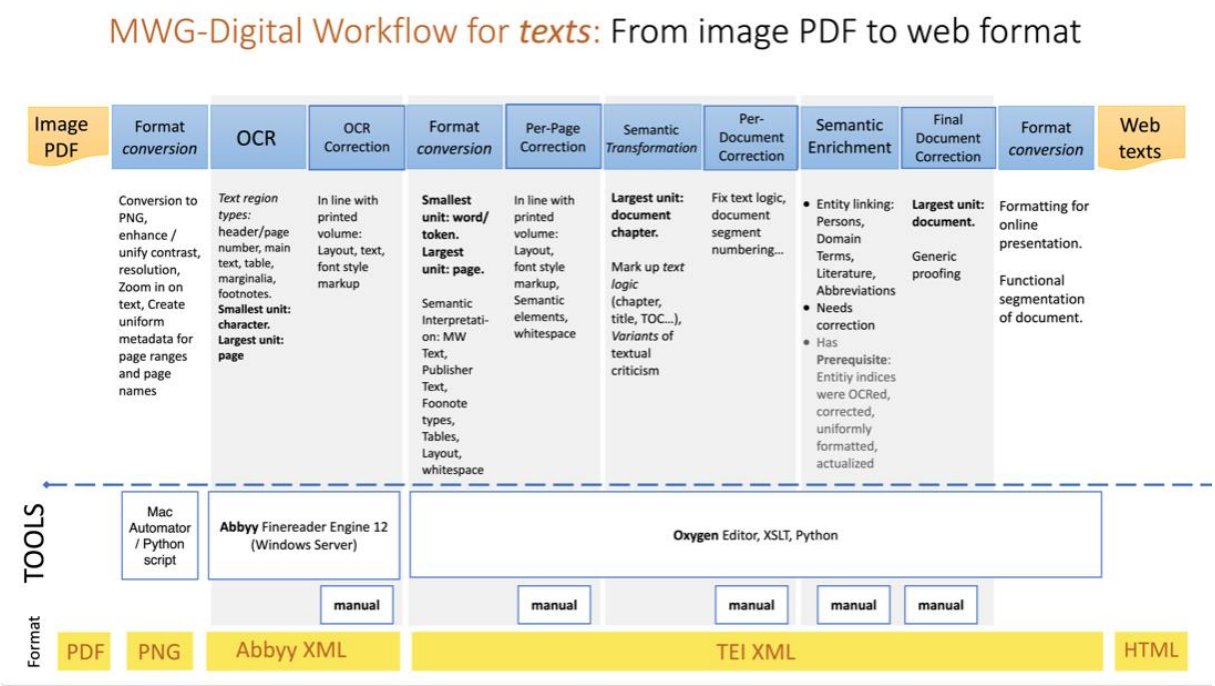
BAdW | BAYERISCHE AKADEMIE DER WISSENSCHAFTEN

Die Transformation der Max-Weber-Texte: Vom Bild-PDF zum Web-Text

Für die Bereitstellung in *MWG digital* durchlaufen die gedruckten Texte der *Max Weber-Gesamtausgabe* (MWG) eine mehrstufige Transformation: vom gescannten Bild-PDF hin zum maschinenlesbaren und semantisch strukturierten HTML-Format. Dieser Web-Text wird allen Interessierten kostenfrei im Internet zugänglich gemacht.

Der gesamte Prozess kombiniert automatisierte Verarbeitungsschritte mit sorgfältiger manueller Qualitätskontrolle, um eine bestmögliche Übereinstimmung mit der gedruckten Edition sicherzustellen.



Überblick über die Transformationsstufen

1. Erste Transformation: Vom Bild-PDF zum Rohtext

Ziel: Umwandlung der MWG-PDFs in bearbeitbare Rohdaten.

- **Bildaufbereitung:**
  Die PDF-Dateien der MWG-Bände werden in einzelne PNG-Bilder pro Seite

konvertiert. Dabei werden Kontrast, Auflösung und Textgröße vereinheitlicht; jede Datei erhält eindeutige Metadaten und einen individuellen Dateinamen.
    - *Werkzeuge:* Mac Automator, Python-Skripte
    - *Format:* PNG
    - *Methode:* Automatisiert

- **Texterkennung (OCR):**
  Mithilfe der ABBYY Finereader Engine wird der Text aus den Bilddateien extrahiert. Dabei erfolgt bereits eine erste Gliederung in Textbereiche wie Haupttext, Kopfzeile, Tabellen, Marginalien und Anmerkungsapparate. Anschließend werden die Ergebnisse manuell nachkorrigiert.
    - *Werkzeuge:* ABBYY Finereader Engine 12 (Windows Server)
    - *Format:* ABBYY XML
    - *Methode:* Automatisiert + manuell

---

2. Zweite Transformation: Strukturierung und semantische Auszeichnung (TEI XML)

Ziel: Formale und inhaltliche Auszeichnung der Texte im TEI-Standard.

- **Einzelauszeichnung:**
  Jede OCR-Einzelseite wird über ein Transformationstool für die Bearbeitung in *Oxygen XML Editor* vorbereitet. Die Textelemente (Haupttext, Anmerkungsapparate, Tabellen, Schrifttypen etc.) werden semantisch ausgezeichnet. Diese Auszeichnungen werden durch Abgleich mit der gedruckten MWG-Seite überprüft und korrigiert.
    - *Werkzeuge:* Oxygen Editor, XSLT, Python
    - *Format:* TEI XML
    - *Methode:* Automatisiert + manuell
- **Zusammenführung zu Gesamtdokumenten:**
  Die Einzelseiten werden zu vollständigen Dokumenten zusammengeführt, die jeweils einer Texteinheit der MWG-Ausgabe entsprechen (inklusive Editorischer Berichte bzw. Vorbemerkungen und Anhänge). Die weiteren Strukturierungen (wie Kapitel, Inhaltsverzeichnisse oder textkritische Varianten) werden automatisiert vorgenommen und anschließend manuell geprüft, insbesondere bei komplexen Elementen wie seitenübergreifenden Anmerkungen.
    - *Werkzeuge:* Oxygen Editor, XSLT, Python
    - *Format:* TEI XML
    - *Methode:* Automatisiert + manuell

- **Semantische Anreicherung (optional):**
  Zusätzlich können Entitäten wie Personennamen, Sachbegriffe, Literaturhinweise und Abkürzungen im Text ausgezeichnet und verlinkt werden. Dieser Schritt wurde im Pilotband I/2 getestet, jedoch aufgrund des hohen Zeitaufwands zunächst nicht weiterverfolgt.
  - *Werkzeuge:* Oxygen Editor, XSLT, Python
  - *Format:* TEI XML
  - *Methode:* Manuell

---

3. Dritte Transformation: Bereitstellung im Webformat (HTML)

Ziel: Umwandlung des strukturierten TEI-Texts in ein webbasiertes Format.

- **HTML-Generierung:**
  Das vollständige TEI XML-Dokument wird mittels eines Transformationstools in HTML überführt, wobei lange Dokumente in sinnvoll gegliederte Abschnitte aufgeteilt werden.
  - *Werkzeuge:* Oxygen Editor, XSLT, Python
  - *Format:* HTML
  - *Methode:* Automatisiert
- **Abschlusskorrektur:**
  Vor der Veröffentlichung wird das HTML-Dokument einer abschließenden Korrektur unterzogen. Die HTML-Ansicht wird eingehend geprüft, um die inhaltliche und strukturelle Übereinstimmung mit der MWG-Druckausgabe sicherzustellen. Die Korrekturen erfolgen im Oxygen Editor direkt am TEI-Dokument.
  - *Werkzeuge:* Oxygen Editor, XSLT, Python
  - *Format:* TEI XML
  - *Methode:* Manuell
- **Feinabstimmung:**
  Die Absatz- und Anmerkungsauszeichnung wird final geprüft, um die korrekte Funktion der Such- und Navigationsfeatures im Webtext zu gewährleisten.

Zusammenfassung

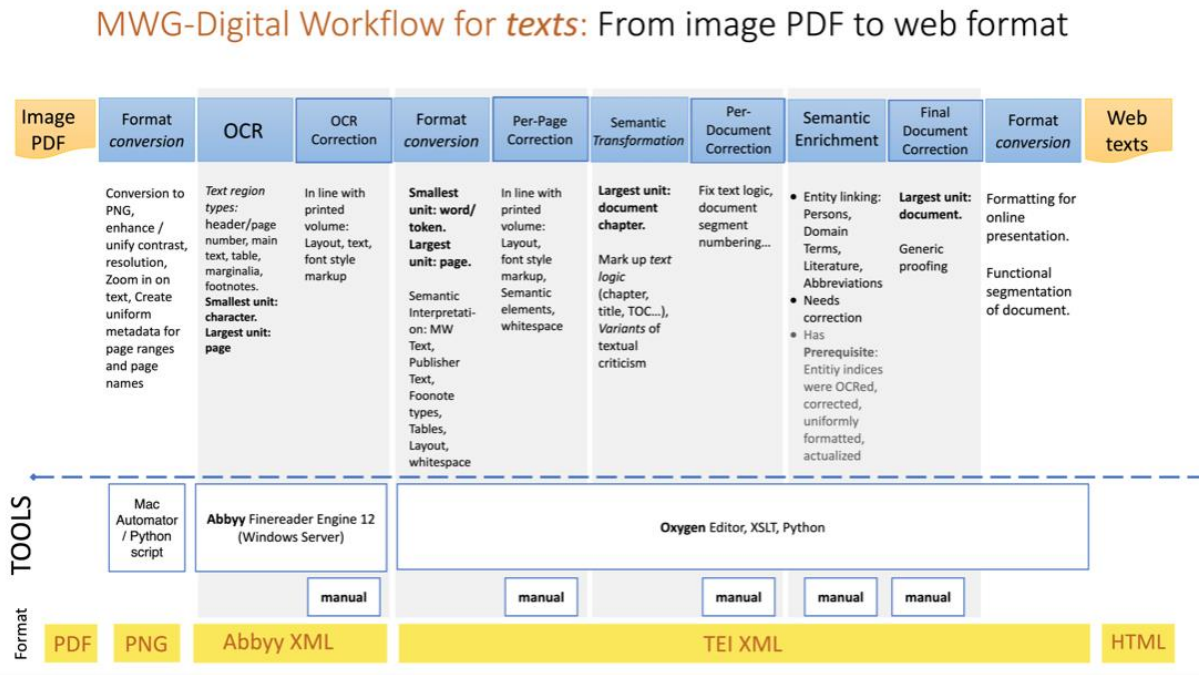
Durch die präzise Kombination von automatisierten Tools und fachkundiger manueller Kontrolle wird der Text der *Max Weber-Gesamtausgabe* in ein modernes, webfähiges Format überführt. Dieses Verfahren stellt sicher, dass die

digitale Ausgabe nicht nur zugänglich und durchsuchbar ist, sondern auch der editorischen Sorgfalt der gedruckten MWG in vollem Umfang entspricht.

Transformation of the Max Weber Texts: From Image-PDF to Web Format

To be made available in *MWG digital*, the printed texts of the *Max Weber-Gesamtausgabe* (MWG) undergo a multi-stage transformation from scanned image-PDFs to machine-readable and semantically structured HTML files. The resulting web texts are made freely available to the public online.

This process combines automated workflows with careful manual review to ensure the highest level of accuracy and fidelity to the printed edition.



Overview of the Transformation Stages

1. First Transformation: From Image-PDF to Raw Text

Objective: Conversion of MWG PDFs into editable raw data.

- **Image Preparation:**
  The PDF files of the MWG volumes are converted into individual PNG images, one per page. Contrast, resolution, and text size are standardized, and each page is assigned consistent metadata and a unique filename.
    - *Tools:* Mac Automator, Python scripts

    - *Format:* PNG
    - *Method:* Automated

- **Optical Character Recognition (OCR):**
  Text is extracted from the images using the ABBYY Finereader Engine, which automatically identifies and distinguishes between text areas such as headers with page numbers, main text, tables, marginal notes, and footnotes. A thorough manual correction follows to ensure accuracy.
    - *Tools:* ABBYY Finereader Engine 12 (Windows Server)
    - *Format:* ABBYY XML
    - *Method:* Automated + manual

---

2. Second Transformation: Structuring and Semantic Tagging (TEI XML)

Objective: Formal and semantic encoding of the texts according to TEI standards.

- **Page-Level Tagging:**
  Each OCR page file is prepared for editing in *Oxygen XML Editor* using a transformation tool. Elements such as main text, annotation apparatus, tables, spacing, and font styles are semantically tagged. This text is manually compared and corrected against the printed MWG page.
    - *Tools:* Oxygen Editor, XSLT, Python
    - *Format:* TEI XML
    - *Method:* Automated + manual
- **Merging into Complete Documents:**
  Individual pages are merged into complete documents that correspond to units in the printed MWG edition, including editorial reports or preliminary remarks and appendices. Titles, tables of contents, chapter divisions, and critical apparatus entries are then automatically tagged and manually verified, especially in cases involving cross-page annotations or textual variants.
    - *Tools:* Oxygen Editor, XSLT, Python
    - *Format:* TEI XML
    - *Method:* Automated + manual
- **Semantic Enrichment (Optional):**
  Additional semantic enrichment, such as linking person names, concepts, bibliographical references, and abbreviations, is possible. This was piloted in volume I/2 but suspended for later volumes due to time constraints.
    - *Tools:* Oxygen Editor, XSLT, Python

    - *Format:* TEI XML
    - *Method:* Manual

---

3. Third Transformation: Web-Ready Presentation (HTML)

Objective: Transformation of structured TEI XML into a web-compatible format.

- **HTML Generation:**
  Using a transformation tool, the full TEI XML documents are converted into an HTML format based on TEI XML. Long documents are divided into functional sections for ease of navigation.
    - *Tools:* Oxygen Editor, XSLT, Python
    - *Format:* HTML
    - *Method:* Automated
- **Final Review:**
  Before publication, each HTML document undergoes a final comprehensive review. It is ensured that the digital text faithfully reproduces the printed MWG edition. Corrections are made directly in the TEI document using the Oxygen Editor.
    - *Tools:* Oxygen Editor, XSLT, Python
    - *Format:* TEI XML
    - *Method:* Manual
- **Final Adjustments:**
  Paragraph and footnote tagging are fine-tuned to ensure that the search and navigation functionalities work reliably in the web presentation.

Summary

By combining automated processing with rigorous manual quality control, the *Max Weber-Gesamtausgabe* is transformed into a modern, web-accessible text format. This ensures that the digital edition remains fully aligned with the scholarly precision and editorial standards of the printed MWG volumes.