A Methodology for Quality Assessment in Collaborative Score Libraries

Vincent Besson, Marco Gurrieri, Philippe Rigaux, Alice Tacaille, Virginie Thion

A Methodology for Quality Assessment in Collaborative Score Libraries

Vincent Besson[1] **Marco Gurrieri**[1] **Philippe Rigaux**[2]
Alice Tacaille[3] **Virginie Thion**[4]
[1] CESR, Univ. Tours, France
[2] CEDRIC/CNAM, Paris, France
[3] IReMus, Sorbonne Universités, Paris, France
[3] IRISA, Univ. Rennes 1, Lannion, France
{vincent.besson,marco.gurrieri}@univ-tours.fr, philippe.rigaux@cnam.fr,
alice.tacaille@paris-sorbonne.fr, virginie.thion@irisa.fr

ABSTRACT

We examine quality issues raised by the development of XML-based Digital Score Libraries. Based on the authors' practical experience, the paper exposes the quality shortcomings inherent to the complexity of music encoding, and the lack of support from state-of-the-art formats. We also identify the various facets of the "quality" concept with respect to usages and motivations. We finally propose a general methodology to introduce quality management as a first-level concern in the management of score collections.

1. INTRODUCTION

There is a growing availability of music scores in digital format, made possible by the combination of two factors: mature, easy-to-use music editors, including open-source ones like MuseScore [10], and sophisticated music notation encodings. Leading formats today are those which rely on XML to represent music notation as structured documents. MusicXML [7] 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 [13, 9], inspired by the TEI, attempts to address the needs of scholars and music analysts with an extensible format [8]. Recently, the launch of the W3C Music Notation Community Group [15] 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 synchronization), *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 [14]. 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 [12]. 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 are in charge of the production, maintenance and dissemination of digital libraries of scores encoded in XML (mostly, MEI). NEUMA is an open repository of scores in various formats, managed by the IReMus [1] , and publicly accessible at `http://neuma.huma-nm.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.

In the rest of the paper we position our work with respect to the field of quality management in databases and Digital Libraries (Section 2) and propose a general methodology to cope with quality issues in the specific area of digital score management. Section 3 exposes a quality management model. We apply this model to represent data quality metrics, usages and goals, as explained in Section 4, which includes our initial taxonomy of data quality metrics for score libraries. Finally, Section 5 recalls the contributions and outlines our perspectives.

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

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

2. QUALITY MANAGEMENT IN DATABASES AND DIGITAL LIBRARIES

Much published data suffers from endemic quality problems. It is now well-recognized that these problems may lead to severe consequences, and that managing the quality of data conditions the success of most existing information systems [5]. 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 [11]. 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 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 meta-data" 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 [4]. 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 according to *fitness for use* of data, meaning that quality measurement involves dimensions and metrics that are relevant to a given user 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 [3]. We focus here on its *assessment*. Roughly speaking, each assessment methodology includes a *quality definition* stage and a *quality measurement* one. In the first stage, the quality definition consists in eliciting data quality requirements. Concretely, this means choosing quality dimensions and metrics of interest, and eventually thresholds associated with. Because data quality is *fitness for use* (depends on the context), defining data quality is not trivial. Dedicated methodological guidelines may be followed like the *Goal Question Metric* [2], which proposes to define quality metrics according to a top-down analysis of quality requirements. For each user (or each user role) and for each of his/her usages of data, conceptual *goals* are identified. Goals specify the intent of measurement according to a usage of data. Each goal is then refined into a set of operational *quality questions*. Each such question is itself expressed in terms of a set of quantitative quality metrics with possible associated thresholds (expected values). Measuring the quality metrics enables to (partly) answer to the quality questions, and consequently enables to decide whether data satisfy the requirements for the given goal (and each usage by extension).

Data quality methodologies 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. We propose such a methodology for an explicit and systematic data quality assessment in DSL. To our knowledge, such a methodology has never been proposed in the MIR literature so far.

3. OUR QUALITY MANAGEMENT MODEL

We assume a very general organization of a DSL, where atomic objects are *scores*, organized in *collections*. We further assume that scores are encoded as structured documents (typically in MusicXML or MEI) that supply a fine-grained representation of all their structural, content, and rendering aspects.

The main components of the model are (i) modelization of metrics at the score level and collection levels, and of their relationships, (ii) definition of usages and goals, expressed with respect to these metrics, and (iii) computation of quality metrics. We present these concepts in order.

3.1 Quality schema

The initial step to address quality issues is to determine the set of relevant indicators, or *metrics*, that support the quality evaluation, and how they are related to each other.

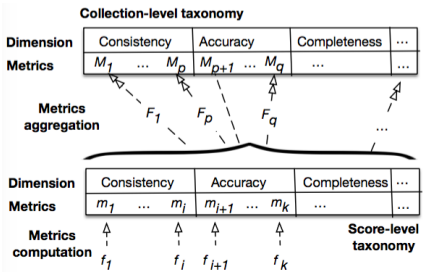

Figure 1. Quality schema: score-level and collection-level metrics

In our context, we consider *score-level metrics*, computed from individual scores, and *collection-level metrics*, essentially computed by aggregation from the score level. We use lowercase/uppercase symbols (e.g., m or M to specifically represent, resp., score-level and collection-level elements (metrics, values, or functions), and small capitals (e.g., $\textsc{m}$) when they do not need to be distinguished. We denote by $\mathcal{M}_{sc}$ the set of score-level quality metrics, $\mathcal{M}_{coll}$ the set collection-level metrics, and as $\mathcal{M}$ their union $\mathcal{M}_{sc} \cup \mathcal{M}_{coll}$.

Metrics are clustered in *quality dimensions*. For simplification reasons, we suppose that (i) a metric belongs to exactly one quality dimension and that (ii) each metric is relevant for every score/collection of the library (the model can easily be extended if these restrictions are too strong). For each metric $m \in \mathcal{M}$, we denote by $dom(m)$ the domain of the metric.

Each DSL has therefore to determine a two-levels organization of dimensions and metrics that constitutes the *quality schema*. Fig. 1 shows its general form. The value of each (score-level) metric m_i is computed from an atomic object (a score) by some function f_i. The domain of a metric can be a Boolean (*"the tempo is/is not missing"*), an integer (*"n measures are complete"*), a rational (*"position is given for x notes out of y"*), etc. In the case of numeric domains, for convenience, we map each value to a predefined scale $\mathcal{S}$ of the form $\{very\_poor(1), poor(2), borderline(3), good(4), very\_good(5), not\_relevant(\perp)\}$ easily adaptable if needed.

The value of a (collection-level) metric M_j is obtained by an aggregation function F_j which operates over the score-level metric vectors. As an illustration, imagine that we aim at representing the syntactic consistency M_s of a collection, defined as a standard variation from the following score-level values: presence of bars m_b, presence of directives m_d, presence of ornamentation m_o. Then the aggregation function F_s takes as input a set of triplets (v_b, v_d, v_o), which denotes values for m_b, m_d and m_o resp., one for each score of the collection. In the general case, an aggregation function F might take into account the whole set of score-level values.

3.2 Usages, goals, and profiles

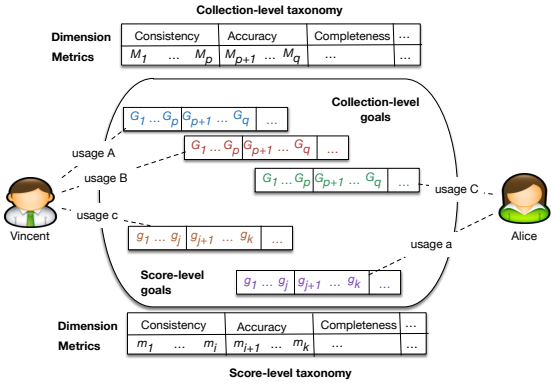

Figure 2. Usages, metric goals, and profiles

Assume now that a quality schema is defined for a DSL managing a set of collections. We are then able to propose to the DSL users a support to express their quality requirements. The main concepts at this level are *usages*, *metric goals* and *profiles* (Fig. 2). A *quality metric goal* assigns an expected quality level for a metric as a threshold $th \in \mathcal{S}$.

A user can express requirements as a set of goals relative to a subset of the metrics, ignoring those which she deems irrelevant in the context of a specific *usage*. For instance a melodic analyst can, *for this usage*, choose to safely ignore the quality metrics that pertain to directives or lyrics. Conversely, for publication purposes, directives and lyrics quality will be required to match a high-quality threshold, whereas analytic annotations are irrelevant. Requirements are therefore usage-related, and take for each usage the form of a set of goals on the metrics specifically relevant for this usage. The specification of such a requirement constitutes what we call a *quality profile*.

From a methodological point of view, each quality profile, including embedded quality dimensions and metrics, is defined by following the *Goal Question Metric* approach (see Section 2). Each metric and dimension appearing in a profile is intrinsically added to the general quality schema. In other words, the set $\mathcal{M}$ of metrics may be defined by union of all metrics appearing in the profiles, which are the relevant metrics identified by the users of the DSL.

3.3 Measurements

A specific module of the DSL is in charge of computing the metrics measurements. As summarized by Fig. 3, this requires to synchronize each score s with the vector $f_1(s), f_2(s), \cdots, f_k(s), \cdots$ representing its quality measurements with respect to the schema.

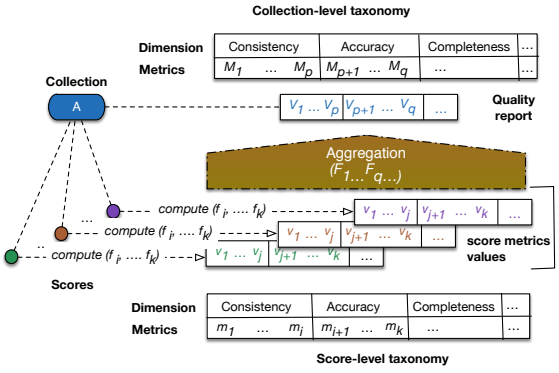

Figure 3. Metrics computation

Likewise, the set of measurements for some collection C must be computed from the quality measurements of the scores that belong to C. One obtains a summary of quality indicators relative to C that we call *quality report*. More formally, the *quality report* of a collection C (resp. of a score s) is a function $QR_C : \mathcal{M}_{coll} \rightarrow \cup_{M \in \mathcal{M}_{coll}} dom(M)$ (resp. $QR_s : \mathcal{M}_{sc} \rightarrow \cup_{m \in \mathcal{M}_{sc}} dom(m)$) that assigns a value for C (resp. score s) to each metric of $\mathcal{M}_{coll}$ (resp. $\mathcal{M}_{sc}$).

Technically, the main issue is to maintain a quality report that faithfully reflects the content of the collection. If the collection is created once for all and never updated, a single batch computation is enough. In general, though, collections are extended, scores are modified, and we have to take those changes into account. At the score level, a trigger seems an appropriate solution: any change of the score results in the execution of the metrics functions $f_1, \cdots, f_k$. Things are more complicated at the collection level. First, a change in some score does necessarily impact the collections metrics, or at least all of them. Second, one has to carefully consider aggregation functions which can be computed incrementally (e.g., *count()*) from those that require a brand new computation from their full input (e.g., *avg()*). Our current implementation adopts a simple recompute-everything strategy, but more sophisticated approaches need to be investigated in the future.

3.4 Matching goals and measurements

Let us now consider how we exploit our two main artifacts: user goals on one side, measurements and reports on the other side. This actually depends on the user's role: *publishers* are responsible for inserting and updating scores and collections, whereas *consumers* can only browse and read. Both a them can define a profile, but for different purposes in the system.

Consumer role and information retrieval. A data consumer may define a profile specifying the expected quality level of data that is needed for a given usage. This profile may be used as a filter by retrieving only appropriate data of the DSL, at the collection or score level, or for recommendation functionalities by suggesting collection or scores of the DSL that respects the profile.

As an example of this filtering facility, one of our DSL supplies a Web front-end interface to browse the collections, some of which exhibit a poor rendering on average, yet are useful for teaching purposes. To limit their access, we can define a usage *browse* with high rendering metric goals, assigned to the *anonymous* user. Anyone accessing to the Web UI without logging in will automatically adopt this default usage and will access only to high-level graphic scores. Connected users can be given access to the teaching collections via another specific usage.

Publisher role and data creation/update. A publisher may define a profile that specifies the needed quality level of data to be achieved before publishing, which may suspend the publishing of the collections or scores that do not respect the profile. This kind of practice goes beyond the control of the quality by the publisher as it also makes it possible to expose a quality certification, for specific usages (profiles) of data available in the platform.

More formally, a quality report QR *satisfies* a profile P iff each quality metric goal of P is satisfied by the values of QR. Given a set (of collections or scores) $\mathcal{S}$ and a profile P, the *filtering* of $\mathcal{S}$ according to a profile P is the set $\{e \in \mathcal{S} \mid QR_e \text{ satisfies } P\}$.

4. APPLICATION: GOALS AND USAGES FOR MEI COLLECTION

We interviewed librarians in charge of the two DSL NEUMA and THE LOST VOICES PROJECT (simply denoted by LOST VOICES in the following). Following the *Goal Question Metric* approach (driven by data use cases, see [2] for details), we exhibited a set of relevant quality questions and metrics for data quality management.

4.1 User requirements for NEUMA

Production of scores. NEUMA is an open repository of scores in MusicXML and MEI. Contributions to NEUMA come either from musicologists for on-line publication purposes, with highly ranked quality standards, or from legacy sources (KernScores for instance, converted to XML formats). For on-line editions, a clean and consistent rendering is required, as well as an homogeneous presentation of the scores of a same collection. The level of details

of meta-data should not strongly vary from one score to the other (in a same collection). Legacy collections are incorporated in NEUMA for teaching or research purposes.

Usage of scores. All the scores submitted to the library are processed with a uniform workflow, which includes production of Lilypond scripts for rendering purposes, conversion from/to Music/MEI, extraction of textual and musical features for indexing, etc. Applying this workflow exacerbates the heterogeneity of the input and reveals a lot of discrepancies and variations in the tolerance levels of the various tools that need to access the score representation. Lilypond for instance hardly accepts incomplete measures, whereas this is tolerated and corrected as much as possible by most MusicXML-based editors. A same meta-data field (e.g., authors, title) can be found in many different places in MusicXML (things are better with MEI), resulting in an erratic rendering with any tool other than the initial editor.

Quality concerns. A common concern met by all users is the need to obtain a decent visualization of a score. Here, "decent" means that any user accessing a score in the library should be able to display it with a desktop tool without a strong readability impact. Unfortunately, this turns out to be difficult to achieve. A score created with a specific engraver $E1$ may contain issues, which are tolerated by $E1$ but result in an awful rendering with $E2$. In general, having a valid MusicXML or MEI document does not guarantee that it can correctly be visualized as a score.

If we leave apart the readability problems, specific usages dictate the quality demands and at which level (score or collections) these demands take place. The Bach chorales for instance are supplied by an external source with accurate music representation, but missing lyrics. This obviously keeps them from any use in a performance perspective, but preserves their interest for music analysis purpose.

4.2 User requirements for LOST VOICES

Production of scores. LOST VOICES is a library of Early European (mostly Renaissance) music scores. It is maintained by a research institution with two main goals: (i) publish (on regular score sheets and on-line) rare collections of Renaissance works, (ii) design and promote advanced editorial practices regarding scholar editions of early music sources, often incomplete or fragmented. All the produced scores are encoded in MEI and must comply to very detailed editorial rules. We cannot of course list them all: they cover usage of ancient and modern clefs, presence of incipits, bars and alterations, signs used for mensural notation, text/music association, etc.

Usage of scores. Scores intended for on-line edition must be submitted to an additional manual process to be compatible with MEI-based rendering tools (VexFlow [3] or Verovio [4]). This includes in particular a specific encoding of variants and missing fragments.

Quality concerns. Most of the editorial rules cannot be automatically checked, and this gives rise to two major issues related to quality management. First, all scores have to be double-checked (i.e., by the person that initially encodes the scores and by a supervisor), a very time-consuming process. Second, the library cannot as such be opened to external contributions, due to the complexity of rules and of the lack of automatic control that would reject inputs falling to match the required encoding requirements.

[3] http://www.vexflow.com

[4] http://www.verovio.org

Syntactic accuracy at the score level
Question – Are measures filled and complete? **Metric** – Proportion of syntactically accurate measures over the total number of measures in the score; 1 for non-measured music.
Question – Do parts have the same length? **Metric** – Proportion of non outliers length parts over (all) the parts of the score.
Question – Is the voice nomenclature correct? **Metric** – Boolean (yes/no).

Table 1. **Syntactic accuracy questions and metrics**

4.3 Quality Metrics

Based on the previous studies, we defined an initial taxonomy of two DSL quality schemas for, resp., NEUMA and LOST VOICES. It is worth noting that the two schemas are significantly distinct, which supports our design choice of a DSL-level modeling of quality requirements.

Due to space restrictions, we illustrate the schemas with a tiny sample of the collected metrics requirements, focusing on consensual quality dimensions of literature: the *consistency*, the *accuracy* and the *completeness*. Other dimensions could be considered if needed. For instance, considering a *provenance* dimension of data (e.g. author, currency, timeliness, volatility) could be relevant.

4.3.1 Score Level

Accuracy is defined as the closeness between data value and their considered correct representation. Classically, two kinds of accuracy are considered: the syntactic accuracy and the semantic one. *Syntactic accuracy* in turn takes two forms. One might first check if the data respect an adequate format (validity). External constraints may also be introduced as goals representing specific editorial rules. For instance LOST VOICES requires a specific voice nomenclature (Superius; Cantus; Altus; Contratenor). In all cases, all the metrics in this dimension can automatically be computed from the score encoding. Table 1 contains a few examples.

Semantic accuracy measures the closeness of a value to a considered true real-world value. Its measurement supposes that there is somewhere a reference for the score content, and cannot thus be evaluated by merely looking at an individual document. For the time being, our schemas do not include semantic accuracy metrics. We defer alternative approaches to future work (see Section 5).

Completeness measures in what extent the score contains all the required information. Table 2 contains some examples. It is worth recalling that defining a metric, and measuring its value, does *not* constitute an *absolute* indicator

Completeness at the score level
Question – Is there a figured bass? **Metric** – Boolean.
Question – Are lyrics present (for vocal music only)? **Metric** – Boolean.
Question – Are meta-data fields present? **Metric** – Proportion of available and syntactically required meta-data fields.

Table 2. Completeness questions and metrics

Consistency at the score level
Question – Are rendering options consistently used in the score encoding? **Metric** – Proportion of rendering options detected among a set of given ones (e.g., note heads, beaming, positioning, spacing, clef changes).
Question – Are performance indications uniformly present? **Metric** – Uniform encoding of slurs, articulation symbols, etc.

Table 3. Consistency questions and metrics

of the DSL quality. Measuring the presence of a figured bass for instance is only important in some usages, and for specific corpora, and its absence does not mean that the corpora are not fit for other usages.

Consistency, at the score level, mostly denotes a uniform encoding of notational features. Positioning information for score elements (notes, chords), fingering, uniform and constant representation of the figured bass are relevant examples for our use cases (see Table 3).

4.3.2 Collection Level

Most of the collection-level metrics are obtained by an aggregation process, which summarizes one or several score-level measurements spread over the collection. In the simplest form, each metric at the score-level allows to define a corresponding metric at the collection-level, which is computed as an average or standard deviation of the score-level metric. Another part of the collection-level metrics are simply not inferred from the collection level. We give a few examples of representative situations.

Accuracy measurements are typically obtained by simple statistical calculation. The (syntactic) accuracy metrics related to measures for instance compute the ratio of scores that contains incomplete measures. Another, less directly computable aspect, is related to the collection structure and presentation. NEUMA for instance requires a fixed ordering of the scores in a collection (Table 4). Note that the later metrics (as many of the same kinds) requires an external information which might be, if available, a public reference of the collection content.

Completeness of a collection (Table 5) measures to what extent the collection is complete enough w.r.t. an expected population size. Here, again, this either requires an external information of reference, or an evaluation by an expert or a group of experts.

Consistency measures in what extend scores of a collection respect a uniform representation (in Table 6).

Collection accuracy
Question – Are measures correctly encoded? **Metric** – Ratio of correct measures.
Question – Are scores ordered as required? **Metric** – Deviation from the required order.

Table 4. Collection accuracy questions and metrics

Collection completeness
Question – Is the collection population complete enough? **Metric** – Proportion of available scores over the expected reference population.

Table 5. Collection completeness questions and metrics

Collection consistency
Question – Are (meta-)data supplied uniformly supplied? **Metrics** – Standard deviations of the collection population for metrics of Tables 1, 2 and 3.

Table 6. Collection consistency questions and metrics

5. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a methodology for assessing data quality in a DSL, based on user preferences. Our approach defines a generic data model that supports the specification of quality schemas, lets users define their goals with respect to the schema of their DSL, and matches usages against quality evaluation. We used this approach for two real digital libraries, and formalized with our model the users' requirements. The implementation is currently in progress for all the metrics that can be evaluated without any external information. This covers syntactic accuracy at the score level, and collection-level aggregated metrics.

Our proposal is a first step that must be completed in several directions. First, several of the metrics identified during our preliminary study cannot be evaluated from the notation itself, but require an external reference. A first solution is a **collaborative** evaluation (some methodologies were proposed e.g. in [6, 1]), for instance based on crowdsourcing. This approach is particularly relevant for the quality dimensions that require external skills like, e.g., semantic accuracy mentioned in 4.3. Another one is to exploit open **semantic web data** by interlinking the DSL collections with other data sources [16].

A second important perspective is to address another aspect of quality management, namely **quality improvement** techniques [4]. Such an improvement can be fully automatic in some specific cases (e.g., filling incomplete measures with rests) but in general, the goal is to help users to identify the insert/update process deficiencies, and to suggest effective improvement strategies.

To our knowledge, no previous work in the literature has proposed metrics for the quality evaluation of music notation. We believe that the topic is important given the lack of constraints of current formats, and the growing production of XML encoded scores.

Acknowledgement: This work has been funded by the French CNRS under the Mastodons project GioQoso.

6. REFERENCES

[1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. Crowdsourcing linked data quality assessment. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 260–276, 2013.

[2] Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach. *Encyclopedia of Software Engineering*, chapter The Goal Question Metric Approach. Wiley, 1994.

[3] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52, July 2009.

[4] Carlo Batini and Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer, 2006.

[5] Martin J. Eppler and Markus Helfert. 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, 2004.

[6] Yolanda Gil and Varun Ratnakar. Trusting information sources one citizen at a time. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 162–176, 2002.

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

[8] Andrew Hankinson, Perry Roland, and Ichiro Fujinaga. The Music Encoding Initiative as a Document-Encoding Framework. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 293–298, 2011.

[9] Music Encoding Initiative. `http://music-encoding.org`, 2015. Accessed Oct. 2015.

[10] MuseScore. Web site. https://musescore.org/.

[11] Thomas C. Redman. *Data Quality for the Information Age*. Artech House Inc., 1996.

[12] Jenn Riley and Constance A. Mayer. Ask a Librarian: The Role of Librarians in the Music Information Retrieval . In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2006.

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

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

[15] W3C Music Notation Community Group. https://www.w3.org/community/music-notation/, 2015. Last accessed Jan. 2016.

[16] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.