Managing a text-critical database of J.S. Bach's 'Well-Tempered Clavier II' with XML and a relational database

Yo Tomita*
School of Music,
Queen's University,
Belfast, GB

Tsutomu Fujinami†
School of Knowledge Science,
Japan Advanced Institute
of Science and Technology

1 Introduction

J. S. Bach's **Well-Tempered Clavier**, Part II (WTC II) is widely regarded as one of the most important works in western music. Many composers of later generations were influenced by this work. The accounts of Beethoven and Mozart are well-documented; both studied this work from a manuscript copy, as it was in this form that it was circulated before its publication at the beginning of the 19th century. Despite the work's exceptional popularity, we are still unable to establish fully the final text intended by the composer, let alone the text that was available to the pupils of subsequent generations. This makes it very difficult to conduct an accurate assessment of its influence. The text-critical database was compiled with the view to ascertaining both the origin and authorship of individual variants found in all the extant manuscript and early printed sources of the work (these are in excess of 130). Currently it contains about 5.5MB of character-based information, which probably makes it the largest database of its kind in existence. In order to use it for analysis, we need to find a way for a computer to process this large corpus of data.

The ultimate goal of our project is to build an intelligent system that is capable of interpreting the data and making the information accessible to the public. To achieve these goals, we need to find answers to the following questions:

1. How can we measure and establish the relationship between sources?
2. How can we describe the relationship between sources?
3. How can we share with other scholars both the data and the result of analysis?

To find an answer to the first question, it is not sufficient to shortlist commonly-found errors and variants; we also need to investigate some physical aspects of the score to account for the scribe's unintentional omissions, misplacement or superfluous symbols that were somehow caused by the appearance of the exemplar, as well as those aspects that are attributable to the scribe's autonomous decision making, such as his concerns about musical matters and his notational preferences that are likely to have been influenced by both historical and geographical settings in which he worked. It is desirable that our system should be capable of handling such complex arrays of qualitative information obtained from a carefully-conducted analysis. The second question concerns with the way we visualize the relationships between sources. Good visualization certainly helps us to capture these relationships, which in turn helps us to revise our theories when analyzing the data. The answer to one question is thus dependent on the other and vice versa.

*y.tomita@qub.ac.uk, http://www.music.qub.ac.uk/~tomita/
†fuji@jaist.ac.jp, http://www.jaist.ac.jp/~fuji

The third question concerns the sharing of the data, which is as important to the public as to the scholars who can refine their theories through cross-referencing their analyses. Feedback from peers is equally important, as it helps us to refine our methods of data presentation and analysis. Data sharing can thus be considered as prerequisite for both data analysis and visualization.

In this paper we address the issue of data sharing as the first step towards our goal. Nowadays, an increasing number of researchers, regardless of their fields of speciality, take the sharing of data on the internet very seriously. We are not an exception, for our collaboration would not be possible without the use of the internet. Our first task is to define and describe the method of data representation, as our data, which are mainly extracted from music manuscripts, are much more complex than other forms of data such as plain text. Owing to this complexity, there is much room for introducing errors when copying out a musical text than there would be with a verbal / literary text. A small number of the errors end up in ambiguous readings or less frequently in valid variant readings. It is thus necessary to permit the recording of some form of descriptive commentary in our data; while the majority of them do not contain this extra information, it is valuable data when evaluating the copyists' scribal activities and other contextual issues.

The paper is organized as follows: we first discuss our data sources. We then present our web application and show how it enables the user to view the data in section 3. Section 4 briefly explains some of our ongoing sub-projects, indicating in which direction our project is developing. We conclude the paper by discussing some of the difficulties we need to address in order to bring our WTC II database project to completion.

2 Data Sources

Our primary data source is the text-critical database for Bach's **Well-Tempered Clavier**, Part II compiled by Tomita. We do not go into the details of the data themselves here, e.g. which manuscripts were examined, what properties of manuscripts were taken into consideration for collecting the data, how these properties are interpreted in assessment, etc., as these are discussed fully in Tomita's book published in 1995.[1] Instead, we shall look at the electronic version of the database. Our secondary data source is the so-called 'London autograph',[2] a partial autograph and partial fair copy in the hand of Bach's second wife, Anna Magdalena. Again, we do not discuss the source in detail but simply use the electronic version created by Tomita as the starting point for our discussion. We shall now explain each electronic resource in turn.

2.1 The Text-Critical Database

The data was collected through the examination of manuscripts using spreadsheet software called WINGZ.[3] The collected data have no strict syntax to describe the data entry itself, allowing the accommodation of such extra information as revisions and ambiguity in the readings. To make the data visually meaningful, some data entries make use of musical symbols that are designed for this purpose.[4] Our first task then is to transcribe them into a form that can be processed by the system that is designed to perform certain analytical tasks. This largely depends on what kind of database we should employ, i.e. relational, object-oriented, and deductive, among others.

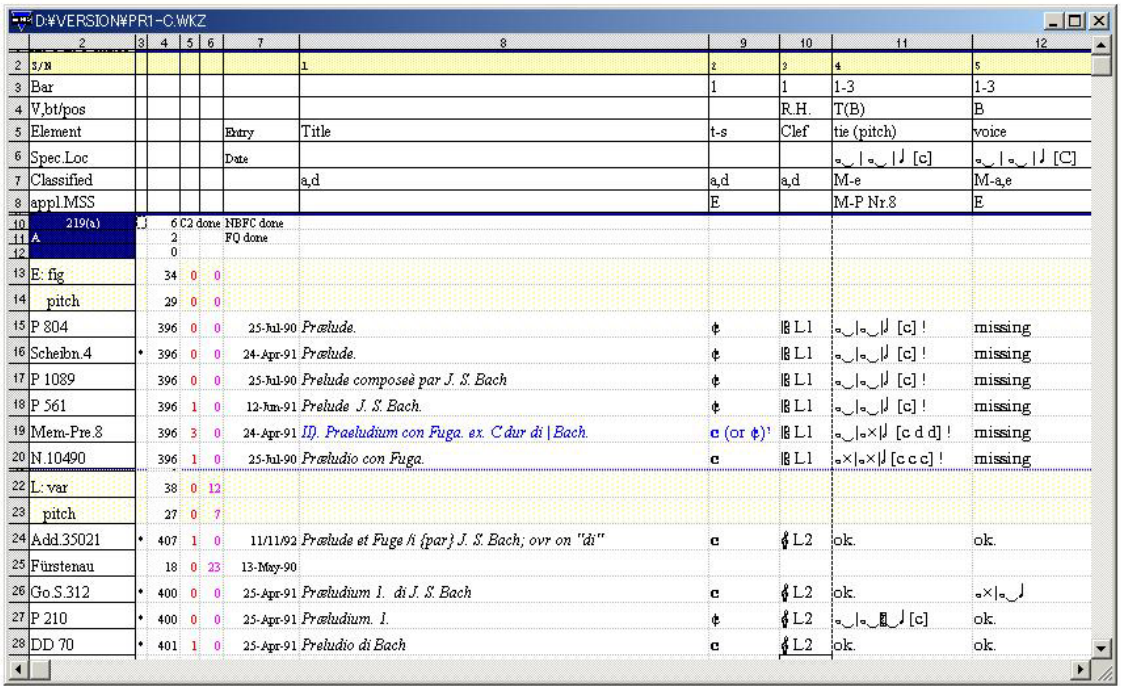
Figure 1 shows a part of the text-critical database of the Prelude No. 1 in C major (BWV 870/1). The upper part of the table describes the points of examination where at least one source is found to contain an error or variant. In other words, this area of the table describes the properties

[1] J. S. Bach's 'Das Wohltemperierte Clavier II'A Critical Commentary. Volume II: All the Extant Manuscripts (Leeds: Household World, 1995) xxx+1001p. ISBN 0-9521516-7-7

[2] The British Library, London: Additional MS 35021

[3] http://www.wingz-us.com/wingz/index.html. It features a large two dimensional sheet of 32K x 32K, which is complemented by a powerful script language called 'Hyperscript'. For further description of how the spreadsheet was used for this project, see Tomita, 'The Spreadsheet in Musicology: An efficient working environment for statistical analysis in text critical study', Musicus iii (1993), pp. 31-37. ISSN 0958-0999

[4] Tomita developed a musicological font 'Bach' for this purpose: see http://www.music.qub.ac.uk/tomita/bach-mf.html



	2	3	4	5	6	7	8	9	10	11	12
2	S/N						1	2	3	4	5
3	Bar							1	1	1-3	1-3
4	V,bt/pos								R.H.	T(B)	B
5	Element					Entry	Title	t-s	Clef	tie (pitch)	voice
6	Spec.Loc					Date				𝅝‿|𝅝‿|𝅗𝅥 [c]	𝅝‿|𝅝‿|𝅗𝅥 [C]
7	Classified						a,d	a,d	a,d	M-e	M-a,e
8	appl.MSS							E		M-P Nr.8	E
10	219(a)		6	C2 done		NBFC done					
11	A		2			FQ done					
12			0								
13	E: fig		34	0	0						
14	pitch		29	0	0						
15	P 804		396	0	0	25-Jul-90	*Præludе.*	𝄵	𝄡 L1	𝅝‿|𝅝‿|𝅗𝅥 [c] !	missing
16	Scheibn.4	•	396	0	0	24-Apr-91	*Præludе.*	𝄵	𝄡 L1	𝅝‿|𝅝‿|𝅗𝅥 [c] !	missing
17	P 1089		396	0	0	25-Jul-90	*Prelude composeè par J. S. Bach*	𝄵	𝄡 L1	𝅝‿|𝅝‿|𝅗𝅥 [c] !	missing
18	P 561		396	1	0	12-Jun-91	*Prelude J. S. Bach.*	𝄵	𝄡 L1	𝅝‿|𝅝‿|𝅗𝅥 [c] !	missing
19	Mem-Pre.8		396	3	0	24-Apr-91	*II). Praeludium con Fuga. ex. C dur di | Bach.*	𝄴 (or 𝄵)'	𝄡 L1	𝅝‿|𝅝×|𝅗𝅥 [c d d] !	missing
20	N.10490		396	1	0	25-Jul-90	*Præludio con Fuga.*	𝄴	𝄡 L1	𝅝×|𝅝×|𝅗𝅥 [c c c] !	missing
22	L: var		38	0	12						
23	pitch		27	0	7						
24	Add.35021	•	407	1	0	11/11/92	*Prælude et Fuge /i {par} J. S. Bach; ovr on "di"*	𝄴	𝄞 L2	ok.	ok.
25	Fürstenau		18	0	23	13-May-90					
26	Go.S.312	•	400	0	0	25-Apr-91	*Præludium 1. di J. S. Bach*	𝄴	𝄞 L2	ok.	𝅝×|𝅝‿|𝅗𝅥
27	P 210	•	400	0	0	25-Apr-91	*Præludium. 1.*	𝄵	𝄞 L2	𝅝‿|𝅝‿|𝅘𝅥𝅯 [c]	ok.
28	DD 70	•	401	1	0	25-Apr-91	*Preludio di Bach*	𝄴	𝄞 L2	ok.	ok.

Figure 1: Part of the text-critical database on WINGZ

that characterize each source. The lower part lists the values (plus comments where necessary) found in each source, i.e. its properties. The eighth column in Figure 1, for example, describes how the title differs among manuscripts and the ninth provides the time signature of each manuscript.

One unique feature of this database is that the attribute of each column is composite, i.e., an attribute is further characterized by the following six properties:

- **S/N** is the serial number given to each examination point (= record number).
- **Bar** indicates in which measure(s) the elements are examined.
- **V, bt/pos** stands for Voice, Beat and Position, respectively, specifying the exact position in which the elements are found within the bar.
- **Element** specifies the target of enquiry, indicating what notational elements are under examination.
- **Spec. Loc** gives graphic representation of the information under examination.
- **Classified** suggests the text-critical significance of the data.

The structure of the table is similar to that adopted in the relational database except that the attribute is composite. What is the best strategy in building a database with this kind of structure? Our solution is to consider the database as a relational database, which is extended with a kind of meta-data of the table, namely, the description of each attribute in the table. Data to fill columns in a table are only defined in terms of data-types in a relational database. In addition to data-types, we introduce an extra layer of data into the table that explains what kind of information each column describes. We format the data using a relational database and use two kinds of tables, one of which stores the data of the WTC II variants and the other the description of attributes.

List 1 shows as an example a part of the table storing the data of errors and variants. The left two columns specify a particular manuscript with an ID (MsID) and its title (Mss). The third

column, SN, standing for the serial number, corresponds to the **S/N** in the source. The fourth column, Content, stores the values of the attribute.

List 1: Part of the table storing the data of variants

MsID	Mss	SN	Content
5	P 804	1	Prælude.
6	Scheibn.4	1	Prælude.
7	P 1089	1	Prelude composeè par J. S. Bach
8	P 561	1	Prelude J. S. Bach.
9	Mem-Pre.8	1	II). Praeludium con Fuga. ex. C dur di | Bach.
10	N.10490	1	Præludio con Fuga.
13	Add.35021	1	Prælude et Fuge /i {par} J. S. Bach; ovr on ''di''
15	Go.S.312	1	Præludium 1. di J. S. Bach
16	P 210	1	Præludium. 1.
72	DD 70	1	Preludio di Bach

List 2 shows a part of the table storing the meta-data of attributes. The leftmost column, SN, lists the serial numbers. The other columns store the same properties as found in the source. These two tables are related to each other by referring to the SN property and the structure found in the source can easily be constructed by combining items of information obtainable from the two tables.

List 2: Part of the table storing the meta-data of attributes

SN	Bar	VbtPos	Element	SpecLoc	Classified
1			Title		a,d
2	1		t-s		a,d
3	1	R.H.	Clef		a,d
4	1-3	T(B)	tie(pitch)	¬_ | ¬_ | ± [c]	M-e
5	1-3	B	voice	¬_ | ¬_ | ± [C]	M-a,e

We replace the decimal character reference for 8-bit characters taken from the character sets the Latin-1 and Bach fonts. The conversion is necessary because the Bach font contains escape sequences, which cause problems in the database management system. The original characters are restored when the user view the data with a web browser. We decided to employ MySQL as our database management system because it processes data very quickly. [We ignore appl MSS.]

2.2 London autograph

For the purpose of analyzing the variants of the WTC II, the text-critical database alone is not sufficient to carry out the task, because it does not contain information about the music itself. The text-critical database is a set of annotations without the content. What is appropriate as the content, then? The ultimate source would be the original, the version written by J. S. Bach himself. Unfortunately we do not have a complete copy in 'Fassung letzter Hand' by Bach. As a compromise, we selected the 'London autograph' as our reference, as it is nearly complete, and there is strong evidence to suggest that it served as a source of reference in Bach's household even when the set had not been fully finalized.

For this reason, we created with Coda Finale each movement of WTC II that can be considered to represent the music of the London autograph version. Figure 2 shows beginning of Prelude No.

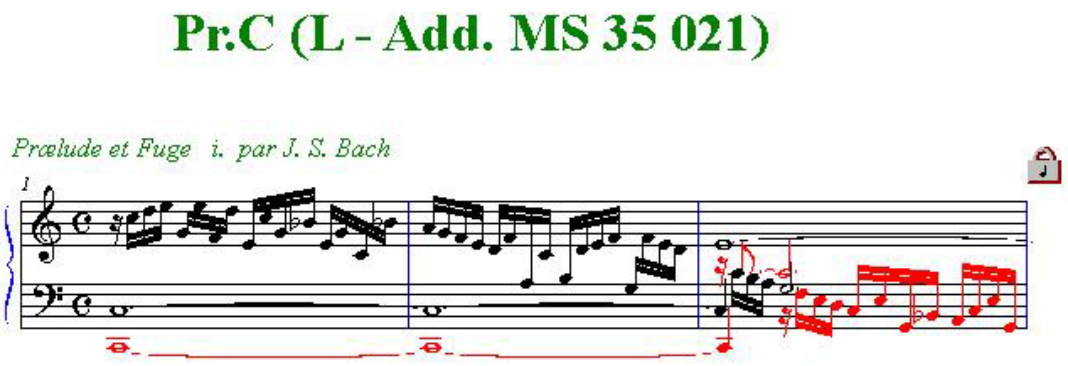


Figure 2: Part of Prelude No. 1 in C major

1 in C major in Finale. The Finale file is then converted to two other formats for data analysis: one in extensible marked language (XML) and the other as a relational database.

There are several reasons why we decided to have the data in XML:

1. The data structure used in Finale Enigma format fits well with the tree structure, the structure for which XML is designed.
2. Data in tree structure are easy to read for us, the developers.
3. We can view the data on the web browser by simply converting the XML data into HTML using extensible style language (XSL).

List 3 shows the beginning of Prelude No. 1 formatted in XML. The part only contains the soprano voice in the first measure.

List 3: Beginning of Prelude No. 1 in XML

```
<?xml version="1.0"?>
<!DOCTYPE wtc SYSTEM "london.dtd" []>
<wtc file="L.new/pr01.ETF">
  <soprano>
    <measure Number="1" barline="normal" beats="4" chg_sys="true" dvibeat="1024" notekey="c">
      <beat Number="1">
        <entry ID="2" Type="rest" duration="256" note="semiquaver"
               pos="1" stem_dir="up" timepoint="0" timespan="256"/>
        <entry ID="3" Type="note" duration="256" note="semiquaver"
               pos="2" stem_dir="down" timepoint="256" timespan="256">
          <noteRecord pitch="c5" staff="upper"/>
        </entry>
        <entry ID="4" Type="note" duration="256" note="semiquaver"
               pos="3" stem_dir="down" timepoint="512" timespan="256">
          <noteRecord pitch="d5" staff="upper"/>
        </entry>
        <entry ID="5" Type="note" duration="256" note="semiquaver"
               pos="4" stem_dir="down" timepoint="768" timespan="256">
          <noteRecord pitch="e5" staff="upper"/>
        </entry>
      </beat>
```

We also converted the data of the London autograph into relational database, as we realised that it takes time to convert XML files into HTML upon request. The table structure describing the London autograph is shown in List 1. We use three separate tables: **Measure** to encode the information of measures; **Entry** to encode the information of entries; and **Note** to encode the information of each note. The data is redundant compared with that in XML, but it is much faster to process them on our server.

List 4: Parts of the tables storing the data of the London autograph

```
mysql> select * from pr01Measure limit 1;
+---------+--------+---------+-------+---------+---------+---------+
| voice   | number | barline | beats | chg_sys | dvibeat | notekey |
+---------+--------+---------+-------+---------+---------+---------+
| soprano |      1 | normal  |     4 | true    |    1024 | c       |
+---------+--------+---------+-------+---------+---------+---------+

mysql> select * from pr01Entry limit 4;
+---------+---------+------+--------+----+------+----------+------------+-----+--
| voice   | measure | beat | record | id | type | duration | note       | pos |
+---------+---------+------+--------+----+------+----------+------------+-----+--
| soprano |       1 |    1 | fals   |  2 | rest |      256 | semiquaver |   1 |
| soprano |       1 |    1 | true   |  3 | note |      256 | semiquaver |   2 |
| soprano |       1 |    1 | true   |  4 | note |      256 | semiquaver |   3 |
| soprano |       1 |    1 | true   |  5 | note |      256 | semiquaver |   4 |
+---------+---------+------+--------+----+------+----------+------------+-----+--

mysql> select * from pr01Note limit 3;
+------+-------+-------+------+------------+------------+
| id   | pitch | staff | tie  | accidental | staffcross |
+------+-------+-------+------+------------+------------+
|    3 | c5    | upper | NULL | NULL       | NULL       |
|    4 | d5    | upper | NULL | NULL       | NULL       |
|    5 | e5    | upper | NULL | NULL       | NULL       |
+------+-------+-------+------+------------+------------+
```

3 WTC II Annotation Server

WTC II Annotation Server allows the user to view the content of both the text-critical database and the London autograph database. Our WTC II Annotation Server is implemented as a web application using Servlet technology, which is available with the Java programming language. The server integrates the two databases into one, and the user can use a web browser to search the database and view the result, although the two have not yet been integrated fully. Here we shall show how the data are presented in a web browser.

Figure 3 describes the search interface of the text-critical database of WTC II. It provides the user with the following search options:

- Title: the user can choose a particular movement from the WTC II.
- Bar: the user can specify a section of the movement with the range of bars.
- Voice: the user can specify which voices are to be shown. The interface allows the combination of different voices.
- Musical: the user can specify the type of musical variants in which he or she is interested. The type of variants are classified according to the patterns of authorship reflected in the sources. These variants are as follows:

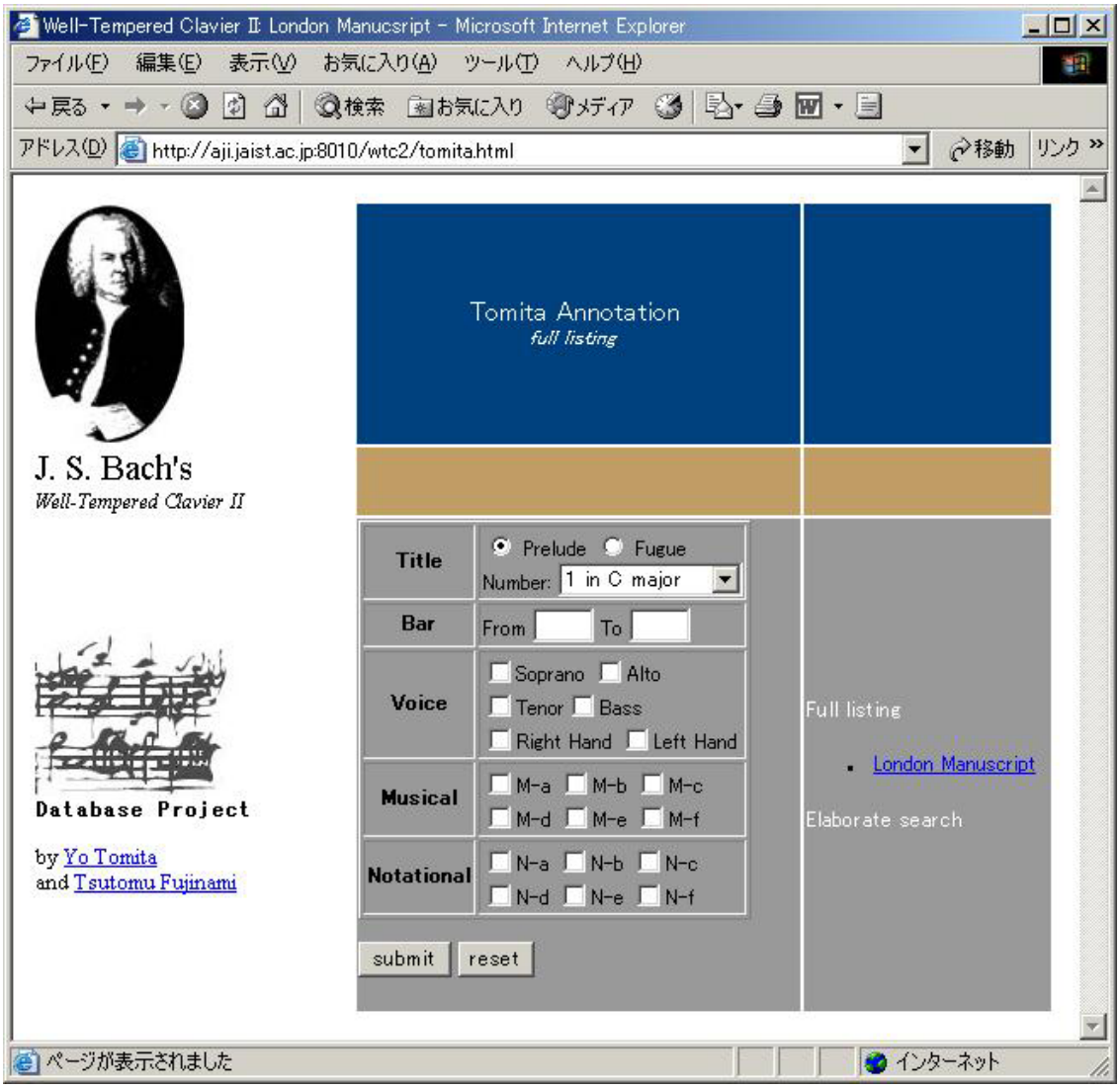


Figure 3: The search interface to the text-critical database

(a) Bach's different versions

(b) minor revision by Bach

(c) Bach's vague notation (unintentionally introduced)

(d) someone other than Bach (deliberately introduced)

(e) error by copyists (due to Bach's ambiguous notation)

(f) error by copyists (not due to Bach's ambiguous notation)

- Notational: the user can specify the type of notational variants in which he or she is interested. The same classification applies as the musical variants described above.

With these options, the user can view only the selected portion of data that interests him or her. The user can, for example, view only the variants that were introduced while Bach was revising his own score, limited to the bars from 1 to 5 of Prelude No. 1 in C major. The search result of this particular example is shown in Figure 4.

Prelude No. 1 in C Major

SN	2	3	5	6	7	9	10	19	33	35
Bar	1	1	1-3	1	1	2	2	3	5	5
Measure	1	1	1	1	1	2	2	3	5	5
Voice		RH	B	S	S	S	S	B	T	B
Beat	0	2	0	3	4	1	2	3	3	4
Position	0	2	0	0	0	0	3	0	0	0
VbtPos		R.H.	B	S,3	S,4	S,1	S,2/3	B,3-4	T,3-4	B,4
Element	t-s	Clef	voice	fig	mel	fig	pitch	fig	NV rest	mel
SpecLoc			[C]	(see below)	(see below)	(see below)	[a]	(see below)	[c]	[c e A c]
Classified	a,d	a,d	M-a,e	M-a	M-a,e	M-a	M-a,e	M-a,e	M-a, b?	M-a, e
Category	M-a M-d N-a N-d	M-a M-d N-a N-d	M-a M-e	M-a	M-a M-e	M-a	M-a M-e	M-a M-e	M-a M-b	M-a M-e
ApplMss	E		E	Tradition	E/F/DD 70	Tradition	F	F/DD 70		M-P Nr.8
P 804		L1	missing	E	E	E	E	E	only	c G A F? (no ledger)
Scheibn.4		L1	missing	E	E	E	E	E	only	c G A F !
P 1089		L1	missing	E	E	E	E	E	only	c G A amb:F/E
P 561		L1	missing	E	E	E	E	E	(beam shrd)	c G A E
Mem-Pre.8	(or)?	L1	missing	E	E	E	E	E	only	c G A E
N.10490		L1	missing	E	E	E	E	E	only	c G A E
L: var										
L: var-pitch				[e′ c″ g′ b♭′]	[e′ g′ c′ b♭′]			[c e G B♭ A c e G]		
Add.35021		L2	ok.	L	L	E	ok. (Lu)	L	only	ok.

Figure 4: A search result of the text-critical database

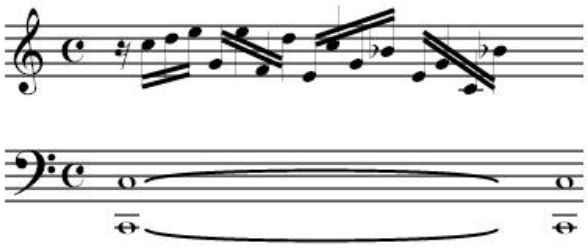


Figure 5: An example of SVG graphics depicting the beginning of Prelude No.1

4 Future Developments

We have so far succeeded in formatting the annotations of WTC II and the data of the London autograph in a relational database and providing the user with a search tool using a web browser. While there is still an outstanding issue of performance, the system currently permits us to share the data. The next steps of our work include the improvement of the data presentation and the mechanism used to analyse the data. Here are some of our on-going sub-projects:

4.1 Visualization of Manuscript Data

One of the problems that we recognize when presenting of text-critical data is the way the table is drawn. It is not very easy to read. We can make it more readable if we present the data with the parts of the movement to which they refer. We are thus trying to generate the score from the data of the London autograph database.

Figure 5 shows the result of our experiment in which the score was generated using Scalable Vector Graphics (SVG) and Bach font. The code is written by hand, which consists of about 150 lines of instructions. The quality of musical representation is not as good as that produced by Finale at the moment, but it is sufficient for us to understand what is stored in the database. We hope to generate the scores of other manuscripts with the information of the variants extracted from the text-critical database.

4.2 Tools for Statistical and Logical Analysis

Quite separately, we are also looking into the ways to analyze these data statistically. One may wish, for instance, to find out what kind of mistakes a particular scribe made from the text-critical data. By using the London autograph data set, one can investigate the inner construction of music, for example, how often a particular tonal structure appears in all the forty-eight movements. These statistical investigations can be implemented in Java.

While statistical analysis may yield certain aspects of the sources successfully, we need further tools that are capable of finding more complex types of answers with logical analysis. A scribe's use of accidentals is one such example, as accidentals are governed not only by the grammar of music but also by the notational conventions that are influenced by the geographical and historical context in which the scribe worked. When combined with the statistical analysis, we may be able to reconstruct the data for the missing manuscript source that links two or more manuscripts, while at the same time extracting those errors and variants that originated from the scribe. As a first step in the logical analysis in music, we have converted the London autograph database into prolog, a logic programming language, in order to see whether or not we can discover structural properties embedded in the autograph. The results of this logical analysis will be stored in tables and used for statistical analysis.

5 Conclusion

What appears at first to be a daunting task to analyse the vast quantity of text-critical data has been made manageable by converting the data into XML and organising the data in relational databases. Although much work is still needed to identify what domain knowledge we need for the computers to perform certain analytical tasks, we estimate that the fundamental framework for this project has now been established.