

[Hintergrund: Im Rahmen des Projekts [Filling the Digital Preservation Gap](#) beschäftigt sich das Borthwick Institut für Archivalien an der Universität zu York viel mit der Bedeutung einer akkuraten Dateiformat-Identifikation und den damit verbundenen Schwierigkeiten bei der Datensuche. Jenny Mitcham, eine Digitalarchivarin der Universität, hat darüber in einem [Blog-Beitrag](#) berichtet, der für die nestor-Kurzartikelreihe leicht gekürzt und auf Deutsch übersetzt wurde.]

03 - Meine erste Dateiformat-Signatur

Ein Blog-Beitrag.

Wie in einem meiner letzten Einträge versprochen, habe ich mir vorgenommen als Teil des Projekts *Filling the Digital Preservation Gap* eine Dateiformatsignatur für PRONOM zu erstellen. Ich wollte herausfinden, ob das auch für weniger technikaffine Menschen schaffbar ist. Bisher hatte ich mich nämlich noch nicht an dieses Thema gewagt. Aber sollte nicht jeder Digitalarchivar wissen, wie eine Dateisignatur entsteht? Ross Spencer's Blog Post [Five Star File Format Signature Development](#) und ein Treffen mit dem PRONOM-Team im britischen Nationalarchiv motivierten mich schließlich dazu, mich selbst einmal an einer Signatur zu versuchen.

Aber wo fängt man da an?

Ich las zunächst „[How to research and develop signatures for file format identification](#)“, einen leicht verständlichen Ratgeber des britischen Nationalarchives, der empfiehlt, Probedateien zu sammeln, im Internet über das Format zu recherchieren und einen „Hex Editor“ zu verwenden, um Muster zu erkennen. Auch ohne über ein großes technisches Wissen zu verfügen, konnte ich den darin vorgeschlagenen Arbeitsschritten folgen. Ich lud mir den [HxD Hex Editor](#) herunter, mit dem man Dateien vergleichen und den Prozess der Mustererkennung teilweise automatisieren kann. Sobald ich ein geeignetes Muster gefunden hatte, arbeitete ich mit [PRONOMS Signature Development Utility](#) weiter, um daraus eine Signatur zu erstellen. Diese testete ich anschließend mit DROID. Dazu muss man in DROID die [Werkzeuge aufrufen](#), dann [Signature files installieren](#) auswählen und die momentane Signatur durch die neu erstellte ersetzen. (Hier ist Vorsicht geboten! Nach dem Testlauf muss unbedingt wieder die alte Signatur eingefügt werden, damit DROID weiterhin fehlerfrei läuft.) Die neue Signatur lässt man mittels DROID über die Probedateien laufen, um zu testen, ob alle Dateien von der erstellten Signatur korrekt identifiziert werden.

Ross Spencer weist darauf hin, dass PRONOM sich am besten zur Erstellung von Signaturen aus binären Formaten eignet, weshalb ich mich auf eben solche konzentrierte. Nach kurzer Untersuchung entschied ich mich schließlich für die .spa Dateien. Ich wusste zwar nichts über dieses Format, fand aber passende Dateien und recherchierte im Internet nach weiteren Informationen und Probedateien. Ich fand schnell heraus, dass .spa Dateien eines von vielen Formaten sind, die für optische Spectroscopy genutzt werden. Sie sind bekannt als Thermo Fisher's OMNIC file format, Thermo Scientific OMNIC oder Nicolet/Thermo OMNIC. Mit Hilfe des Hex Editors sah ich sofort, dass es jeweils ein übereinstimmendes Muster in der Startsequenz der Dateien gibt - den „Spectral Data File“, bestehend aus 53 70 65 63 74 72 61 6C 20 44 61 74 61 20 46 69 6C 65 in Hexadezimalen. Eigentlich dachte ich, das Muster sei länger, aber das PRONOM-Team empfahl mir, es zu kürzen.

Ich schaute mir außerdem das Ende jeder Datei an und dachte zunächst, dass auch hier ein Muster zu finden sei. Diese Hypothese stellte sich aber schnell als falsch heraus, als ich eine der Dateien, die ich online gefunden hatte, mit meinen eigenen verglich. Die Online-Datei wies dieselbe Startsequenz auf, endete aber anders. Diese Entdeckung macht deutlich, wie wichtig es ist, Probedateien aus verschiedenen Quellen zu benutzen. Ein Datensatz eines einzelnen Forschers könnte falsche Muster aufweisen, die lediglich aufgrund einer ähnlichen Arbeitsweise und nicht als Teil eines eigenen Formats entstanden sind.

Als ich schließlich alle 33 Dateien durchsucht und mich von meiner Annahme überzeugt hatte, nutzte ich das [Entwicklungstool für Signaturen](#), das vom Nationalarchiv zur Verfügung gestellt wird, um meine Signatur zu

erstellen. Das Tool war insgesamt relativ leicht zu bedienen, in einigen Teilschritten brauchte ich aber weitere Unterstützung, die ich problemlos vom PRONOM-Team erhielt. Ich hoffe, dass die Dokumentation des Tools in Zukunft noch weiter ausgebaut wird, damit Nutzer wissen, wie sie ihre Signatur am besten nennen sollen, wo sie eine vollständige Liste von Mimetypes finden und was das „Versions“-Feld enthalten sollte (hier geht es um die Version des Datenformats, nicht um die Version der erstellten Signatur).

Nach dem Speichern meiner Signaturdatei erhielt ich folgende XML Datei:

https://3.bp.blogspot.com/-qJ20lchr7WE/V7bjDpHq23I/AAAAAAAAAuk/6cqG6fEIJvEDo3HIJ_wJAAvpNKK0qetwgCLcB/s1600/spa_sig_xml.jpg

Ich bewunderte meine Handarbeit kurz und schickte sie anschließend für ein Feedback an das Nationalarchiv. Insgesamt hat mich die Arbeit, einschließlich Lektüre der Anleitungen, Download des Hex Editors und die Suche nach einem Format, von dem ich mir zutraute, es bearbeiten zu können, ungefähr einen Tag beschäftigt. Die nächste Signatur werde ich deutlich schneller erstellen können.

Meine größten Schwierigkeiten waren:

Ich habe eine Weile nach einem binären Format gesucht, mit dem ich arbeiten konnte. Viele Dateien die ich in Erwägung zog, waren im ASCII Format, was für die digitale Langzeitarchivierung zwar sinnvoll ist, für meinen Versuch aber zu komplex war.

Ich habe das bearbeitete Format nicht wirklich verstanden. Ich bin keine Chemikerin, hatte vorher noch nie vom .spa-Format gehört und kann „Spectroscopy“ kaum aussprechen. Auch meine Recherche hat mich eher verwirrt. Ich denke, wenn ich mehr über das Format gewusst hätte, wäre mir die Arbeit insgesamt leichter gefallen.

Die Metadaten, die wir von Wissenschaftlern erhalten, wenn sie ihre Daten in „*Research Data York*“ hinterlegen, sind limitiert. Die kurzen Beschreibungen zu den vorhandenen Datensätzen enthielten keine genauen Angaben zu den einzelnen Dateien oder zu den verwendeten Soft- und Hardwares.

Ich konnte die Dateiformat-Beschreibung nicht online finden, weshalb ich bei meinem nächsten Versuch direkt den Anbieter der Software um Hilfe bitten werde.

Die vorhandenen Unterlagen für die Erstellung und Überprüfung von Signaturen sollten erweitert werden.

Während des Arbeitsprozesses ergaben sich mir noch einige Fragen.

Das PRONOM-Team konnte mir zwar immer schnell Auskunft geben, die Informationen online zu finden würde aber sicherlich auch anderen Neulingen helfen.

Fazit

Abgesehen von meinen Schwierigkeiten, fand ich meinen Versuch erfreulich und nützlich. Der offensichtliche Gewinn, der aus meiner Arbeit resultiert ist, dass .spa Dateien nun von DROID und anderen PRONOM Identifizierungstools erkannt werden. Darüber hinaus konnte ich meinen Wissenshorizont erweitern und verstehe nun den Prozess der Signaturvergabe.

Ich rate anderen Digitalarchivaren, Repository Managern und Forschungsdatenmanagern unbedingt dazu diesen Versuch einmal selbst zu wagen.

Jenny Mitcham, Digital Archivist, Borthwick Institute for Archives

University of York, Heslington, York YO10 5D, Tel: 01904 321170,

Borthwick Institute website: <http://www.york.ac.uk/borthwick/>

Digital archiving blog: <http://digital-archiving.blogspot.co.uk/>

Twitter: @Jenny_Mitcham Skype: jenny_mitcham

Weitere Kurzartikel aus der Reihe „nestor Thema“ finden Sie auf www.langzeitarchivierung.de - der Webseite von **nestor – Kompetenznetzwerk Langzeitarchivierung**.