09 - Formatvalidierung mit JHOVE

Was ist JHOVE?

JHOVE wurde 2005 als gemeinsames Projektergebnis von JSTOR und der Harvard University Library veröffentlicht. Seit März 2014 wird JHOVE von der Open Preservation Foundation (OPF) gehostet, einer international tätigen Organisation, die für die Nachhaltigkeit von Technologien und Wissen für die Langzeitverfügbarkeit von digitalen Materialien Sorge trägt.

JHOVE ist open source und wird von der LZA-Community gepflegt und weiterentwickelt. Das geschieht durch:

- Arbeiten am Quellcode selber via GitHub
- Analyse der möglichen JHOVE-Fehlermeldungen bei invaliden Dateien
- JHOVE Hack Days (2016, 2017)

Was ist Validierung?

Eine valide Datei meint, dass die Datei der Dateispezifikation (Regelwerk, wie die Datei aufgebaut sein muss, was enthalten sein muss und was nicht enthalten sein darf) entspricht und somit standardkonform ist. Invalide meint, dass es die Spezifikation in mindestens einem Punkt verletzt.

Für die Langzeitarchivierung ist es wichtig, dass die archivierten Dateien valide sind, da die Viewer sich oftmals nach der Spezifikation richten und invalide Dateien daher möglicherweise später oder auch schon heute nicht mehr korrekt lesbar und verwendbar sind.

Wie kann JHOVE verwendet werden?

JHOVE bietet drei verschiedene Nutzungsmöglichkeiten:

- GUI (Graphical User Interface)
- via Kommandozeile (somit als Batch-Programm)
- Java library

Die GUI ist einfach zu installieren und zu bedienen (siehe Beginners Guide). Sofern man eine große Anzahl an Dateien zeitgleich analysiert, ist die Ausgabe allerdings umständlich zu lesen, da JHOVE viele Informationen über die Dateien liefert, die über die reine Information über die Validität und eventueller Fehlermeldungen hinausgehen.

Für das Einbinden in Workflows und Langzeitarchivierungsumgebungen ist das Batch-Programm gut nutzbar, hier können die mittels JHOVE gewonnenen Metadaten sowie Informationen über die Validität und eventuelle Fehlermeldungen mit der Datei gemeinsam als weiteres Metadatum gespeichert werden. Je nach Umgebung lässt sich so eine gezielte Suche nach invaliden Dateien realisieren.

Für das Einbinden von JHOVE in eigene Java-Programme gibt es eine gut verständliche Anleitung von Gary McGath, der während seiner Zeit an der Harvard University Library und auch danach JHOVE entwickelt und weiterentwickelt hat.

Wie verlässlich ist JHOVE?

Die 14 Module von JHOVE unterscheiden sich in der Qualität und Tiefe der Validitätsprüfung, was nicht zuletzt auch mit der unterschiedlichen Komplexität der Dateiformate zusammenhängt. Insbesondere das PDF-Format ist sehr komplex. Immerhin bietet JHOVE nach wie vor die umfangreichste Validierung des Standard-PDF-Formats bis zu Version 1.6.

Für PDF/A hingegen gibt es eine Vielzahl an Tools, JHOVE bietet nur scheinbar eine PDF/A-Validierung an (Artikel der PDF Association).

Für die Module TIFF, JPEG und WAVE gibt es jeweils OPF-Blog-Beiträge, die mittels Benchmarking untersuchen, ob JHOVE fehlerhafte Dateien entgehen (false negatives). Hierbei entdeckte JHOVE-Mängel werden an die OPF zurückgemeldet, damit diese bei der Weiterentwicklung berücksichtigt werden.

Das PDF-Modul kann nicht mittels Tool-Benchmarking analysiert werden, da es keine anderen Tools gibt, die in der Tiefe eine PDF-Validierung ermöglichen. In den Präsentationen „How valid is your validation", „Wahrheit oder Pflicht" und im iPRES Paper „A Test-Set for Well-Formedness Validation in JHOVE – The Good, the Bad and the Ugly" wird ein anderer Ansatz untersucht: Hier wurden 90 sehr simple PDF-Dateien erstellt, die jeweils gegen einen Punkt der PDF-Spezifikation verstoßen und im Anschluss getestet, ob JHOVE die Fehler bemerkt. Auch die hierbei festgestellten Mängel wurden umgehend an die JHOVE-Verantwortlichen zurückgemeldet, so dass diese in kommenden Versionen behoben werden können.

Fazit

Obwohl die Analysen zur Qualität von JHOVE zeigen, dass JHOVE fehlerhafte TIFF-, WAVE-, JPEG-, und vor allem PDF-Dateien nicht immer erkennt, hat sich herausgestellt, dass JHOVE häufige Fehler wie unvollständige Dateien auch bei komplexen Formaten durchaus bemerkt.

Außerdem meldet die Community entdeckte Fehler und Schwächen von JHOVE regelmäßig zurück, so dass diese behoben werden können. Für einige Dateiformate empfiehlt es sich, andere Validatoren hinzuzuziehen oder auch zu einem späteren Zeitpunkt die Validierungsprüfung mit JHOVE zu wiederholen, sobald eine neue, verbesserte Version von JHOVE zur Verfügung steht.

JHOVE ist seit mehr als zehn Jahren das meistgenutzte Tool zur Validierung von Dateien in der Langzeitarchivierung. Durch den Einsatz der Community ist zu erwarten, dass die Qualität sich in den folgenden Jahren noch verbessern wird.

Yvonne Tunnat
Langzeitarchivierung
ZBW – Deutsche Zentralbibliothek für Wirtschaftswissenschaften
Leibniz-Informationszentrum Wirtschaft, Düsternbrooker Weg 120, D-24105 Kiel
Tel: +49 431. 88 14-610
y.friese@zbw.eu

Yvonne Tunnat ist Mitglied der nestor AG Formaterkennung