RISIKOMANAGEMENT FÜR DIE LANGZEIT-ARCHIVIERUNG VON NETZPUBLIKATIONEN

Wiebke Abel, Stefan Hein

Die Prozesse zur Verarbeitung und Archivierung von Netzpublikationen an der Deutschen Nationalbibliothek beinhalten seit Ende 2012 das Konzept der Ingest-Level, das seither vor allem in Hinblick auf die Sicherung der langfristigen Nutzbarkeit digitaler Objekte einen wesentlichen Bestandteil des Risiko-Managements bildet.

Motivation

Innerhalb der digitalen Langzeitarchivierung (LZA) gehört der Umgang mit Risiken zum täglichen Geschäft. In jedem Bereich der LZA stellt sich stets die Aufgabe, Risiken frühzeitig zu erkennen, deren mögliche Auswirkungen zu bewerten, Gegenmaßnahmen zu entwickeln und bei Bedarf auch einzuleiten. Dieses Risikomanagement muss in Organisationen institutionalisiert werden, damit eine stetige Überwachung potenzieller Risikoquellen und die Minimierung von Auswirkungen gewährleistet sind.

Level	DI	ID	BF	MD	V
0	X	O	O	O	O
1	X	X	O	O	O
2	X	X	X	O	O
3	X	X	X	X	O
4	X	X	X	X	X

Tabelle 1: Ingest-Level und Kriterien

Das Ingest-Level-Konzept

Risiken liegen häufig in den zu archivierenden digitalen Materialien selbst. So ist die **technische Qualität** der digitalen Materialien oftmals sowohl unbekannt als auch fehlerhaft, so dass die Erhaltung ihrer langfristigen Nutzbarkeit bereits nach heutigem Kenntnisstand fraglich ist.

Ein **Ingest-Level** ist das Ergebnis eines mehrstufig aufeinander aufbauenden **Prüfverfahrens** für Dateien und Dateiformate, welches teilweise in kooperativer Weise zwischen der Deutschen Nationalbibliothek und den abliefernden Partnern durchgeführt wird. Mit der Zuweisung eines Ingest-Levels (vgl. Tabelle 1) werden somit qualitative Aussagen über bestimmte technische Gegebenheiten eines digitalen Objekts getroffen.

Prüfkriterien:
- Dateiintegrität (DI)
- Identifizierbarkeit (ID)
- Beschränkungsfreiheit (BF)
- Generierung formatspezifischer technischer Metadaten (MD)
- Format-Validität (V)

Umsetzung

Die einzelnen Prüfschritte sind Bestandteil des Import-Verfahrens für Netzpublikationen. Die **Checksummenprüfung** ist hierbei eine der ersten Prüfroutinen zur Sicherung der **Dateiintegrität** (DI).

Unter dem Namen diagnose digital objects (didigo) werden die **Generierung technischer Metadaten** (MD) und das Ableiten eines Ingest-Levels zwischen den Werten 0 und 4 zusammengefasst. Didigo meldet hier bspw. für jede Datei einer Publikation, um welches **Dateiformat** (ID) es sich hierbei handelt und ob diese **Nutzungsbeschränkungen** (BF), wie z. B. einen Passwortschutz, aufweist.

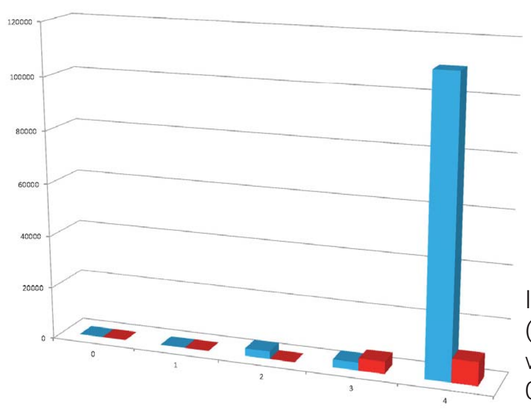
Für viele gängige Dateiformate wird zudem eine **Validitätsprüfung** (V) gegen die Dateiformatspezifikation durchgeführt.

Die **Generierung technischer Metadaten** erfolgt im Kern mithilfe des File Information Tool Set (FITS), das eine Reihe einschlägiger Metadatentools (JHOVE, DROID, ExifTool, usw.) zum Einsatz bringt[1].

Die **Analyse** umfasst vor allem die Berechnung des endgültigen Ingest-Levels. Dabei wird der FITS-Output anhand der Prüfkriterien in der oben genannten Reihenfolge untersucht.

Jede erfolgreich bestandene Prüfung erhöht den Ingest-Level um den Wert 1 bis hin zu seinem maximalen Wert 4.

Eine für jeden Ablieferer und jedes Dateiformat individuell definierbare **Format-Policy** gilt als minimale Erwartung an die Objektqualität im Sinne der genannten Prüfkriterien. Sie legt somit für jedes Dateiformat den Mindestwert des zu erreichenden Ingest-Levels fest und dient im letzten Schritt »**Policy-Vergleich**« als Entscheidungskriterium, ob ein Objekt angenommen wird oder nicht.



Ingest-Level-Verteilung (PDF ■, ePub ■) von 12/2012 bis 04/2013

Erfahrungen und Fazit

Die Inbetriebnahme erfolgte im Dezember 2012. Verglichen mit der Gesamtzahl der PDF-Dateien in Höhe von 116.138 bewegt sich die deutliche Mehrheit (110.222) im Bereich von Ingest-Level 4. Auch wenn aktuell noch verhältnismäßig wenig PDF-Objekte (3.014) ein Validitätsproblem aufweisen (Ingest-Level 3) und für 2.890 Objekte (Ingest-Level 2) keine technischen Metadaten generiert werden konnten, ist diese absolute und aller Wahrscheinlichkeit nach auch weiter steigende Größe nicht zu unterschätzen. In nur wenigen Jahren kann sich ein Bestand von mehreren Tausend problem behafteten Objekten ansammeln, die für Erhaltungsstrategien, wie z. B. Formatmigration, besonders in Betracht gezogen werden müssen.

Das Ingest-Level-Konzept liefert ein **praktikables Steuerungsinstrument**, welches Handlungsgrenzen und -regeln greifbar werden lässt. Es gibt den abliefernden Partnern zudem die Möglichkeit, die eigenen Anforderungen und **Erwartungen** an Objektqualität und Risikoanalyse zu formulieren, und schafft ein **Bewusstsein** unter den Verlagen, die »Qualität« ihrer Objekte im Auge zu behalten. Das Konzept dient sowohl als internes **Kommunikationsinstrument** als auch der Absprache von zu treffenden Service-Vereinbarungen zwischen LZA-Dienstnehmer und Dienstleister[2].

Links

[1] File Information Tool Set (FITS): http://code.google.com/p/fits/

[2] Langzeitarchivierung – Ein Handlungsleitfaden für Dienstleister und Dienstnehmer: http://dp4lib.langzeitarchivierung.de/index_downloads.php.de

Adickesallee 1
60322 Frankfurt am Main

Kontakt: Wiebke Abel, Tel. 069/1525-1727, w.abel@dnb.de
Stefan Hein, Tel. 069/1525-1722, s.hein@dnb.de

www.dnb.de