SLUB
Wir führen Wissen.

Handreichung Publikationen HTML/JATS

SLUB Dresden

Version 1.0.1, 2022-09-23

Inhaltsverzeichnis

Vorwort . . . 1
Allgemeines . . . 1
Abgrenzung . . . 2
Nutzungsszenarien und signifikante Eigenschaften . . . 2
Anforderungen . . . 3
- Allgemein . . . 3
- Ressourcentyp Text . . . 4
- Ressourcentyp Bild . . . 5
- Ressourcentyp kodierte Noten . . . 6
- Ressourcentyp Geoinformationsdaten . . . 7
- Ressourcentyp Formeln . . . 8
- Ressourcentyp Tabellen . . . 9

Quellenverweise . . . 10

Release Note 1.0.1 vom 2022-09-23

- Anpassung Struktur des Dokumentes

Vorwort

Dieses Dokument richtet sich an Produzenten, die digitale Objekte in das SLUBArchiv.digital einliefern und diese langfristig benutzbar erhalten wollen.

Allgemeines

Die SLUB strebt eine Erweiterung der OpenAccess Publikationen basierend auf HTML/JATS in den Bereichen musiconn.publish und Fachinformationsdienst Mobilitäts-und Verkehrsforschung an.

Die Quellformate dieser Publikationen sollen langzeitgesichert werden.

Die SLUB hat für jede der Publikationen folgende Ressourcentyp-Gruppen ermittelt:

- Text
- Bild
- kodiert Noten
- Geoinformationsdaten
- Formeln
- Tabellen

Jede Publikation (digitales Objekt) kann dabei aus mehreren Ressourcen bestehen. Ressourcen können dabei durch die Ressource Text

- innerhalb eines Objekts verlinkt
- integriert

werden.

Für jede der Ressourcentypgruppe wurden die Nutzungsszenarien für die Zielgruppen

- Urheber
- Forscher
- Studenten
- interessierte Öffentlichkeit

bestimmt und die im Langzeitarchiv dauerhaft zu erhaltenden (d.h. signifikanten) Eigenschaften der digitalen Resourcentypen abgeleitet.

Diese bilden die Grundlage für die Auswahl der geeigneten lza-fähigen Dateiformate und den zugehörigen verbindlichen Regelsätzen.

Vorbehaltlich gesonderter Absprachen mit dem SLUBArchiv sind von dieser Handreichung abweichende Änderungen **nicht** gestattet. Informationspakete, die diese Anforderung nicht erfüllen, sind nicht langzeitarchivfähig und werden vom SLUBArchiv zurückgewiesen.

Abgrenzung

Dieses Dokument beschreibt die Anforderungen des SLUBArchiv an OpenAccess Publikationen für die gehosteten HTML-JATS Repositorien.

Nicht von diesem Dokument erfasst sind Publikationen, die retrodigitalisiert wurden, oder die originär als PDF vorliegen.

Nutzungsszenarien und signifikante Eigenschaften

Die Nutzungsszenarien und signifikante Eigenschaften werden im Dokument jeweils zum entsprechenden Ressourcentyp (Abschnitt Eigenschaften) behandelt. Grundlage dafür bildet die Datei https://git.slub-dresden.de/digital-preservation/significantproperties/-/blob/master/sigprops_publications_JATS.xml.

Anforderungen

Allgemein

Der Ressourcentyp Text ist Basisresource einer Publikation, dh. jede Publikation besteht mindestens aus dem Ressourcentyp Text.

Für jede Publikation gilt:

- die Interoperabilität und Verlinkbarkeit der Ressourcentypen untereinander ist gesichert
- es kann nur innerhalb eines digitalen Objektes auf andere Ressourcentypen verlinkt werden
- Nicht genannte Script- und Auszeichnungsprachen, sowie andere Ressourcentypen sind nicht zu verwenden!

Ressourcentyp Text

Eigenschaften

Für die Überprüfbarkeit und Zitationsfähigkeit des Textes müssen folgende Eigenschaften erhalten bleiben:

- Der Text ist menschenlesbar
- Der Text ist strukturiert
- Der Lesefluss des Textes bleibt erhalten

Um eine automatisierte Textauswertung zu gewährleisten muss folgende Eigenschaft erhalten bleiben:

- Der Text ist maschinenlesbar

Da Publikationen in der Regel nicht nur aus Text bestehen, müssen andere Bestandteile der Publikation miteinander verknüpft werden können. Es muss daher folgende Eigenschaft erhalten bleiben:

- Der Text enthält auflösbare Links zu anderen, innerhalb der Publikation bestehenden Ressourcen

Verpflichtend, LZA-fähiges Dateiformat und Regeln

Zulässig sind die XML-basierten Dateiformate

- JATS 1.3 („Standardized Markup for Journal Articles: Journal Article Tag Suite JATS“, 2020)
- XHTML1.0 („XHTML 1.0 strict“, o. J.)
- XHTML1.1 („XHTML 1.1“, o. J.)
- Docbook (Walsh, 2016)
- TEI (*TEI P5: Guidelines for Electronic Text Encoding and Interchange.*, 2021)

> HTML5 („HTML, Living Standard“., 2022) kann aufgrund fehlender offline-Validatoren (siehe *WHATWG - HTML Conformance Checkers*, o. J.) zur Zeit nicht verwendet werden!

Nicht statthaft sind verschlüsselte Inhalte, zB. nach „Encrypted Media Extension“, o. J..

Als Zeichenkodierung ist UTF-8 ohne BOM zu verwenden.

Ressourcentyp Bild

Eigenschaften

Bild ist Ressource mit Belegcharakter und dienen der Abbildung und Veranschaulichung von Sachverhalten und sonst unzugängliche Materialien.

Um den Belegcharakter zu wahren, ist

- die optische Darstellung
- der kontextuelle Zusammenhang zum Text

zu erhalten.

Dient das Bild als Referenz für historische Vorlagen oder gar zur Reproduktion, so ist

- die farbgetreue Darstellung
- die geeignete Bild-Auflösung

zu erhalten.

> Im Fall, dass das Bild als Referenz für historische Vorlagen oder gar zur Reproduktion dient, so sollte die vom SLUBArchiv herausgegebene Handreichung retrodigitalisierte Monographien [SLUBArchiv_Handreichung_RetroMonographie.pdf] vorrangig angewandt werden.

Verpflichtend, LZA-fähiges Dateiformat und Regeln

Vektorgrafiken

Zulässig ist das XML-basierte Dateiformat

- SVG1.1 („Scalable Vector Graphics SVG 1.1“, 2011), Pronom ID: fmt/92

Pixelgrafiken

Zulässig sind die in Handreichung retrodigitalisierte Monographien [SLUBArchiv_Handreichung_RetroMonographie.pdf] beschriebenen Dateiformate, sowie:

- PNG (*Portable Network Graphics PNG*, 2011), Pronom ID: fmt/11, fmt/12, fmt/13 [1]

Ressourcentyp kodierte Noten

Eigenschaften

Kodierte Noten ist eine Resource mit Belegcharakter un dienen als Bezugsstelle für den Text zB. im Rahmen musikalischer Analysen.

Für die Überprüfbarkeit müssen folgende Eigenschaften erhalten bleiben:

- Die Noten sind menschenlesbar
- Die Noten sind vorlagengetreu

Um eine Nachnutzung und eine automatisierte Auswertung zu gewährleisten muss folgende Eigenschaft erhalten bleiben:

- Die Noten sind maschinenlesbar
- Die Noten sind reproduzierbar / kopierbar

Verpflichtend, LZA-fähiges Dateiformat und Regeln

Zulässig sind die XML-basierten Dateiformate:

- MusicXML („MusicXML 4.0“, 2021)
- MEI (*Music Encoding Initiative 4.0 Schemas*, 2022)

Ressourcentyp Geoinformationsdaten

Eigenschaften

Geoinformationsdaten dienen zur Georeferenzierung von Ressourcen, sowie zur Vermittlung und Veranschaulichung raumbezogener Informationen.

Für die Überprüfbarkeit müssen folgende Eigenschaften erhalten bleiben:

- Die Geoinformationsdaten sind darstellbar

Um eine Nachnutzung und eine automatisierte Auswertung zu gewährleisten muss folgende Eigenschaft erhalten bleiben:

- Die Geoinformationsdaten sind maschinenlesbar
- Die Geoinformationsdaten sind reproduzierbar / kopierbar

> Im Fall, dass Geoinformationen für Bildressourcen verwendet werden, so sollte die vom SLUBArchiv herausgegebene andreichung retrodigitalisierte Monographien mit Geoinformationen [SLUBArchiv_Handreichung_RetroGeoMonographie.pdf] vorrangig angewandt werden.

Verpflichtend, LZA-fähiges Dateiformat und Regeln

Zulässig sind die in Handreichung retrodigitalisierte Monographien mit Geoinformationen [SLUBArchiv_Handreichung_RetroGeoMonographie.pdf] beschriebenen Dateiformate, sowie:

- GML (*Geography Markup Language 3.3*, o. J.), Pronom ID: x-fmt/227
- KML (*Keyhole Markup Language 2.2*, o. J.), Pronom ID: fmt/244

Ressourcentyp Formeln

Eigenschaften

Formeln sind integraler Bestandteil naturwissenschaftlicher Publikationen. Für den Erhalt und Prüfung der Korrektheit müssen folgende Eigenschaften erhalten bleiben:

- Die Formeln sind menschenlesbar
- Die Formeln sind überprüfbar
- Die Formeln enthalten die beabsichtigten Symbole (Darstellung)

Um eine Nachnutzung und eine automatisierte Auswertung zu gewährleisten müssen folgende Eigenschaften erhalten bleiben:

- Die Formeln sind maschinenlesbar
- Die Formeln sind reproduzierbar / kopierbar

Verpflichtend, LZA-fähiges Dateiformat und Regeln

Zulässig sind die folgenden Dateiformate:

- MathML 1 („Mathematical Markup Language MathML™“, 1999)
- MathML 2 („Mathematical Markup Language MathML™“, 2003)
- MathML 3 („Mathematical Markup Language MathML™“, 2014)

Ressourcentyp Tabellen

Eigenschaften

Tabellen sind Ressourcen einer in Zeilen und Spalten angeordnete Zusammenstellung von Daten.

Um einzelne Daten zuordnen und referenzieren zu können müssen folgende Eigenschaften erhalten bleiben:

- jedes Datum einer Tabelle (Tabellenfeld) ist durch Angabe von Spalte und Zeile referenzierbar
- vorhandene Tabellenköpfe und Vorspalten, sowie Tabellenfelder bleiben unterscheidbar
- informationstragende Hervorhebungen von Tabellenfeldern sind optisch und strukturell unterscheidbar

Für den Erhalt und Prüfung der Korrektheit müssen folgende Eigenschaften erhalten bleiben:

- Tabellen sind menschenlesbar
- Tabellen sind überprüfbar
- Tabellenstrukturen sind optisch erkennbar

Um eine Nachnutzung und automatisierte Auswertung zu gewährleisten müssen folgende Eigenschaften erhalten bleiben:

- Tabellen sind maschinenlesbar
- Tabellen sind reproduzierbar / kopierbar

Nicht erfasst sind:

- Datenbanken
- dynamische Tabellen [2]

Diese werden bei Bedarf in einer separaten Handreichung spezifiziert werden.

Verpflichtend, LZA-fähiges Dateiformat und Regeln

Einfache Tabellen können durch geeignete und eigens zu dem Zweck zur Verfügung stehende Elemente der in Ressource Text vorgegebenen Dateiformate verwendet werden, zB. *<table>* in XHTML.

Unzulässig ist die Verwendung von CSV-Dateien, da dieses Format mehrere Unzulänglichkeiten aufweist (Bollig, 2019) und nicht langzeitarchivfähig (Sachse & Romeyke, 2020) ist.

Quellenverweise

XHTML 1.0 strict. *W3C Recommendation*. https://www.w3.org/TR/xhtml1/#strict

XHTML 1.1. *W3C Recommendation*. https://www.w3.org/TR/xhtml11/

WHATWG - HTML Conformance Checkers. Web Hypertext Application Technology Working Group. Abgerufen 8 Februar 2022, von https://whatwg.org/validator/

Encrypted Media Extension. *W3C Recommendation*. https://www.w3.org/TR/encrypted-media/

Geography Markup Language 3.3. Open Geospatial Consortium. Abgerufen 9 Februar 2022, von https://www.ogc.org/standards/gml

Keyhole Markup Language 2.2. Open Geospatial Consortium. Abgerufen 9 Februar 2022, von https://www.ogc.org/standards/kml

Mathematical Markup Language (MathML™). (1999). *W3C Recommendation*. https://www.w3.org/1999/07/REC-MathML-19990707/

Mathematical Markup Language (MathML™). (2003). *W3C Recommendation*. https://www.w3.org/TR/2003/REC-MathML2-20031021/

Scalable Vector Graphics (SVG) 1.1. (2011). *W3C Recommendation*. https://www.w3.org/TR/SVG11/

Portable Network Graphics (PNG). (2011). http://www.libpng.org/pub/png/spec/

Mathematical Markup Language (MathML™). (2014). *W3C Recommendation*. https://www.w3.org/TR/MathML3/

Standardized Markup for Journal Articles: Journal Article Tag Suite (JATS). (2020). *National Information Standards*, *Z39.96-2021*. http://www.niso.org/publications/z3996-2021-jats

TEI P5: Guidelines for Electronic Text Encoding and Interchange. (2021). http://www.tei-c.org/Guidelines/P5/

MusicXML 4.0. (2021). *W3C Recommendation*. https://www.w3.org/2021/06/musicxml40/

HTML, Living Standard. (2022). *WHATWG Standard*. https://html.spec.whatwg.org/

Music Encoding Initiative 4.0 Schemas. (2022). https://music-encoding.org/resources/schemas.html

Bollig, A. (2019). *Why You Don't Want to Use CSV Files*. https://haveagreatdata.com/posts/why-you-dont-want-to-use-csv-files/

Sachse, J., & Romeyke, A. (2020). *Why it is a stupid idea to consider CSV as a valid long-term preservation file format*. https://kulturreste.blogspot.com/2020/07/why-it-is-stupid-idea-to-consider-csv.html

Walsh, N. (Hrsg.). (2016). DocBook Version 5.1. *OASIS Standards*. http://docs.oasis-open.org/docbook/docbook/v5.1/os/docbook-v5.1-os.html

[1] Alle Versionen von 1.0 bis 1.2 sind möglich
[2] Tabellen mit integrierten Formeln, die bei Änderung von Werten eine Neuberechnung vornehmen, wie sie in Tabellenkalkulationsprogrammen vorgenommen werden