# Regeln zur Konvertierung der Docx-Dateien der Editionsreihe "Korrespondenz Österreichischer Herrscher" zu XML/TEI

## 1. Regeln von Docx zu XML/TEI

Diese Regeln transformieren einen Text, der als Docx-Datei vorliegt, in einen anderen Text, der in XML/TEI vorliegt. Sie sind die Basis der Datenbereitstellung und bilden daher die Grundlage einer digitalen Edition. Diese Regeln sollen möglichst früh feststehen, um spätere Ausbesserungen zu verhindern.

1. Alle leeren Zeilen werden automatisch gelöscht.
2. Die Kommentare über Kommentier-Funktion des MS-Words werden so konvertiert:
3. Fängt das Kommentar mit P. und, so wird ein <persName> daraus, und das, was nach "P." kommt, wird als @key eingetragen. (Alle Stellen, welche dieselbe @key haben, verweisen auf dieselbe Person.)
4. Ähnliches geschieht mit O., nur hier wird ein <placeName> daraus.
5. Mit "S." wird ein <index> daraus. Ich muss leider zugeben, dass ich vergessen habe, was "S." bedeutet.
6. Trifft nichts von i. bis iii. zu, so wird eine "Einzelstellenkommentar" daraus. Und das Kommentar bleibt als <note> bestehen.
7. Gibt es zwei (oder mehrere) Kommentare, die sich überlappen, so wird überprüft, ob sich die Überlappung auch hierarchisch abbilden lässt. Ist das der Fall, werden die Regeln i-iv angewendet; ist das nicht der Fall, werden <anchor/> gesetzt, um die Anfangs- und Endpunkte der Kommentare zu markieren.
8. Die Metadaten werden "erraten":
9. Die erste Zeile ist die ID des Briefs (z.B. 370215KFa). Diese ID wird in <sourceDesc> innerhalb eines <idno> eingetragen.
10. Die erste Zelle der ersten Tabelle beinhaltet Informationen über Absender (vor "an") und Empfänger (nach "an"). Diese Information wird verwendet, um den Titel zu generieren.
11. Die zweite Zelle der ersten Tabelle beinhaltet die Datierung und das Absende-Ort. Absende-Ort von der Datierung trennen.
12. Das <title> in <titleStmt> wird so generiert: <persName role="sender">$sender</persName> an <persName role="recipient">$recipient</persName>, von <date>$date</date>
13. Die Strukturen der Briefe werden erraten:
14. Bis zur ersten Tabelle: Header. Dieser Bereich wird in <div type="header"> zusammengefasst.
15. Die Absätze nach der Tabelle bis zu "Druck:..." sind Regesten. Diese werden in <div type="regesta"> zusammengefasst.
16. Die Absätze nach "Druck:..." bis zur Kursivschreibung eines ganzen Absatzes gehören zum Transkript. Dieser Bereich wird in <div type="transcript"> zusammengefasst.
17. Alles, was nach dem Transkript kommt, sind Kommentare. Sie werden in <div type="notes"> zusammengefasst.
18. Entspricht ein Brief dieser Struktur nicht, oder wird die einzelnen "Hinweis-Absätze" nicht erkannt, wird der ganze Text in ein <div type="notRecognized"> gefasst.
19. Interne Verlinkungen herstellen. Folgende Verlinkungen werden automatisch hergestellt:
20. Die markierten Stellen (zu Personen und Ortsnamen etc.) werden als Link dargestellt. Wenn man darauf klickt, erscheint das Key.
21. Die Stellen im Transkript, die mit hochgestellten Buchstaben markiert sind, werden mit den Kommentaren verlinkt.

## 2. Regeln von XML/TEI zu HTML

Diese Regeln sind frei wählbar und werden je nach der verwendeten Technologie verändert. Im Moment kommen folgende HTML-CSS-Technologien zum Einsatz: Twitter Bootstrap für das Layout, JQuery für das Laden und das Darstellen der Briefe.

1. Gestaltung eines Briefs
2. Jeder Brief bekommt einen Titel, einen Block zum Herunterladen der XML- bzw. Docx-Dateien, einen Block mit Metadaten und einen Block mit Text.
3. Metadaten-Block listet folgende Daten auf: Absender, Empfänger, Datierung, ID des Briefs und die Verweise zu Personen, Ortsnamen und Index-Einträge. Die Verweise werden automatisch generiert. (Link-Funktion überprüfen, vielleicht einkreisen der markierten Stellen?)
4. Text-Gestaltung:
5. Der Text soll weitgehend mit dem Text in der entsprechenden Docx-Datei übereinstimmen. Prinzipiell als Blocksatz formatiert.
6. Die Hervorhebungen sollen beibehalten werden. Übernommene Hervorhebungen sind bis jetzt italic, bold, underline, background(#color), subscript, superscript, left, right, center und justify.
7. Ein <div> wird in <div class="panel"> dargestellt mit entsprechender Überschrift, um die einzelnen Textteile voneinander besser unterscheiden zu können.
8. Ein Absatz wird als <p> dargestellt, entsprechenden Tags für <table>, <row> und <cell>.