Computer Vision Center

A Unified Representation Framework for the Evaluation of Optical Music Recognition Systems

Pau Torras – `ptorras@cvc.uab.cat`
Sanket Biswas
Alicia Fornés
Computer Vision Center, Universitat Autònoma de Barcelona

Table of contents

1 Motivation

2 Design

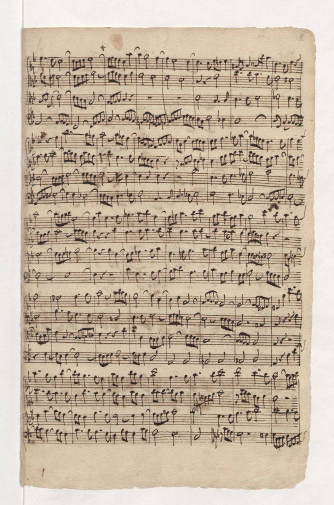
3 Proof-of-concept

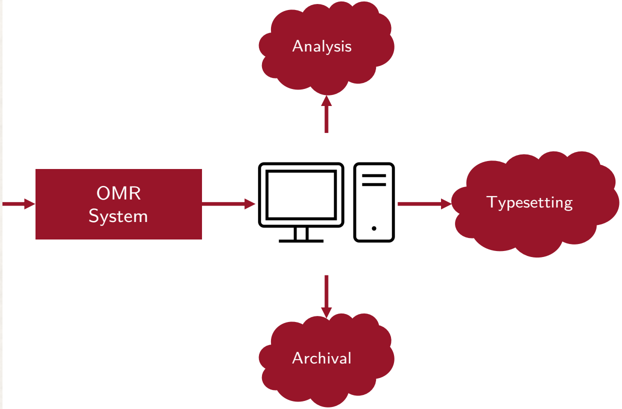
4 Conclusions

Computer Vision Center

Motivation

Optical Music Recognition (OMR)





OMR System

Analysis

Typesetting

Archival

The Requirements of OMR

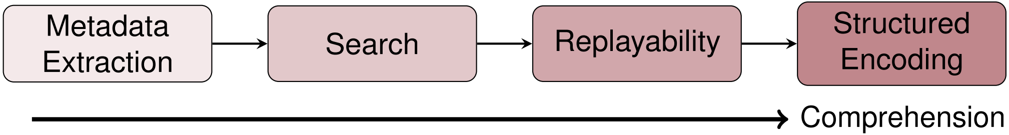
We identify a few key requirements that have not yet been addressed in OMR

- **A standardised output:** Most OMR systems assume different collections of objects and semantics as output.
- **Evaluation Metrics:** Measuring the quality of OMR systems fairly and equitably is currently not possible.

Both of these issues are intertwined: Unless defined on the same representation, fair evaluation metrics are not possible

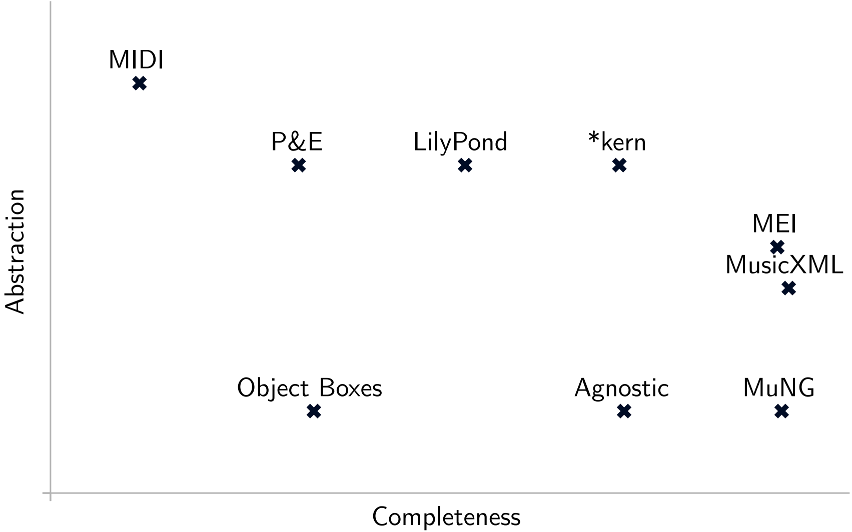
An important side note

The focus of this work is on **fully structured encodings** of music, as defined in [Calvo-Zaragoza et al., 2021].

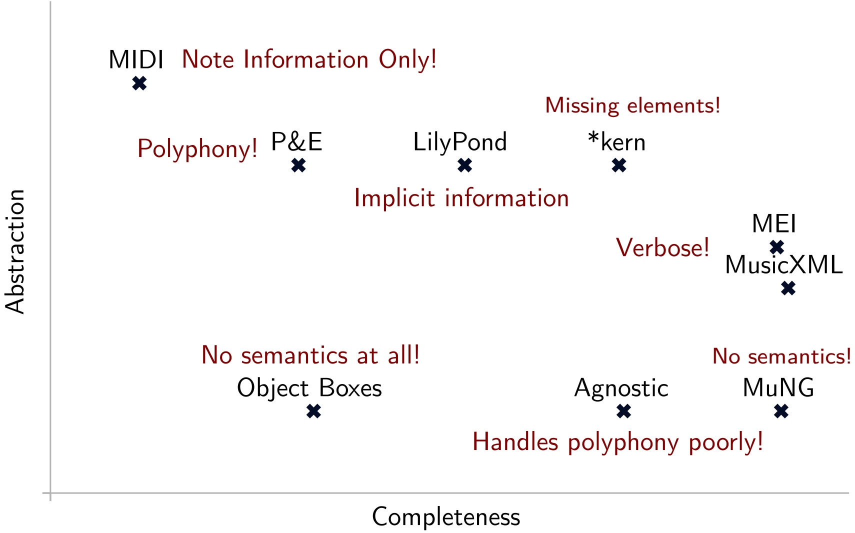


We only consider the application on **Common Western Music Notation.**

A Taxonomy of Music Representations



A Taxonomy of Music Representations



A Taxonomy of Music Representations

Outside of OMR, complex scores are usually engraved using

- **MusicXML:** Particularly in large repository sites and most musicological applications
- **MEI:** It is lately getting traction in mainstream applications as well as its original archive-focused domain.

A Taxonomy of Music Representations

Outside of OMR, complex scores are usually engraved using

- **MusicXML:** Particularly in large repository sites and most musicological applications
- **MEI:** It is lately getting traction in mainstream applications as well as its original archive-focused domain.

Nevertheless...

It is not trivial to work with these formats for OMR. It is possible, but some simplifying assumptions must be made [Mayer et al., 2024].

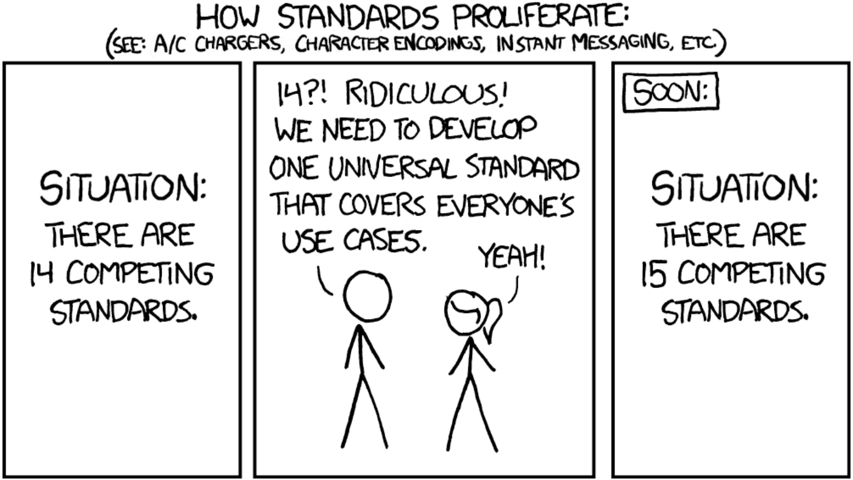
OMR Datasets

Publication	Dataset Name	Score Type	Document Type	Annotation Type
[Hajič and Pecina, 2017]	MUSCIMA++	CWMN	Modern Handwritten	MuNG
[Tuggener et al., 2018, Tuggener et al., 2020]	DeepScores	CWMN	Typeset	Bounding Boxes
[Tuggener et al., 2023]	RealScores	CWMN	Scanned Typeset	Bounding Boxes
[Shatri and Fazekas, 2021]	DoReMi	CWMN	Typeset	Multiple
[Parada-Cabaleiro et al., 2017]	SEILS	Mensural	Scanned Typeset	Agnostic + MEI
[Baró et al., 2020]	Baró Synthetic	CWMN	Typeset	Agnostic
[Baró et al., 2020]	Pau Llinás	CWMN	Historical Handwritten	Agnostic
[Calvo-Zaragoza and Rizo, 2018b]	PRIMuS	CWMN	Typeset	Agnostic + MEI
[Calvo-Zaragoza and Rizo, 2018a]	Camera PRIMuS	CWMN	Typeset (Scan-like)	Agnostic + MEI
[Ríos-Vila et al., 2023]	GrandStaff	CWMN	Typeset	*kern
[str, 2023]	OpenScore Quartets	CWMN	Typeset	MuseScore
[Gotham and Jonas, 2022]	OpenScore Lieder	CWMN	Typeset	MuseScore

So, what should we do?

- We have to try and find a way to connect all of these isolated datasets and formats and find a way of representing them using the same language.
  - The main goal is **evaluation**, but having such a tool could improve researchers' QOL by standardising an output for all of OMR, **making it possible to share tools and efforst more easily**.
- None of the existing formats are completely suitable for a majority of use cases within OMR.
- Therefore, **we built one with these in mind.**

Why design a new format?


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

"Standards" by XKCD, CC-BY-NC 2.5

Why design a new format?

Reason #1

Fair evaluation must be possible regardless of the original annotation format of the material.

The point of agreement of all sources is the presentation of the score.

"Engraving that can be processed"

Example

The SVG output of the Verovio engraving software.

Why design a new format?

Reason #1

Fair evaluation must be possible regardless of the original annotation format of the material.

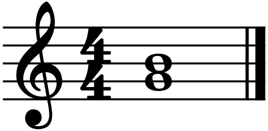
- This has the effect of **standardising how to deal with each format's assumptions.**
- ... but a notation like this does not really exist yet!

Why design a new format?

Reason #2

There must only be **one** way to represent each score.

Most formats have semantic constructs that allow equivalent representations of a score, which can make evaluation metrics meaningless in some contexts.

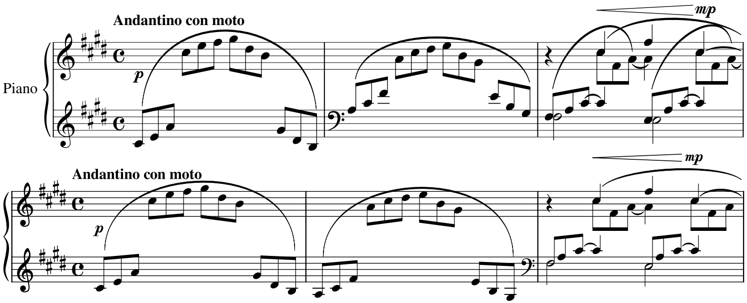


Which note is defined first?

Why design a new format?

Reason #3

The representation must be **faithful** and **exhaustive.**



Why design a new format?

Reason #3

The representation must be **faithful** and **exhaustive.**

- These issues can sometimes be sidestepped by using optional features or modifying file structures → **Need to be standardised anyway**

Why design a new format?

Reason #4

Semantics and presentation must be separate.

From [Calvo-Zaragoza et al., 2021]

> *In particular, there is no known meaningful edit distance between two scores [...] However, this does not necessarily provide a good measure of quality, because it is unclear how to weight the costs of different edit operations, e.g., getting a note duration wrong vs. missing an articulation mark.*

Why design a new format?

Reason #5

We need to communicate with the rest of the ecosystem.

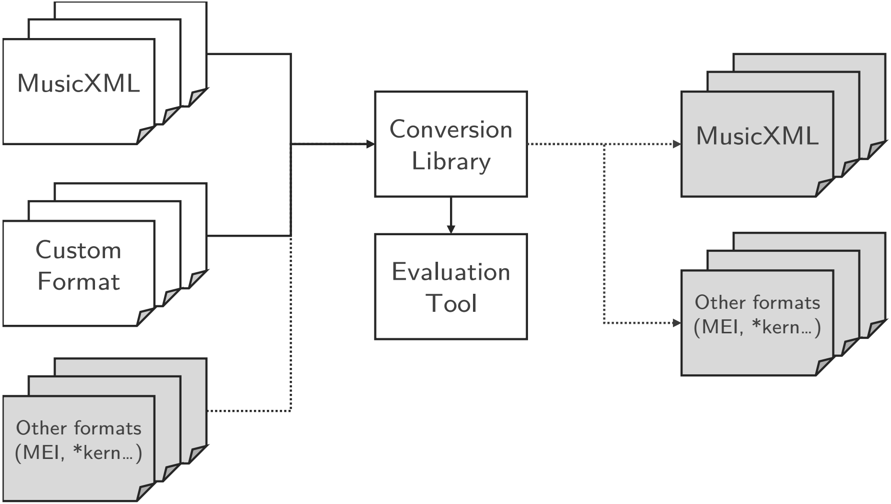
We need tools that facilitate using external sources and exporting our results.

Wrapping up

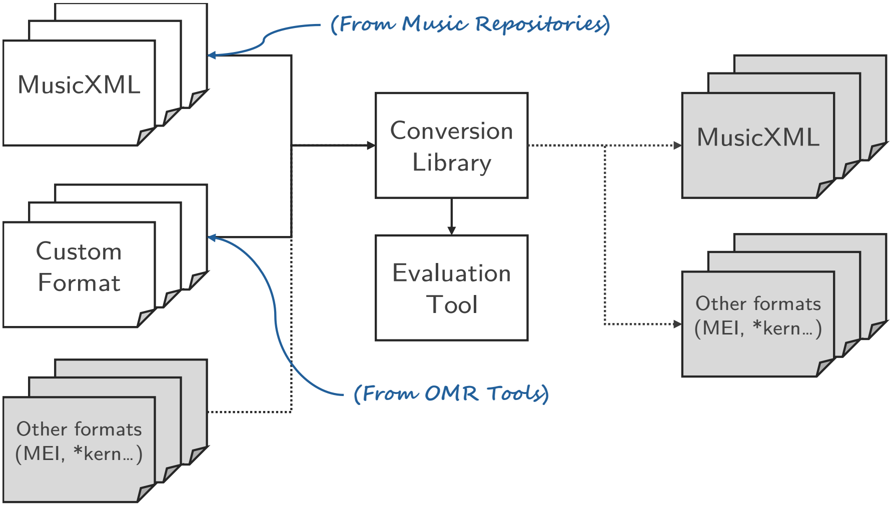
We decided to prototype a representation that encapsulates all of these requirements. We call it Music Tree Notation, or **MTN**.

Design

An Overview


MusicXML
Custom Format
Other formats (MEI, *kern...)
Conversion Library
Evaluation Tool
MusicXML
Other formats (MEI, *kern...)

An Overview



MusicXML

Custom Format

Other formats (MEI, *kern...)

(From Music Repositories)

(From OMR Tools)

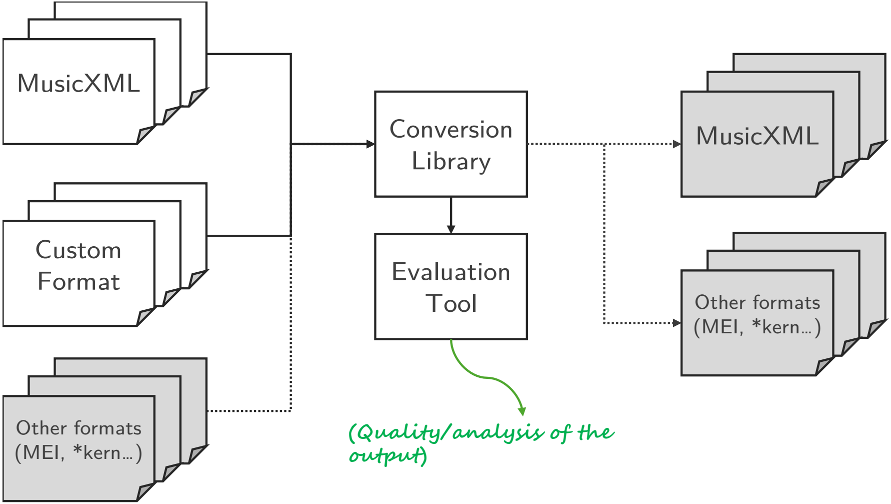
Conversion Library

Evaluation Tool

MusicXML

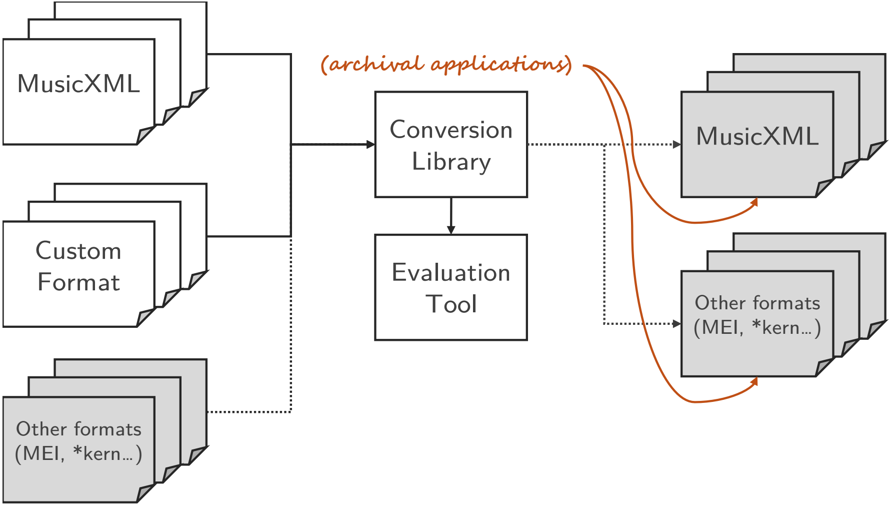
Other formats (MEI, *kern...)

An Overview

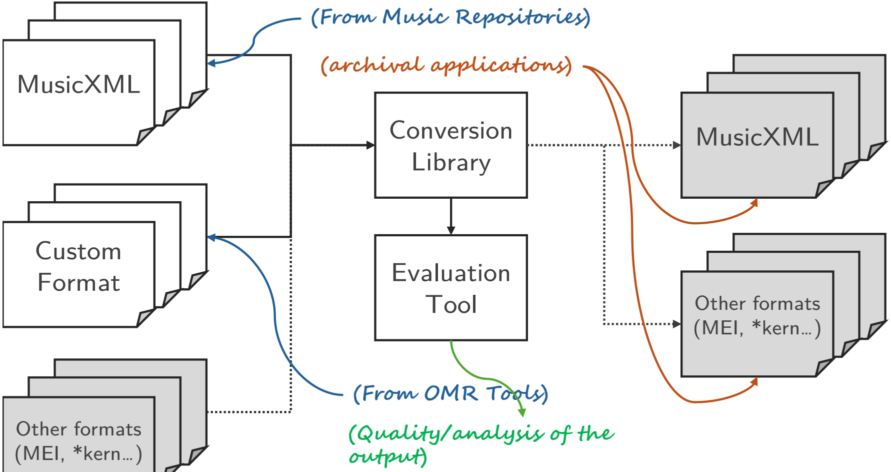


(Quality/analysis of the output)

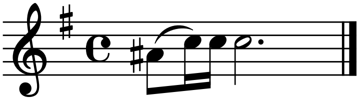
An Overview

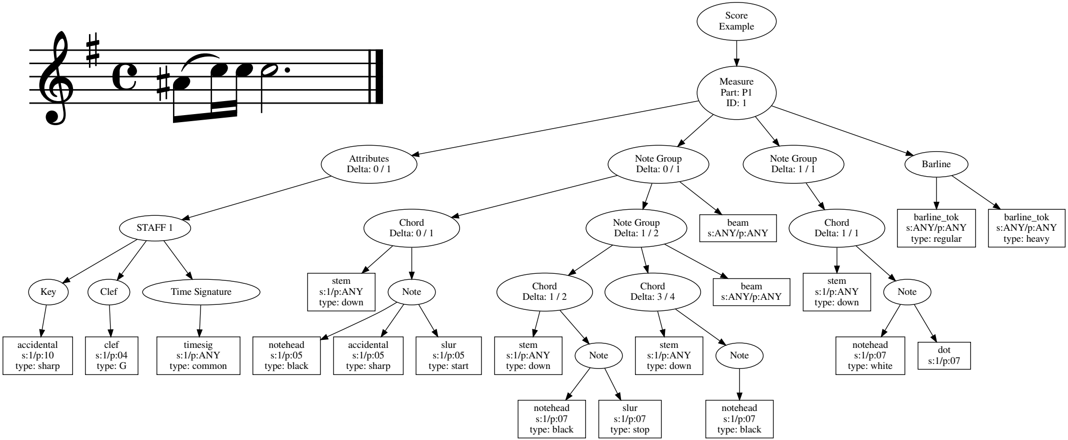

MusicXML
Custom Format
Other formats (MEI, *kern...)
(archival applications)
Conversion Library
Evaluation Tool
MusicXML
Other formats (MEI, *kern...)

An Overview


MusicXML
Custom Format
Other formats (MEI, *kern...)
(From Music Repositories)
(archival applications)
Conversion Library
Evaluation Tool
MusicXML
Other formats (MEI, *kern...)
(From OMR Tools)
(Quality/analysis of the output)

The format



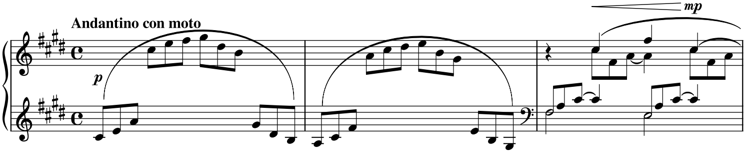

Score
Example
Measure
Part: P1
ID: 1
Attributes
Delta: 0 / 1
Note Group
Delta: 0 / 1
Note Group
Delta: 1 / 1
Barline
STAFF 1
Chord
Delta: 0 / 1
Note Group
Delta: 1 / 2
beam
s:ANY/p:ANY
Chord
Delta: 1 / 1
barline_tok
s:ANY/p:ANY
type: regular
barline_tok
s:ANY/p:ANY
type: heavy
Key
Clef
Time Signature
stem
s:1/p:ANY
type: down
Note
Chord
Delta: 1 / 2
Chord
Delta: 3 / 4
beam
s:ANY/p:ANY
stem
s:1/p:ANY
type: down
Note
accidental
s:1/p:10
type: sharp
clef
s:1/p:04
type: G
timesig
s:1/p:ANY
type: common
notehead
s:1/p:05
type: black
accidental
s:1/p:05
type: sharp
slur
s:1/p:05
type: start
stem
s:1/p:ANY
type: down
Note
stem
s:1/p:ANY
type: down
Note
notehead
s:1/p:07
type: white
dot
s:1/p:07
notehead
s:1/p:07
type: black
slur
s:1/p:07
type: stop
notehead
s:1/p:07
type: black

Key aspects of the format

- **Tree data structure:** Intuitively follows music rules, many algorithms available, mimicks abstract parse tree.

Key aspects of the format

- **Tree data structure:** Intuitively follows music rules, many algorithms available, mimicks abstract parse tree.
- **Time information:** Required, added mostly for sync. **Columns**
- **Placement Information:** Objects that have placement rules incorporate staff and line.



Key aspects of the format

- **Tree data structure:** Intuitively follows music rules, many algorithms available, mimicks abstract parse tree.
- **Time information:** Required, added mostly for sync. **Columns**
- **Placement Information:** Objects that have placement rules incorporate staff and line.
- **Measures are self-contained:** Semantics are parsed *a posteriori*.

Key aspects of the format

- **Tree data structure:** Intuitively follows music rules, many algorithms available, mimicks abstract parse tree.
- **Time information:** Required, added mostly for sync. **Columns**
- **Placement Information:** Objects that have placement rules incorporate staff and line.
- **Measures are self-contained:** Semantics are parsed *a posteriori*.
- **Exchange formats:** We packaged it through XML.

Metrics

- **Tier 0:** Methodology-specific metrics
  - *Whatever is defined for an existing approach*

Metrics

- **Tier 0:** Methodology-specific metrics
- **Tier 1:** Primitive detection

$$\text{precision} = \frac{\|P \cap G\|}{\|P\|}$$

$$recall = \frac{\|P \cap G\|}{\|G\|}$$

Metrics

- **Tier 0:** Methodology-specific metrics
- **Tier 1:** Primitive detection
- **Tier 2:** Structure reconstruction

$$TER = \frac{S + D + I}{\|G\|}$$

Metrics

- **Tier 0:** Methodology-specific metrics
- **Tier 1:** Primitive detection
- **Tier 2:** Structure reconstruction
- **Tier 3:** Semantic reconstruction

$$MNR = \frac{\| \{ n_g \in G : (n_p, n_g) \notin M, \forall n_p \in P \} \|}{\|G\|}.$$

$$\text{Pitch Precision} = \frac{\| \{ (n_p, n_g) \in M : p_{n_p} = p_{n_g} \} \|}{\|M\|}.$$

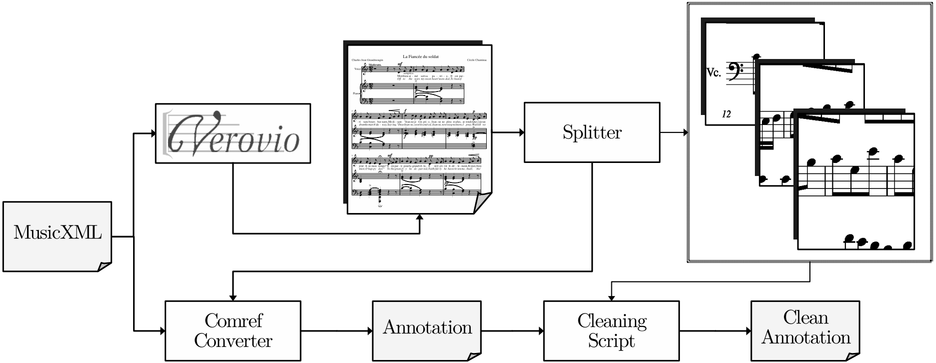
$$\text{Avg. Pitch Shift} = \frac{1}{\|M\|} \sum_{\forall (n_p, n_g) \in M} p_{n_p} - p_{n_g}.$$

Proof-of-concept

A simple baseline: Dataset

- Lieder Corpus, String Quartet Corpus [Gotham et al., 2018] and miscellaneous OpenScore project CC0 transcriptions, totalling 894 MusicXML files.
- Creation of a dataset at measure and page-level and testing it on an OMR system.
- **88 Classes.**

A simple baseline: Dataset



MusicXML

Verovio

Splitter

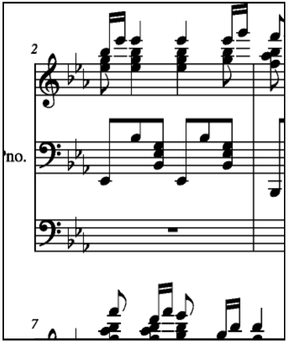
Comref Converter

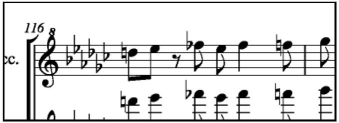
Annotation

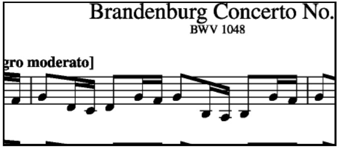
Cleaning Script

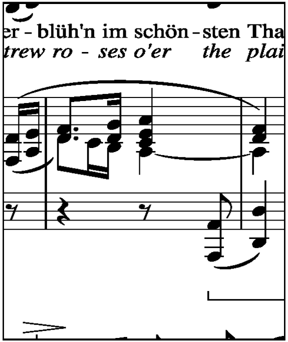
Clean Annotation

A simple baseline: Dataset

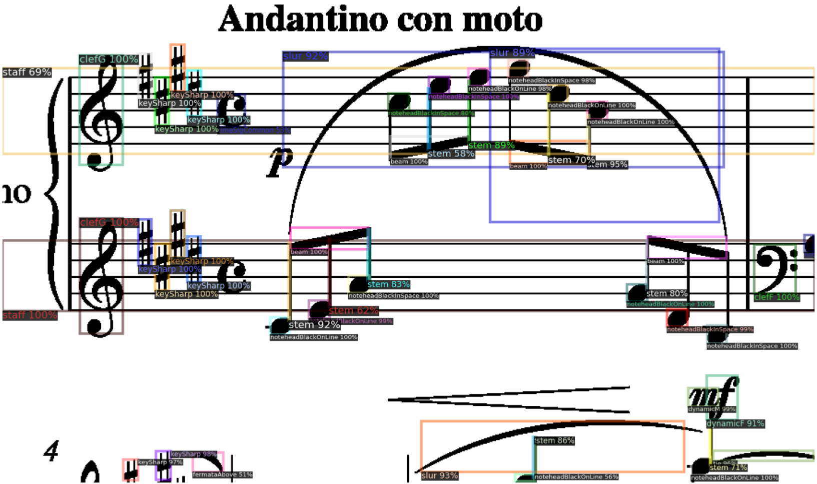








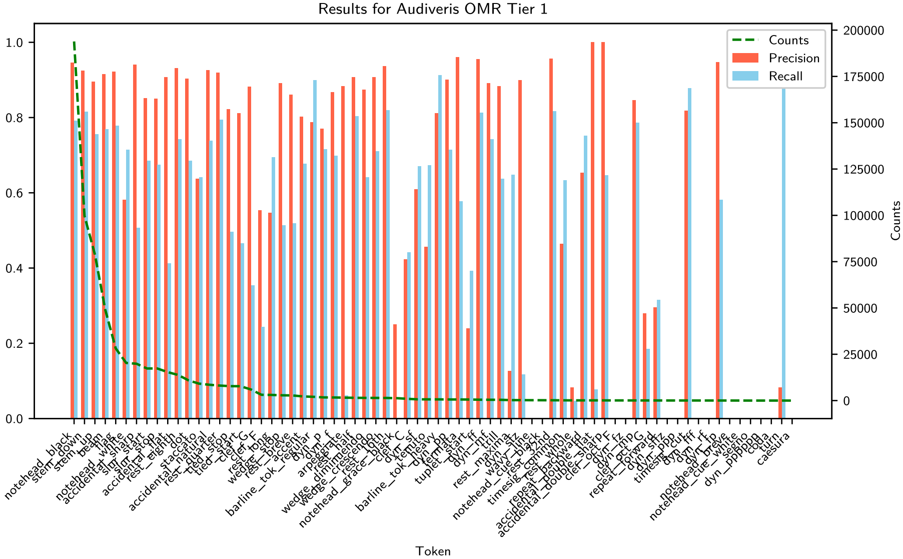
A simple baseline: First approach



A simple baseline: Audiveris

- Off-the-shelf Open Source OMR system for typeset scores - Page Level.
- 45822 predicted measures from the 52884 present on the test set.
- Out of these, 40622 measures from both sets could be matched together, corresponding to a coverage of 76.9%.
- **Simple Matching algorithm based on page order.**

A simple baseline: Audiveris



A simple baseline: Audiveris

TER	Time Shift	Pitch Shift	Staff Shift	Time Prec.	Pitch Prec.	Staff Prec.	FPR	MNR
0.372	-0.096	-0.091	0.022	0.802	0.749	0.963	0.097	0.216

Conclusions

Conclusions & Future Work

- We have proposed a notation format for OMR.
- We have proposed an accompanying set of metrics.
- We have tested this approach with an off-the-shelf OMR system.

Conclusions & Future Work

- We have proposed a notation format for OMR.
- We have proposed an accompanying set of metrics.
- We have tested this approach with an off-the-shelf OMR system.

As future work:

- Support more formats
- Accomodate widespread adoption
- Construct a dataset with object bounding boxes as well.
- Adjust abstractions to simplify processing.

Repository

Acknowledgements

We gratefully thank the participation of Carles Badal and Jan Hajič Jr. in discussions that led to improvements on this work.

This work has been partially supported by the Spanish projects PID2021-126808OB-I00 (GRAIL) and CNS2022-135947 (DOLORES). Pau Torras is funded by the Spanish FPU Grant FPU22/00207 and developed part of this work while funded by the AGAUR Joan Oró FI grant 2023 FI-1-00324. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities.

References I

(2023).
String quartet corpus.
Accessed: 2023-10-10.

Baró, A., Badal, C., and Fornés, A. (2020).
Handwritten Historical Music Recognition by Sequence-to-Sequence with Attention Mechanism.
In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 205–210.

Calvo-Zaragoza, J., Hajič Jr., J., and Pacha, A. (2021).
Understanding Optical Music Recognition.
ACM Comput. Surv., 53(4):1–35.

Calvo-Zaragoza, J. and Rizo, D. (2018a).
Camera-PrIMuS: Neural End-to-End Optical Music Recognition on Realistic Monophonic Scores.
In *19th International Society for Music Information Retrieval Conference*, pages 248–255, Paris, France.

Calvo-Zaragoza, J. and Rizo, D. (2018b).
End-to-End Neural Optical Music Recognition of Monophonic Scores.
Applied Sciences, 8(4):606.
Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Gotham, M., Jonas, P., Bower, B., Bosworth, W., Rootham, D., and VanHandel, L. (2018).
Scores of Scores: An Openscore Project to Encode and Share Sheet Music.
In *5th International Conference on Digital Libraries for Musicology*, pages 87–95, Paris, France. ACM.

References II

Gotham, M. R. H. and Jonas, P. (2022).
The OpenScore Lieder Corpus.
In Münnich, S. and Rizo, D., editors, *Music Encoding Conference Proceedings 2021*, pages 131–136. Humanities Commons.

Hajič, J. and Pecina, P. (2017).
The MUSCIMA++ Dataset for Handwritten Optical Music Recognition.
In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 39–46.
ISSN: 2379-2140.

Mayer, J., Straka, M., Hajič jr., J., and Pecina, P. (2024).
Practical End-to-End Optical Music Recognition for Pianoform Music.

Parada-Cabaleiro, E., Batliner, A., Baird, A., and Schuller, B. (2017).
The SEILS Dataset: Symbolically Encoded Scores in Modern-Early Notation for Computational Musicology.
In *18th International Society for Music Information Retrieval Conference*, Suzhou, China.

Ríos-Vila, A., Rizo, D., Iñesta, J. M., and Calvo-Zaragoza, J. (2023).
End-to-end optical music recognition for pianoform sheet music.
International Journal on Document Analysis and Recognition (IJDAR), 26(3):347–362.

Shatri, E. and Fazekas, G. (2021).
DoReMi: First glance at a universal OMR dataset.
In Calvo-Zaragoza, J. and Pacha, A., editors, *Proceedings of the 3rd International Workshop on Reading Music Systems*, pages 43–49, Alicante, Spain.

References III

Tuggener, L., Elezi, I., Schmidhuber, J., Pelillo, M., and Stadelmann, T. (2018).
DeepScores - A Dataset for Segmentation, Detection and Classification of Tiny Objects.
In *24th International Conference on Pattern Recognition*, Beijing, China. ZHAW.

Tuggener, L., Emberger, R., Ghosh, A., Sager, P., Satyawan, Y. P., Montoya, J., Goldschagg, S., Seibold, F., Gut, U., Ackermann, P., Schmidhuber, J., and Stadelmann, T. (2023).
Real world music object recognition.
Transactions of the International Society for Music Information Retrieval.
Accepted: 2023-09-08T13:52:37Z Publisher: Ubiquity Press.

Tuggener, L., Satyawan, Y. P., Pacha, A., Schmidhuber, J., and Stadelmann, T. (2020).
The DeepScoresV2 Dataset and Benchmark for Music Object Detection.
In *Proceedings of the 25th International Conference on Pattern Recognition*, Milan, Italy.