A COMPREHENSIVE ONLINE DATABASE OF MACHINE-READABLE LEADSHEETS FOR JAZZ STANDARDS

François Pachet
Sony CSL
pachetcsl@gmail.com

Jeff Suzda
Sony CSL
jeff@jeffsuzda.com

Daniel Martín
Sony CSL
daniel.martin@csl.sony.fr

ABSTRACT

Jazz standards are songs representative of a body of musical knowledge shared by most professional jazz musicians. As such, the corpus of jazz standards constitutes a unique opportunity to study a musical genre with a "closed-world" approach, since most jazz composers are no longer in activity today. Although many scores for jazz standards can be found on the Internet, no effort, to our knowledge, has been dedicated so far to building a comprehensive database of machine-readable scores for jazz standards. This paper reports on the rationale, design and population of such a database, containing harmonic (chord progressions) as well as melodic and structural information. The database can be used to feed both analysis and generation systems. We report on preliminary results in this vein. We get around the tricky and often unclear copyright issues imposed by the publishing industry, by providing only statistical information about songs. The completeness of such a database should benefit many research experiments in MIR and opens up novel and exciting applications in music generation exploiting symbolic information, notably in style modeling.

1. MOTIVATION

Building a *reference* database for music information retrieval is a complex issue. Many databases of audio content have been made available with some success to the research community, raising essential annotation issues [25]. For scores and symbolic information in general, the situation is more problematic. There is a large amount of this information on the net, and many illegal scans of scores (e.g. in pdf format) but, to our knowledge, there is no machine-readable online *reference* database for well-defined corpora, such as jazz standards.

A difficulty when defining a reference database is to define its boundary. In the case of jazz, most composers are no longer active, so it is relatively easy to define such a boundary. For instance, Pepper Adams composed exactly 43 songs; most of Charlie Parker's compositions are known and available in various formats, and the same holds for almost all composers of jazz standards. Such a closed-world approach to jazz standards is key to scholarly and academic work, in particular for *evaluating* operational music systems. Ideally, research experiments involving analyzing and generating jazz compositions should exploit, or apply to, all jazz tunes ever composed, but the absence of such information makes it impossible in practice. As a consequence, many research papers dealing with jazz compositions are based on *ad hoc* databases which are not publicly available ([2], [11-12], [20-21], [23]).

An obvious option to build such a reference database would be to use automatic chord recognition and melodic extraction software on existing audio repositories. There are two problems with this approach. Most importantly, unlike many other musical genres, scores in jazz, called *leadsheets*, play a central role as they represent the "essence" of a tune, harmony- and melody-wise. As a consequence, jazz musicians rarely play the chords as they are written, and part of the game of jazz is precisely to take liberty and interpret the score: unlike classical music, the leadsheet, in general, cannot be deduced from actual performances. Second, the accuracy of chord recognition software is not sufficient to enable fully automatic processes. State of the art methods such as [3], [7] report accuracies in the order of 70%, which is insufficient for our task.

There are numerous attempts at building databases of scores in various genres. For instance, the International Music Score Library Project (IMSLP) assembles scores for classical music composers, but only those in public domain. UCLA's score library proposes many popular music scores, including jazz but it is by no means complete.

2. A REFERENCE CORPUS OF STANDARDS

The notion of jazz standards is ubiquitous in jazz, although not completely well-defined: Jazz standards and pieces that are routinely performed by jazz musicians and widely known to listeners. Most of these songs were composed from the 20s up to the 80s. In practice, jazz standards are often thought of as the songs which appear in the so-called "Fake Books". The most well-known of these is probably the "Real Book", published by Berklee students in the 70s as a reaction to previous Fake Books, which were considered as over simplified to be used by jazz musicians [13]. This book, still widely used today, contains 460 hand-written songs with the melody, the chord sequence, and basic editorial information (composer, style, tempo, and a reference recording of the song).

Since the 70s however, Real Books have evolved significantly. The original Real Book being illegal, several publishers subsequently released other songbooks containing sets of songs for which they obtained or cleared copyrights. The most important publishers are Sher (New Real Books, Volume I to III [26] and Hal Leonard (the Real Book Sixth edition, and the Real Book Volume II, III, IV and V [15]). However, other sources of jazz standards are commonly available through various channels (printed, online as well as illegal). Other notable sources are composer-specific songbooks, which often contain yet different versions of songs, such as the Charlie Parker Omnibook [24] or Michel Legrand song book [14]. As a result, songs appear usually in several song books, with sometimes significant differences. For instance, Figure 2 and Figure 3 show several versions of the song Solar by Miles Davis (or rather, Chuck Wayne, see [1]). Subtle differences are visible concerning chords. In some cases more significant differences appear, including mistakes or different harmonizations.

Finally it is important to observe that, in our experience at least, some songs (such as Body and Soul) are played in almost every jam session, but many others are hardly played at all: all songs are not equally "standard".

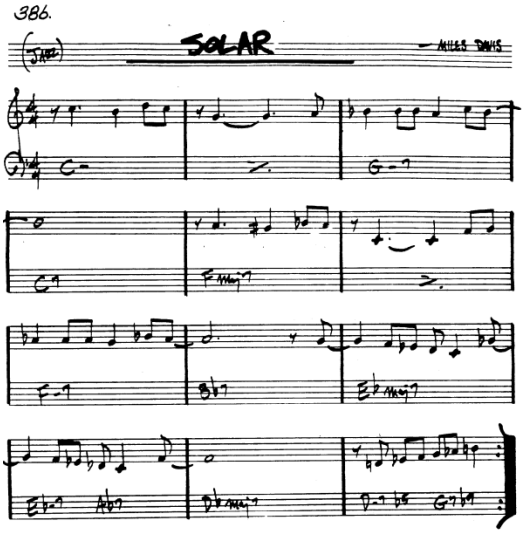

Figure 1. The original Real Book version of Solar.

To summarize, we can point out two important facts about jazz standards that provide us with guidelines:

1) There is no *official version* of any given score unless directly from the author's personal collection, and even then, composers often "update" their compositions afterwards. There are indeed significant differences between scores, depending on the publisher. Differences affect the chord notation used as well as the chords themselves (e.g. their various enrichments) as well as the song structure.
2) The very notion of a standard relies on the existence of songbooks. These books are the medium by which musicians learn and play songs, maintain and evolve the repertoire. The publication of new volumes or new editions of existing volumes impacts the evolution of standards, though on a slow time pace.

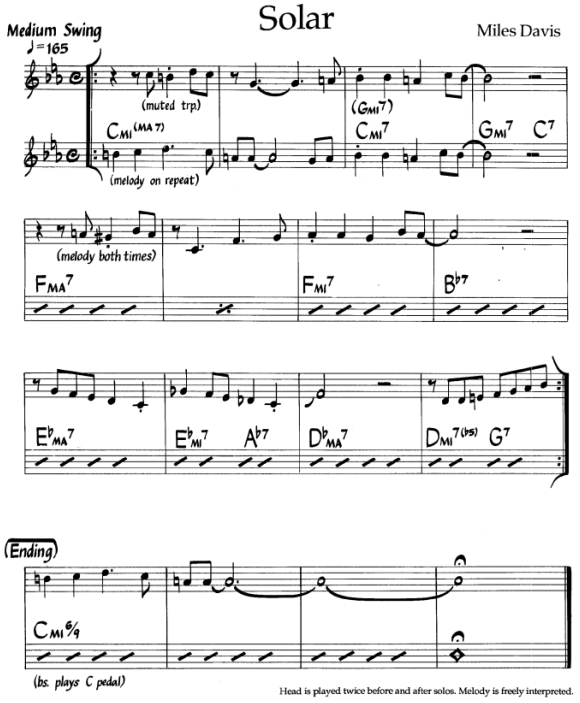

Figure 2. The New Real Book version (Sher) of Solar. Note the different chords (e.g. first chord is C min maj7 instead of C minor), the different chord, and the different structure (ending).

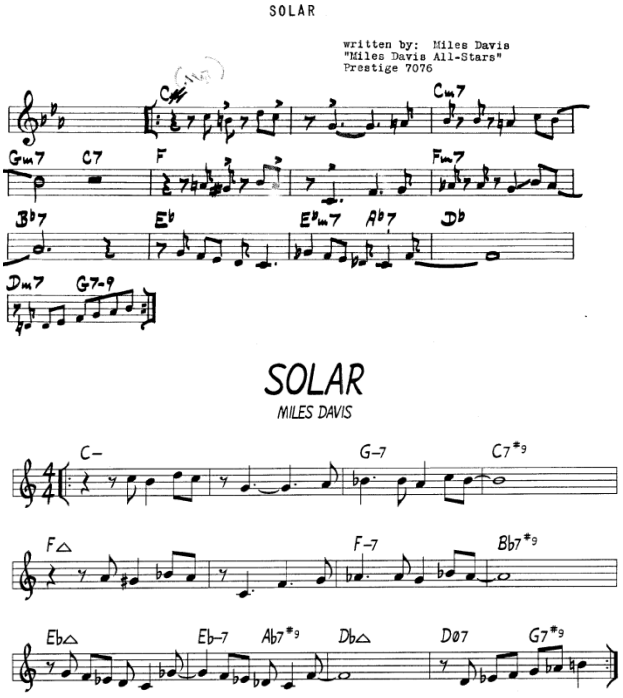

Figure 3. Two other versions of Solar found in popular fake books. Note that none of them can be considered as the official version.

3. AN ONLINE DATABASE

There is a wealth of information about jazz standards on the Internet, but no online database of machine-readable jazz standards exists, to our knowledge. Har-

monic information (chord progressions) is known to be copyright-free so several collections can be found on the web, notably the smartphone application iRealB [10]. But this database does not contain melodies, because of copyright issues, and their content is determined in part by users through a social, collaborative process, with no guarantee on coverage and quality.

3.1 Design: Sources and Songsets

Our database is a web service based on two concepts: *sources* and *songsets*. We define the scope of jazz standards by referring to the already substantial body of work one by reference publishers (such as Sher or Hal Leonard). The primary concept in the database is therefore the "source", which contains the list of songs of a given, published corpus. Figure 4 shows a list of currently entered sources. Sources already contain implicit editorial information concerning the *choice of songs* (publishers want to publish songs that people will actually play), as well as their notation (they try to propose an accurate and consistent notation for musicians). Of course, there are many redundancies in sources, as a popular song will typically appear in various published collections. This redundancy in itself is informative, and can be used, to some extent, to derive automatically information about the popularity of a title, from the viewpoint of publishers. A preliminary analysis of occurrence of songs within 10 sources shows that only one song, *Body and Soul* appears in 8 sources (out of 10), a fact that is confirmed, e.g. by the site jazz-standards.com in which Body and Soul appears as the most popular song to record among jazz musicians. Only 3 compositions occur 7 times (*Here's that rainy day*, *In a Sentimental Mood*, *Bye Bye Blackbird*), and, like *Body and Soul*, they are all famous and routinely performed. More precise information will be enabled as the repository grows, and many analysis can be performed, e.g. on the distribution of popularity in relation with composers, eras, styles, etc.

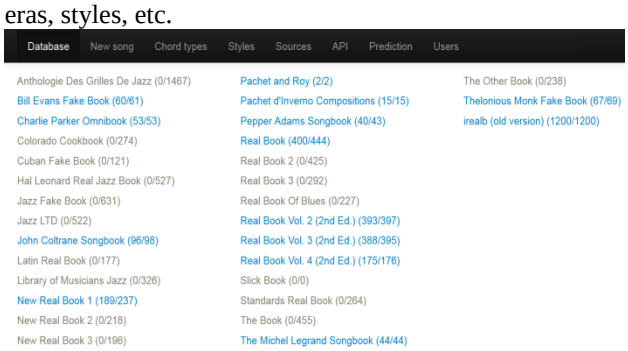

Figure 4. A snapshot of the interface showing the list of currently entered sources (number of completed songs between parenthesis).

Songsets are defined by users, and contain meaningful collections of songs, taken from various sources. Typical songsets are: *all* (the list of all songs in all versions), *bebop* (the complete collection of all compositions by bebop composers such as Charlie Parker or Dizzy Gillespie), *Charlie Parker blues*, the list of all Charlie Parker compositions which are 12-bar blues (see Section 4), *ternary*, the list of all standards in 3/4, etc.

Users define songsets by selecting sources, authors or individual songs, and by filtering them using the information in the database. Information about the redundancy can also be used for specifying songsets (e.g. all songs that appear only once in a given source, or at least 3 times, etc.). Songsets are stored in the database cloud, and can be shared and reused by other users.

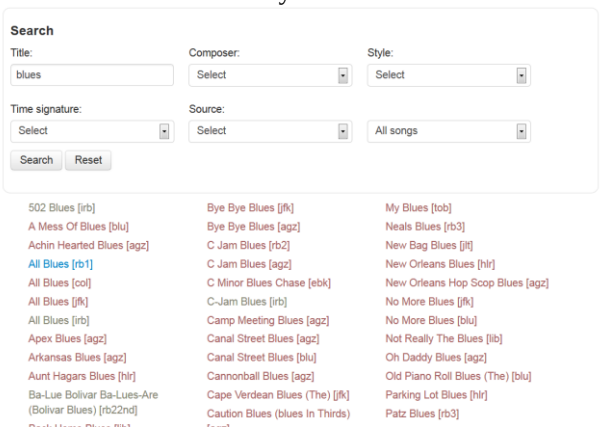

Figure 5. A search tool, here all songs with the word "blues" in the title.

3.2 Song entering

Songs are entered by professional musicians (including the second author), source by source. For each song, a specific online song editor is used, that enables the musician to enter the structure, chords and then melody, as well as basic editorial information (composer, tempo, style, metrics). Average time to enter a song is 3 minutes, but this varies greatly from about 2 to 15 minutes, for complex songs. Note that only basic information about the melody is entered (pitch, quantized position and duration). For instance, the melody of the song Solar, from the Real Book (original) source is illustrated in Figure 6. It can be noted that no typographic information is saved, only the basic MIDI data. This melody is then synchronized to the structure (organization in sections) and chord sequences of the song.

Song enterers do not "copy" the source, but reinterpret it to be stored in the database. Interpretation concerns *chord notation* (see next section) and *structure*. Indeed, one of the problems with extrapolating musical information from a leadsheet is the "folding" problem: Many leadsheets are published in a condensed, folded format - usually a one page leadsheet - of musical information, which is very practical for use in performance situations. However, this is not always the best solution for a machine-readable format. For this reason, some of the compositions are "unfolded" in terms of their form so that there is no ambiguity with regards to repeats, codas, or melodic variations. Of course, such transformations preserve the semantics as both versions describe the same sequence of events (chords and notes).

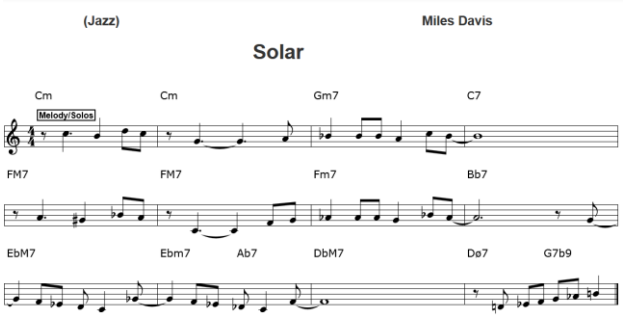

Figure 6. The melody and chord sequence of Solar [Real Book, 5th edition] entered with our online editor.

Figure 7. The song entering process: interpreting a published leadsheet to enter it in a machine-readable format.

Finally a few songs are ignored, either because they contain no melody (*Domino Biscuit* by Steve Swallow) have no time signature (*And now, the Queen* or *Batterie* by Carla Bley), or because the melody is too polyphonic (*Ay Arriba* by Stu Balcomb), and therefore outside the scope of our target (all examples from the original Real Book).

Error checking is performed using two means. First, automatic checks are performed to ensure that the durations of melodies in each bars and section are the same as the corresponding durations of chord sequences. Second, song enterers periodically manually check about 5% random songs entirely (melodies and chords) entered by other song enterers. Manual checking has revealed so far that very little errors are encountered (less than 1% of songs contain errors).

3.3 API and Implementation

The API is a delicate matter. Because we do not own copyrights to the compositions, melodies in particular, we provide an API that only delivers statistical information. The API provides, for a given songset, the following information:

- The chords prior probabilities for songset with id *s*:

http://.../api/getChords.php&songset_id=s returns the list of chords in *s* with their probability:

```
{{"prob": 0.217634, "chord": "Am7"},
 {"prob": 0.119352, "chord": "CM7"},
 {"prob": 0.112842, "chord": "G7"}...
```

- The prior probabilities for pitches occurring in a songset. For instance, query *http://../api/getPitches.php&songset_id=s* would return:

```
{{"prob": 0.251634, "pitch": "G"},
 {"prob": 0.250932, "pitch": "C"},
 {"prob": 0.247842, "pitch": "D"}...
```

- For any prefix of chords, the probabilities of all possible continuation chords, at the order equal to the prefix length. For instance, to get the continuations of Gm7, the query *http://.../api/chords.php?method=getTransitions&chord=Gm7&songset_id=s* would return:

```
{"+5/7": {"prob": 0.537634, "chord": "C7"},
"+5/m7": {"prob": 0.071774,"chord": "Cm7"},
"+5/7b9": {"prob": 0.028494...
```

where for each continuation, we have the distance in semitones between G and the continuation's root (+5 between G and C), type (7, minor7 and 7b9), probability and actual chord name.

- For any prefix of pitches, the list of probabilities of all possible continuations, at the order corresponding to the length of the prefix. For instance, *http://.../api/chords.php?method=getTransitions&pitch=A&songset_id=s* would return:

```
{"-2": {"prob": 0.064516, "pitch": "G"},
 "+5": {"prob": 0.043709, "pitch": "D"}...
```

Additionally, the API provides, for each song in a songset, the histogram of chords and pitches, as well as the joint probabilities of chord and pitches.

To our knowledge, such an API does not violate copyright, as it is, in general, impossible to completely reconstruct a melody or even a chord sequence from this statistical information. This API will, however, evolve, to adapt to the needs of applications and the evolution of copyright policies of the music publishing industry. Songs for which copyright has ceased will be made progressively available to users in their entirety. Chord sequences, in principle not copyrighted, are provided entirely in text format.

Current implementation uses standard web technology HTML/CSS and Javascript in the client side, PHP in the server side with a noSQL database in JSON format. Melodies are stored in musicXML format [19].

3.4 Chord notation and substitution rules

As can be seen by the example, there is no common, reference notation for jazz chords, and sources use different notations [6]. Some works in MIR have addressed the problems of chord notation ([8], [16-17], [28]) but these notations are mostly used for automatic audio chord extraction tasks.

Additionally, within a given notation, there are differences in precision. For instance, a dominant seventh chord can be written simply as "7", or, in other sources, with additional notes (e.g. "9", or "dim9"). In order to preserve as much as possible the data accuracy we have chosen to enter sources with chord names that are as close as possible to the chord written in the source, and adding them when the score enterer considered it is not in the current list (we have reached currently a total of 86 chord names, see Figure 8): no effort at consistency or uniformity has been conducted at this step.

Such an approach is obviously not sufficient when several sources are mixed together to form a coherent songlist. In order to cope with this problem (seen here as a *sparsity* problem), we use sets of substitution rules, that

transform chords from their original formulation (e.g. *C 7#4#5*) into a sparser formulation that is significant for the task at hand. For instance, some applications may need to distinguish only between, say, 4 chord types (major, minor, dominant 7^{th}, diminished), while other may need more.

To address this issue, we introduce *transformers:* sets of substitution rules that transform a chord in a source into the most relevant chord name in a given vocabulary. For instance, *C 7#4#5 => C7*, or *DM7#11 => D M*.

Such a use of chord substitution rules can be extended to cope not only with lexical redundancy, but also with some form of semantic equivalence. This problem has been well studied in computer music ([20], [27]) and accepted sets of rules can be easily identified. For instance, many forms of "ii-V7-I" can be considered as more or less equivalent: a dominant chord such as *C 7* can be rewritten as *G min7 / C7*, or even as *G min7 / F# 7*, depending on the degree of precision requested and the task. Such application-dependent considerations can all be handled through sets of substitution rules, defined once and for all by users and shared, like songsets.

(empty)	2	5	6	m
+	7	9	11	13
+7	m6	69	M7	m9
m7	M9	7b9	7#11	aug
Alt	m13	m#5	m69	m11
Dim	7#5	7#9	9b5	7b5
mb6	9#5	7#4	M13	7b6
#11	Sus	7b13	add9	11b9
7alt	6#11	m7#5	M7b9	+7#9
+7b9	m9M7	(b5)	7sus	13b9
9#11	mM7	dim7	9sus	4sus
M7#5	M7#4	m9b5	M9#5	13b5
sus2	sus4	M9b5	M7#9	7#9b5
7#5b5	7#4#5	13#11	M9#11	13sus
7b9#9	7#5#9	pedal	+add9	7b5#9
(#11)	m(M9)	dimM7	7#9#5	7b9b5
M7#11	7b9#5	aug#4	+(b9)	6sus4
m11b5	madd9	5add9	7#5#11	7b9#11
Lydian	7#9#11	7b9b13	Dorian	M7#9b9
m7add4	m7b5#5	(add9)	m7sus4	7b9sus
dim7M7	add9b5	mM7#11	mM7b13	13b9b5
add#11	M13#11	7omit5	Aeolian	m(add9)
13b9sus	+(add9)	m7b5b13	(no3rd)	m(m7M7)
(b9b13)	7b13#11	7b13sus	13b9#11	M7add13
m9add13	m7addM7	Phrygian	M7(?4)	m7(b5b2)
(9, #11)	halfdim7	7susadd3	13(b9b5)	m(omit5)
sus4add9	7b9b13sus	13(b9#11)	m7(omit5)	7susomit5
13(add11)	6#9	M7b5	13#9	m9#11
m7#11	7#5b9	69#11	mb5b13	m13#11
M7#5#11	M7#9#11	add9addb13	madd9add11	halfdim7b9
m7add11add13	halfdim7add11			

Figure 8. The current chord names used in about 12 reference sources.

4. APPLICATIONS

Our database is developed in the context of a large-scale project about the representation of musical style, in particular for popular music. In this context, songsets are considered as concrete representations of a user-defined style. Various style analysis and generation mechanisms, e.g. using the technology of *Markov constraints* [22] can be implemented to generate sequences "in the style of", that also satisfy arbitrary user constraints. An example was exhibited in [22] with the so-called *Boulez Blues*: a 12-bar Blues chord sequence in the style of Charlie Parker blues (the *Parker Blues* songset) that satisfies an "All different" constraint (hence the Boulez label), and is optimally Parkerian, i.e. maximizes its probability w/r the Parker Blues corpus.

Other applications can be developed to exploit this database. Generation algorithms based on statistical information, in particular using *random walk* algorithms can be trivially implemented with our API. Indeed, random walk consists in selecting at random the "next" event (chord or note) using the transition probabilities, given a prefix (the sequence already created), which is exactly what our API provides.

The database is also used for analysis studies. To our knowledge, few studies attempt to assess to what extent composers are recognizable through their chord sequences only, or through their melodies, or both. Attempts to address these issues (e.g. [18-19]) are not comprehensive, nor easily reproducible. Such studies are under way [9], and its results will be made credible only the comprehensive nature of this database.

5. CONCLUSION

We described the motivation and rationale for a comprehensive online database of machine-readable leadsheets of jazz standards[1]. The specification of the database is simple because its goals are very clear: provide a machine-readable representation of melodies and chord progressions as found in reference, published fake books, and following a "closed-world" approach. The database is already being used by several projects dealing with analysis and generation of jazz compositions.

The closed-world approach does not mean that this database effort is to be stopped soon. First, new compositions are regularly been published, such as the European Real Book [5], though not at a pace comparable to that of the Fake Books of the 1970s and after. The contents of such books will be added progressively to the database, which will enable interesting experiments, for instance, regarding the evolution of compositional styles.

We do not infringe on copyrights, because 1° our database does not contain typographical information specific to publishers and 2° we provide an API that prevents reverse engineering to the original sources.

Other sources of editorial information will be progressively added, such as the list of official recordings for each standard, with the audio content when possible, or the exact date of composition, when available.

Our effort can be generalized to other music genres, notably for which leadsheets play such a central role. This concerns for instance large chunks of the Brazilian popular music repertoire such as Bossa Nova or Choros: like jazz, these repertoire are somewhat closed but rich enough musically to deserve such a treatment. Several works have already addressed analysis tasks on partial databases [4]. Most importantly our approach applies to songs that can be reduced to their leadsheet representation without losing their essence.

[1] www.flow-machines.com/lsdb

Our jazz database targets a total of 15 sources (see Figure 4) and 8000 songs (4000 of them unique) by the date of presentation of this paper, obtained through a steady song entering process. With such a consistent mass of information, the first comprehensive style-based jazz composition and analysis systems will, at last, see the light of day. The corresponding research will be easily reproducible. Hopefully, more genres will follow.

6. ACKNOWLEDGEMENTS

This research is conducted within the Flow Machines project which received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 291156.

7. REFERENCES

[1] L. Appelbaum: "Performing Art Blog", http://blogs.loc.gov/music/2012/07/chuck-wayne-sonny-solar, 2012.

[2] J. Biles: "GenJam: A Genetic Algorithm for Generating Jazz Solos", *International Computer Music Conference*, pp. 131-137, 1994.

[3] J. A. Burgoyne, J. Wild, and I. Fujinaga: An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis, *ISMIR*, pp. 633-638, 2011.

[4] G. Cabral and R. Willey: "Analyzing Harmonic Progressions with HarmIn: the Music of Antonio Carlos Jobim", *11th Brazilian Symposium on Computer Music*, São Paulo, 2007.

[5] Europe: *European Real Book*, Sher music, 2012.

[6] M. Granroth-Wilding and M. Steedman: "Statistical Parsing for Harmonic Analysis of Jazz Chord Sequences", *International Computer Music Conference*, pp. 478–485, 2012.

[7] B. de Haas, J. P. Magalhães, F. Wiering: Improving Audio Chord Transcription by Exploiting Harmonic and Metric Knowledge, *ISMIR*, pp. 295-300, 2012.

[8] C. Harte et al: "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations", *ISMIR*, pp. 66-71, 2005.

[9] T. Hedges, P. Roy and F. Pachet: Predicting the Composer and Style of Jazz Chord Progressions, *submitted*, 2013.

[10] iRealB, smartphone application, http://www.irealb.com, 2013.

[11] R. Keller and D. Morrison: "A Grammatical Approach to Automatic Improvisation", *Fourth Sound and Music Computing Conference*, Greece, 2007.

[12] R. Keller et al.: "Jazz Improvisation Advisor", http://www.improvisor.com, 2009.

[13] B. Kernfeld: *The Story of Fake Books: Bootlegging Songs to Musicians*, Scarecrow Press, 2006.

[14] M. Legrand, *The Michel Legrand Songbook*, Warner Bros. Publications, 1997.

[15] H. Leonard: *The Real Book, Volume I, II, III, IV and V*, Hal Leonard, 2012.

[16] M. Mauch, S. Dixon, C. Harte, M. Casey, and B. Fields: "Discovering Chord Idioms through Beatles and Real Book Songs", *International Symposium on Music Information Retrieval*, 2007.

[17] M. Mauch et al.: "Can Statistical Language Models be used for the Analysis of Harmonic Progressions?" *International Computer Music Conference*, Japan, 2008.

[18] L. Mearns, D. Tidhar, and S. Dixon: Characterisation of composer style using high-level musical features, In *3rd ACM Workshop on Machine Learning and Music*, 2010.

[19] MusicXML 3.0 Specification, MusicXML.com. MakeMusic, Inc. Retrieved 26 February 2013.

[20] M. Ogihara and T. Li: N-Gram Chord Profiles for Composer Style Representation, *ISMIR*, pp. 671-676, 2008.

[21] F. Pachet: "Surprising Harmonies", *International Journal of Computing Anticipatory Systems*, Vol. 4, 1999.

[22] F. Pachet and P. Roy: "Markov constraints: steerable generation of Markov sequences", *Constraints*, 16(2):148-172, 2011.

[23] G. Papadopoulos, G. Wiggins: "A genetic algorithm for the generation of jazz melodies", *STeP, 8th Finish Conference on Artificial Intelligence*, Jyväskylä, 1998.

[24] C. Parker: *Charlie Parker Omnibook*, Atlantic Music Corp, 1978.

[25] G. Peeters, K. Fort: "Towards A (Better) Definition Of The Description Of Annotated M.I.R. Corpora", *ISMIR*, pp. 25-30, Porto, 2012.

[26] Sher Music, *The New Real Book, Volume I, II and III*. Sher Music Co, Petaluma, USA, 2012.

[27] M. J. Steedman: "A Generative Grammar for Jazz Chord Sequences", *Music Perception* 2(1):52–77, 1984.

[28] C. Sutton, Y. Raimond, M. Mauch, and C. Harte: "The Chord Ontology", http://purl.org/ontology/chord, 2007.