Watermarks and Where to Find Them: Digitisation, Recognition, and Automated Clustering of Watermarks in the Music Manuscripts of Franz Schubert

MEC 2022 PEOPLE'S CHOICE – BEST PAPER AWARD

Paul Gulewycz, Clemens Gubsch, Marlene Peterlechner, Katharina Loose-Einfalt
Austrian Center for Digital Humanities and Cultural Heritage (ACDH-CH)
paul.gulewycz@oeaw.ac.at, clemens.gubsch@oeaw.ac.at, marlene.peterlechner@oeaw.ac.at, katharina.loose@oeaw.ac.at

Günther Koliander
Acoustics Research Institute (ARI)
guenther.koliander@oeaw.ac.at

Andrea Lindmayr-Brandl
Paris-Lodron-University Salzburg (PLUS)
andrea.lindmayr-brandl@plus.ac.at

Abstract

Our project focuses on watermarks found in the music manuscripts of Franz Schubert. The endeavour incorporates thermography, machine learning and signal processing to produce digitized watermarks for databases and manuscript descriptions as well as to curtail the approximate dating of some undated autographs. By applying fingerprint recognition software to the acquired thermographic watermark images, a new method for automatic clustering will be established. MEI will be used as the system's foundation for data presentation online and to guarantee long-term archiving and open source access. As MEI currently does not provide a best practice for encoding watermark information, a standardized form will be developed in collaboration with the community.

Introduction

Until the mid-nineteenth century, the production of paper was a manual process. A mixture of cotton fibers and water called pulp was poured into a wire mesh, then pressed and dried. To mark their products, the workers of the paper mills formed letters and symbols such as crowns, flowers and crescents out of copper wire and attached them onto the mesh. The thickness of the paper would be reduced by the copper wire and the symbols could then be seen as watermarks when the paper was held up against the light (Gerardy, 1972).

By analyzing watermarks scholars working on historical documents can ascertain vital information on the provenance of a particular document and on the date of creation of the manuscripts in the source description. The identification of the symbols can aid in the temporal categorization of music manuscripts and thus the written musical works as information about the chronology of the papers can be derived from the watermarks. When the Internet offered better access and distribution of knowledge, various online databases for watermarks were established (WZMA, 1999; WZIS, 2010) on the basis of watermark collections from Charles-Moïse Briquet, Gerhard Piccard and others (Eineder, 1960), to facilitate the comparison and subsequently the dating of different or similar watermarks.

The relatively recent dawn of digital tools for watermark identification also promised new insights on already analyzed manuscripts, so that the provenance or dating of a sheet of paper, and subsequently the written information found on it, could possibly be adjusted and corrected. Furthermore, these findings could lead to an exact standardized workflow, whose results can be accepted as cogent evidence, especially when compared to the hitherto existing strategies of watermark documentation. In the course of this project, these techniques will be developed and tested on autograph music manuscripts by the Austrian composer Franz Schubert (1797–1828), which are held in two libraries in Vienna.

1 Watermarks in Music Manuscripts

1.1 Identification

For decades, the main method for documenting watermarks was tracing the symbols by hand. Although the accuracy of the tracings is sometimes remarkably high, the results obtained by this method leave room for interpretation regarding the positioning of the watermark on the sheet as well as any deformations of the mark. In the absence of alternatives that ensured the documents remained unharmed, a large catalogue of watermark tracings developed over time. Although they were produced in great numbers, naturally, the quality of the tracings and the information content regarding meta-information varied greatly.

In the past fifteen years, two new techniques have been developed for imaging watermarks digitally: thermography and image-subtraction by means of transmitted light. Both methods are safe in terms of conservation and do not affect the manuscripts. Image-subtraction is significantly cheaper than thermography, but the filtering of anything written or printed on the paper in ink or pencil is problematic. In the case of music manuscripts, staves and music notes remain visible in the transmitted light procedure, which is unfavourable for further digital processing. Additional pixels would be left on the images and would have to be extracted in a separate work sequence in order to further process only the depiction of the watermark. Also, the higher recording quality of thermography simplifies further processing of the data (Meinlschmidt et al., 2011), which is why the thermographic method is preferable.

By using a system based around a thermographic camera, it is possible to precisely record and process the watermarks as well as minute structures of the wire mesh, which could possibly also be used for comparison to identify paper types. As watermarks are gradually deformed through extensive use and eventually replaced by new metal objects, the thermographic system helps to discern minimal variations of the shape of a watermark, which are impossible to distinguish with the naked eye.

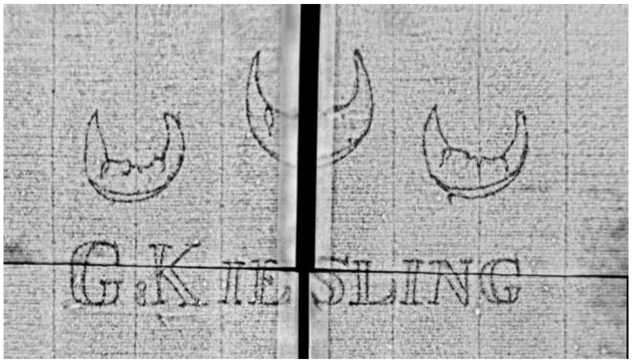

Figure 1: “Watermark DE0960-Schubert16_1”. https://www.wasserzeichen-online.de/?ref=DE0960-Schubert16_1 (accessed April 19, 2022).

1.2 Signal processing

Based on the high-quality digitized thermographic watermark images, signal processing methods will be used to group manuscripts that contain the same watermark. To achieve this, the difference of the watermarks of two digitized manuscripts will be quantified. Small values indicate that the watermarks are almost identical, while large values indicate that they are completely different. However, a direct comparison of the images is impossible when defining such a distance measure. First, the image data needs to be preprocessed in order to obtain a compact and robust representation of the watermarks. Inspired by ideas from fingerprint recognition (Peralta et al., 2015), it is possible to extract characteristic points-of-interest from a grayscale image, such as intersections of lines or corner points by using established image processing methods. This turns an image into a dot pattern, which is characteristic for a watermark and at the same time can be processed well mathematically. Fingerprint identification algorithms require the images to be processed further due to their complexity, whereas black and white (b/w) versions of the watermark images can already be used for identification. In both point patterns and b/w-images, the so-called ‘Earth Mover’s Distance’ (or ‘Wasserstein metric’) will be used to measure how closely two watermarks match. Roughly speaking, this distance evaluates how far the points of one pattern (or the black parts of a b/w-image) have to be moved to match the second pattern (or image), respectively. Although rotation invariance is difficult to deal with, this distance can easily be adapted to be translation invariant. Based on this or a similar distance, various clustering algorithms will be used to identify which images share the same watermark. In particular, k-means clustering, which is a frequently applied algorithm for vector quantization and cluster analysis, will be used for point patterns. This method is closely related to another algorithm for vector quantization, the so-called Linde-Buzo-Gray algorithm, that was already successfully applied to point patterns by Günther Koliander, one of the principal investigators. The precise choice and adaptation of the algorithms to detect the characteristic dots is one of the first technical challenges of the project.

Once this procedure is established, we will use classical or Bayesian nonparametric clustering techniques to group the watermarks. In doing so, we must consider that watermarks have often been fragmented when large papers were divided and cut. Watermarks would then be positioned not in the centre but on the margins of the paper, which in the worst case could be exposed to further cuttings to fit certain formats. These pieces must then be assembled from different pages to form a more or less complete watermark. Although it is often clear which pages originally belonged together, correctly matching the pages should also be automated. This process will be based on the comparison of the imprints of the wire mesh, the so-called laid lines. Subsequently, however, the clustering should also be able to correctly assign partial prints, possibly with additional information on where the end of the sheet is located in relation to the dot pattern. In summary, the goal in this essential part of our project is to enable an objective, quantitative comparison between watermarks and, based on this, to merge the digitized images into groups with identical watermarks.

In order to also make the aforementioned existing watermark tracings usable in our procedure, they will be integrated in further processing. As explained previously, hand-made tracings are qualitatively quite different from thermographic images: there are usually additional notes, no mesh imprint or laid lines except for stylized chain lines, and parts of different pages are usually merged without paying attention to millimeter-precise alignment, especially in the spacing between papers. In addition, the tracings may have been idealized, e.g. lines may be drawn through, although gaps are actually visible in the watermark. Nevertheless, it should be possible to quantify the similarity between a watermark in a tracing and a thermographic image. Since we cannot rely on the same level of detail in tracings as in thermographic images, we plan to harness tracings through style transfer methods from the field of machine learning (Isola, 2016). An algorithm can be trained to convert tracings into thermographic images of the same watermark. These converted images can then be further processed in our method. As training data for the algorithm, we will use pairs of existing tracings and associated thermographic images that we have created. The whole procedure can be validated on pairs that are not used for training by computing the distance between the converted tracing and the true thermographic image that we use. Only if we get consistently good results will we actually use this workflow to integrate tracings. Alternatively, with manual pre-processing (deleting the notes and chain lines, splitting the watermarks into individual sheets), we can directly use an adapted algorithm to detect the characteristic points. In any case, what we expect from the comparison of tracings and thermographic images is an analysis of potential errors in the recording of tracings, which further reduce the quality compared to thermographic images and make automated use more difficult.

2 Music Manuscripts by Franz Schubert

2.1 Schubert's Autographs in Vienna

Franz Schubert's compositions form an integral part of the genre-historical repertoire of the early nineteenth century and are of significant importance for music history. The oeuvre comprises almost all genres: German Lieder, song cycles, choral works, symphonies, chamber and piano music, dance music and marches as well as stage works and church compositions. All these compositions document the rich musical practice in Vienna in the first decades of the nineteenth century. Although the composer died young, his works had a considerable impact on later generations. During the editorial work of individual compositions, the relevant music manuscripts have been studied and documented in the critical reports of the New Schubert Edition. These reports provide rich codicological information in close detail. A base of this source documentation is a stock of about 1,300 watermark tracings that were produced by several different editors over approximately 50 years. As mentioned above, the method to produce such tracings was by pencil and paper (Shen et al., 2019; La Rue, 1961; Koliander et al., 2018), partly for technical reasons, partly for conservation reasons. The results of this research were primarily used to verify the date of works (Winter, 1982), but not to draw conclusions about the compositional process (Litschauer, 2001). Moreover, it was not part of the editorial work to analyze these findings across different works or sources, so that a comparative overview of Schubert's manuscripts and oeuvre is a task still to be accomplished.

Two thirds of Schubert's music manuscripts are located in Vienna. Most of them are held by the Music Departments of the *Wienbibliothek im Rathaus* (WB) and the *Österreichische Nationalbibliothek* (ÖNB). The WB holds the larger stock of Schubert manuscripts for the project, with approximately 492 manuscripts and 890 complete or fragmented watermarks, which are presented on the website Schubert Online (Schubert Online, 2010). Both institutions are collaborative partners of the project and will allow us to use a thermographic system in their premises. The watermarks can then be precisely recorded and processed by a thermographic camera, catalogued in a database and displayed in an IIIF-viewer. To supply the two libraries with a thermographic system, a portable version will be assembled by the Fraunhofer Institute for Wood Research, Wilhelm-Klauditz-Institut (WKI). The fact that the system is easily moveable opens up the possibility for future projects to carry out watermark research in smaller archives and monastery collections in accordance with preservation standards.

2.2 Difficulties

Apart from the acquisition, transportation and transfer of an expensive device such as a thermographic system, we anticipate that the manuscripts themselves will pose certain problems in thermographic imaging and in analysis. It must be kept in mind that, naturally, the manuscripts are exceedingly valuable and must be handled with great care. Access to the manuscripts is very restrictive, which will not easily facilitate working on them, either. As well, many of the manuscripts are bound like a book and are not available as individual sheets or leaves. This may make it difficult to capture the digital images in some circumstances, e.g., when the watermark is located near the inner margins. The biggest problems, however, will probably arise when connecting the thermographic images. We have to assume that on the majority of the pages only parts of the watermark, instead of the complete symbols, will be found. This is due to the aforementioned fact that the papers were either already cut when Schubert bought them or divided by Schubert himself after the purchase in order to get two bifolia, four folia or eight writable pages out of one large sheet. Since the watermarks were positioned centrally on the original sheet, they too were divided when the paper was cut. These fragments are located at the top or bottom of the sheet, as is to be expected. Further problems can arise at this point, because the edges of the paper may have been trimmed, so that up to one centimetre of one page could be missing completely.

So-called twin watermarks present another difficulty. Not only one, but two moulds were usually used during paper production. A team of two workers would manufacture paper in a rhythmic procedure: the so-called vatman would plunge the first mould into the pulp, let excessive pulp run over the edges and then pass the mould to his helper, the so-called coucher. After allowing the sheet to dry for a moment, the coucher would then turn it out on felt and return the mould back to the vatman. The vatman would take up the second mould in the meantime and hand it to the coucher at the same time as he received the empty mould from the coucher (Stevenson, 1951/1952, p. 60–61). These pairs of moulds would include identical watermarks, which skilful workers sewed onto the meshes. The watermarks are usually only slightly different, but can also be more easily distinguishable by intentionally mirrored features. However, it is important to record the watermarks from the correct side, that is the one with a slightly rougher paper surface, as instead of a twin a variant watermark could be erroneously recognized. For this reason, we currently have to assume that some of the watermarks recorded so far have been incorrectly classified as new independent watermarks and are actually twin watermarks.

3 Data Modelling and Representation

3.1 Data basis

The data basis for the project comprise MEI files that contain a variety of different information that is used for online presentation on the one hand, and for cataloguing on the other. For each of Franz Schubert's manuscripts that are located in the two libraries, a single file is created. Information about the source is stored in the `<meiHead>`, specifically, the shelf mark, the provenance, the repository, etc. In addition, the `<meiHead>` provides work classifications such as the so-called D-number, which refers to an entry in the catalog of works by the Austrian musicologist Otto Erich Deutsch (1883–1967). Furthermore, the works are divided into genres such as stage works, chamber music, German Lieder, etc. The physical order of the pages in `<foliaDesc>` is also relevant for the inclusion of watermarks. Based on collation diagrams and source descriptions found in the critical reports, which document the order of the sheets, the information can be transferred to MEI using the elements `<folium>`, `<bifolium>`, and, if there are any pieces of paper attached to a page, `<patch>`. At this point the watermarks will also be included and connected to the page, in which they appear by referencing their IDs. In `<music>` all necessary structures of `<surface>` and `<graphic>` elements are nested in the correct order in a `<facsimile>` element for display in an IIIF viewer. The IDs of the `<graphic>` elements are referenced in the physical order of the pages in `<foliaDesc>`.

Concrete descriptions of the paper mills must also be encoded. Relevant information includes: from when to when the mill was operated, at which location, by which persons, and which watermarks were produced there. The watermarks are to be encoded in another container that needs to include information about the description of the watermark symbols and whether they are complete or fragmentary or twin watermarks. This information is then to be linked to databases and images.

3.2 Encoding Watermarks in MEI

The encoding of the watermarks themselves is a particularly important part of our work, because a standardised form or best practice of entering watermark information in MEI does not currently exist. This has to be established as a standard within the project, in collaboration with the MEI community. At MEC 2021, one of the authors, Clemens Gubsch, already proposed solutions and presented an elaborate encoding concept. Since the verbal description of watermarks requires a high degree of uniform terminology, the typology established by the Bernstein project (Bernstein-Portal, 2006) will be used. This guarantees the searchability of the data in the environment of the aforementioned watermark database WZIS.

Next to this verbal information, we will establish an image-based characterization of the watermarks. The details of this characterization will be developed in the course of the project. Possible representations include complete black and white images of the watermarks in a predefined resolution or describing merely a point pattern that summarizes the most distinctive points within the watermark. The project aims at finding a representation that enables the efficient comparison of new watermarks with the existing catalogue. Such a comparison should also be possible when only a small part of the watermark is visible, or the watermark is in some way distorted.

Summary and Conclusion

The main goals of the project DRACMarkS are to digitize watermarks found in the music manuscripts of Franz Schubert, which are located in two libraries in Vienna, via a thermographic camera system, and to develop an extension for watermark encodings to the MEI schema. The thermographic process ensures high-quality watermark images that will be used to develop an algorithm-based automatic analysis software. By distinguishing lighter and darker dots on the thermographic images, characteristic dot patterns are generated which, when compared in a manner similar to fingerprint recognition software, enable rapid and unambiguous assignment. The resulting data will be integrated into the larger context of the future source database Schubert digital, which will be formed by MEI files that contain encoded information on persons, institutions, musical works and watermarks.

For the encoding of the watermarks, extensions of the existing MEI elements are necessary to allow a concrete description of the characters, which is essential for the identification of different watermark symbols. There have already been similar initiatives in TEI (Müller, 2020) and drafts for ontology-based languages (Eichenberger, 2019), and an inclusion of new watermark-related elements could also improve MEI encodings of music manuscripts and source descriptions. This would follow MEI's holistic approach to ensure the representation of a broad range of musical documents and structures.

References

Bernstein-Portal (2006). www.memoryofpaper.eu

Eichenberger, Nicole (2019). Ein Ontologie-Entwurf für die Klassifikation von historischen Wasserzeichen. In *Informationspraxis* 5(1). https://doi.org/10.11588/ip.2019.1.56833

Eineder, Georg (1960). *The Ancient Paper-Mills of the Former Austro-Hungarian Empire and Their Watermarks*. Hilversum.

Gerardy, Théo (1972). Die Techniken der Wasserzeichenuntersuchung. [Conference presentation]. In *Les techniques de laboratoires dans l'étude des manuscrits. Actes du Colloques Internationaux du C.N.R.S. organise par J. Glenisson et L. Hay*, *Paris (September 13-15, 1972)*, 143–157. https://bernstein.oeaw.ac.at/twiki/pub/Handbook/WebHome/gerardy_techniken.pdf

Isola, Phillip, et al. (2016). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. https://arxiv.org/abs/1611.07004

Koliander, Günther, Schuhmacher, Dominic and Hlawatsch, Franz (2018). Rate-distortion theory of finite point processes. In *IEEE Transactions on Information Theory* 64.8, 5832-5861*.*

La Rue, Jan (2001). Watermarks and Musicology. In *The Journal of Musicology* 18(2): 313–343. Reprinted from *Acta Musicologica* 33 (1961).

Litschauer, Walburga (2001). Wasserzeichen als Datierungshilfe. Neue Erkenntnisse zu Schuberts Werken vor 1823. In D. Berke, W. Dürr, W. Litschauer and C. Schumann (Eds.), *Schubert-Jahrbuch 1999* (= Bericht über den Internationalen Schubert-Kongreß Duisburg 1997, Teil III), Bärenreiter, 155–162.

Meinlschmidt, Peter, Kämmerer, Carmen and Märgner, Volker (2011). Thermographie – ein neuartiges Verfahren zur exakten Abnahme, Identifizierung und digitalen Archivierung von Wasserzeichen in mittelalterlichen und frühneuzeitlichen Papierhandschriften, -zeichnungen und -drucken. In *Codicology and Palaeography in the Digital Age* 2, 209–226.

Müller, Ermenegilda (2020). A TEI Customization for Paper and Watermarks Descriptions. In *Digital Medievalist* 13(1), 1–24. https://doi.org/10.16995/dm.91

Peralta, Daniel, et al. (2015). A survey on fingerprint minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation. In *Information Sciences* 315, 67–87.

Schubert Online (2010). www.schubert-online.at

Shen, Xi, Pastrolin, Ilaria, Bounou, Oumayma, Gidaris, Spyros, Smith, Marc, Poncet, Olivier and Aubry, Mathieu (2019). Large-Scale Historical Watermark Recognition: Dataset and a New Consistency-based Approach. https://arxiv.org/abs/1908.10254

Stevenson, Allan H (1951/1952). Watermarks Are Twins. In *Studies in Bibliography* 4, 57–91.

WZMA: Wasserzeichen des Mittelalters (1999). www.wzma.at/

WZIS: Wasserzeichen-Informationssystem (2010). www.wasserzeichen-online.de/

Winter, Robert (1982). Paper studies and Schubert research. In E. Badura-Skoda and P. Branscombe (Eds.), *Schubert Studies. Problems of style and chronology*, Cambridge University Press, 209–27.