Henrike Berthold, Sabine Krug, Andreas Romeyke, Jörg Sachse

Das digitale Langzeitarchiv der SLUB Dresden

Möglichst hoher Automatisierungsgrad soll steigendem Datenvolumen gerecht werden

In einem zweijährigen Projekt hat sich die SLUB Dresden für die Langzeitarchivierung digitaler Dokumente fit gemacht. Anfang 2015 wurde für die Daten aus der hauseigenen Retrodigitalisierung der produktive Betrieb des SLUBArchivs aufgenommen. Zeit für einen Zwischenbericht.

Die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek (SLUB) ist eine der leistungsstärksten wissenschaftlichen Bibliotheken im deutschsprachigen Raum. Laut »Gesetz über die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SächsLBG)« hat die SLUB den Auftrag zur Sammlung »der wertvollen Bestände der sächsischen, nationalen und internationalen Literatur und Buchkultur« sowie zur »Archivierung von Literatur, Bild- und Tonträgern über Sachsen sowie der in Sachsen erscheinenden […] Publikationen«. Daraus ergibt sich auch die Verpflichtung zur digitalen Langzeitarchivierung (dLZA). Sie erlaubt es, digitale Dokumente, ob born-digital oder retrodigitalisiert, benutzbar und verfügbar zu halten.

Anders als man zunächst denken würde, ist ein Langzeitarchiv mehr als die Summe seiner Hard- und Software. Für den Erhalt der Korrektheit der Daten (Bitstream Preservation) und der Interpretier- und Nutzbarkeit (Content Preservation) werden Mitarbeiter benötigt, die die Daten und Systeme pflegen und entwickeln, Wissen über Datenformate, Medientypen und Datenerzeugungsprozesse aufbauen und ständig aktuell halten, mit Dienstleistern und Ablieferndenkommunizieren und die Nutzergruppen und deren Bedürfnisse kennen. Sie entwickeln Strategien, um angesichts der begrenzten Lebensdauer von Hardware, Software und Dateiformaten sowie des steten technologischen und personellen Wandels das Archivgut langfristig verfügbar zu halten.

Das im ISO-Standard 14721:2012 spezifizierte OAIS-Referenzmodell (engl. Open Archival Information System)[1] beschreibt die Konzepte und definiert die Fachbegriffe der digitalen Langzeitarchivierung. Es bildet den Rahmen heutiger Langzeitarchivsysteme. Ein Ablieferndender (engl. Producer) übergibt dem digitalen Langzeitarchiv ein Einlieferungspaket (engl. Submission Information Package, SIP). Nach der erfolgreichen Bearbeitung des SIPs (engl. Ingest) wird ein Archivpaket erzeugt (engl. Archival Information Package, AIP) und dem Permanentspeicher (engl. Archival Storage) übergeben. Dort ist es Gegenstand ständiger Pflege (engl. Data Management) und Bewahrungsmaßnahmen (engl. Preservation Planning). Nutzer, die vom Langzeitarchiv Daten anfragen (engl. Access), bekommen diese als Auslieferungspaket (engl. Dissemination Information Package, DIP).

Schwerpunkt

Themenschwerpunkte in BuB

Heft 05/2016
Bestandsaufbau

Heft 06/2016
Konflikte in Bibliotheken

Heft 07/2016
Digitalisierung

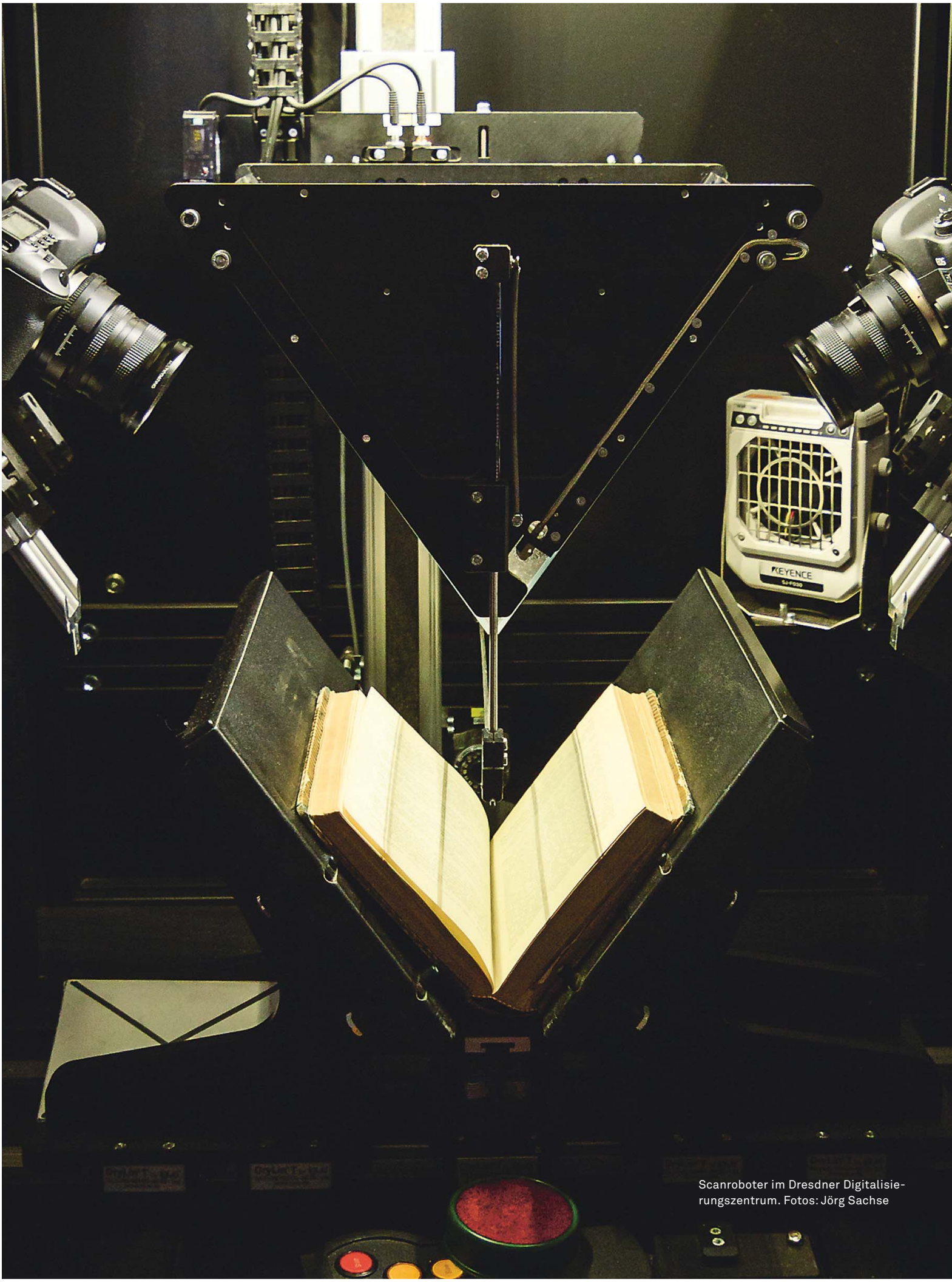
Heft 08-09/2016
Bibliotheksbau

Heft 10/2016
Frankfurter Buchmesse

Heft 11/2016
Mobile Bibliotheksarbeit

SLUBArchiv: Aufbau und Betrieb

Das SLUBArchiv basiert auf der Software Rosetta von ExLibris[2]. Die SLUB betreibt es in Kooperation mit dem ZIH (Zentrum für Informationsdienste und Hochleistungsrechnen) der TU Dresden. Das ZIH ist für den Betrieb der



Scanroboter im Dresdner Digitalisierungszentrum. Fotos: Jörg Sachse

Beeindruckend: die Magnetspeicherbänder in der Bandbibliothek.

IT-Infrastruktur verantwortlich, die aus mehreren Servern, Festplattenspeichersystemen und Bandbibliotheken besteht. Die Infrastruktur und die darauf gespeicherten Daten sind auf zwei räumlich getrennte Rechenzentren des ZIH verteilt.

Der Hauptvorteil in der Entwicklung automatisierter Lösungen besteht in der damit einhergehenden Definition von Schnittstellen zum Langzeitarchiv. In der Praxis zeigt sich, dass dazu bestehende Arbeitsabläufe hinterfragt werden müssen.

Das SLUBArchiv erhält die Datenpakete der produzierenden Systeme (beispielsweise nach der erfolgreichen Verarbeitung mit der Digitalisierungssoftware Kitodo – ehemals Goobi)[3] über ein Austauschverzeichnis. Diese asynchrone Übergabe entkoppelt die Produktionssysteme vom SLUBArchiv und erlaubt so unter anderem unabhängige Wartungsarbeiten. Jedes Datenpaket wird automatisch von der Submission Application in ein SIP umgeformt, wie es die Software Rosetta erwartet. Bei diesem Vorbereitungsschritt (engl. Pre-Ingest) werden die Daten, die zu einem digitalen Dokument gehören, aus dem Austauschverzeichnis gelesen, von der Submission Application geprüft, in die richtige Verzeichnisstruktur umgeformt, eine beschreibende Metadatendatei beigelegt und das so gebildete SIP schließlich an die Ingestverarbeitung von Rosetta übergeben.

Je nach Projektanforderungen kommt es vor, dass Archivpakete nur mit einem rudimentären Metadatensatz archiviert werden und die tiefere Erschließung später erledigt wird oder einzelne Master ausgetauscht werden müssen. In diesen Fällen führt die Submission Application ein sogenanntes AIP-Update durch, bei dem die geänderten (Meta-)Daten durch eine neue Version ersetzt werden.

Eine Automatisierung ist nur für korrekte Daten möglich. Fehler erfordern immer ein manuelles Eingreifen.

Die Submission Application enthält außerdem Funktionen zur Fehlerdiagnose, für statistische Auswertungen, zur Abfrage von DIPs und zur Durchführung der Exit-Strategie im Falle eines Systemwechsels oder einer Katastrophe.

Rosetta kann über Plugins (das heißt anwenderspezifischen Codes) erweitert werden. Die Ingestverarbeitung im SLUBArchiv wurde um eine Virenprüfung erweitert. Die Ablage im Permanentspeicher wurde so angepasst, dass alle Dateien, die zu einem Objekt gehören, in einem Verzeichnis abgelegt werden.

Für regelmäßige Integritätsprüfungen auf dem Permanentspeicher entwickelte die SLUB eine Software, die mehrstufig arbeitet und Rückschlüsse auf die Binärdatenqualität auf dem Speicher und dessen Alterungsprozess zulässt. Seit Juli 2015 ist der Retrodigitalisierungsworkflow des SLUBArchivs mit dem Data Seal of Approval zertifiziert, das auf einer Selbsteinschätzung[4] beruht. Damit ist ein erster wichtiger Schritt getan, um Nutzern und der Community gegenüber die Vertrauenswürdigkeit des Archivs nachzuweisen. Mittelfristig strebt die SLUB auch die Zertifizierung für weitere Workflows sowie höhere Zertifikatsstufen an. Seit Januar 2015 wurden etwa 48 000 IEs mit 2,2 Millionen Dateien und einem Datenvolumen von 31 Terabyte in das SLUBArchiv übernommen (Stand 31. März 2016).

Die SLUB ist Partner in nestor und dort zum Beispiel in der AG Media aktiv. Selbstentwickelte Software wird frei zur Verfügung gestellt.[5] Die Mitarbeit in der dLZA-Community stellt sicher, dass die Mitarbeiter des SLUBArchivs stets über die neuesten Entwicklungen informiert sind und die SLUB selbst Einfluss auf die Entwicklung von Technologien und Best Practices nehmen kann.

Erfahrungen mit der digitalen Langzeitarchivierung

Um dem stetig steigenden Datenaufkommen im SLUBArchiv gerecht zu werden, setzt die SLUB durchweg auf einen möglichst hohen Automatisierungsgrad. Dadurch sinkt der Aufwand für manuelle Arbeiten und die Fehlerrate durch

menschliche Fehler. Die Entwicklung geeigneter Automatisierungslösungen erfordert zwar den Einsatz teils erheblicher Entwicklerressourcen, weil die nötige Software für jeden Workflow der Einrichtung maßgeschneidert werden muss. Sie zahlt sich jedoch auf lange Sicht aus.

Voraussetzung dafür ist die Aufnahme des IST-Standes der bestehenden Workflows und deren Anpassung an die bestehenden automatisierten Workflows. So wird sichergestellt, dass zwischen den Arbeitsstationen exakt spezifizierte Schnittstellen etabliert sind und die Metadaten in standardisierten Formaten gespeichert und übergeben werden. Dazu kommen Verfahren, die in Fehlerfällen angewendet werden, und Maßnahmen zur Wahrung der Versionskonsistenz der IEs. Diese Anpassungen können je nach Produktionsverfahren langwierig und aufwendig in der Umsetzung sein.

Der Hauptvorteil in der Entwicklung automatisierter Lösungen besteht in der damit einhergehenden Definition von Schnittstellen zum Langzeitarchiv. In der Praxis zeigt sich, dass dazu bestehende Arbeitsabläufe hinterfragt werden müssen. So ist eine Forderung aus Sicht der digitalen Langzeitarchivierung die Kenntnis der Nutzerzielgruppen und signifikanten Eigenschaften, die unbedingt erhalten bleiben müssen – für die verschiedenen Abteilungen die Chance, ihre Sicht auf die Nutzer und Daten zu aktualisieren.

Lösbar sind die Probleme, wenn die Community gemeinsam Open Source-Software voranbringt und gegenüber den Herstellern kommerzieller Software mit gemeinsamer Stimme auftritt.

Eine Automatisierung ist nur für korrekte Daten möglich. Fehler erfordern immer ein manuelles Eingreifen. Um den Automatisierungsgrad auf ein akzeptables Niveau zu heben, ist es unerlässlich, die Datenqualität sehr früh im Produktionsworkflow zu prüfen und zu verbessern. Nur dadurch bleibt der Aufwand für manuelle Fehlerbearbeitung, Kommunikation, erneute Ingestversuche oder sogar vollständige Neuerzeugung der digitalen Daten überhaupt beherrschbar. Abliefernde können so auch selbst den Erfolg ihrer Reparaturoperationen einschätzen und helfen, die Datenqualität nachhaltig zu steigern.

Dazu wurde unter anderem eine Werkzeugsuite für das TIFF-Format entwickelt und als Open Source veröffentlicht[6], mit der sich Bilddateien gegen die TIFF-Spezifikation validieren und Fehler in den Dateien beheben lassen.

Problematisch ist oft die für die Produktion digitaler Dokumente eingesetzte Hard- und Software. Meist wurden bei der Anschaffung und Implementierung der Systeme die Erfordernisse der Langzeitarchivierung nicht mitgedacht oder die Geräte gar auf ihre Eignung dahingehend getestet. Softwarelösungen für die Erzeugung oder Bearbeitung digitaler Daten (zum Beispiel Scannersoftware) setzen die Spezifikationen der erzeugten Datenformate teilweise fehlerhaft um oder lassen sich nur unzureichend konfigurieren. Zudem fehlen oft geeignete Datenexport- und -importschnittstellen, die

internationalen Standards folgen und valide Daten ausliefern oder entgegennehmen.

Lösbar sind die Probleme, wenn die Community gemeinsam Open Source-Software voranbringt und gegenüber den Herstellern kommerzieller Software mit gemeinsamer Stimme auftritt. Letzteres geschieht teilweise schon, zum Beispiel im deutschen Kompetenznetzwerk für Langzeitarchivierung nestor oder im Rahmen der Internationalen Rosetta-Anwendergruppe. Ein Beispiel für die Open Source-Entwicklung ist das PREFORMA-Projekt[7], in dem aktuell die Software veraPDF[8] zur Validierung von PDF/A entwickelt wird.

Ausblick

Im Rahmen des »Landesdigitalisierungsprogrammes Wissenschaft und Kultur« (LDP) in Sachsen wird seit 2015 die Digitalisierung und die digitale Langzeitarchivierung wissenschaftlich und kulturell wertvoller Literatur in sächsischen Bibliotheken gefördert. Die digitale Langzeitarchivierung erfolgt im SLUBArchiv. Parallel dazu baut die SLUB ein Dienstleistungsangebot für die digitale Langzeitarchivierung auf. Dafür werden derzeit alle nötigen Vorbereitungen getroffen. Es werden Ablieferungsschnittstellen spezifiziert, Softwareanpassungen vorgenommen und die Automatisierung vorangetrieben.

Mit dem Ziel eines verteilten Leistungsnetzwerks soll zwischen den Einrichtungen, die ein digitales Langzeitarchiv betreiben, eine Kooperation und eine Schwerpunktsetzung vereinbart werden.

Auch in der SLUB werden nach Abschluss der Vorbereitungen weitere Produktionsworkflows automatisiert an das SLUBArchiv angebunden. Das betrifft E-Publikationen sowie audiovisuelle Inhalte.

Die SLUB baut die akzeptierten Datenformate bedarfsgerecht aus. Mit dem Ziel eines verteilten Leistungsnetzwerks soll zwischen den Einrichtungen, die ein digitales Langzeitarchiv betreiben, eine Kooperation und eine Schwerpunktsetzung vereinbart werden. Die SLUB und die TIB befinden sich aktuell in der Abstimmung. Bedarfe entstehen in der SLUB selbst, aus dem Landesdigitalisierungsprogramm und durch Dienstnehmer. Die Einführung eines neuen akzeptierten Datenformats ist ein Projekt, in dem Formaterkennung, Formatvalidierung und Metadatenanreicherung sichergestellt werden müssen.

Wenn die Server und Festplattenspeicher in den kommenden Jahren ihr geplantes Lebensende erreichen, wird in Zusammenarbeit mit dem ZIH der Austausch geplant werden. Die Architektur wird gemäß den aktuellen und absehbaren technischen Entwicklungen und Anforderungen neu geplant und ein leistungsfähiges Nachfolgesystem beschafft. Damit ist das SLUBArchiv auch für die Herausforderungen der nächsten Jahre gut gerüstet.

1 http://public.ccsds.org/publications/archive/650x0m2.pdf

2 http://www.exlibrisgroup.com/de/category/Rosetta

3 http://www.kitodo.org

4 http://blog.slub-dresden.de/beitrag/2015/07/04/zertifizierter-langzeitarchivierungsservice

5 http://www.langzeitarchivierung.de

6 https://github.com/SLUB-digitalpreservation

7 http://www.preforma-project.eu

8 http://verapdf.org

Die Autoren

Jörg Sachse hat 2012 seine Ausbildung zum Fachinformatiker/Systemintegration an der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB) abgeschlossen und ist seitdem am Aufbau und Betrieb des SLUBArchivs beteiligt. Dort beschäftigt er sich vorrangig mit technischer Administration, Formatvalidierung und Fehleranalyse. Außerdem betreut er die virtuelle Infrastruktur und das Datennetz der SLUB.

Henrike Berthold ist promovierte Informatikerin und leitet die IT-Abteilung der SLUB. Seit 2011 beschäftigt sie sich mit der digitalen Langzeitarchivierung und hat diesen Kompetenzbereich an der SLUB mit aufgebaut.

Andreas Romeyke ist als Diplom-Informatiker (FH) seit 2012 an der SLUB Dresden im Team digitale Langzeitarchivierung tätig und beschäftigt sich dort vorrangig mit der Analyse von Datenformaten und Automatisierung von Prozessworkflows. Um auch seine bibliothekarischen Kenntnisse zu vertiefen, studiert er berufsbegleitend im dritten Fachsemester Master für Bibliotheks- und Informationswissenschaften in Leipzig.

Sabine Krug ist Diplom-Informatikerin und seit 1993 an der Sächsischen Landesbibliothek, später SLUB Dresden, in der Abteilung IT tätig. Ihr fachlicher Schwerpunkt liegt auf Storagesystemen, Datensicherheit und Langzeitverfügbarkeit.