Data Quality Rules for Digital Score Libraries

David Fiala, Philippe Rigaux, Alice Tacaille, Virginie Thion, Gioqoso Members

Data Quality Rules for Digital Score Libraries

David Fiala[1], Philippe Rigaux[2], Alice Tacaille[3], Virginie Thion[4], and the GioQoso members*

[1] Univ. Tours, CESR, France
[2] CNAM, Cedric, Paris, France
[3] IReMus, Paris, France
[4] Univ. Rennes, IRISA, Lannion, France

February 2018

Abstract

This document deals with data quality in Digital Libraries of Scores (DSL). Data quality management means assessing and possibly improving quality of data. It is a major concern of the information system lifecycle. The first and fundamental step of the data quality management process consists in eliciting data quality requirements. Because data quality is defined as being the *fitness for use* of data (meaning that the notion of data quality depends on the context), it is a conceptually complex notion, whose implementation for a given use case is not trivial. So context-dependant guidelines are needed in order to help users to define data quality in their context. This is the problem that we tackle here, by proposing a set of quality rules specific to DSL, which can serve as a basis in order to elaborate users' quality requirements.

*Léo Bechet (ENSSAT, Lannion), Vincent Besson (CESR, Tours), David Fiala (CESR, Tours), Francesco Foscarin (CNAM, Paris), Alban Framboisier (IReMus, Paris), Adrien Guillaud-Rollin (ENSSAT, Lannion), Marco Gurrieri (CESR, Tours), Fayçal Hamdi (CNAM, Paris), Olivier Pivert (IRISA, Lannion), Samira Si-Said Cherfi (CNAM, Paris), Nicolas Travers (CNAM, Paris).

1 Introduction

There is a growing availability of music scores in digital format, produced by numerous individuals and institutions, and often publicly accessible from web sites and social media. This has been made possible by the combination of two factors: mature, easy-to-use music editors, including open-source ones like MuseScore [MuseScore, 2017], and sophisticated music notation encodings. Leading formats today are those which rely on XML to represent music notation as structured documents. MusicXML [Good, 2001] is probably the most widespread one, due to its acceptance by major engraver softwares (Finale, Sibelius, and MuseScore) as an exchange format. The MEI initiative [Rolland, 2002, MEI, 2015], inspired by the TEI, attempts to address the needs of scholars and music analysts with an extensible format [Hankinson et al., 2011]. Recently, the launch of the W3C Music Notation Community Group [W3C, 2015] confirms that the field tends towards its maturity, with the promise to build and preserve large collections of scores encoded with robust and well-established standards. We are therefore facing emerging needs regarding the storage, organization and access to potentially very large Digital Libraries of Scores (DSL).

It turns out that building such a DSL, particularly when the acquisition process is collaborative in nature, gives rise to severe quality issues. In short, we are likely to face problems related to *validity* (measure durations, voices and parts synchronisation), *consistency* (heterogeneous notations, high variability in the precision of metadata, undetermined or inconsistent editorial rules), *completeness* (missing notes, directives, ornamentation, slurs or ties), and *accuracy* (music, lyrics).

There are many reasons for this situation. First, encoding formats have changed a lot during the last decades. We successively went through HumDrum and MIDI to finally come up with modern XML formats such as MusicXML and MEI [Selfridge-Field, 1997]. A lot of legacy collections have been converted from one encoding to the other, losing information along the way. Given the cost and time to edit scores, incorporating these collections in a modern repository is a strong temptation, but requires to accept, measure, and keep track of their quality shortcomings.

Second, the flexibility of music notation is such that it is extremely difficult to express and check quality constraints on the representation. Many of the formats we are aware of for instance do not impose that the sequence of events in a measure exactly covers the measure duration defined by the metrics. As another example, in polyphonic music, nothing guarantees that the parts share the same metric and same duration. So, even

with the most sophisticated encoding, we may obtain a score presentation which does not correspond to a meaningful content (the definition of which is context-dependent), and will lead to an incorrect layout (if not a crash) with one of the possible renderers.

Third, scores are being produced by individuals and institutions with highly variables motivations and skills. By "motivation", we denote here the purpose of creating and editing a score in digital format. A first one is obviously the production of material for performers, with various levels of demands. Some users may content themselves with schematic notation of simple songs, whereas others will aim at professional editing with high quality standards. The focus here is on rendering, readability and manageability of the score sheets in performance situation. Another category of users (with, probably, some overlap) are scientific editors, whose purpose is rather an accurate and long-term preservation of the source content (including variants and composer's annotations). The focus will be put on completeness: all variants are represented, editor's corrections are fully documented, links are provided to other resources if relevant, and collections are constrained by carefully crafted editorial rules. Overall, the quality of such projects is estimated by the ability of a document to convey as respectfully as possible the composer's intent as it can be perceived through the available sources. Librarians are particularly interested by the searchability of their collections, with rich annotations linked to taxonomies [Riley and Mayer, 2006]. We finally mention analysts, teachers and musicologists: their focus is put on the core music material, minoring rendering concerns. In such a context, part of the content may be missing without harm; accuracy, accessibility and clarity of the features investigated by the analytic process are the main quality factors.

Finally, even with modern editors, qualified authors, and strong guidelines, mistakes are unavoidable. Editing music is a creative process, sometimes akin to a free drawing of some graphic features whose interpretation is beyond the software constraint checking capacities. A same result may also be achieved with different options (e.g., the layer feature of Finale), sometimes yielding a weird and convoluted encoding, with unpredictable rendering when submitted to another renderer.

The authors of the present paper participate in the production, maintenance and dissemination of digital libraries of scores encoded in XML (mostly, MEI). One of these DSLs is the Neuma platform. Neuma is an open repository of scores in various formats, managed by the IReMus[1], and

[1] *Institut de Recherche en Musicologie*, `http://iremus.cnrs.fr`.

publicly accessible at `http://neuma.huma-num.fr`. The CESR[2] publishes rare collections of Renaissance music for scholars and musicians (see, e.g., the "Lost voices" project, `http://digitalduchemin.org`). Both institutions have been confronted with the need to address issues related to the consistent production of high-level quality corpora, and had to deal with the poor support offered by existing tools. The current, ad-hoc, solution adopted so far takes the form of editorial rules. The approach is clearly unsatisfying and unable to solve the above challenges. Even though we assume that the scores are edited by experts keen to comply with the recommendations, nothing guarantees that they are not misinterpreted, or that the guidelines indeed result in a satisfying encoding. Moreover, rules that are not backed up by automatic validation safeguards are clearly non-applicable in a collaborative context where un-controlled users are invited to contribute to the collections. Managing data quality in such a context is then a major issue.

The document is organised as follows. In Section 2, we briefly introduce the process of data quality management and the motivation for designing the catalog of quality rules presented in this document. In Section 3, we present the multidimensional conceptual model, which is specific to DSL data, that we defined in order to classify the quality rules. Sections 4, 5 and 6 compose the catalog of data quality rules, categorized in these three sections according to the conceptual model above mentioned. The concrete implantation and the evolution of the catalog are discussed in Section 7. We conclude this work in Section 8.

2 Data quality management

Much published data suffers from quality problems [Zaveri et al., 2016]. It is now well-recognised that these endemic problems may lead to severe consequences, and that managing the quality of data conditions the success of most existing information systems [Eppler and Helfert, 2004]. The last two decades have then witnessed an increasing interest in data quality management, from both a theoretical and a practical point of view.

Data quality is a complex concept, which embraces different semantics depending on the context [Redman, 1996]. It is described through a set of quality *dimensions* aiming to categorize criteria of interest. Classical quality dimensions are *completeness* (the degree to which needed information is

[2] *Centre d'Etudes Supérieures de la Renaissance*, `http://cesr.univ-tours.fr`.

present in the collection), *accuracy* (the degree to which data are correct), *consistency* (the degree to which data respect integrity constraints and business rules) and *freshness* (the degree to which data are up-to-date). Data quality over a dimension is measured according to a set of *metrics* that allow a quantitative definition and evaluation of the dimension. Examples of metrics are "the number of missing metadata" for the evaluation of the *completeness*, and "the number of conflicting duplicates" for *consistency*. These are simple examples but the literature proposes a large range of dimensions and metrics, conceptualized in quality models [Batini and Scannapieco, 2016]. Of course, not all the existing dimensions and metrics may be used for evaluating data quality in a given operational context. An important property concerning data quality is that it is defined as being *fitness for use* of data, meaning that quality measurement involves dimensions and metrics that are relevant to a given (set of) user(s) for a given *usage*. User u_1 may be concerned by some quality metrics for a specific usage, by some other metrics for another one, and they can be completely different than those needed by user u_2.

The literature proposes general methodologies for managing data quality [Batini et al., 2009]. We focus here on *quality assessment*. Roughly speaking, each assessment methodology includes a *quality definition* stage and a *quality measurement* one.

The first stage, the quality definition, consists in eliciting data quality requirements of interest. Concretely, this means choosing a set of quality metrics, and eventually thresholds associated with, that allows to measure in what extend the data fit the quality requirements according to data usages.

Because data quality is *fitness for use* (depends on the context), defining data quality is not trivial. Dedicated methodological guidelines can be followed like the *Goal Question Metric* paradigm [Basili et al., 1994], which proposes to define quality metrics according to a top-down analysis of quality requirements, whose stages are defined hereafter.

1. For each user (or each user role) and for each of his/her usage of data, conceptual *business goals* are identified. A business goal specifies the intent of a quality measurement according to a usage of data.

   (Example) We make this process more concrete by illustrating it on a simple example. Let us assume that a business user retrieves music scores in order to *Perform a given algorithm that searches for similar patterns in the parts of a music score*. This is a business goal.

2. Each goal is then refined into a set of operational *quality questions*, which are a first step towards eliciting the quality requirements.

   (Example) For the example, the user may express that the results of his/her study is relevant provided that data is complete enough and that the used algorithm computes relevant results provided that data is accurate enough. Quality questions associated with this use case could then be *(QQ1) Does the data contain all needed information?* and *(QQ2) Are the notes accurate?*

3. Each quality question is then itself expressed in terms of a set of quantitative quality metrics with possible associated thresholds (expected values).

   (Example) The quality question *(QQ1)* could be refined into two more precise quality questions. A first "quantitative" quality question could be *Is the figured bass available?*, measured by the quality metric ($QQ1/M1$) defined below.

   (QQ1/M1) Availability of the figured bass. (Boolean result).

   A second quantitative quality question associated with *(QQ1)* could be *Does each measure cover exactly the expected number of beats?*, measured by the quality metric ($QQ1/M2$) defined below.

   (QQ1/M2) Number of satisfactory measures over the total number of measures.

   Assuming that the algorithm is robust up to 10% of malformed measures, then the threshold 0.9 could be associated with the quality metric *(QQ1/M2)*.

   Concerning the quality question *(QQ2)*, it could be refined into a quality metric that measures the syntactic accuracy of the notes, meaning that each note should be an existing one (which belongs to the usual range of notes). A third quality metric could then be ($QQ2/M3$) defined below.

   (QQ2/M3) Number of syntactically accurate notes over the total number of notes.

Measuring the quality metrics enables to (partly) answer to the quality questions, and consequently enables to decide whether the data satisfy the

requirements for the given business goal. As soon as data quality metrics are defined, one can consider different processes for their computation, including collaborative ones if the information system makes it possible.

Data quality methodologies of literature are designed at a generic level, leading to difficulties for their implementation in a specific context (operational context and available information system and data). Additional context-dependent quality methodologies are then needed [Barrau et al., 2016]. In particular, the literature proposes a large range of quality metrics [Batini and Scannapieco, 2016, Zaveri et al., 2016] but such metrics are general ones. Quality metrics that are specific to the data of the considered domain are still needed, more specifically in the context of DSL data for which, to our knowledge, only few quality metrics were proposed in the literature (the only work of the literature that proposes some quality metrics for DSLs is [Besson et al., 2016]).

In the following of this document, we propose a catalog of quality rules specific to DSL data, which was elaborated according to the authors experience in maintaining and using DSLs. The idea is that this catalog can serve as a basis in order to elaborate users' quality requirements, by "picking" relevant quality rules according to specific use cases.

3 Overview of quality rules for DSL

Identifying quality metrics started with the work presented in [Besson et al., 2016], where the authors propose a DSL-specific framework for data quality management. Some DSL-specific quality metrics are attached to this framework. The catalog that we propose in the following generalizes and extends this proposal.

Based on the authors' practical experience and skills, we identified a set of quality rules specific to DSL data. A data quality rule expresses a possible quality requirement. It may be used either (i) in order to tag the data where a quality problem occurs, or (ii) in order to compute a quality metric associated with a score or a corpus. For instance, the quality rule "*Each note is syntactically correct, meaning that it is an existing one (which belongs to the usual range of notes)*" expresses the fact that having syntactically accurate notes is a data quality requirement. Such quality rule can lead to *tag* syntactically inaccurate notes that appear in music scores of interest. It

can also lead to *compute a quality metric* in order to assess the quality of a music score according to the rule, like the number of syntactically correct notes over the total number of notes appearing in the score. By extension, quality metrics at the corpus level may easily be defined, for instance the average and standard deviation of the corresponding metric at the score level, computed over the set of scores that belong to the corpus.

In order to improve the understandability and the usability of the quality rules (not only in the catalog but also for the rendering of quality reports designed for the end-users), we classified them into a multidimensional conceptual model. Such a model contains two analysis axes:

- a *DSL-specific* analysis axis reflects the DSL-specific business point of view of the data, and
- a *data quality* analysis axis reflects the classical data quality point of view organizing the quality issues according to quality dimensions.

Each axis offers different levels of abstraction. The proposed model is illustrated in Figure 1.

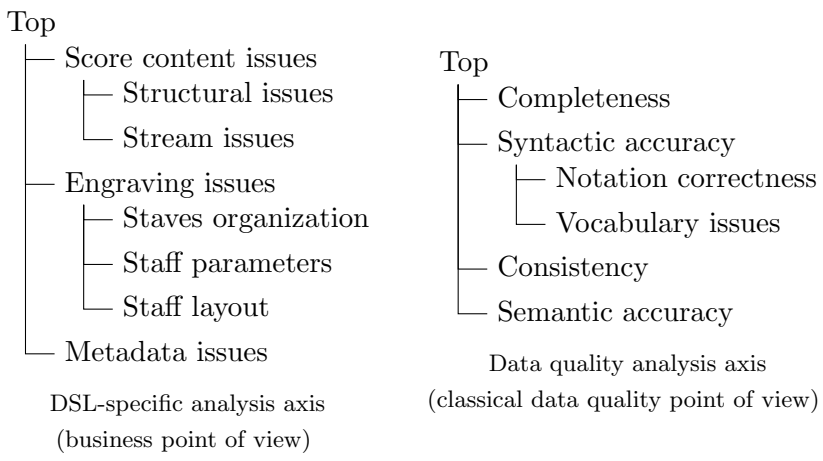

Figure 1: Two-dimensional model for classifying quality rules

The DSL-specific analysis axis (see Figure 1) allows to classify quality rules, according to the DSL specific business point of view. It contains the following elements:

Score content issues This part covers all aspects related to the *score content*, independently from any encoding or rendering concern. Essentially, it captures the *structural* organization of a score in parts and streams.

Engraving issues Score engraving denotes the mapping of the score content into a set of staves, which will be used for the rendering of the music score.

Metadata issues Metadata is data about data, *i.e.*, in our case, any content that annotates either the score content or the score engraving. The title, subtitle, composer are metadata that annotate a score as whole.

This analysis axis is detailed in [Foscarin et al., 2018].

The *data quality* analysis axis (see Figure 1) models the simplest way proposed in the literature in order to classify generic quality issues [Batini and Scannapieco, 2016]. More complex classifications, which possibly contain several dimensions (like e.g. the one proposed in [Peralta et al., 2009]), could be used. In this case, the DSL specific business axis would just have to be added to the set of axes that belong to the quality model.

Each quality rule is classified according to the axes of the model. For instance, the quality rule 4 (Available key signature) belongs to the *Completeness* position of the *data quality* analysis axis as it aims to measure the completeness of data, and to the *Score content / Stream issues* position of the *DSL-specific* analysis axis as its concerns the voices of the music score content.

Such a model allows to perform, if needed, a multidimensional analysis of the results issued from the data quality assessment, for instance using OLAP-based warehouse technologies [Jarke et al., 2001].

In the following, according to the classification above-defined, the quality rules are primary classified into three classes: the rules that concern the content of the music score presented in Section 4, those that concern the engraving of the music score presented in Section 5 and those that concern the metadata attached to the music score presented in Section 6. We now present a review of the quality rules that we identified for each of these categories, refined according to the data quality dimensions above-given,

that is to say the completeness, the consistency and the accuracy (syntactic and semantic).

4 Quality rules concerning the score content

Quality rules over the content only concern the encoded content of the score, independently from its encoding or rendering. Content issues concern either the structural level or the stream level of the score.

4.1 Quality rules concerning the structure of the score content

The structural level concerns the organisation of the score in parts. Structural issues may be studied from the perspective of the completeness, the consistency and the accuracy of the information.

Completeness of the structure The completeness of the structure concerns the availability of the expected parts.

QUALITY RULE 1 (Available parts). *Each expected part appears in the music score.*

Consistency of the structure The following rule 2 consists in checking the consistency of the parts' length with each other.

QUALITY RULE 2 (Aligned parts). *The parts are aligned.*

Semantic accuracy of the structure Even if the structure of a music score is consistent, this does not mean that this information is semantically correct, meaning that the information contained in the music score corresponds to the semantically accurate real world information (for instance, a part that does not belong to the real work music score appears, or the parts are aligned but their length is not accurate). Additional quality rules are defined in order to express such a requirement (if computable).

For each relevant information of the structure:

QUALITY RULE 3 (Semantically accurate structure). *The information is semantically accurate.*

The semantic accuracy rule concerning the structure is usually rather complex to implement because its necessitates to compare the encoded value with a trustable accurate reference like another reliable source or with a real world value given by a business expert. (Another reliable source or a business expert resources that are rarely available...)

4.2 Quality rules concerning the stream of the score content

For each part, the stream (notes) level is considered. The quality issues concerning the streams deal either with the pitch or the rhythm (presented together here but can be separately considered if needed), or the lyrics.

4.2.1 Quality rules concerning the pitch and the rhythm

The content of a music score is a complex information. Much quality rules may be associated to such data, that refer to quality problems according to the completeness, the syntactic accuracy, the consistency and the semantic accuracy of data. These quality dimensions are respectively addressed hereafter.

Completeness according to the pitch and the rhythm According to the completeness of data, several rules may the thought of, concerning the figured bass and the notes.

QUALITY RULE 4 (Available key signature). *The key signature is defined.*

QUALITY RULE 5 (Available time signature). *The time signature is defined.*

QUALITY RULE 6 (No missing beat). *Each measure is complete, meaning that it covers at least the number of beats defined by the time signature (if not then a note could be missing).*

Of course, the rule 6 can be computed only if the time signature is available[3]. In the following, we consider that the availability of the elements mentioned in a quality rule is an obvious prerequisite to the computation of the rule.

QUALITY RULE 7 (Ornaments). *The performance indications (appoggiaturas, slurs, articulation symbols, ...) are uniformly present.*

[3]The rule also appears in the consistency dimension as it measures the consistency of the information available in the time signature with the number of beats that belong to each measure.

Syntactic accuracy according to the pitch and the rhythm Even if an information is available, checked by the quality rules dealing with the completeness dimension, this does not mean that the provided information is syntactically accurate. The following rules go further in the inspection of the quality of this information, by checking if each provided information belongs to its respective usual ranges (belong to the intended vocabulary).

A first rule consists in checking if the document respects the encoding format that it is supposed to, meaning that it satisfies the schema associated with its encoding format. For instance, if the document is declared as being a MEI document then it has to respect the MEI standard.

QUALITY RULE 8 (Validity w.r.t. the encoding format). *The music score respects the encoding format.*

The encoding formats still offer a large degree of freedom. If they relatively constraint the structure of the document (tags and their arrangement), they do not precisely control the domain of range of the information embedded in the tags, known as the syntactic accuracy of the information. Then complementary quality rules must be defined in order to control the syntactic accuracy of the voices.

QUALITY RULE 9 (Syntactic accurate key signature). *The key signature belongs to the usual range of key signatures.*

QUALITY RULE 10 (Syntactic accurate time signature). *The time signature belongs to the usual range of time signatures.*

QUALITY RULE 11 (Syntactic accurate of the notes). *Each note is syntactically correct, meaning that it is an existing note (which belongs to the usual range of notes).*

QUALITY RULE 12 (Syntactic accurate slur). *Each slur is syntactically accurate, at least with a start note and an end one that belong to the same part. Slurs do not overlap on a same voice.*

QUALITY RULE 13 (Syntactic accurate appoggiatura). *Each appoggiatura is syntactically accurate, meaning that, at least, it forms a second interval with the note it prefixes.*

QUALITY RULE 14 (Syntactic accurate chord). *A chord contains at least two notes.*

Let us note that the above rules, which can be relatively easily computed, only control the syntax of the information but do not check the semantic accuracy that would consist, for instance for an instrument associated with a voice, in checking that the provided instrument is the real world accurate one associated with the voice. Consistency rules, given below, are a step towards the detection of semantically inaccurate data. They make it possible to identify suspect data.

Consistency according to the pitch and the rhythm The consistency rules check the consistency of the voices w.r.t. other information available in the music score, for instance the instrument associated with the voices, the key or the time signatures.

A first set of rules consists in expressing the adequacy of the number of beats that appear in a measure w.r.t. the time signature defined for the part the measure belongs to. The following rules may express quality requirements according to this adequacy. These versions may be contradictory or overlapping because they express different points of view for the consistency. The relevant version has to be chosen by the user according to its use case.

QUALITY RULE 15 (Complete measure). *Each measure is complete (covers at least the number of beats defined by the time signature).*

QUALITY RULE 16 (Non-overflowing measure). *No measure overflows (each measure covers at most the number of beats defined by the time signature).*

QUALITY RULE 17 (Accurate number of beats in the measure). *Each measure covers exactly the number of beats defined by the time signature.*

The rule 17 (Accurate number of beats in the measure) expresses the conjunction of the rule 15 (Complete measure) satisfaction and the rule 16 (Non-overflowing measure) satisfaction.

In some cases, more especially in the case of research-based study of early music, the intent of the composer has to be encoded as faithfully as

possible. This means that the encoding must reflect as possible the rendering of the initial music score handwritten by the composer. In this situation, it is not uncommon to see a music note whose beats belong to two adjacent maesures. When musicologists analyse early music, this is not considered as being a quality problem (on the contrary). In order to fit this requirement, a relaxed version of the rule 17 is defined in the rule 18.

QUALITY RULE 18 (Accurate number of beats w.r.t. a frame of measures). *Each frame of $\mathcal{N}$ measures respects the number of beats defined in the time signature (where $\mathcal{N}$ is given as a parameter of the quality rule). More formally, for each measure $\mathcal{M}$, if the measure $\mathcal{M}$ does not strictly cover the number of beats defined in the time signature (i.e. the measure $\mathcal{M}$ does not satisfy the rule 17) then there is a frame of $\mathcal{N}$ adjacent measures including the frame $\mathcal{M}$ such that the number of beats of the frame is $\mathcal{N}$ times the number of beats defined by the time signature (i.e. the global frame respects the time signature).*

More complex variants of these rules, tolerant to appoggiaturas and acciaccatura (which may not be taken into account in the number of beats of the measure), could be defined.

Each note that appears in a part is associated with an musical instrument or a voice by extension. The following rules expresses the consistency of the note w.r.t. the instrument that has to play it.

QUALITY RULE 19 (Notes in instrument tessitura). *Each note of a part belongs to the tessitura of the instrument or voice that is associated with the part.*

If the rule 19 is not satisfied then this means that either some notes are inaccurate (those that do not belong to the scope defined by the instrument tessitura), or (non-exclusive) that the instrument is inaccurate.

The information of the instrument associated with a part allows to define some other interesting consistency rules.

QUALITY RULE 20 (Chords composition w.r.t. the instrument). *If a chord contains two occurrences of the same note, then the instrument allows to play it (a string instrument can play such a chord while most of the wind instruments cannot).*

If the rule 20 is not satisfied then this means that either the chord is inaccurate, or (non-exclusive) the instrument is inaccurate.

Semantic accuracy according to the pitch and the rhythm Even if the information of a music score is consistent, this does not mean that this information is semantically correct, meaning that the information contained in the music score corresponds to the semantically accurate real world information. Additional quality rules are defined in order to express such a requirement (if computable).

QUALITY RULE 21 (Semantically accurate key signature). *The key signature is semantically accurate.*

QUALITY RULE 22 (Semantically accurate time signature). *The time signature is semantically accurate.*

QUALITY RULE 23 (Semantically accurate voices). *Each element of the voices (note, slurs, articulations) is semantically accurate.*

Semantic accuracy rules are usually rather complex to implement because they necessitate to compare the encoded value with a trustable accurate reference like another reliable source or with a real world value given by a business expert. (Another reliable source or a business expert resources that are rarely available...)

It is worth noticing that some rules are interdependent. For instance, if an element respects the rule 21 (Semantically accurate key signature) then, by definition, it respects the rule 9 (Syntactic accurate key signature), and then intrinsically respects the rule 4 (Available key signature). By contraposition, if an element does not respect the rule 4 (Available key signature) then it does not respect the rule 9 (Syntactic accurate key signature) and by transition does not respect the rule 21 (Semantically accurate key signature).

4.2.2 Quality rules concerning the lyrics

Like the other quality rules, the rules concerning lyrics apply only if needed and, of course, only concern vocal music.

Completeness of the lyrics The completeness rule consists in checking the availability of lyrics.

QUALITY RULE 24 (At least a lyric per note). *A lyric sequence is associated with each (expected) voice.*

QUALITY RULE 25 (At least a lyric per note). *For each available lyric sequence, a lyric element is associated with each note.*

The rules 24 and 25 concern only the availability of lyrics. The lyrics' values are controlled by other quality rules.

Syntactically accuracy of the lyrics The following rules go further in the inspection of lyrics quality, by checking their syntax.

QUALITY RULE 26 (Single lyrics). *There is at most one lyric element associated with each note.*

A more constrained version of the rule 26 is the rule 27.

QUALITY RULE 27 (Singable lyrics). *Each lyric element associated with a note is singable (each lyric element is a syllable).*

QUALITY RULE 28 (Consistent numbering of the verses). *The numbering the verses is consistent. More formally, the rule is defined by a recursive definition composed of the two (sub)rules (R1) the first verse has number* 1 *and (R2) for each other verse, if the verse has number* n *(with* $n \geq 2$*) then it occurs after the verse* $n - 1$*.*

Consistency of the lyrics Even if a lyric is syntactically correct, this does not implies that it is semantically accurate. Like for the voices discussed before, the semantic accuracy is difficult to control. A first step towards controlling accuracy is to express consistency rules that allow to identify semantically suspect values.

QUALITY RULE 29 (Lyrics associated with note). *Each lyric element is associated with a note (the lyrics are consistent with the notes).*

Semantic accuracy of the lyrics Even if the information of the lyrics is consistent, this does not mean that this information is semantically correct, meaning that the information contained in the music score corresponds to the semantically accurate real world information. An additional quality rule is then defined in order to express such a requirement (if computable).

For each relevant information of the lyrics:

QUALITY RULE 30 (Semantically accurate lyric elements). *The information of the lyrics is semantically accurate.*

Such a semantic accuracy rule is usually rather complex to implement because its necessitates to compare the encoded value with a trustable accurate reference like another reliable source or with a real world value given by a business expert.

5 Quality rules concerning the engraving

Score engraving denotes the mapping of the score content into a set of staves according to engraving rules. The engraving rules take a score content, determine the number of staves, allocate parts to staves, and develop the stream representation on each staff. The score engraving issues concern either the organization of the staves, or the staff parameters or the staff layout.

We consider that this quality facet concerns the consistency of data (the result produced by the engraving step may be consistent according to the score content and the engraving rules).

Consistency of the engraving Concerning the *organization of the staves*, two quality rules may the thought of.

QUALITY RULE 31 (Validity of the staff order). *The staff order is valid.*

QUALITY RULE 32 (Number of parts per staff). *Each staff contains the expected number of parts.*

Concerning the *staff parameters*, two quality rules may the thought of.

QUALITY RULE 33 (Validity of the key signature). *The key signature is valid.*

Quality rule 34 (Validity of the clef). *The clef signature is valid.*

Concerning the *staff layout*, two quality rules may the thought of.

Quality rule 35 (Validity of the duration). *The duration is valid.*

Quality rule 36 (Validity of the beaming). *The beaming is valid.*

These rules can be checked by automatic procedures based on the analysis of the music score content [**?**].

Semantic accuracy of the engraving Of course, the semantic accuracy may also be checked.

For each relevant information of the engraving issue:

Quality rule 37 (Semantically accurate engraving). *The engraving information is semantically accurate.*

6 Quality rules concerning the metadata

Much metadata may be attached to a music score (see for instance the MEI guidelines describing the MEI header of a document [Music Encoding Initiative Board, 2016]). The quality rules below deal with some classical relevant metadata. Quality requirements could concern the availability of metadata information, and possibly their accuracy (rule 49 if computable).

Completeness of the metadata

Quality rule 38 (Available title). *The title of the music score is available.*

Quality rule 39 (Available composer). *The composer of the music score is available.*

Quality rule 40 (Available date). *The date of creation of the music score is available.*

QUALITY RULE 41 (Available opus reference). *Each music score belongs to an opus (a corpus). This reference is available in metadata.*

QUALITY RULE 42 (Available copyright). *The copyright of the music score document is available.*

QUALITY RULE 43 (Available author). *The author (producer) of the numeric music score document is available.*

QUALITY RULE 44 (Available date). *The date of production of the numeric music score is available.*

QUALITY RULE 45 (Available instruments). *An instrument is associated with each part.*

Syntactic accuracy of the metadata

QUALITY RULE 46 (Syntactic accurate instrument/voice). *Each instrument/voice associated with a part is syntactically correct (is an existing instrument/voice).*

Such a rule may be automatically computed by checking the occurrence of each instrument/voice in a predefined dictionary or nomenclature (a local repository or a distant source like in an open encyclopedia). Accepted instrument/voices and their number may by restricted to a given subset for specific works.

Consistency of the metadata

QUALITY RULE 47 (Known instrument at the creation). *Each instrument existed at the date of creation of the music score.*

If this rule is not satisfied, then this means that either the date of creation is inaccurate or (non-exclusively) the instrument is inaccurate. This rule then also appears in the quality rules concerning the content for checking the instrument.

QUALITY RULE 48 (Known instrument by the composer). *Each instrument existed during the period of life of the composer.*

If this rule is not satisfied, then this means that either the composer is inaccurate or (non-exclusively) the instrument is inaccurate. This rule then also appears in the quality rules concerning the content for checking the instrument.

Semantic accuracy of the metadata For each relevant metadata information:

QUALITY RULE 49 (Semantically accurate metadata). *The metadata information is semantically accurate.*

7 Evolution of the catalog, the NEUMA platform

The NEUMA platform [Rigaux et al., 2012] is a digital library devoted to the preservation and dissemination of symbolic music content (scores). The corpora of NEUMA are publicly available, on open access at `http://neuma.huma-num.fr`. Some of the quality rules presented in the previous sections are currently being implemented in the NEUMA platform in the form of a quality module [Besson et al., 2016] that detects quality problems in the data and tags them [Si-Said Cherfi et al., 2017a, Si-Said Cherfi et al., 2017b]. A graphical user interface allows their visualisation, as illustrated in Figure 2. In such an interface, the user chooses a music score whose quality has to be checked, her/his data problems of interest (in the right frame in Figure 2). After the quality module processing, graphical elements appear in the form of an overprinting layer on the layout of the music score (the coloured points in Figure 2) in order to report identified quality problems.

The set of quality rules that we proposed is obviously not exhaustive. New quality rules are regularly discovered and added to the framework. Refinements of the quality model that allows to classify the quality rules are also discussed. The catalog is then subject to evolutions and enrichments.

An up-to-date version of the quality metrics implemented in the NEUMA platform is available at `http://neuma.huma-num.fr/quality` (in the right frame of the interface, see Figure 2).

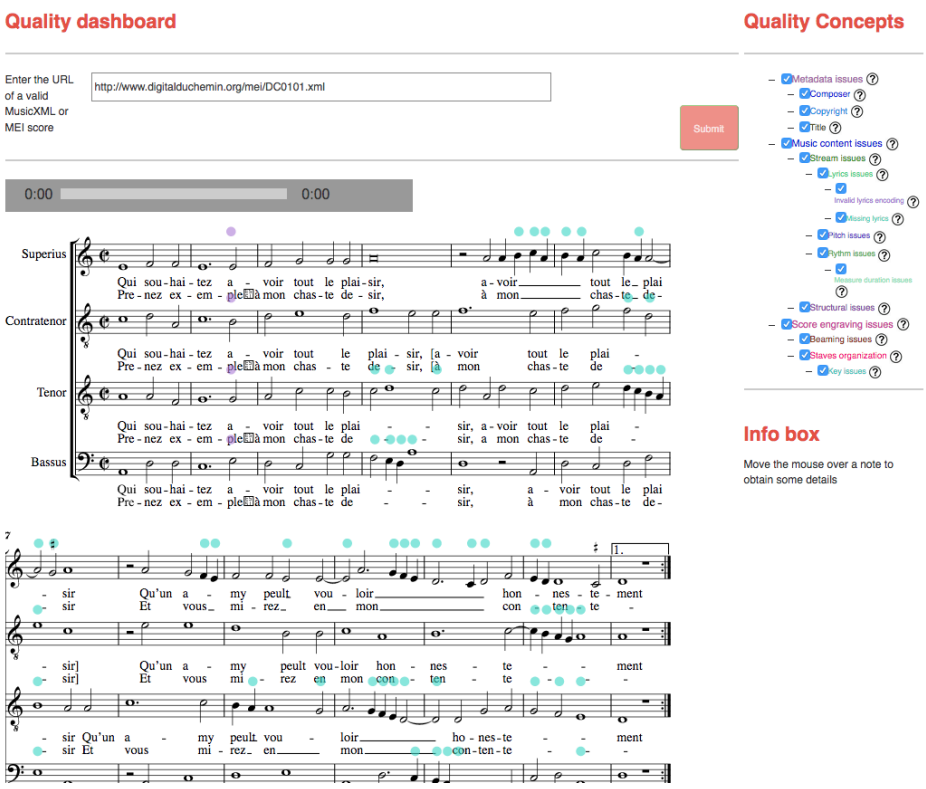

Figure 2: Visualization of quality problems in the NEUMA platform

8 Conclusion

In this document, we consider the problem of data quality management of encoded music scores in DSL.

After a brief state of the art of general data quality management methodologies, we highlighted the fact that context-dependent data quality concepts are missing for the real live implementation of these methodologies.

We then proposed a catalog of quality rules that are specific to DSL data. The quality rules were exhibited according to the authors experience in maintaining and using DSL. The catalog can serve as a basis in order to elaborate users'quality requirements, by choosing relevant quality rules according to specific use cases. In order to classify the rules, we proposed a specific taxonomy that mixes the classical data quality point of view that organises quality rules/metrics according to quality dimensions and a business vision that introduces supplementary DSL-dependant levels in the taxonomy. Some rules of the catalog are implemented in the Neuma platform [Rigaux et al., 2012], in the form of an open access quality module (see [Si-Said Cherfi et al., 2017a], [Si-Said Cherfi et al., 2017b] and [Foscarin et al., 2018] for details).

This catalog is a basis. We believe that it is not static as other quality rules can be thought of. New contributions should enrich this framework soon.

Acknowledgements

This work has been funded by the French National Center for Scientific Research (CNRS) under the *défi Mastodons GioQoso* [GioQoso, 2016].

References

[Barrau et al., 2016] Barrau, D., Barthélémy, N., Kedad, Z., Laboisse, B., Nugier, S., and Thion, V. (2016). Gestion de la qualité des données ouvertes liées - État des lieux et perspectives. *Revue des Nouvelles Technologies de l'Information*.

[Basili et al., 1994] Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). *Encyclopedia of Software Engineering*, chapter The Goal Question Metric Approach. Wiley.

[Batini et al., 2009] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3):16:1–16:52.

[Batini and Scannapieco, 2016] Batini, C. and Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer.

[Besson et al., 2016] Besson, V., Gurrieri, M., Rigaux, P., Tacaille, A., and Thion, V. (2016). A Methodology for Quality Assessment in Collaborative Score Libraries. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

[Eppler and Helfert, 2004] Eppler, M. J. and Helfert, M. (2004). A framework for the classification of data quality costs and an analysis of their progression. In *Proceedings of the International Conference on Information Quality*, pages 311–325.

[Foscarin et al., 2018] Foscarin, F., Fiala, D., Jacquemart, F., Rigaux, P., and Thion, V. (2018). GioQoso, an On-line Quality Assessment Tool for Music Notation. In *Proceedings of the International Conference on Technologies for Music Notation and Representation (TENOR)*.

[GioQoso, 2016] GioQoso (2016). The GioQoso project web site. `http://gioqoso.irisa.fr/`.

[Good, 2001] Good, M. (2001). *MusicXML for Notation and Analysis*, pages 113–124. W. B. Hewlett and E. Selfridge-Field, MIT Press.

[Hankinson et al., 2011] Hankinson, A., Roland, P., and Fujinaga, I. (2011). The Music Encoding Initiative as a Document-Encoding Framework. In *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, pages 293–298.

[Jarke et al., 2001] Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P. (2001). *Fundamentals of Data Warehouses*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2nd edition.

[MEI, 2015] MEI (2015). Music Encoding Initiative. `http://music-encoding.org`. Accessed March 2017.

[MuseScore, 2017] MuseScore (consulted in 2017). MuseScore. Web site. https://musescore.org/.

[Music Encoding Initiative Board, 2016] Music Encoding Initiative Board (2016). Music Encoding Initiative Guidelines. `http://music-encoding.org/documentation`. Accessed december 2017.

[Peralta et al., 2009] Peralta, V., Thion, V., Kedad, Z., Berti-Équille, L., Comyn-Wattiau, I., Nugier, S., and Sisaid-Cherfi, S. (2009). Multidimensional Management and Analysis of Quality Measures for CRM Applications in an Electricity Company. In *Proceedings of the International Conference on Information Quality (ICIQ)*, Potsdam, Germany.

[Redman, 1996] Redman, T. C. (1996). *Data Quality for the Information Age*. Artech House Inc.

[Rigaux et al., 2012] Rigaux, P., Abrouk, L., Audéon, H., Cullot, N., Davy-Rigaux, C., Faget, Z., Gavinet, E., Gross-Amblard, D., Tacaille, A., and Thion, V. (2012). The Design and Implementation of NEUMA, a Collaborative Digital Score Library - Requirements, architecture, and models. *Intl. Journal On Digital Libraries (IJODL)*, pages 1–24.

[Riley and Mayer, 2006] Riley, J. and Mayer, C. A. (2006). Ask a Librarian: The Role of Librarians in the Music Information Retrieval. In *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*.

[Rolland, 2002] Rolland, P. (2002). The Music Encoding Initiative (MEI). In *Proc. Intl. Conf. on Musical Applications Using XML*, pages 55–59.

[Selfridge-Field, 1997] Selfridge-Field, E., editor (1997). *Beyond MIDI : The Handbook of Musical Codes*. Cambridge: The MIT Press.

[Si-Said Cherfi et al., 2017a] Si-Said Cherfi, S., Guillotel-Nothmann, C., Hamdi, F., Rigaux, P., and Travers, N. (2017a). Ontology-Based Annotation of Music Scores. In *Proceedings of the International Conference on Knowledge Capture (K-CAP)*.

[Si-Said Cherfi et al., 2017b] Si-Said Cherfi, S., Hamdi, F., Rigaux, P., Thion, V., and Travers, N. (2017b). Formalizing Quality Rules on Music Notation ? an Ontology-based Approach. In *Proceedings of the International Conference on Technologies for Music Notation and Representation (TENOR)*.

[W3C, 2015] W3C (2015). W3C Music Notation Community Group. https://www.w3.org/community/music-notation/.

[Zaveri et al., 2016] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.

List of quality rules