converted to RDF using the same process described in Section 6.1. The resulting alignment RDF is then posted (via HTTP) into a Linked Data Platform container (e.g., a user’s Solid POD), where it becomes available for display in MELD clients.

7 Conclusion

In a project relating to digital music libraries and archives of the scope we are targeting in TROMPA, alignment tasks are necessarily multimodal as many different kinds of information resources - timed (e.g., performance audio), spacial (e.g., digital score images), and structural (e.g., score encodings) - need to be interconnected. In this document, we have provided an enumeration of the different resource types subject to alignment through TROMPA's data infrastructure; we have catalogued a number of different algorithmic solutions to problems of audio-audio (MIDI-MIDI) and audio-score (MIDI-score) alignment and we have presented an alignment data model, extending the Contributor Environment's model and building on previous work around Music Encoding and Linked Data, to provide a framework for unified exploration, discovery, and retrieval of aligned musical resources. We have also identified an initial musical corpus for alignment - Beethoven's piano compositions - upon which our current development efforts have been focused. Finally, we have described alignment workflows around two tools chosen from the catalogued set, capable of real-time and offline alignment of MIDI performances to MEI score-encodings, which we have implemented for use in demonstrations and pilot user studies.

8 References

8.1 Written references

[1] Nakamura, E., Ono, N., Saito, Y., & Sagayama, S. (2014). Merged-output hidden markov model for score following of midi performances with ornaments, desynchronized voices, repeats and skips. In *Proceedings of the 2014 Joint SMC/ICMC Conference (pp. 1185–1192).*

[2] Cancino-Chacón, C., Bonev, M., Durand, A., Grachten, M., Arzt, A., Bishop, L., Goebl, W., & Widmer, G. (2017). The ACCompanion v0.1: An expressive accompaniment system. Late Breaking / Demos session, *18th International Society for Music Information Retrieval Conference*, Suzhou, China.

[3] Dixon, S. & Widmer, G. (2005). MATCH: A music alignment tool chest. In Proc. *6th International Conference on Music Information Retrieval*, London, UK.

[4] Cannam, C., Ladone, C, & Sandler, M. (2010) Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia 2010 International Conference*.

[5] Arzt, A., Böck, S., Flossmann, S., Frostel, H., Gasser, M., Liem, C., & Widmer, G. (2014). The Piano Music Companion. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence,* pp. 1221–1222

[6] Niedermayer, B., & Widmer, G. (2010). A Multi-pass Algorithm for Accurate Audio-to-Score Alignment. In Proc. *11th International Society for Music Information Retrieval Conference,* pp. 417–422.

[7] Fields, B., Page, K., De Roure, D., & Crawford, T. (2011). The Segment Ontology: Bridging music-generic and domain-specific. In Proc. *2011 IEEE International Conference on Multimedia and Expo (ICME 2011)* (pp. 1–6). IEEE.

[8] Weigl, D. M. & Page, K. R. A framework for distributed semantic annotation of musical score: "Take it to the bridge!". In Proc. *18th International Society for Music Information Retrieval Conference*, Suzhou, China.

[9] Kallionpää, M., Greenhalgh, C., Hazzard, A., Weigl, D. M., Page, K. R., & Benford, S. (2017). Composing and realising a game-like performance for disklavier and electronics. In Proc. *New Interfaces for Musical Expression*, Copenhagen, Denmark.

[10] Lewis, D., Weigl, D. M., Bullivant, J., & Page, K. R. (2018). Publishing musicology using multimedia digital libraries: creating interactive articles through a framework for linked data and MEI. In Proc *5th International Conference on Digital Libraries for Musicology,* pp. 21–25, Paris, France.

[11] De Roure, D., Klyne, G., Pybus, J., Weigl, D. M., & Page, K. (2018). Music SOFA: An architecture for semantically informed recomposition of digital music objects. In Proc. *1st International Workshop on Semantic Applications for Audio and Music*, pp. 33–41, California, USA.

8.2 Web references

[a] Source code available at https://code.soundsoftware.ac.uk/projects/match

[b] Source code available at https://code.soundsoftware.ac.uk/projects/match-vamp

[c] Humdrum encodings available at https://github.com/craigsapp/beethoven-piano-sonatas and converted to MEI using Verovio

[d] Available from https://github.com/trompamusic-encodings/

[e] Source code available at https://github.com/trompamusic/trompa-align/blob/master/scripts/convert_to_rdf.py

[f] See https://www.w3.org/TR/eventsource/

[g] Source code available at https://github.com/trompamusic/trompa-align/blob/master/scripts/trompa-align.R

8.3 List of abbreviations

Abbreviation	Description
DTW	Dynamic Time Warping
HMM	Hidden Markov Model
MAPS	Matcher for Alignment of Performance and Score
MATCH	Music Alignment Tool Chest
MDW	University of Music and Performing Arts Vienna
MEI	Music Encoding Initiative
MIR	Music Information Retrieval

OMR	Optical Music Recognition
POD	Personal Online Datastore
SMAT	Symbolic Music Alignment Tool
SO	Segment Ontology