



IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften

Herausgeber:
IANUS

Version 1.0
15. Dezember 2016

Koordination



Förderung



Autoren: siehe Autorenverzeichnis

Titel: IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften

Sprache: Deutsch

Zitierhinweis:

IT-Empfehlungen V1.0 IANUS – Forschungsdatenzentrum Archäologie & Altertumswissenschaften (2016) [Version 1.0] Hrsg. von IANUS. (doi:10.13149/000.y47clt-t)

Kontakt:

ianus-fdz@dainst.de

www.ianus-fdz.de

Lizenz:



Dieses Werk bzw. Inhalt ist lizenziert unter einer Creative Commons Namensnennung – Weitergabe unter gleichen Bedingungen 4.0 International Lizenz.

<http://creativecommons.org/licenses/by-sa/4.0/>

Versionshistorie

Version	Autor	Datum	Beschreibung
0.9	IANUS	18.02.2014	Erste veröffentlichte Fassung.
0.9	IANUS	11.03.2014	Änderung der Lizenzbedingungen.
0.95	IANUS	05.02.2015	Kapitel „3D und Virtual Reality“ und „Textdokumente“ neu. Anpassungen im Layout und kleinere Korrekturen.
0.96	IANUS	27.05.2015	Kapitel „Datenmanagement“ neu, Ergänzung Versionskontrolle und Korrekturen.
0.97	IANUS	09.12.2015	Kapitel „Tabellen“ neu, Ergänzungen im Glossar und Korrekturen.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Version	Autor	Datum	Beschreibung
0.97.5	IANUS	21.07.2016	Kapitel „Video“, „Audio“ und „RTI“ neu, Ergänzungen im Glossar, Anpassungen im Kapitel „Archivierung bei IANUS“ und Korrekturen.
0.97.5	IANUS	13.09.2016	Einarbeitung Reviews für „Video“, „Audio“ und „RTI“, Ergänzungen im Glossar und Korrekturen.
0.98	IANUS	22.11.2016	Überarbeitung „Datenmanagement“ und „3D und Virtual Reality“, Ergänzung Bild in „RTI“.
0.99	D. Hagmann	09.12.2016	Kapitel „Webseiten“ neu.
1.0	IANUS	15.12.2016	Kapitel „Vektorgrafiken - Übersicht“, Kapitel „Datenbanken - Übersicht, Vertiefung“ neu. Korrekturen.

Inhaltsverzeichnis

1 Einleitung	5
1.1 Benutzung der IT-Empfehlungen	8
1.2 Der Lebenszyklus von Forschungsdaten	11
2 Projektphasen	14
2.1 Datenmanagement	16
2.1.1 Übersicht der Aufgaben in den Projektphasen	16
2.1.2 Datenmanagementplan	17
2.1.3 Weiterführende Informationen	26
2.2 Dokumentation	28
2.2.1 Dokumentation und Metadaten	28
2.2.2 Arten von Metadaten	30
2.2.3 Kontrollierte Vokabulare	31
2.2.4 Struktur von Metadaten	32
2.2.5 Speicherung von Metadaten	34
2.2.6 Metadaten in der Anwendung	36
2.2.7 Weiterführende Informationen	41
2.3 Dateiverwaltung	43
2.3.1 Dateiablage	43
2.3.2 Dateibenennung	45
2.3.3 Versionskontrolle	46
2.4 Dateispeicherung und -sicherung	48
2.4.1 Kurzfristige Speicherung	49
2.4.2 Mittelfristige Sicherung	49
3 Dateiformate	52
3.1 PDF-Dokumente	53
3.2 Textdokumente	65
3.3 Bilder – Rastergrafiken	78
3.4 Bilder – Vektorgrafiken und CAD-Daten	90
3.5 Tabellen	97
3.6 Datenbanken	109
3.7 GIS	126
3.8 Video	132
3.9 Audio	149
3.10 3D und Virtual Reality	159
3.11 Webseiten	173
4 Forschungsmethoden	194
4.1 Geodäsie	194
4.2 Georeferenzierung	199
4.3 Oberflächen- (DOM) und Geländemodellierung (DGM)	200
4.4 Rekonstruktion	205
4.5 Reflectance Transformation Imaging (RTI)	210
4.6 Satellitenmessungen	218

5 Archivierung bei IANUS	222
A Autorenverzeichnis	231
B Glossar	232

1 | Einleitung

F. Schäfer, M. Trognytz

Für die Durchführung archäologischer und altertumswissenschaftlicher Forschungsprojekte ist die Anwendung von Informationstechnologien verschiedener Art immer geläufiger und selbstverständlicher. Dennoch fehlt es bislang an etablierten Strukturen, Standards und Verfahrensweisen, die allen Projekten den erprobten und effizienten Einsatz von IT ermöglichen, Vergleichbarkeit gewonnener Datenbestände erlauben, übergreifend gleichwertige Qualität digitalen Materials sicherstellen und die langfristige Les- und Nutzbarkeit von Forschungsergebnissen in digitaler Form garantieren. Durch die Einhaltung von Standards können nicht nur redundante Entwicklungsarbeiten mit zusätzlichen Finanzierungs-, Zeit- und Arbeitsaufwänden vermieden werden, sondern auch der Austausch von wissenschaftlichen Inhalten vereinfacht werden.

In zunehmendem Maße fordern auch wissenschaftspolitische Gremien (DFG¹, Wissenschaftsrat², Schwerpunktinitiative "Digitale Informationen"³) die Nachhaltigkeit und die Vernetzbarkeit digitaler wissenschaftlicher Informationen ein, was eine Auseinandersetzung mit Dateiformaten, Datenmanagement, Normen, Mindeststandards, semantischen Referenzmodellen, Dokumentations- und Austauschverfahren erforderlich macht – sowohl bei kleineren Projekten eines Einzelnen als auch bei größeren disziplinen- und institutionenübergreifenden Forschungen. Einheitliche und in den jeweiligen Fachcommunities abgestimmte Vorgehen eröffnen für die aktuelle und zukünftige Forschung neue Möglichkeiten, um auf einer größeren, qualitätvoller und homogeneren Datengrundlage komplexe Fragestellungen besser als bisher beantworten zu können.

Die IT-Empfehlungen sind als Hilfestellungen für alle Akteure in den Altertumswissenschaften konzipiert, um die langfristige Erhaltung und Benutzbarkeit des wissenschaftlichen Quellenmaterials zu gewährleisten sowie die Interoperabilität von Daten unterschiedlichen Ursprungs zu verbessern. Durch die Veröffentlichung von vollständigen digitalen Datensammlungen kann der Prozess der traditionellen Publikation von Forschungsergebnisse befördert und ergänzt werden, bspw. durch die zusätzliche Online-Bereitstellung von nicht in Druckform publizierten Katalogen, Analyseserien, Rekonstruktionen o. ä. m. (sog. Datenpublikationen).

Die hier zusammengetragenen IT-Empfehlungen befassen sich vor allem mit der Frage, wie digitale Daten besser ausgetauscht, für die Zukunft aufbewahrt und dauerhaft verstanden werden können. Sie fokussieren also auf technische Formate und inhaltliche Beschreibung von Daten und Methoden, die für die langfristige Archivierung, Bereitstellung und Nachnutzung von Daten von zentraler Bedeutung sind. Spezifische Vorgaben zu fachlichen Methoden, Fragestellungen, Workflows oder Anwendungen sind dagegen nicht Gegenstand der Empfehlungen.

Weitere Informationen zur Relevanz und Kuratierung von digitalen Forschungsdaten in den Altertumswissenschaften sind in den "Guides to Good Prac-

¹http://www.dfg.de/foerderung/rechtliche_rahmenbedingungen/gwp/index.html

²<http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf>

³<http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaezze>

tice"⁴ des Archaeology Data Service zu finden, wo ebenfalls auf weiterführende Publikationen verwiesen wird.

Umsetzung

In der Regel wird es in bestimmten Phasen eines Projektes oder aufgrund institutioneller Rahmenbedingungen notwendig sein, von diesen IT-Empfehlungen abzuweichen, vor allem während der Generierung und der Auswertung von Daten. Da im Folgenden nur minimale Standards beschrieben werden, kann es für deren praktische Anwendung hilfreich sein, darüber hinausgehende Spezifizierungen festzulegen, z. B. Wege zur konkreten Umsetzung von Vorgaben zu definieren, Zuständigkeiten beim Forschungsdatenmanagement zu verteilen oder zusätzliche forschungs- bzw. organisationsspezifische Bedingungen zu diskutieren. Diese Spezifizierungen sowie Abweichungen von Standards sollten immer sorgfältig abgewogen, hinsichtlich der digitalen Nachhaltigkeit und Austauschbarkeit hinterfragt und ausreichend dokumentiert werden. Zudem gilt es jeweils zu prüfen, ob dadurch zusätzliche Arbeitsschritte erforderlich werden.

Da Forschungsdaten in hohem Maße von einem zyklischen Charakter geprägt werden, wie beim Lebenszyklus von Daten auf Seite 11 beschrieben, sind die Archivierung und Nachnutzung von Daten keine Aspekte, die erst am Ende eines Projektes zu behandeln sind, sondern bereits vor bzw. bei der Erhebung der ersten digitalen Dateien berücksichtigt werden müssen.

Wird z. B. bei Fotografien nicht von Anfang an festgehalten, wer an welchem Ort und zu welchem Zeitpunkt ein Bild aufgenommen hat, oder bei Messungen gerätespezifische Eigenschaften, die die Qualität von Daten bestimmen, nicht dokumentiert, so können diese wichtigen Angaben nach dem Projektende fehlen oder mangels Personal, Finanzmittel oder Möglichkeiten nachträglich nicht mehr erhoben werden. Forschungsdaten sind dadurch potentiell in ihrer Qualität für eine wissenschaftliche Nachnutzung durch Dritte stark eingeschränkt. Ähnlich kann es sich bei der Verwendung proprietärer Formate verhalten, bei denen einerseits eine Migration auf neuere Programmversionen nur mit inhaltlichen und/oder funktionalen Verlusten einhergeht und andererseits kein Export der Inhalte in anwendungsunabhängige Formate unterstützt wird. In diesem Fall ist ein realer Datenverlust vorhersehbar.

Zur Verwirklichung der hier aufgeführten IT-Empfehlungen sollten Einrichtungen entsprechende Strukturen aufbauen, die Forschungsprojekte bei der Formulierung von IT-Konzepten beraten und diese begutachten, die bei der Implementierung eines Forschungsdatenmanagements unterstützen, die die Qualitätskontrolle von Daten übernehmen sowie deren langfristige Verfügbarkeit und Archivierung gewährleisten.

Für laufende Projekte und „Altdaten“ gilt Bestandsschutz. Da aber auch diesen Daten ein hoher wissenschaftlicher Wert zukommt, den es in der Regel zu erhalten gilt, ist auch hier eine Anpassung bzw. Migration auf aktuelle Standards und Empfehlungen in jedem Fall erstrebenswert und es sollte immer geprüft werden, ob ein solcher Schritt möglich ist.

Weiterhin ist der Einsatz von Informationstechnologien und der Umgang mit digitalen Daten an die Berücksichtigung des Urheberrechts und des Datenschutzes gebunden.

⁴<http://guides.archaeologydataservice.ac.uk/g2gp/GuideAim>

Wann sollten Sie diese Empfehlungen berücksichtigen?

- In der Antrags- und Planungsphase eines neuen Projektes, um frühzeitig technische, personelle und finanzielle Anforderungen und Lösungen berücksichtigen zu können
- Während der Durchführung von Forschungsprojekten als Referenz für spezifische Fragen und Probleme
- Vor der Übergabe von Daten an ein institutionelles Archiv oder Datenrepositorium

Autoren und Quellen der Empfehlungen

Die hier präsentierten Empfehlungen und Hintergrundinformationen beruhen im Wesentlichen auf drei Quellen.

1. **Texte von Fachspezialisten**, die aus unterschiedlichen Institutionen, Disziplinen und Projekten stammen und Vorschläge für die jeweiligen Fachcommunities formuliert haben. Entsprechend werden am Ende der einzelnen Kapitel die maßgeblichen Autoren genannt. Dieser Prozess wurde seit 2007 von der Deutschen Forschungsgemeinschaft angestoßen und unterstützt (u. a. durch die Finanzierung vorbereitender Arbeitstreffen) sowie durch das Deutsche Archäologische Institut in seiner Gesamtheit koordiniert. IANUS hat für die Ergebnisse die technische Plattform bereitgestellt und die redaktionelle Qualitätssicherung übernommen. Die weitere Pflege und Aktualisierung der IT-Empfehlungen wird durch entsprechende Mitarbeiter übernommen sowie ihre Umsetzung durch entsprechende Angebote unterstützt.

2. Der vom DAI entwickelte "Leitfaden zur Anwendung von Informationstechniken der archäologischen Forschung"⁵, der sich im Downloadbereich befindet, wird vollständig durch die von IANUS betreuten IT-Empfehlungen abgelöst. Inhalte dieses IT-Leitfadens sind volumnäßig in die hier formulierten Mindeststandards eingeflossen und wo notwendig aktualisiert bzw. ergänzt worden.

3. **Standards und Veröffentlichungen anderer Einrichtungen**, um die hier formulierten Mindeststandards in den nationalen und internationalen Kontext zu stellen und länderübergreifende Übereinstimmung in Kernfragen zu erreichen. Hervorzuheben sind in diesem Kontext besonders die "Guides to Good Practice"⁶ des Archaeology Data Service in York, England und der Digital Antiquity in Tempe, Arizona, USA, die "How-to Guides"⁷ des Digital Curation Center in London, England und der noch nicht finalisierte "Ratgeber zur Archivierung Digitaler Daten"⁸ des Verbandes der Landesarchäologen in Deutschland.

⁵<http://www.ianus-fdz.de/it-empfehlungen/downloads#dai-leitfaden>

⁶<http://guides.archaeologydataservice.ac.uk/g2gp/Main>

⁷<http://www.dcc.ac.uk/resources/how-guides>

⁸http://www.landesarchaeologen.de/fileadmin/Dokumente/Dokumente_Kommissionen/Dokumente_Archaeologie-Informationssysteme/Dokumente_AIS_Archivierung/Ratgeber-Archivierung_Vorab-V0.07.pdf

1.1 Benutzung der IT-Empfehlungen

F. Schäfer, M. Trognitz

Die IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften liegen sowohl in statischer Form als verschiedene PDF-Dateien vor, als auch in dynamischer Form auf <http://www.ianus-fdz.de/it-empfehlungen>. Beide Formen sind inhaltlich identisch und können unabhängig voneinander verwendet werden kann. Die Webseite bietet neben vielfältigen Verlinkungen und Downloads weitere Funktionalitäten, zu denen auch eine Suchfunktion und ein integriertes Glossar gehören. Die PDF-Dokumente eignen sich dagegen z. B. als Hilfsmittel im Feld, das individuell ausgedruckt werden.

Die IT-Empfehlungen umfassen neben diesem Einleitungskapitel die drei Kapitel **Projektphasen**, **Datenformate** und **Forschungsmethoden**.

Aufbau der einzelnen Kapitel

In dem Kapitel **Projektphasen** sind allgemeine Hinweise hinterlegt, die sich mit den konkreten Arbeitweisen und Phasen in Forschungsprojekten befassen. Es enthält Hinweise darauf, was bei der Planung und dem Beginn eines Projektes zu beachten ist, welche Aspekte während der Projektdurchführung, also vor allem bei der Erhebung und Verwaltung von Forschungsdaten, wichtig sind und welche Fragen beim Abschluss eines Forschungsvorhabens beantwortet werden sollten.

Das Kapitel **Dateiformate** beinhaltet einerseits konkrete Hinweise zu den technischen Dateiformaten, die nach heutigem Wissen für die Langzeitarchivierung geeignet sind. Andererseits werden die minimalen Anforderungen an die Dokumentation von Forschungsdaten genannt, damit diese nicht nur heute, sondern auch in Zukunft von Dritten verstanden und nachgenutzt werden können. Diese Abschnitte sind weiter unterteilt in: **Übersicht**, **Vertiefung**, **Praxis** und **Quellen**.

Der Unterabschnitt **Übersicht** bietet eine Zusammenfassung über die jeweils geeigneten technischen Formate, die erforderlichen Metadaten und eventuelle Fehlerquellen. Er richtet sich sowohl an den erfahrenen Anwender als auch an Nutzer mit konkreten, schnell zu beantwortenden Fragen. Die Informationen aus allen Übersichts-Abschnitten werden in einer eigenen PDF-Datei als Kurzfassung der IT-Empfehlung zusammengefasst und zum Download bereit gestellt. Zu beachten ist, dass die angegebenen Metadaten nur ein Minimum darstellen und gegebenenfalls abhängig vom Projektkontext und der verwendeten Methode ergänzt werden müssen.

Der interessierte Nutzer findet im Unterabschnitt **Vertiefung** weiterführende Hintergrundinformationen über die in der jeweiligen Übersicht angesprochenen Teilelemente und im Abschnitt **Praxis** werden praktische Anleitungen, Vorlagen, Formulare, Best-Practice-Beispiele und andere nützliche Werkzeuge gesammelt, die bei der praktischen Umsetzung der beschriebenen Mindeststandards helfen sollen. Ergänzende Informationen werden in dem Unterabschnitt **Quellen** gelistet.

Das Kapitel **Forschungsmethoden** befasst sich mit den verschiedenen in den Altertumswissenschaften angewandten Methoden und Disziplinen. Auch wenn langfristig eine möglichst hohe Vollständigkeit angestrebt wird, ist es nicht Ziel und Zweck der IT-Empfehlungen, ein umfassendes Handbuch über

die Geschichte, Fragestellungen, Anwendung oder Vor-/Nachteile aller relevanten Methoden und Arbeitsweisen zu sein. Stattdessen stehen Dokumentation und Metadaten im Vordergrund, die für das Verständnis, die Nachvollziehbarkeit und Nachnutzbarkeit der generierten digitalen Daten durch Dritte zentral sind. Wie bei den Dateiformaten sind auch hier die einzelnen Abschnitte in **Übersicht**, **Vertiefung**, **Praxis** und **Quellen** untergliedert. Sofern bei einzelnen Disziplinen spezifische Dateiformate eine besondere Relevanz besitzen, sind diese entweder hier beschrieben oder es wird mittels Links auf die entsprechenden Abschnitte im Kapitel „Dateiformate“ verwiesen.

Häufig genutzte oder zentrale Fachbegriffe werden in einem **Glossar** erklärt und sind auf der Webseite unterstrichen. Wenn der Mauszeiger auf einem markierten Wort liegt, wird die Erklärung mittels Tooltip eingeblendet.

Im **Download-Bereich** können sowohl die verschiedenen längeren und kürzeren Varianten der aktuellen IT-Empfehlungen als auch älteren Versionen heruntergeladen werden. Außerdem sind hier zentrale Dokumente für die Archivierung von Forschungsdaten in IANUS enthalten.

In der **rechten Spalte** findet man Materialien zum Download, die auf den gerade ausgewählten Abschnitt abgestimmt sind.

Zitation der IT-Empfehlungen

Die gesamten IT-Empfehlungen können als eine einzige Publikation betrachtet werden, deren unterschiedliche Kapitel von unterschiedlichen Autoren verfasst wurden. Auf die einzelnen Kapitel kann unter Nennung der Autoren entweder auf die Online-Version oder auf die Vollversion als PDF-Dokument verwiesen werden:

- P. Grunwald - M. Trognitz (2015), „Rastergrafiken“, in: IT-Empfehlungen. IANUS - Forschungsdatenzentrum Archäologie & Altertumswissenschaften, www.ianus-fdz.de/it-empfehlungen/rastergrafiken (27.05.2015)
- P. Grunwald - M. Trognitz, „Rastergrafiken“, in: IT-Empfehlungen V0.96. IANUS - Forschungsdatenzentrum Archäologie & Altertumswissenschaften (2015) 77 ff.

Für die PDF-Version muss zusätzlich die Versionsnummer angegeben werden, da diese von Zeit zu Zeit erzeugt und als Download zur Verfügung gestellt werden. Die Online-Version der IT-Empfehlungen ist immer die aktuellste Fassung und hat für die dauerhafte Referenzierung einen DOI:

- IANUS (Hrsg.) IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften. doi: 10.13149/000.111000-a

Ausgewählte frühere Versionen der IT-Empfehlungen sollen ebenfalls einen DOI erhalten.

Mitarbeit und Diskussion

Jeder Abschnitt in den Kapiteln **Dateiformate** und **Forschungsmethoden** enthält einen Unterabschnitt **Diskussion**, der angemeldeten Nutzern die Möglichkeit bietet, Hinweise und Kommentare, z. B. zu nützlichen Softwaretools oder neuen Standards, zu hinterlegen. Ebenso können hier noch fehlende oder

unzureichend formulierte Aspekte benannt werden oder gemachte Empfehlungen offen diskutiert werden. Daneben kann natürlich auch eine E-Mail an it-empfehlungen.ianus-fdz@dainst.de geschickt werden, um individuelle Anmerkungen zu formulieren oder die Bereitschaft zur Mitarbeit bei neuen oder zu aktualisierenden Kapiteln zu signalisieren. Alle Rückmeldungen werden regelmäßig von Mitarbeitern in IANUS geprüft und gegebenenfalls in die IT-Empfehlungen eingearbeitet.

1.2 Der Lebenszyklus von Forschungsdaten

M. Trognitz

Die Anforderungen an das Management von Forschungsdaten gehen weit über die Langzeitarchivierung, einer Kernaufgabe von IANUS, hinaus. Daher soll im Folgenden kurz auf das Konzept des Datenlebenszyklus eingegangen werden.

Forschungsdaten haben im Allgemeinen eine deutlich längere Lebensdauer als die Projekte, in denen sie entstanden sind. Auch nach Ablauf von Förderungsphasen kann es sein, dass Forscher weiterhin an den Daten arbeiten oder sie werden in Folgeprojekten weiter verarbeitet und ergänzt. Die Daten können auch von anderen Wissenschaftlern unter Berücksichtigung neuer Forschungsfragen wiederverwendet werden. Somit können aus bereits gesammelten Forschungsdaten neue Daten entstehen. Dieser Kreislauf wird mit dem sogenannten Datenlebenszyklus beschrieben.

In der gegenwärtigen Praxis ist der Datenlebenszyklus oft unterbrochen, da nur einige wenige Daten ihren Weg in eine Publikation finden. Die übrigen Daten geraten auf den lokalen Speichermedien der Forscher in Vergessenheit oder werden komplett gelöscht und stehen somit einer Nachnutzung nicht mehr zur Verfügung. Diese Tatsache gilt es für die Zukunft zu verbessern und im Idealfall vollständig zu vermeiden.

Der hier vorgestellte Datenlebenszyklus beruht auf dem des „UK Data Archive“⁹. Er beschreibt sehr vereinfacht und idealtypisch die Abfolge der einzelnen Phasen im Zyklus aus einem eher praktischen Umgang mit Forschungsdaten. Es gibt auch andere komplexere Modelle, die einen anderen Blickwinkel beschreiben, wie etwa das „Curation Lifecycle Model“¹⁰, das verschiedene Tätigkeitsfelder bei der Erhaltung und Pflege der Daten beschreibt, und das „Data Curation Continuum“¹¹, das vor allem auf die technischen Bedingungen eingeht und Daten unter dem Gesichtspunkt der Nutzergruppen, von privat bis öffentlich, beschreibt.

In jeder Phase des Datenlebenszyklus gelten unterschiedliche Schwerpunkte mit ihren eigenen Implikationen, die im Folgenden umrissen werden. In dem Kapitel Projektphasen ab Seite 14 werden die wichtigsten Aspekte dann ausführlich behandelt.



Abb. 1.1: Datenlebenszyklus

⁹<http://data-archive.ac.uk/create-manage/life-cycle>

¹⁰<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

¹¹<http://www.dlib.org/dlib/september07/treloar/09treloar.html>
und http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Erstellung Vor der eigentlichen Erstellung von neuen Daten muss schon die Projektplanung erfolgt und die Fragestellung formuliert sein, die maßgeblich bestimmen, wie die gesammelten und erzeugten Daten letztendlich aussehen sollen. Ein Bestandteil der Planung muss auch das Datenmanagement sein, das ab Seite 16 beschrieben wird, um vorab festzulegen, welche Formate und Benennungsregeln verwendet werden sollen, wie die Speicherung und Sicherung der Daten und die Vernetzung von Daten und Projektmitarbeitern untereinander aussehen soll. Bereits vorhandene relevante Daten sollten lokalisiert und berücksichtigt werden.

Schon während der Erstellung der Forschungsdaten erfolgt idealerweise die Beschreibung der Daten, da diese besonders im Hinblick auf die spätere Nachnutzung für die Verständlichkeit eine elementare Rolle spielt. Beispielsweise müssen die technischen Parameter der verwendeten Aufnahmegeräte so früh wie möglich dokumentiert werden.

Verarbeitung Sobald die Daten erhoben wurden, können sie verarbeitet werden. Dazu gehören verschiedene Abläufe, wie etwa Digitalisierung, Übersetzung, Überprüfung, Validierung, Bereinigung und andere Verarbeitungsformen mittels Programmen. Für einige Anwendungen oder im Hinblick auf eine spätere Veröffentlichung, kann eine Anonymisierung notwendig sein. Auch gilt es festzuhalten, wie Ausgangsdaten verändert und bearbeitet wurden.

Außerdem ist die Beschreibung von Forschungsdaten ein wichtiger Aspekt für die Archivierung, um die Verständlichkeit für die spätere Nachnutzung deutlich zu erhöhen. Während der Verarbeitung von Forschungsdaten spielt die Verwaltung und Speicherung der Daten eine große Rolle, da gewährleistet werden muss, dass mit der richtigen Version gearbeitet wird und notfalls auch Sicherheitskopien zur Verfügung stehen.

Analyse Durch die Analyse und Interpretation der Forschungsdaten werden die Forschungsergebnisse gewonnen, die dann in Publikationen veröffentlicht werden. In der Regel sind die den neuen Erkenntnissen zugrunde liegenden Daten aber nur ein kleiner Bestandteil einer solchen Publikation. Jedoch werden alle Daten benötigt, um die Analyseergebnisse vollständig nachvollziehen zu können.

Archivierung Die Archivierung von Forschungsdaten beinhaltet die Auswahl und Umwandlung der einzelnen Dateien in geeignete Formate und deren Speicherung auf einem für die Archivierung geeigneten Medium. Die Erstellung und Speicherung von Backups gehört ebenfalls dazu.

Eine wesentliche Rolle bei der Archivierung spielt sowohl die strukturierte als auch die freie Beschreibung der Forschungsdaten, da diese Informationen darüber enthalten, wie und durch wen die Daten gewonnen wurden, welche Geräte dabei Verwendung fanden, wie diese konfiguriert waren und was die Daten bedeuten.

Zugang Da das Ziel einer Datenarchivierung immer die Nachnutzbarkeit von Inhalten ist, sollten archivierte Daten im Rahmen rechtlicher Rahmenbedingungen verbreitet und Dritten ein Zugriff auf diese gewährt werden. Den Zugriff auf die Daten kann man mit verschiedenen Zugriffsrechten steuern, die

es beispielsweise nur einer eingeschränkten Gruppe von Nutzern erlauben auf die Daten zuzugreifen. Bevor die Daten zugänglich gemacht werden, sollte man dafür sorgen, dass das Urheberrecht geklärt und gekennzeichnet ist.

Durch die Zugriffsgewährung wird die Sichtbarkeit von eigenen Forschungsleistungen erhöht.

Nachnutzung Archivierte Daten, die zugänglich gemacht wurden, können für eigene Forschungsvorhaben kostenlos wiederverwendet und neuen Analysen unterzogen werden. Durch die Nachnutzung wird die Nachprüfbarkeit von Ergebnissen im Sinne der guten wissenschaftlichen Praxis erleichtert.

2 | Projektphasen

S. Jahn, F. Schäfer, M. Trognitz

Im Umgang mit digitalen Daten gelten für jede Phase eines Forschungsprojektes unterschiedliche Anforderungen und Bedingungen. Da Forschungsdaten einem Lebenszyklus (siehe Seite 11) unterliegen, haben Entscheidungen und Arbeitsschritte, die in einer bestimmten Phase getroffen werden, auch Auswirkungen auf die anderen Phasen des Datenkreislaufs. Bereits bei Entwurf und Planung eines neuen Forschungsprojektes muss überlegt werden, welche Informationen bereits digital existieren, welche Arten von Dateien neu erzeugt werden müssen, welche Informationstechnologien zum Einsatz kommen sollen und wie das Management der Forschungsdaten gestaltet sein wird.

Insofern gilt es einige Aspekte so früh wie möglich, häufig noch vor der Erstellung der ersten Datei, zu adressieren:

- Informationen über existierende Standards und Richtlinien einholen
- Entscheidungen über die zu verwendenden Dateiformate und Softwareprogramme fällen
- Verantwortliche für das Datenmanagement bestimmen
- Kriterien für die laufende Dokumentation entwickeln
- Konventionen für die Ablage, Benennung und Versionierung von Dateien festlegen
- Strategien zur Speicherung und Sicherung definieren
- Geeignete Infrastrukturen für die Archivierung auswählen und kontaktieren

Diese und weitere Fragen zur Vorbereitung, Durchführung und Nachbereitung eines Projektes werden im Folgenden anhand eines sogenannten Datenmanagementplans thematisiert. Der Datenmanagementplan bildet einen wichtigen Baustein zur Festlegung der Arbeitsabläufe im Umgang mit Forschungsdaten und für eine Kostenkalkulation.

Je mehr und je früher in den Einzelentscheidungen die Nachnutzung der Daten durch Dritte auch über die Lebensdauer eines Projektes hinaus berücksichtigt wird und die Langzeitarchivierung als kontinuierliche Maßnahme und nicht als letzter Schritt eines bereits abgeschlossenen Projektes begriffen wird, umso leichter gelingt es, einmal erhobene Daten auch für die Zukunft zu erhalten. Die Archivierung der Daten nach Abschluss eines Projektes wird erleichtert, wenn schon während der Durchführung eines Projektes auf die konsequente Erfüllung der im Datenmanagementplan genannten Aufgaben geachtet wird.

Das Finanzierungskonzept eines Forschungsprojektes sollte neben Hard- und Softwareausstattung ausdrücklich angemessene Anteile für Personalkosten und Dienstleistungen berücksichtigen. Im Zweifelsfall ist es ratsam, sich hierbei beraten zu lassen. Ab einer bestimmten Projektgröße bzw. Umfang der zu prozessierenden Daten sollten eine oder mehrere Personalstellen eingeplant werden, die überwiegend oder ausschließlich für IT-Belange zuständig sind.

Vor oder kurz nach Beginn des Projektes erstellte und kommunizierte Richtlinien, ermöglichen es allen Projektmitarbeitern diese auch einzuhalten. Auf diese Weise wird die Handhabung der Projektdaten erheblich erleichtert, da geregelte Ordnerstrukturen und vorgegebene Namenskonventionen das Auffinden der Daten vereinfachen, die parallele Dokumentation das Verstehen und die Nachnutzung ermöglicht und Sicherungskopien das Risiko eines Datenverlustes reduzieren.

Abweichungen von den hier aufgeführten Empfehlungen sind nicht auszuschließen, doch sollten diese bewusst diskutiert und in Abwägung der jeweiligen Vor- und Nachteile entschieden werden.

2.1 Datenmanagement

M. Trognitz, S. Jahn, D. Hagmann, J. Räther

Bereits vor Beginn eines Forschungsvorhabens steht die Konzeption und Planung des Projektes. Dazu gehört auch eine Beschreibung über den Umgang mit den resultierenden Forschungsdaten, der sogenannte Datenmanagementplan. Ein vollständiger Datenmanagementplan berücksichtigt alle Phasen des auf Seite 11 beschriebenen Lebenszyklus von Forschungsdaten. Er dient zunächst als Mittel zur strukturierten Reflexion über datenrelevante Aspekte eines Projektes und beantwortet grundsätzliche Fragen zu Verantwortlichkeiten, Maßnahmen zur Pflege und Verarbeitung der Daten, Standards und bereits vorhandenen Daten. Außerdem bietet er die Grundlage für Arbeitsabläufe (Workflows) im Umgang mit Forschungsdaten und für eine Kostenkalkulation.

Eine vollständige Dokumentation des Datenmanagements in einem Datenmanagementplan spart Zeit und Kosten, wenn beispielsweise Zusammenhänge beim Wechsel von Mitarbeitern hergestellt werden sollen, und beugt einem Datenverlust vor. Da die Nachnutzung von Forschungsdaten zunehmend an Bedeutung gewinnt, setzen Geldgeber oft einen Datenmanagementplan als Teil eines Förderantrags voraus.

Die konsequente Einhaltung aller im Datenmanagementplan gemachten Vorgaben während der gesamten Projektlaufzeit stellt sicher, dass die Daten auch von Dritten interpretiert und nachgenutzt werden können. Wird das Überführen der Daten in ein Archiv schon von Beginn des Projektes vorbereitet und einkalkuliert ist zum Projektabschluss nur noch ein geringer Aufwand für die Übergabe in ein Archiv erforderlich, weil die Umformatierung, Neustrukturierung oder nachträgliche Dokumentation von Daten wegfällt.

Ein aktives Datenmanagement beugt insbesondere während der Planung und Durchführung eines Projektes späteren Zeit- und Budgetverlusten vor und stellt sicher, dass Daten am Ende in nachhaltigen Formaten, gut dokumentiert und gut strukturiert vorliegen.

2.1.1 Übersicht der Aufgaben in den Projektphasen

Mit Hilfe des zu Beginn eines Projektes erstellten Datenmanagementplans können sämtliche Prozesse während des Projekts strategisch umgesetzt werden, wobei die Umsetzung des Plans als laufender Vorgang zu verstehen ist. Wenn während eines Projektes Änderungen an dem Datenmanagementplan notwendig werden, sollten diese begründet, dokumentiert sowie Arbeitsprozesse und Ergebnisse angepasst werden. Außerdem muss dokumentiert werden wie die Änderungen sich auf bereits bestehende Daten auswirken.

In den unterschiedlichen Phasen des Datenlebenszyklus und des Projektes sind die Aspekte des Datenmanagementplans unterschiedlich stark zu berücksichtigen. Die folgende Tabelle veranschaulicht, wann im Lebenszyklus von Forschungsdaten welchen Aufgaben aus dem Datenmanagement besondere Aufmerksamkeit zuteil werden muss.

In Abhängigkeit der Projektgröße kann der Umfang der Aufgaben des Datenmanagementplanes und der Plan selbst skaliert werden, wobei jedoch Minimalanforderungen einzuhalten sind, um zuallererst die Verfügbarkeit der Daten im laufenden Projekt zu gewährleisten. Um die Planungen zu erleichtern und zu beschleunigen, sollten bereits vorhandene institutionelle Vorgaben und In-

frastrukturen genutzt werden. Besonders wichtige, als Minimalanforderung zu betrachtende Aspekte des Datenmanagementplans sind in der Tabelle mit einem Stern gekennzeichnet.

	Vorbereitung	Erstellung	Verarbeitung	Analyse	Archivierung	Zugang	Nachnutzung
Rahmendaten*	●	○	○	○	●	○	○
Verantwortlichkeiten*	●	○	○	○	○	○	○
Rechtliche Aspekte*	●	○	○	○	●	●	●
Methoden	●	●	●	●	○	○	○
Vorgaben und Standards*	●	●	●	○	●	●	○
Kosten & Ressourcen*	●	●	●	○	●	○	○
Externe Partner	●	●	●	○	○	●	●
Hard- und Software*	●	●	●	●	○	○	○
Datentypen und Datenformate	○	●	●	○	●	○	○
Nachnutzung vorhandener Daten*	○	●	●	●	○	●	○
Datenerzeugung & -prozessierung*	○	●	●	○	●	●	○
Datenmenge	●	●	●	●	●	○	○
Dateispeicherung und -sicherung*	○	●	●	●	○	○	○
Dateiverwaltung	○	●	●	●	●	○	○
Dokumentation*	○	●	●	●	●	●	○
Qualitätssicherung	○	●	●	●	●	○	●
Datenaustausch	○	●	●	●	○	●	○
Mittelfristige Datenaufbewahrung	○	●	●	●	○	○	○
Langfristige Archivierung*	○	●	●	●	●	○	●
Zugänglichkeit und Nachnutzung	●	●	●	●	●	●	●
Projektabschluss	○	○	○	○	●	●	○

Tabelle 2.1: Tabellarische Übersicht über die verschiedenen zu berücksichtigenden Aspekte eines Datenmanagementplans während unterschiedlicher Projektphasen und unter Beachtung des Lebenszyklus von Forschungsdaten.

2.1.2 Datenmanagementplan

Die einzelnen in einem Datenmanagementplan zu berücksichtigenden Aspekte werden im Folgenden kurz skizziert, wobei diese Liste keinen Anspruch auf Vollständigkeit erhebt, sondern vor allem als Anregung für verschiedene Themenfelder gedacht ist. Zu einigen Aspekten folgen ausführlichere Abschnitte im Anschluss an diese Übersicht.

Die Aufgaben können grob in drei Phasen gegliedert werden:

- Planungsphase (ab Seite 18)
- Durchführungsphase (ab Seite 20)

- Abschlussphase (ab Seite 24)

Planungsphase

In der Planungsphase eines Projektes ist es im Hinblick auf den Datenmanagementplan insbesondere notwendig, die hierfür erforderlichen Ressourcen einzuplanen. Diese stehen in direkter Abhängigkeit von dem allgemeinen Rahmendaten und den anzuwendenden Methoden. Sie umfassen sowohl den technischen, als auch den personellen Aufwand. Neben der allgemeinen Festlegung von Verantwortlichkeiten empfiehlt es sich, für den Datenmanagementplan als ganzes oder spezifischer Teile und Datenarten Datenverantwortliche zu benennen.

Rahmendaten und administrative Angaben

- Welche allgemeinen Informationen und Rahmenbedingungen des Projektvorhabens kontextualisieren das Vorhaben und die Daten? (z.B. Eckdaten des Projektes wie Titel, Namen der Verantwortlichen, Partner, Methoden und Laufzeit)
- Was sind, zusammengefasst, die Ziele und das Vorhaben?
- Wer sind die Projektträger und Finanzgeber?

Dazu kann auch die Liste für Projektbezogene Metadaten ab Seite 36 konsultiert werden.

Verantwortlichkeiten

- Wie werden die Rollen und Zuständigkeiten beim Datenmanagement eingeteilt?
- Wer beaufsichtigt die Einhaltung des Datenmanagementplans und der daraus resultierenden Vorgaben?
- Wer ist für Hard- und Software zuständig?
- Wer kümmert sich um die Sicherung der Daten und Backups?
- Wer sind institutionelle Ansprechpartner?
- Wer sind sonstige Ansprechpartner?
- Wer kümmert sich um die Integrität der Daten?
- Wie werden die Verantwortlichkeiten kommuniziert?
- Ist für eine Kontinuität bei den Verantwortlichkeiten gesorgt?
- Wer bekommt welche Berechtigungen für welche Daten?
- Gibt es für bestimmte Datenarten Datenverantwortliche?

Rechtliche Aspekte

- Welche Daten sind urheberrechtlich geschützt?
- Welche Daten fallen unter Datenschutz?
- Wie werden die Rechte am geistigen Eigentum für die Daten von Beginn an dokumentiert?
- Gibt es Anforderungen und Einschränkungen für eine Veröffentlichung der Daten?
- Mit welchen Lizzenzen sollen die Daten für Dritte zur Verfügung gestellt werden?

Methoden

- Welche Methode oder Grabungsmethode wird angewandt?
- Hat die Fundstelle oder das Projektvorhaben Einfluss auf die zu verwendenden Methoden?
- Gibt es Richtlinien oder Best-Practices für die eingesetzten Methoden?
- Welche Dokumentationsmethoden kommen zum Einsatz?
- Beeinflusst die zu verwendende Methode die Datenmenge?

Vorgaben, Richtlinien und Standards

- Gibt es Gesetze, Vorschriften der Institution, der Projektträger, der Geldgeber, der externen Partner, der zuständigen Landesämter, die eingehalten werden müssen?
- Kann eine bestehende institutionelle Infrastruktur zur Organisation, Verwaltung und Speicherung der Daten genutzt werden?
- Sind externe Standards und Richtlinien für den Umgang mit den Daten bekannt?
- Welche Qualitätsvorgaben sind für die verschiedenen Datenarten notwendig?
- Müssen eigene Vorgaben definiert werden?
- Gibt es Checklisten zur Kontrolle der Einhaltung von Vorgaben?

Kosten und Ressourcen

- Wie hoch werden die anfallenden Kosten für Personal und Technik eingeschätzt?
- Berücksichtigt die Kostenkalkulation die Kosten in Abhängigkeit der zu erwartenden Datenmenge?
- Fallen weitere Kosten für externe Partner oder Dienstleister an?

- Muss mit zusätzlichen Kosten für spezielle Anwendungen, Werkzeuge, Systeme etc. gerechnet werden?
- Wie sind die Kosten für die Publikation der Ergebnisse zu erwarten?
- Welche Folgekosten sind nach Projektende zu erwarten, beispielsweise für die Archivierung der Forschungsdaten?
- Welche weiteren Ressourcen werden benötigt?
- Bei reproduzierbaren Daten: Sind die Kosten für eine Aufbewahrung höher als für eine Wiederbeschaffung?
- Welche Kosten fallen einmalig an, welche regelmäßig?
- Können Kosten durch regelmäßige Aufgaben oder durch eine frühzeitige Berücksichtigung von bestimmten Aufgaben verringert werden? (z.B. Dokumentation, Auswahl für das Archiv)
- Wer trägt welche Kosten?
- Kann der Standard während der gesamten Laufzeit gehalten werden?

Weitere Hinweise zur Planung von Kosten und Ressourcen sind in der Broschüre "Einstieg ins Forschungsdatenmanagement in den Geowissenschaften" des EWIG-Projektes ab Seite 11 zu finden. (<http://doi.org/10.2312/lis.14.01>)

Durchführungsphase

Während der Durchführung eines Projektes werden die in der Planungsphase gesammelten und erstellten Vorgaben umgesetzt und deren Einhaltung aktiv kontrolliert, sowie bei Bedarf angepasst.

Eine zeitnah nach Plan durchgeführte Erstellung, Verarbeitung, Analyse und Dokumentation von Daten und Arbeitsabläufen vermeidet eine lange und kostenintensive Abschlussphase.

Externe Partner oder Dienstleister

- Mit welchen externen Partnern soll kooperiert werden?
- Welche Dienstleister sollen in Anspruch genommen werden?
- Welche Auflagen entstehen dadurch und sind diese mit den eigenen Vorgaben vereinbar?
- Wie erfolgt der Datenaustausch?
- Bei wem verbleiben die Rechte an den Daten?
- Wie werden die verschiedenen Vorgaben und Richtlinien an externe Partner kommuniziert?

Hard- und Software

- Welche Hard- und Software steht zur Verfügung?
- Werden spezielle Geräte oder Programme benötigt?
- Erfüllen die Systeme die vorgegebenen Auflagen, etablierte Standards und die Anforderungen an ein nachhaltiges Datenmanagement?
- Können kostenpflichtige Programme durch frei verfügbare Programme (so genannte Open-Source-Software) ersetzt werden?

Datentypen und Datenformate

- Welche Daten werden verwendet oder erzeugt? (z.B. Beobachtungs- und Messdaten, prozessierte Daten etc.)
- In welchen Formaten werden die Daten erzeugt und in welchen sollen sie gesichert werden?
- Welches Dateiformat ist für die Archivierung geeignet?
- Gibt es verbreitete Standards, die bei der Wahl des Formats zu beachten sind?
- Können offene Formate verwendet werden oder müssen proprietäre verwendet werden und hat das Implikationen für die verwendete Hard- und Software?

Weitere Informationen sind in den Kapiteln „Dateiformate“ ab Seite 52 und „Forschungsmethoden“ ab Seite 194 zu finden.

Nachnutzung vorhandener Daten

- Gibt es bereits vorhandene Daten, die nachgenutzt werden können?
- Wurde nach Datenbeständen im Besitz der eigenen Institution oder von Dritten recherchiert?
- Wie sind deren Zugriffsmöglichkeiten und Urheberrechte?
- Ist die Qualität der Daten ausreichend? (z.B. geeignete Formate, ausreichend Dokumentation)
- Wie wird die Integration zwischen bereits bestehenden und neuen Daten organisiert?
- Sind auch analoge Quellen zu berücksichtigen?

Erzeugung und Prozessierung von Daten

- Welche Daten müssen neu erzeugt werden oder können bestehende nachgenutzt werden, um das Projektziel zu erreichen?
- Handelt es sich um einmalige Daten oder können sie reproduziert werden?
- Welche Rolle spielen die Daten innerhalb des Projektes? (z.B. Dokumentation, Publikation, Nachnutzung etc.)
- Gibt es sensible oder besonders schützenswerte Daten?
- Kann man Anforderungen potentieller Nachnutzer bei der Datenerzeugung mitberücksichtigen?

Datenmenge

- Wie groß ist die zu erwartende Datenmenge für die Gesamtdauer des Projekts?
- Hat das Folgen für die Speicherung und Sicherung der Daten? (z.B. längere Backup-Zeiten)
- Ergeben sich daraus besondere Anforderungen an die technische Infrastruktur? (z.B. mehr Speicherplatz)
- Fallen verschiedene Bearbeitungsstufen mit verschiedenen Versionen an, die gegebenenfalls ein Versionierungssystem erfordern?

Weitere Informationen zur Versionierung sind in dem Abschnitt „Versionskontrolle“ ab Seite 46 zu finden.

Dateispeicherung und -sicherung

- Welche Maßnahmen zur Dateispeicherung und -sicherung sind während des Projekts notwendig?
- Auf welcher Hardware sollen die Daten gesichert werden? (z.B. Server oder Festplatten)
- Wie oft, womit und durch wen werden Backups durchgeführt?
- Wer ist für die Datenspeicherung und -sicherung verantwortlich?
- Gibt es ein Disaster Management?
- Wurden die Maßnahmen zur Dateiwiederherstellung geprobt?

Weitere Informationen sind im Abschnitt „Dateispeicherung und -sicherung“ ab Seite 48 zu finden.

Dateiverwaltung

- Wie sollen Dateien benannt, geordnet und versioniert werden?
- Gibt es Benennungsregeln für die Dateien?
- Kann auf bestehende Vorgaben und Systeme für die Dateiverwaltung zurückgegriffen werden?
- Sind Verzeichnisstrukturen logisch nachvollziehbar und selbsterklärend?
- Ist die Datenablage dokumentiert?
- Wie ist der Umgang mit verschiedenen Dateiversionen geplant?
- Ist der Einsatz eines Versionierungssystems notwendig?
- Werden Daten, die für eine künftige Nachnutzung geeignet sind oder archiviert werden sollen, gesondert verwaltet?

Weitere Informationen sind im Abschnitt „Dateiverwaltung“ ab Seite 43 zu finden.

Dokumentation

- Wie sollen die Daten beschrieben werden, damit sie kurz- und langfristig lesbar und verständlich sind?
- Welche Informationen sind zur Dokumentation der Forschungsdaten notwendig?
- Zu welchem Zeitpunkt muss die Dokumentation geschehen?
- Gibt es Vorgaben oder Standards dafür?
- Werden Veränderungen und Aktualisierungen der Daten dokumentiert?
- Wie sollen Metadaten abgelegt und gespeichert werden?
- Werden Anpassungen an der Projektstruktur und dem Datenmanagementplan dokumentiert?
- Werden Ausnahmen dokumentiert?
- Gibt es Werkzeuge, die den Dokumentationsprozess unterstützen?

Weitere Informationen sind im Abschnitt „Dokumentation“ ab Seite 28 zu finden.

Qualitätssicherung

- Welche Kriterien sind hinsichtlich vorhandener Standards zur Qualitäts-sicherung zu beachten?
- Wie werden Vorgaben zu Formaten und zur Datenbearbeitung eingehal-ten?
- Sind die Daten genau, konsistent und authentisch?
- Sind die Daten vollständig?
- Wie steht es um die Integrität der Daten?
- Sind die Daten verständlich Dokumentiert und geht aus der Dokumenta-tion hervor: *Wer* hat *wann*, zu welchem *Zweck*, *was* und *womit* gemacht?
- Wird eine Qualitätskontrolle durchgeführt?
- Gibt es Checklisten zur unterstützung der Qualitätskontrolle?
- Welche Maßnahmen gegen ein versehentliches Löschen oder eine Manipu-lation der Daten werden getroffen?

Datenaustausch

- Wie ist der Datenaustausch zwischen den Projektbeteiligten geplant?
- Welche technische Infrastruktur ist für den Datenaustausch erforderlich?
- Sind gesetzliche Vorgaben oder andere Einschränkungen zu beachten?
- Wie soll auf die Daten zugegriffen werden?
- Welche Nutzungsrechte liegen für die Daten vor?

Abschlussphase

In der das Projekt abschließenden Phase müssen besonders Entscheidungen so-wohl zur mittelfristigen Datenaufbewahrung, als auch zur langfristigen Archivie-rung der im Projekt generierten Daten und Ergebnisse getroffen werden. Neben der Klärung der organisatorischen und rechtlichen Rahmenbedingungen muss hierbei spezielles Augenmerk auf die Regelung der Zugänglichkeit zu den lang-zeitarchivierten Daten und der Nachnutzung gelegt werden. So wird sicherge-stellt, dass über den Projektabschluss hinaus die Daten langfristig zur Verfü-gung stehen.

Mittelfristige Datenaufbewahrung

- Welche Gründe gibt es für eine Aufbewahrung der Daten?
- Liegen Vorgaben zur Aufbewahrungs dauer der Daten vor?
- Liegen Vorgaben zu Aufbewahrungsorten der Daten vor?

- Welche Daten sollen oder müssen aufbewahrt werden, welche Daten sollen oder können gelöscht werden?
- Sind sie selbst Hersteller der Daten?
- Ist die Aufbewahrung von Fremddaten zwingend notwendig?
- Wie lange sollen die Daten aufbewahrt werden?
- Wie und wo sollen die Daten aufbewahrt werden?
- Wer ist für die Aufbewahrung der Daten verantwortlich?
- Sind Kosten zu erwarten?

Vorbereitung für die langfristige Archivierung

- Welche Daten sollen archiviert werden?
- Liegen Kriterien für die Auswahl der Daten vor?
- Gibt es eine passende Archivlösung?
- Wurde bereits Kontakt mit dem Archiv aufgenommen?
- Welche Besonderheiten gilt es zu beachten? (z.B. eine gesonderte Aufbereitung der Daten)
- Wann und durch wen werden die Daten übergeben?

Zugänglichkeit und Nachnutzung

- Wie sollen die Daten zugänglich sein?
- Welche Zusatzinformationen sind für das Verständnis der Daten notwendig?
- Wer darf die Daten nutzen, welche Lizizenzen sollen verwendet werden?
- Gibt es Einschränkungen für den Zugriff auf und die Nutzung der Daten?
- Auf welche Weise sollen die Daten zur Verfügung gestellt werden?

Projektabchluss

- Liegt die Dokumentation vollständig und den Vorgaben entsprechend vor?
- Ist eine gesonderte abschließende Dokumentation erforderlich?
- Ist der Datenmanagementplan in die Dokumentation integriert?
- Ist die Nachnutzbarkeit der Forschungsdaten auch nach Projektende gewährleistet?
- Wie ist die mittelfristige Aufbewahrung der Daten geregelt?
- Wie ist die langfristige Archivierung der Daten geregelt?
- Wie ist die Zugänglichkeit der Daten geregelt?

2.1.3 Weiterführende Informationen

Quellen

- Archaeology Data Service, Planning for the Creation of Digital Data
http://guides.archaeologydataservice.ac.uk/g2gp/CreateData_1-0
- Archaeology Data Service, Data Management and Sharing Plans
<http://archaeologydataservice.ac.uk/advice/DataManagementPlans>
- Archaeology Data Service, DataTrain. Open Access Post-Graduate Teaching Materials in Managing Research Data in Archaeology
<http://archaeologydataservice.ac.uk/learning/DataTrain>
- S. Büttner – H.-C. Hobohm – L. Müller (Hrsg.) Handbuch Forschungsdatenmanagement (Bad Honnef 2011)
- DCC (Hrsg.) Data Management Plans
<http://www.dcc.ac.uk/resources/data-management-plans>
- Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum GFZ – Bibliothek und Informationsdienste – Institut für Meteorologie der FU Berlin – Konrad-Zuse-Zentrum für Informationstechnik Berlin (Hrsg.) Einstieg ins Forschungsdatenmanagement in den Geowissenschaften (2014), DOI: 10.2312/lis.14.01
<http://doi.org/10.2312/lis.14.01>
- HU Berlin (Hrsg.) Datenmanagementplan. Anleitung zur Erstellung eines Datenmanagementplans (DMP)
<https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/dmp>
- U. Jakobsson, Data Management Planning. What it is and how to do it (2013)
http://www.ariadne-infrastructure.eu/fre/Events/ARIADNE-Workshop-EAA-2013/Agenda-Presentations/DataManagementPlanning_SND_Ujakobsson_04092013
- H. Neuroth – A. Oßwald – R. Scheffel – S. Strathmann – M. Juhn (Hrsg.) nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.0 (2009) Kap.3:15
- K. Perrin – D. H. Brown – G. Lange – D. Bibby – A. Carlsson – A. Degraeve – M. Kuna – Y. Larsson – S. U. Pálsdóttir – B. Stoll-Tucker – C. Dunning – A. Rogalla von Bieberstein, A Standard and Guide to Best Practice for Archaeological Archiving in Europe, EAC Guidelines 1 (2013)
<http://archaeologydataservice.ac.uk/arches/Wiki.jsp?page=The%20Standard%20and%20Guide%20to%20Best%20Practice%20in%20Archaeological%20Archiving%20in%20Europe>
- UK Data Archive (Hrsg.) Create & Manage Data. Planning for Sharing
<http://data-archive.ac.uk/create-manage/planning-for-sharing>
- Uni Marburg (Hrsg.) Projekt "Kompetenzzentrum Forschungsdatenmanagement und -archivierung"
<https://www.uni-marburg.de/projekte/forschungsdaten>
- V. van den Eynden – L. Corti – M. Woppard – L. Bishop – L. Horton, Managing and Sharing Data. Best Practice for Researchers (Essex 2011)
<http://ukdataservice.ac.uk/manage-data/handbook.aspx>
- WissGrid (Hrsg.) Leitfaden Forschungsdaten-Management (2011)

<http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-oeffentlicher-Entwurf-Leitfaden-Forschungsdaten-Management.pdf>

Beispiele für Datenmanagementpläne

DataONE (Hrsg.) Data Management Planning

<https://www.dataone.org/data-management-planning>

DCC (Hrsg.) Data plan guidance and examples

<http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples>

MITLibraries (Hrsg.) Write a Data Management Plan

<http://libraries.mit.edu/data-management/plan/write/#Resources>

Checklisten und Tools für Datenmanagementpläne

DCC (Hrsg.) Checklist for a Data Management Plan (2014)

http://www.dcc.ac.uk/webfm_send/1279

DCC (Hrsg.) DMPOnline

<https://dmponline.dcc.ac.uk/>

J. Richards, Online Resources for Data Management Planning (2013)

http://www.ariadne-infrastructure.eu/fre/Events/ARIADNE-Workshop-EAA-2013/Agenda-Presentations/DataManagementPlanning_Online_Resources_ADS_JRichards_04092013

WissGrid (Hrsg.) Checkliste zum Forschungsdaten-Management (2011)

<http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-oeffentlicher-Entwurf-Checkliste-Forschungsdaten-Management.pdf>

2.2 Dokumentation

S. Jahn

Neben technischen Aspekten ist eine vollständige Dokumentation das wichtigste Kriterium für die Archivierung und Nachnutzbarkeit von Daten. Damit Forschungsdaten von Dritten gefunden und sinnvoll verwertet werden können, müssen sie verständlich und strukturiert beschrieben werden. Ohne zusätzliche Informationen sind die meisten digitalen Dateien, die in einem Forschungsprojekt entstehen, für die Wissenschaft verloren und der Aufwand einer langfristigen Datenkuratierung ist vergeblich. Die Dokumentation und Beschreibung von Daten ist als ein kontinuierlicher Prozess zu begreifen, der in allen Phasen eines Datenlebenszyklus adressiert werden sollte.

Der folgende Abschnitt vermittelt einen Überblick über die Thematik. Es wird sowohl auf die Dokumentation von Projektdaten in freier, beschreibender Form eingegangen als auch der Unterschied zu strukturierten, maschinell verarbeitbaren Metadaten erklärt. Auch werden die verschiedenen Arten von Metadaten erläutert, die Relevanz von einheitlichen Begriffen und Vokabularen hervorgehoben, Hinweise zur Strukturierung von Metadaten gegeben sowie unterschiedliche Arten der Metadaten-Speicherung skizziert.

Ein weiteres Kapitel widmet sich Metadaten in der Anwendung. Hier sind Vorlagen für Metadaten zu finden, die sich einmal auf die Beschreibung ganzer Projekte oder Dokumentensammlungen beziehen und einmal auf einzelne Dateien. Die hier aufgeführten Kriterien bilden auch die Grundlage für die Beschreibung von Daten, die über IANUS archiviert werden sollen.

2.2.1 Dokumentation und Metadaten

Bei der Durchführung eines Projekts werden Daten erhoben, analysiert, ausgetauscht und gespeichert. Dies geschieht in der Regel in zyklischen oder sich wiederholenden Prozessen, wie die Ausführungen zum Lebenszyklus von Forschungsdaten auf Seite 11 veranschaulichen. Zu festgelegten Zeitpunkten sowie am Ende von Projekten werden üblicherweise Berichte und Publikationen erzeugt, die die Fragestellung, die angewandte Methode, den untersuchten Gegenstand und die Ergebnisse beschreiben und dokumentieren. Diese können in den Altertumswissenschaften ergänzt werden durch weitere Dokumente wie beispielsweise Tagebücher, Kataloge, Listen, Fotos, Abbildungen, Zeichnungen, Pläne, Karten, Messreihen, Tabellen, Diagramme, Modelle, Filme, Animationen und andere Informationsobjekte. Die Dokumente können dabei nur in analoger Form, in analoger und digitaler oder ausschließlich digital vorliegen. Bei Daten, die direkt mit Hilfe elektronischer Geräte erzeugt wurden und daher zumeist nur digital existieren, wird von *digital born data* gesprochen.

Viele Dokumente und Dateien sind nicht aus sich selbst heraus verständlich. Sie stehen immer in einem gewissen Forschungs- oder Projektkontext, der für die Nachnutzung dokumentiert werden muss. Welche Informationen sind hierbei erforderlich? Wichtig sind alle Angaben, die den Entstehungsprozess und -kontext sowie die Konventionen von Inhalten und Daten beschreiben oder zumindest skizzieren, damit andere Personen die Daten auffinden, die Inhalte richtig verstehen und sinnvoll für weitere Forschungen verwenden können. So ist es beispielsweise üblich, bei Fotografien das Aufnahmedatum, eine Kurzbeschreibung des abgelichteten Objektes, den Fotografen und eine Inventarnummer anzuge-

ben, Zeichnungen mit einem Maßstab, der geographischen Ausrichtung, dem Zeichner und einem Kurztext zu beschriften oder bei Texten einen Autor, einen Titel und ein Datum zu nennen.

Noch stärker als bei einer rein analogen Dokumentation sind bei digitalen Daten zusätzliche, spezialisierte Informationen erforderlich, die über die rein deskriptiven Angaben zu den Inhalten und zum Forschungsinteresse hinausgehen und nicht nur die primären Fragen zu „*Wer, Wie, Was, Wo, Wann und Warum?*“ beantworten. So sind etwa technische und administrative Angaben über die Vorgehensweise der Datenerhebung und die eingesetzten Programme, mit denen die Daten erzeugt oder digitalisiert wurden, für ein späteres Auslesen, Auswerten und Interpretieren von Forschungsdaten unverzichtbar. Auch wie die Daten strukturiert wurden und in welcher Beziehung Dateien zueinander stehen, muss in der Regel explizit erklärt werden. Angaben zu Qualitätssicherungsverfahren, Änderungshistorie und Versionierung von Daten erlauben es das empirische Vorgehen im Forschungsprozess nachzuprüfen. Zudem ist wichtig zu erklären, wie Dritte zukünftig auf die Daten zugreifen oder sie nutzen dürfen.

Damit unbeteiligte Personen den größeren Zusammenhang einer einzelnen Information oder Datensets verstehen und nachvollziehen können, sollte jede Dokumentation eines Projektes mindestens folgende Punkte enthalten:

- Angaben zur Fragestellung und zur Auswahl des Untersuchungsgegenstandes
- Angaben zu den Projektverantwortlichen für Nachfragen
- Zusammenfassung der wissenschaftlichen Ergebnisse
- Beschreibung von Arbeitsabläufen und Methoden, vor allem bezüglich der Datenerhebung, -verarbeitung und Qualitätssicherung
- Auflistung der erzeugten unterschiedlichen Dokumentarten (z. B. Tagebücher, Berichte, Listen, Fotos etc.)
- Beschreibung von verwendeten Handbüchern, Standards, projektspezifischen Konventionen, Nummernsystemen etc.
- Auflistung der eingesetzten technischen Geräte und Programme
- Relevante Publikationen und wichtige Sekundärliteratur
- ggf. wichtige Korrespondenzen, Verträge, Anträge etc. (in anonymisierter Form)
- Bestimmungen zur weiteren freien oder eingeschränkten Verwendung von Daten

Ein spezifischer Bestandteil einer jeden Dokumentation sind die sog. Metadaten. Während die Dokumentation die Summe aller Angaben zu einem Projekt und der zugehörigen Daten umfasst und sie in Form, Inhalt, Länge und Struktur je nach fachlichen Gegebenheiten völlig frei formuliert sein kann, sind die Metadaten nach festen formalen Kriterien strukturiert (z. B. als Formulare oder Eingabemasken) und geben nur die wichtigsten Kontextinformationen in knapper Form (z. B. als Stichworte) wieder. Aufgrund der Art der Speicherung von

Metadaten, beispielsweise als Listen, Tabellen, Datei-Header oder Textauszeichnungen, können sie maschinell verarbeitet werden.

Dadurch dass Metadaten die Eigenschaften von anderen Daten beschreiben, ohne diese selbst zu enthalten, und die fachlich oder organisatorisch notwendigen Kontextinformationen explizit machen, dienen sie dazu, relevante Informationen zu finden, zu identifizieren, auszuwählen und zu verwalten. Beispielsweise sind Fotos nur anhand der mit ihnen assoziierten beschreibenden Metadaten effizient durchsuch- und auffindbar oder Karten nur dank einer Maßstabsangabe und mit einer aussagekräftigen Legende wissenschaftlich nutzbar. Die National Information Standards Organization (NISO) definiert Metadaten als Werkzeuge, um Forschungsdaten nachhaltig zu managen, zu erschließen, zu verstehen und zu benutzen.

Die Dokumentation eines Projektes und die Erzeugung von Metadaten über Forschungsdaten sollte als ein wichtiger, kontinuierlicher Prozess verstanden werden, der von Anfang an berücksichtigt und während des gesamten Lebenszyklus von Daten umgesetzt wird – und nicht erst nach Abschluss von Arbeiten oder bei Übergabe von Beständen an ein Archiv. Insofern bedarf es innerhalb von größeren Projekten eines Verantwortlichen oder einer Organisationsform aller Beteiligten, der bzw. die die Struktur und Pflege der Dokumentation konzeptionell und koordinierend begleitet. Gemeinsam muss definiert werden, welcher Gegenstand dokumentiert wird, in welchem Umfang und mit welcher Gliederung, welche Form geeignet ist (z. B. Wiki versus PDF), wie die Aktualität der Inhalte erhalten wird, wie Nutzer über Änderungen informiert werden usw. Fehlende Strategien für den Aufbau und die Pflege einer Dokumentation führen häufig zu Unübersichtlichkeit und dadurch zu einer Ablehnung der beteiligten Akteure. Von umfassenden, detaillierten und strukturierten Angaben zu digitalen Daten profitieren nicht nur die Archive und zukünftige Generationen von Forschern, sondern auch die Wissenschaftler selbst bereits während der Durchführung eines Projektes – genauso wie es auch bei analogen Dokumenten und Objekten der Fall ist.

Nur wenn Daten ausreichend fachlich wie technisch dokumentiert werden, können sie durch andere Personen wissenschaftlich nachgenutzt werden und besitzen einen Wert für die Zukunft. Nur dann lohnt sich auch der Aufwand einer analogen oder digitalen Langzeitarchivierung.

2.2.2 Arten von Metadaten

Da Metadaten immer von unterschiedlichen Informationsbedürfnissen und Anwendungskontexten abhängig sind, lassen sie sich unter verschiedenen Blickwinkeln betrachten und nach Arten zusammenfassen.

Metadaten können, egal ob sie ein Forschungsvorhaben in seiner Gesamtheit oder den Inhalten eines einzelnen Dokuments beschreiben, sich auf folgende Aspekte beziehen:

- Deskriptive Angaben, die ein Projekt, ein Objekt oder eine Methode fachlich und inhaltlich näher beschreiben (z. B. Kurzbeschreibung eines Fotos, Laufzeit eines Forschungsvorhabens, Autor eines Textes)
- Strukturelle Angaben, welche die physischen oder logischen Beziehungen zwischen komplexen Objekten beschreiben (z. B. referenzierte Dateien in Zeichnungen, die Reihenfolge von Fotos)

- Administrativ-rechtliche Angaben, die Rechteinhaber, Lizenzbedingungen und Zugangsregeln benennen und für das Management eines Objektes relevant sind (z. B. Fristen für die Veröffentlichung von Ergebnissen)
- Administrativ-erhaltungsbezogene Angaben, welche die Geschichte eines digitalen Objektes, also die vorhergehenden und gegenwärtigen Zustände, nachvollziehen lassen und Erhaltungsmaßnahmen beschreiben (z. B. Angabe zur Herkunft einer Karte, Konvertierungsmaßnahmen)
- Technische Angaben, die Informationen zu Software und Hardware liefern (z. B. ein Dateiformat eines Bildes, die Zeichenkodierung eines Textes oder technische Parameter eines Messgerätes)

Ein anderes Unterscheidungskriterium von Metadaten ist der Umfang der Daten, die durch sie beschrieben werden. So können sie sich beziehen auf:

- Eine größere zusammenhängende Datensammlung, sog. Kollektionen (z. B. der gesamte Datenbestand eines Projektes)
- Eine einzelne Datei
- Eine einzelne Information innerhalb eines Systems (z. B. ein Datensatz in einer Datenbank)

Während für ein laufendes Projekt eventuell nur ein Ausschnitt dieser Metadaten eine praktische Relevanz besitzt, sind für die digitale Langzeitarchivierung von Dokumenten alle Aspekte gleichberechtigt wichtig. Einige der Metadaten können ausschließlich von beteiligten Personen frühzeitig während des Prozesses der Datenerzeugung erfasst werden und sollten es auch, da nur sie den Inhalt, den Charakter, die Struktur, den Kontext und die Quellen der Daten kennen. Andere dagegen lassen sich auch erst zu einem späteren Zeitpunkt bei der Archivierung und Veröffentlichung von Daten generieren. Technische Metadaten zum ursprünglichen Dateiformat und zur Hardware- und Softwareumgebung werden benötigt, um Daten in veralteten Dateiformaten in neuere umwandeln oder um die ursprüngliche Programmumgebung mittels aktueller Technik nachbauen zu können.

Außerdem kann bei Metadaten unterschieden werden zwischen solchen, die eine manuelle Eingabe erfordern, und anderen, die durch Systeme und Programme automatisch erzeugt werden können. Erstere können beispielsweise eine Kurzbeschreibung oder Schlagworte sein. Letztere sind Angaben wie Erstellungsdatum, Dateiname oder Einstellungen der Digitalkamera. In dem Abschnitt Speicherung von Metadaten ab Seite 34 sind nähere Informationen zu finden.

2.2.3 Kontrollierte Vokabulare

Damit Metadaten möglichst sinnvoll genutzt und maschinell verarbeitet werden können, ist es notwendig möglichst einheitliche Begriffe und homogene Beschreibungen zu verwenden. Nur wenn gleiche Dinge auch mit den gleichen Begriffen benannt werden, ist es möglich vollständige und präzise Suchergebnisse zu erhalten oder vergleichbare Daten richtig miteinander zu verknüpfen. Die Vorgabe und Definition von festen Begriffen und Regeln hilft zudem, Mehrdeutigkeiten und Redundanzen zu vermeiden, etwa wenn eine Zeichenkette verschiedene

Bedeutungen besitzen kann (z. B. Abakus als Rechenhilfsmittel oder als architektonischer Abschluss eines Kapitells), ein identischer Sachverhalt durch unterschiedliche Worte erfasst werden kann (z. B. Survey und Oberflächenbegehung) oder die Form der Angabe variieren kann (z. B. das Datum 12.03.2012 oder 2012-03-12).

Das geeignete Mittel zur Vereinheitlichung der sprachlichen Vielfalt sind sogenannte kontrollierte Vokabulare, die entweder einfache Wortlisten oder strukturierte Thesauri sein können, in denen Wörter zusammen mit ihrem semantischen Kontext verwaltet werden. Diese „terminologische Kontrolle“ kann in unterschiedlicher Weise systematisiert und implementiert sein. Beispielsweise kann sie durch projektinterne, kontrollierte Vokabulare erfolgen. Dazu werden alle verwendeten Worte innerhalb eines Projektes unter allen Beteiligten abgestimmt, klar fachlich definiert bzw. voneinander abgegrenzt und in strukturierter Form dokumentiert. Diese Absprachen können dann in einem zentralen Textdokument als Projektleitfaden abgelegt oder in einer Datenbank als Felder umgesetzt werden, die nur eine begrenzte Auswahl von Begriffen zur Beschreibung eines spezifischen Sachverhaltes zulassen (z. B. für das Attribut Filmart nur die Werte Diapositiv (Farbe) – Negativfilm (Farbe) – Negativfilm (SW) – Digital erlauben).

Daneben existieren etablierte, standardisierte, projektunabhängige Thesauri und Vokabulare, die häufig eine fachspezifische oder institutionelle Ausprägung aufweisen und von maßgeblichen Einrichtungen kontinuierlich gepflegt und verbreitet werden. So ist es etwa im Bibliothekssektor üblich, Publikationen mit Schlagworten aus einem gemeinsamen Wörterkatalog zu versehen, damit verschiedene Werke einheitlich beschrieben werden und die Auffindbarkeit von Werken zu einer bestimmten Thematik überall nach den gleichen Kriterien erfolgen kann. Bereits existierende Thesauri und Vokabulare sollten bei der Vergabe von Metadaten möglichst berücksichtigt und angewendet werden. Insofern gilt die Empfehlung, sich frühzeitig zu informieren, welche einschlägigen kontrollierten Vokabulare vorhanden sind und in einem Projekt genutzt werden können. Abzuwägen ist dabei, ob die Verwendung bestehender Standardvokabulare dem Detailinteresse oder der neuen Erkenntnis in der Forschung gerecht werden, da eine Aktualisierung sich in der Regel nur sehr langsam vollzieht und sehr vom organisatorischen Rahmen der Institution bzw. eines Gremiums abhängt, in dessen Verantwortung die Weiterentwicklung liegt. Trotz dieser Einschränkung überwiegen langfristig aber die Vorteile, wenn existierende Thesauri und Vokabulare verwendet werden, da dies auch den elektronischen Austausch, die Interoperabilität, der eigenen Daten mit anderen Systemen erheblich vereinfacht.

Wenn Ressourcen in mehreren Sprachen vorliegen und entsprechend multilingual beschrieben werden sollen, müssen die genutzten Wörterbücher, Thesauri und Schlagwortsysteme äquivalente Begriffe in mehreren Sprachen abbilden.

2.2.4 Struktur von Metadaten

Neben der Verwendung einheitlicher, möglichst weit verbreiteter Begriffe stellt die Struktur und der Umfang von Metadaten einen wesentlichen Faktor dar, um diese sinnvoll nutzen zu können. Die Art und Weise, wie die verschiedenen Informationen erfasst und organisiert werden, spielt eine wesentliche Rolle, ob sie von einer unbeteiligten dritten Person richtig verstanden werden und ob Computer die Angaben automatisiert maschinell verarbeiten können. Wie bei den

Vokabularen existieren auch hierfür verschiedene Standards, die als sogenannte Metadatenschemata den Inhalt und die Gliederung von Metadaten verbindlich beschreiben und die bei der Vergabe von eigenen Metadaten berücksichtigt werden sollten. Für viele Bereiche, wie etwa Medienarchive, Bibliotheken oder Museen, gibt es langjährige Erfahrungen und entsprechende eigene Vorgaben.

Ein Metadatenschema definiert die verwendeten Metadaten-Kategorien. Beispielsweise kann eine Publikation mit den Elementen Autor – Jahr – Titel – Reihe – Verlag – Erscheinungsort und Schlagworte beschrieben werden, die dann mit den entsprechenden Informationen gefüllt werden. Diese Minimalbeschreibung kann je nach Bedarf erweitert werden, z. B. um die Elemente Sprache – Seitenzahl – Abbildungszahl – Auflage und ISBN-Nummer. In einem Metadatenschema wird zudem für einzelne Informationen die Genauigkeit (die Granularität) der erwarteten Information festgelegt, also z. B. ob für die Angabe Autor eine freie Texteingabe ausreichend ist oder ob eine Aufteilung in die Attributte Anrede – Titel – Vorname(n) und Nachname sowie eine Referenz zu einer Personennormdatei erforderlich ist.

Für den Austausch von Metadaten zwischen verschiedenen technischen Systemen besitzen die zugrunde liegenden Schemata eine zentrale Bedeutung. Sollen Informationen aus unterschiedlichen Quellen miteinander in Beziehung gesetzt werden, so dass sie gemeinsam ausgewertet werden können, müssen die Informationen auf der Ebene ihrer semantischen Bedeutung miteinander verknüpft und ggf. abgeglichen werden. Das heißt neben dem eigentlichen Wert (z. B. der Zeichenkette „Pompeji“) muss auch die Eigenschaft angegeben werden, die durch diesen Wert beschrieben wird (z. B. das Metadatum Fundort oder Aufbewahrungsort oder Publikationsort).

Je internationaler und standardisierter ein verwendetes Schema ist, desto eher ist die Austauschbarkeit von Metadaten mit anderen Systemen gewährleistet. Sofern sich unterschiedliche Metadaten auf ein gemeinsames Metadatenschema abbilden lassen, können sie in ein drittes System importiert und dort gemeinsam geöffnet werden.

Das am weitesten verbreitete und ISO-zertifizierte Metadatenschema ist *Dublin Core*. Es verfügt über folgende 15 Kernfelder, die jeweils individuell definiert sind: Mitwirkender – Geltungsbereich – Urheber – Zeitangabe – Beschreibung – Format – Identifikator – Sprache – Verleger – Beziehung – Rechte – Quelle – Thema – Titel – Typ.

Eine besondere Relevanz im archäologischen und altertumswissenschaftlichen Kontext haben zudem folgende Standards erlangt:

- ADeX – Archäologischer Datenexport-Standard für den Austausch archäologischer Fachdaten, erarbeitet von der Kommission „Archäologie und Informationssysteme“ im Verband der Landesarchäologen (VLA).
- CIDOC Conceptual Reference Modell (CIDOC CRM) als erweiterbare Struktur für Begriffe und Informationen im Bereich des Kulturerbes, erstellt von der Arbeitsgruppe „Dokumentationsstandards“ im Internationalen Komitee für Dokumentation (CIDOC) des internationalen Museumsverbandes (ICOM).
- Lightweight Information Describing Objects (LIDO) und museumdat als Austauschformate für bewegliche Objekte in Museen, herausgegeben von der Fachgruppe Dokumentation des Deutschen Museumsbundes und von

der Arbeitsgruppe „Data Harvesting and Interchange“ im Internationalen Komitee für Dokumentation (CIDOC).

- Connecting Archaeology and Architecture in Europeana (CARARE) Metadaten Schema als archäologiespezifisches Datenmodell unter Berücksichtigung verschiedener europäischer Standards, herausgegeben von Projekt-partnern innerhalb von Europeana.

2.2.5 Speicherung von Metadaten

Für eine strukturierte Speicherung von Metadaten gibt es verschiedene Möglichkeiten und Wege, die einerseits von dem Zeitpunkt (synchron mit der Datenerhebung oder im Nachhinein) und andererseits von der Art der Metadatenerfassung (manuell oder automatisiert) abhängen. Grundsätzlich sollte man sich bewusst machen, dass Metadaten so früh wie möglich, also bereits bei der Generierung von neuen digitalen Objekten oder am Anfang eines neuen Forschungsvorhabens vergeben werden sollten, selbst wenn diese noch nicht vollständig gemacht werden können. Die Erfahrung zeigt, dass viele Informationen sich im Nachhinein (z. B. am Ende eines Projektes oder am Ende eines Arbeitsverhältnisses) nicht mehr ermitteln lassen oder in Vergessenheit geraten. Je nach Art der Metadaten und der Veränderlichkeit des beschriebenen Objektes ist eine regelmäßige Aktualisierung und kontinuierliche Metadatenpflege erforderlich.

Der einfachste Fall liegt dann vor, wenn Metadaten sich auf eine einzelne Datei beziehen und von dem Gerät oder dem Programm, dass die Datei erzeugt, automatisch generiert werden. Beispielsweise können bei digitalen Fotos von der Kamera automatisch technische Angaben zum Dateiformat, Verschlusszeit, Blendenöffnung, Farbinformationen usw. ermittelt und direkt in der Bilddatei hinterlegt werden. Je nach Gerät können über die Voreinstellungen auch deskriptive Informationen wie Fotograf, Aufnahmedatum, Land etc. hinzugefügt werden. Ähnliche Verfahren zur automatischen Metadatenerzeugung sind auch bei Vermessungsgeräten weit verbreitet.

Die durch Geräte generierten Metadaten werden unmittelbar in einem besonderen Bereich und auf standardisierte Weise in der resultierenden Datei gespeichert, etwa bei Fotos üblicherweise als Exif im Dateikopf. Diese Informationen können über die Eigenschaften einer Datei angezeigt und teilweise verändert und ergänzt werden. In Abbildung 2.1 sind beispielsweise die Dateieigenschaften eines PDF-Dokumentes zu sehen.

Abhängig vom verwendeten Dateiformat können Metadaten direkt in eine Datei integriert werden. Beispielsweise können mit den entsprechenden Textverarbeitungsprogrammen für Textdokumente zusätzliche Angaben in den Dateieigenschaften von DOCX- oder ODT-Dokumenten gemacht werden. Hierbei ist in der Regel eine manuelle Eingabe in entsprechende software- und dateitypspezifische Formulare erforderlich. Bei der Auswahl von Anwendungen sollte darauf geachtet werden, ob diese spezifische Metadaten entweder als separate Datei exportiert oder ob sie mit den Dateien, in denen sie enthalten sind, auch unabhängig von der ursprünglichen Software geöffnet werden können.

Bei Textdokumenten, die auf expliziten Auszeichnungssprachen basieren (z. B. HTML-, XML-, TEI-, SVG-Dateien), lassen sich die Metadaten ebenfalls einfach in die Datei integrieren, da sie an den Anfang des Dokumentes mit speziellen Elementen (sog. Tags) platziert werden können.

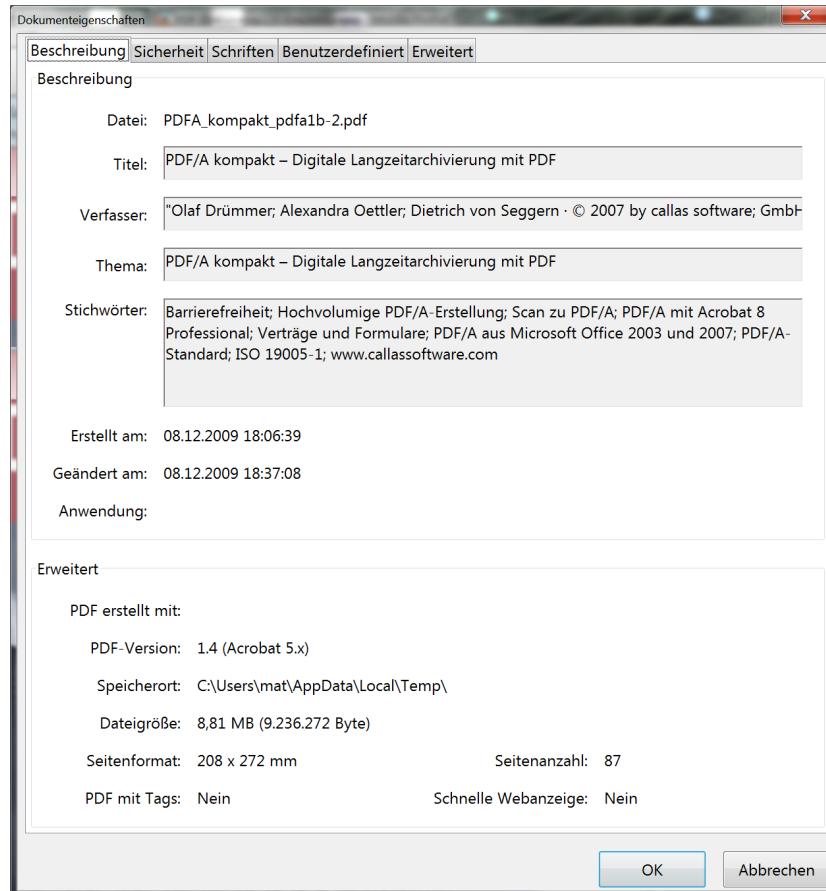


Abb. 2.1: Beispiel für Metadaten für ein PDF im Programm Adobe Reader. Die Metadaten können unter „Datei > Eigenschaften“ angezeigt werden.

Konkrete Hinweise für die Bearbeitung und Ergänzung von Metadaten für bestimmte Dateiformate sind in den entsprechenden Abschnitten im Kapitel Dateiformate ab Seite 52 zu finden.

Eine andere Form der Metadatenspeicherung besteht darin, die Angaben in separaten Dateien oder Systemen zu speichern und zu verwalten. Die häufigste Form sind dabei Tabellen und Datenbanken, da dort eine manuelle Erfassung, zum Teil mithilfe spezifischer Eingabemasken, eine gezielte Suche und eine strukturierte Sicherung von Metadaten möglich ist. Auch hier besteht prinzipiell die Möglichkeit, Metadaten (teil-)automatisiert erzeugen zu lassen, z. B. das Datum der letzten Bearbeitung eines Datensatzes und den Namen des Bearbeiters. Ein weiterer Vorteil besteht in der effizienten Bearbeitung, wenn ein einzelnes Merkmal für mehrere Dokumente identisch ist, z. B. wenn alle Zeichnungen eines Projektes den gleichen Maßstab besitzen.

Hinweise für die Extraktion von Metadaten für bestimmte Dateiformate sind ebenfalls in den entsprechenden Abschnitten im Kapitel Dateiformate ab Seite 52 zu finden.

Da bei dieser Form der Metadatenspeicherung die Metadaten von den digitalen Objekten, die sie beschreiben, physisch getrennt sind, muss in besonderer Weise darauf geachtet werden, dass die Beziehungen zwischen beiden Einheiten eindeutig und aktuell sind. Wird etwa bei den Metadaten der Name einer Datei zur Identifizierung verwendet, muss sichergestellt werden, dass dieser einmalig ist (ggf. in Kombination mit seinem Speicherort) und dass im Falle der Umbenennung der Datei der Eintrag auch bei den Metadaten aktualisiert wird.

Die Erfassung von Metadaten in einer Freitextdatei sollte vermieden werden, da diese häufig nicht automatisiert durch Computer überprüft und verarbeitet werden können.

Bei Dokumenten mit Textinhalt können beschreibende Metadaten auch automatisch erzeugt werden, indem nach charakteristischen Zeichenketten gesucht wird und Schlag- und Stichworte extrahiert werden. Allerdings reicht die Qualität einer automatischen Erschließung bislang nicht an eine manuelle und intellektuelle Metadatenvergabe heran. Da eine Maschine ein Dokument nicht verstehen kann, ist sie auch nicht in der Lage, ausschließlich sinnvolle Deskriptoren herauszufiltern.

Für die Übergabe von Dateien und Metadaten an ein Langzeitarchiv gilt die Empfehlung, die Metadaten sowohl in den Originaldateien selbst als auch in einem separaten, strukturierten und textbasierten Dokument vorzuhalten: in den Originaldateien selbst, um keine verwaisten, d. h. undokumentierten Werke zu erzeugen, und als separate Textdatei, um eine automatisierte Verarbeitung und den Austausch von Referenzangaben zu vereinfachen.

2.2.6 Metadaten in der Anwendung

Projektbezogene Metadaten

Die wichtigsten Metadaten, die für die Beschreibung eines Projektes oder einer Dokumentensammlung erforderlich sind, werden in der folgenden Tabelle abgebildet und knapp definiert. Sie geben einen Überblick über einen größeren, zusammenhängenden Datenbestand und beschreiben den fachlichen Kontext in dem dieser entstanden ist. Vergleichbar einem Bibliothekskatalog liegt die Hauptfunktion dieser Metadaten darin, dass externe Personen ein Projekt oder eine Dokumentensammlung über Web-Portale und Suchmaschinen finden und einordnen können. Darüber hinaus enthalten sie rechtliche Informationen, die für den weiteren Umgang mit den Daten wichtig sind.

Die hier zusammengestellten Eigenschaften basieren auf dem Dublin Core Metadata Schema und den Angaben, wie sie vom ADS in UK und tDAR in USA erhoben werden, um einen zukünftigen Austausch zu vereinfachen. Ein ausführliches Metadatenschema, das auch die Grundlage für die Archivierung von Projektdaten bei IANUS bildet, sowie ausgefüllte Beispielformulare sind in dem Kapitel Archivierung von Forschungsdaten in IANUS ab Seite 222 zu finden.

Bezeichnung	Kurzdefinition
Identifizierung – Projekttitle	Verbindliche Kurzbezeichnung des Projektes.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Bezeichnung	Kurzdefinition
Identifizierung – Alternativtitel	Ggf. alternative Titel für ein Projekt.
Identifizierung – Projektnummer(n)	Nummern oder Kennungen, die z.B. innerhalb der durchführenden Organisation oder von Mittelgebern verwendet wird, um das Projekt eindeutig identifizieren zu können.
Kurzbeschreibung	Knappe Angaben zur Fragestellung, zum Verlauf und Ergebnis des Projektes sowie Skizzierung der Datensammlung (insgesamt ca. 100-300 Worte)
Schlagworte – Fachdisziplinen	Stichworte, die die beteiligten Disziplinen und Fächer benennen. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.
Schlagworte – Inhalt	Stichworte, die den Inhalt der Datensammlung benennen., z. B. zu Materialgruppen, Fundstellen-Klassifizierung, Quellenarten, Kulturgruppen etc. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.
Schlagworte – Methoden	Stichworte, die die eingesetzten Forschungsmethoden beschreiben. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.
Ausdehnung – Geografisch-1	Detaillierte Angaben zur räumlichen Ausdehnung oder zum Fundort des untersuchten Gegenstandes mittels geografischer Koordinaten. Die maximale Ausdehnung kann als Bounding Box angegeben werden.
Ausdehnung – Geografisch-2	Sprachliche Beschreibung des untersuchten Gegenstandes mittels Ortsangaben mit Land, Stadt, Kreis, Straße, Gemarkung etc. Sofern Namen sich im Lauf der Zeit geändert haben, dies gesondert vermerken. Sofern eine Referenz zu einer Geo-Ressource oder einem Gazetteer existiert, sollte diese ebenfalls angegeben werden.
Ausdehnung – zeitlich	Chronologische Angaben zum untersuchten Gegenstand, entweder als Periodenbezeichnung und/oder mit groben/genauen Jahresangaben. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Bezeichnung	Kurzdefinition
Primärforscher – Person	Personen, die entweder für das Projekt als Ganzes, für das Datenmanagement oder für die Erzeugung bestimmter Datenarten zentral bzw. verantwortlich sind. Hier ist eine Kontaktadressen erforderlich und die aktuelle/letzte institutionelle Zugehörigkeit, damit die Personen bei Rückfragen erreicht werden kann.
Eigentümer – Organisation	Organisation, der die unter „Primärforscher“ genannten Personen angehören, oder die nach Ausscheiden derselben für die Daten verantwortlich ist, im weitesten Sinne also Eigentümer der Daten ist. Hier ist eine Kontaktadresse erforderlich, damit die Organisation bei Rückfragen erreicht werden kann.
Finanzierung	Nennung der Organisation(en) / (Dritt-)Mittelgeber, durch die das Projekt finanziert wurde. Es sollte jeweils der Zeitraum der Finanzierung angegeben werden.
Veröffentlichung – Projektdaten	Wenn die hier beschriebene Datensammlung des Projektes bereits an anderer Stelle veröffentlicht / online gestellt wurde, bitte entsprechende Angaben machen, z. B. durch Nennung der Organisationen, Datenarchive, Online-Ressourcen etc.
Veröffentlichung – Ergebnisse	Analoge oder digitale Publikationen zu Ergebnissen des Projektes oder zur Datensammlung des Projektes, ausführliche bibliographische Angaben (ohne fachspezifische Abkürzungen) unter Nennung des Verlages erforderlich.
Dauer – Projekt	Anfangs- und Enddatum des Projektes.
Dauer – Datenbestand	Anfangs- und Enddatum der Erzeugung oder Verarbeitung digitaler Daten im Rahmen des Projektes.
Rechtliches – Urheberrechte	Name des Inhabers der Urheber-, Nutzungs- und Verwertungsrechte; i. d. R. die Organisation, an der der Primärforscher beschäftigt war.
Rechtliches – Lizenzgeber	Angabe der Person, die i. d. R. als Vertretung für eine Organisation für die Lizenzierung von Daten zur Nachnutzung verantwortlich und berechtigt ist, einen Datenübergabevertrag abzuschließen.
Rechtliches – Datenschutz	Angaben, ob in der Datensammlung datenschutzrelevante Informationen enthalten sind. Wenn ja, in welchem Umfang.
Quellen – Ältere	Ältere Quellen oder existierende Ressourcen, auf denen die Daten aufbauen.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Bezeichnung	Kurzdefinition
Quellen – Zugehörige	Sofern während des Projektes Informationen, Datensammlungen, (un-)publizierte Dokumente, Online-Ressourcen etc. verwendet oder erzeugt wurden, die nicht Teil der hier beschriebenen Datensammlung sind, aber für deren Verständnis wichtig sind, bitte entsprechende Angaben zu Art und Umfang dieser Quellen machen.
Sprache	Die in den Dokumenten und Dateien verwendete(n) Sprache(n). Sprachkennungen nach ISO 639 angeben.
Art der Daten	Kurzcharakterisierung der Daten, z. B. ob es sich um Rohdaten, verarbeitete Daten, Interpretationen, Ergebnisse, Abschlussberichte etc. handelt.
Vollständigkeit	Aussagen zur Vollständigkeit der Projektdaten, z. B. ob bestimmte Datenarten noch fehlen und warum.
Dateiformate	Auflistung der Dateiformate, die in der Datensammlung vorkommen, ggf. unter Nennung der verwendeten Programme und Zeichenkodierungen.
Zugriffsrechte	Festlegung der gewünschten Zugriffsrechte für die Daten, sofern diese für den gesamten Projekt-Datenbestand gelten sollen; differenzierte Regelungen müssen auf Dateiebene vorgenommen werden.
Signatur Metadaten	Angabe darüber, wer die o. g. Metadaten wann ausgefüllt hat.

Dateibezogene Metadaten

Bei dieser Art von Metadaten handelt es sich um Informationen, die Nutzern verständlich machen, wie einzelne Dateien innerhalb eines Projektes oder einer Datensammlung beschaffen sind und welche Möglichkeiten der Nachnutzbarkeit sie beinhalten. Dateibezogene Metadaten sind abhängig von dem Dateiformat, dem Inhalt und der Methode, mit der sie erzeugt wurden. Beispielsweise sind für Rasterdateien, die durch digitale Fotografie entstanden sind, andere Angaben erforderlich (Fotograf, Aufnahmedatum, Aufnahmeort, abgelichtetes Objekt etc.) als für Rasterdateien, die durch geophysikalische Messungen erzeugt wurden (Koordinaten, Messgerät, Genauigkeit, Datum etc.). Die Angaben, die für bestimmte Dateiformate empfohlen werden, sind in den verschiedenen Kapiteln zu den jeweiligen Formaten beschrieben.

Daneben existieren aber auch dateibezogene Metadaten, die unabhängig von Format, Inhalt und Methode für alle Einzeldateien gleichermaßen relevant und notwendig sind. Sie sind eher administrativer und technischer Natur und spielen für ein laufendes Forschungsprojekt möglicherweise nur eine untergeordnete Rolle. Für die Langzeitarchivierung von Daten, besonders für die Migration in neuere Dateiformate oder für die Bereitstellung für Dritte, enthalten sie dagegen

zentrale Informationen. Zu diesen Metadaten gehören neben den Angaben zu Dateiname, Dateiformat und Dateiversion auch detaillierte Angaben zur verwendeten Soft- und Hardware, Hinweise zur Speicherung, Versionierung und Referenzen sowie rechtliche Aspekte.

Bezeichnung	Kurzdefinition
Dateiname	Eindeutiger Name der Datei.
Dateiformat	Format, in der die Datei abgespeichert ist.
Software – Dateierstellung	Software, mit der die Datei erstellt wurde.
Hardware – Dateierstellung	Hardware, mit der die Datei erstellt wurde, v. a. bei technischen Geräten wie Kameras, GPS-Geräten, Vermessungsinstrumenten, Laserscanner etc.
Betriebssystem – Dateierstellung	Betriebssystem, das verwendet wurde als die Datei erstellt wurde.
Erstellungsdatum	Datum, an dem die Datei erstellt wurde. Datum und Zeit in UTC nach ISO 8601.
Letzte Aktualisierung	Datum, an dem die Datei zuletzt bearbeitet wurde. Datum und Zeit in UTC nach ISO 8601.
Dateiversion	Angabe der Versionsnummer der Datei.
Referenzdateien	Referenzen auf Dateien, die für das Verständnis einer anderen Datei zentral sind, v. a. bei zusammenhängenden, komplexen Dateien oder wenn auf eine Ursprungsdatei verwiesen werden soll.

Auch wenn es in der Theorie wünschenswert ist, diese Angaben sowie die zugehörigen spezifischen Angaben zu Methoden und Dateiformaten für jede Datei einzeln zu erfassen, so zeigt die Praxis, dass es häufig ausreichend ist, einen Metadatensatz für Gruppen von Dateien anzulegen, wenn diese das gleiche Format und/oder die gleichen inhaltlichen Eigenschaften aufweisen.

Methodenbezogene Metadaten

Jede Fachdisziplin innerhalb der Altertumswissenschaften verfügt über individuelle Forschungsmethoden. Auch diese haben Einfluss auf Umfang, Art, Struktur und Inhalt digitaler Objekte, da sie bei verschiedenen Arbeitsweisen und technischen Geräten unterschiedlich ausfallen. Daher variieren die methodenbezogenen Metadaten von Fachgebiet zu Fachgebiet und werden individuell in den einzelnen Kapiteln des Abschnittes zu den „Forschungsmethoden“ beschrieben.

Eine Gemeinsamkeit ist den vielen Dateien allerdings, dass sie in einer methodenspezifischen Prozesskette stehen, die es ebenfalls zu dokumentieren gilt — vor allem dann, wenn mehrere Zwischenstände eines solchen Prozesses archiviert und Dritten zur Verfügung gestellt werden sollen. Je nach Genauigkeit der Methodenbeschreibung kann sie sich sowohl auf eine Einzeldatei als auch auf mehrere Dateien gleichen Typs beziehen.

Bezeichnung	Kurzdefinition
Prozessnummer	Eindeutige Nummer eines Prozesses oder einer Methode.
Ausgangsformat(e)	Format der Dateien, die am Anfang eines gesamten Prozesses stehen und den Ausgangspunkt bilden.
Zwischenformat(e)	Format der Dateien, die im Verlauf eines Prozesses erzeugt werden und den Ausgangspunkt für weitere Prozesse bilden.
Zielformat(e)	Format der Dateien, die am Ende eines gesamten Prozesses erzeugt werden und den Zielpunkt definieren.
Durchführende(r)	Person(en), die den Prozess durchgeführt hat (haben).
Prozessbeschreibung	Beschreibung des Prozesses/der Methode, vor allem der Ausgangssituation und der Zielvorstellung.
Prozessbeginn	Datum, an dem der Prozess begonnen wurde. Datum und Zeit in UTC nach ISO 8601.
Prozessende	Datum, an dem der Prozess beendet wurde. Datum und Zeit in UTC nach ISO 8601.
Software	Software, mit der der Prozess durchgeführt wurde.
Hardware	Hardware, auf der der Prozess durchgeführt wurde.

2.2.7 Weiterführende Informationen

Metadaten allgemein

Gesis: Dokumentation und Metadaten
<http://www.gesis.org/archive-and-data-management-training-and-information-center/forschungsdatenmanagement/dokumentation-und-metadaten/>

Understanding Metadata: National Information Standards Organization
<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

Archaeology Data Service: Guides to Good Practice - Project Metadata
http://guides.archaeologydataservice.ac.uk/g2gp/CreateData_1-2

Kontrollierte Vokabulare

J. M. Lill, Kontrolliertes Vokabular. Wieso? Weshalb? Warum?, Fachgruppe Dokumentation im DMB: Terminologie – das Schweizer Messer"der Dokumentation (9. Mai 2012, Kunstmuseum Stuttgart)
http://swop.bsz-bw.de/volltexte/2012/1002/pdf/Lill_Textfassung_VortragDMB2012.pdf

J. Lindenthal, Normen und Standards für Thesauri (2015)
http://www.ianus-fdz.de/it-empfehlungen/sites/default/files/NormenEmpfehlungenEntwicklungThesauri_Lindenthal.pdf

W3C (Hrsg.) Vocabularies
<http://www.w3.org/standards/semanticweb/ontology>

Metadatenstandards

nestor (Hrsg.) Standardisierung – Metadaten

<http://www.langzeitarchivierung.de/Subsites/nestor/DE/Standardisierung/Metadaten.html>

DARIAH-DE (Hrsg.) Daten- und Metadatenformate in den Fachdisziplinen. Archäologie (2015)

<http://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=20058856>

Deutsche Initiative für Netzwerkinformation e. V. (Hrsg.) Kompetenzzentrum Interoperable Metadaten (KIM)

<http://www.dini.de/ag/standards/>

U. Jensen – A. Katsanidou – W. Zenk-Möltgen, Metadaten und Standards, in: S. Büttner – H.-C. Hobohm – L. Müller (Hrsg.) Handbuch

Forschungsdatenmanagement (Bad Honnef 2011) 83-100

http://opus4.kobv.de/opus4-fhpotsdam/files/198/2.4_Metadaten_und_Standards.pdf

J. Riley – D. Becker, Glossary of Metadata Standards (2010)

http://www.dlib.indiana.edu/~jenrlile/metadata-map/seeingstandards_glossary_pamphlet.pdf

Metadatenschemata

Dublin Core Metadata Initiative

<http://dublincore.org/>

ADeX – Standard für den Austausch archäologischer Fachdaten

<http://www.landesarchaeologen.de/verband/kommissionen/archaeologie-und-informationssysteme/projektarbeitsgruppen/adex/>

ARIADNE: C. Papatheodorou – D. Gavrilis – K. Fernie – H. Wright – J. Richards – P. Ronzino – C. Meghini, D3.1 Initial Report on the project registry (2013)

<http://ariadne-infrastructure.eu/content/download/1781/9956/file/D3.1%20Initial%20report%20on%20standards%20and%20on%20the%20project%20registry.pdf>

CIDOC Conceptual Reference Model (Vers. 5.0.1)

<http://cidoc-crm.gnm.de/wiki/Hauptseite>

LIDO Lightweight Information Describing Objects

<http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/lido-technical/specification/>

CARARE metadata schema

<http://www.carare.eu/eng/Resources/CARARE-Documentation/CARARE-metadata-schema>

2.3 Dateiverwaltung

M. Trognitz

Der tägliche Umgang mit digitalen Daten wird durch eine effiziente Dateiverwaltung erheblich erleichtert. Aussagekräftige Dateinamen, die auch für Dritte verständlich sind, sorgen dafür, dass die Dateien gefunden und deren Inhalt verstanden wird. Einheitliche Dateinamensstrukturen erhöhen die Lesbarkeit. Die konsequente Einhaltung von Versionierungsangaben sorgt dafür, dass immer mit der richtigen Dateiversion gearbeitet wird und eine selbsterklärende Ordnerstruktur hilft dabei auch in großen Projekten bestimmte Dateien wiederzufinden.

Die Konzipierung der Dateiverwaltung muss schon zu Beginn des Projektes erfolgen, damit die Regeln von Anfang an angewendet werden können. Die Niederschrift der durch Beispiele ergänzten Benennungsregeln und deren Weitergabe dient im laufenden Betrieb als wichtiges Nachschlagewerk für die konsistente und konsequente Einhaltung derselben. Dies ermöglicht dann eine effizientere Arbeitsweise.

Im laufenden Projektbetrieb sollte die Einhaltung der Benennungsregeln kontrolliert und gegebenenfalls angepasst werden.

2.3.1 Dateiablage

Mit Dateiablage ist hier vor allem die Ordnerstruktur gemeint. Für die Benennung der Ordner gelten die gleichen Regeln, wie für die Dateibenennung, die im nächsten Abschnitt thematisiert werden. Lediglich die Dateinamenserweiterung wird bei Ordnernamen nicht verwendet. Die Dateiablage sollte selbsterklärend sein und unpräzise Namen wie etwa „*In Arbeit*“ sollten unbedingt vermieden werden.

Wichtig ist, dass die Dateiablage logisch und hierarchisch aufgebaut ist. Dies bedeutet unter anderem, dass die Inhalte und Unterordner auch tatsächlich thematisch in den übergeordneten Ordner hineinpassen. Beispielsweise erwartet man in einem Ordner mit dem Namen „*Fotos*“ diverse Fotos, die bei einer entsprechenden Menge vielleicht noch auf verschiedene Unterordner, wie zum Beispiel „*Plana*“ und „*Profile*“, verteilt sind.

Ist aber in dem Ordner „*Fotos*“ ein Unterordner mit dem Namen „*Zeichnungen*“ enthalten, in dem verschiedene digitalisierte Zeichnungen abgelegt sind, führt das zu Schwierigkeiten. Der Unterordner wird wahrscheinlich nur schwer wiedergefunden werden, da der Name des übergeordneten Ordners einen anderen Inhalt verspricht.

Tritt der Fall ein, dass eine Datei oder ein Ordner thematisch in mehrere verschiedene übergeordnete Ordner passen würde, können verschiedene Lösungswege zur Anwendung kommen. Beispielsweise könnte eine Kopie abgelegt werden, was allerdings problematisch ist, wenn eine der Kopien verändert wird und diese nicht mit der anderen abgeglichen wird.

Alternativ könnte eine Dateiverknüpfung angelegt werden. Dateiverknüpfungen verschiedener Betriebssysteme sind meist nicht zu anderen Systemen kompatibel und funktionieren im schlimmsten Fall sogar nur auf dem einen Rechner, auf dem sie erstellt wurden. Dies liegt zum einen daran, dass jedes Betriebssystem unterschiedliche Methoden zur Erstellung verwendet und zum

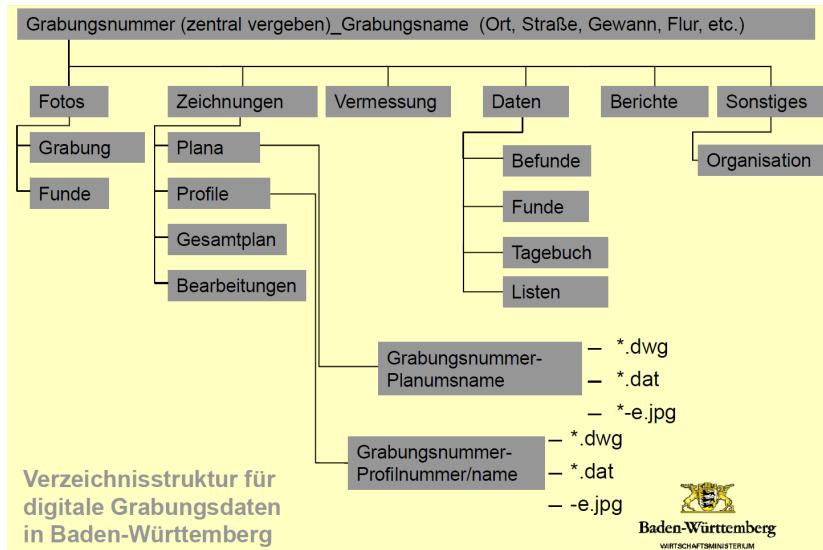


Abb. 2.2: Datenstruktur Baden-Württemberg

anderen können schon die Laufwerke unterschiedlich benannt sein. Außerdem kann die Dateiverknüpfung ins Leere führen, wenn die originale Datei verschoben oder gelöscht wurde. Daher ist von Dateiverknüpfungen abzuraten.

Eine bessere Methode ist das Anlegen einer Textdatei mit einem Hinweis auf den Ablageort der Datei oder des Ordners. In diesem Fall muss aber darauf geachtet werden, dass die gesuchte Datei sich auch tatsächlich an dem angegebenen Ort befindet. Bei Verschieben oder Löschen müsste die Textdatei also auch immer berücksichtigt und angepasst werden.

Am besten ist es, nur eine Instanz einer Datei oder eines Ordners zu haben. Bei Zuordnungsschwierigkeiten kann es helfen, die Datei oder den Ordner in der hierarchischen Struktur höher anzusiedeln.

Eine Textdatei kann auch verwendet werden, um die vorhandene Ordnerstruktur zu erklären oder auf Besonderheiten hinzuweisen. Beispielsweise kann in einem Ordner mit Fotos, deren Metadaten in einer Datenbank abgespeichert sind, mit Hilfe einer Textdatei verdeutlicht werden, wo die Metadaten zu finden sind. Damit die Textdatei auch als eine Hilfdatei erkannt wird, kann man ihr den Namen „*README*“ oder „*LIESMICHT*“ geben. Die Schreibung mit Großbuchstaben erhöht die Auffälligkeit der Datei. Durch eine vorangestellte Null kann sie dann auch in der alphabetischen Sortierung ganz an den Anfang gestellt werden.

Praxisbeispiele

Abhängig von dem Umfang und der Dauer des Projektes, kann sich die Ordnerstruktur beispielsweise am Thema, dem Material, dem Jahr, dem Bearbeiter oder den einzelnen Arbeitsschritten orientieren.

Das Landesdenkmalamt von Baden-Württemberg verwendet eine Dateiablage, die an der analogen Ordnerstruktur einer Grabung orientiert ist. Sie wird in Abbildung 2.2 veranschaulicht.

„Diese Struktur ist in den oberen Ebenen rigide genug, um für Ordnung und Datendisziplin zu sorgen und gleichzeitig weiter unten flexibel genug, um den unausweichlichen Eigenarten der einzelnen Archäologen Rechnung tragen zu können.“

Die Grabungs- oder Vorgangsnummer, die zentral vergeben wird, ist wichtig. Sie beschreibt nicht nur die Ausgrabung, sie ist auch die Inventar-Stammnummer aller Funde, die auf dieser Grabung gefunden werden. Insofern ist sie die Schnittstelle zum zentralen Fundarchiv des Archäologischen Landesmuseums. Die Ausgräber werden aber nicht allein gelassen. Mit der Datenstruktur werden Anleitungen geliefert, die beschreiben, was wohin gehört. Damit wird tatsächlich eine gewisse Einheit im Land erreicht.“¹

Auch die nachträgliche Strukturierung von bereits vorhandenen digitalen Dateiablagen ist relativ einfach, da Klartextinhaltsverzeichnisse in der obersten Ebene des Verzeichnisses abgelegt werden können, welche die Eigenheiten der Grabung und die strukturellen Abweichungen beschreiben.

Anwendungsgebiete mit mehreren voneinander abhängigen Dateikomplexen, wie zum Beispiel 3D-Scanning, erfordern eigene Dateiablagen. Diese werden in den entsprechenden Abschnitten im Kapitel „Forschungsmethoden“ ab Seite 194 beschrieben.

2.3.2 Dateibenennung

Üblicherweise besteht der Dateiname einer Datei aus dem eigentlichen Namen und der durch einen Punkt getrennten Dateinamenserweiterung, die das Format der Datei angibt. Die Erweiterung wird in der Regel von dem Programm, mit dem die Datei gespeichert wurde, automatisch an den Dateinamen gehängt.

Ein Beispiel: *IT-Empfehlungen.pdf* gibt an, dass es sich um eine Datei mit dem Namen „*IT-Empfehlungen*“ handelt, die in dem PDF-Format vorliegt.

Da die Erweiterung automatisch erzeugt wird, muss die Datei mit einem geeigneten Programm konvertiert werden, wenn man das Dateiformat ändern möchte. In dem Kapitel „Dateiformate“ ab Seite 52 sind ausführliche Informationen darüber zu finden.

Im Folgenden geht es nur noch um den reinen Dateinamen, ohne die Dateinamenserweiterung.

Erlaubte Zeichen Moderne Betriebssysteme können mit Sonderzeichen umgehen, zu denen Umlaute und Leerzeichen gehören. Das war aber nicht immer so und kann auch heute noch zu Problemen führen. Webserver lesen zum Beispiel das Leerzeichen als die Zeichenfolge „%20“ ein und da Umlaute nicht immer gleich kodiert werden, kann auch das zu Schwierigkeiten führen.

Für die Langzeitarchivierung sollten also nur die alphanumerischen Zeichen des englischen Alphabets, also a-z, A-Z und 0-9 verwendet werden. Zusätzlich kann der Bindestrich (-) und bei Bedarf auch der Unterstrich (_) verwendet werden.

Zu vermeidende Zeichen Es gibt eine ganze Reihe von Zeichen, die für besondere Aufgaben von Betriebssystemen verwendet werden. Der Punkt

¹D. Bibby, Digitale Datenstruktur auf Ausgrabungen und Archivierung digitaler Grabungsdaten. Praxis und Praxisversuche aus Baden-Württemberg, ANachr 14 2009, 159-163.

dient zur Trennung des Dateinamens von der Dateinamenserweiterung und der Schrägstrich dient in Windows um Ordnerstrukturen zu kennzeichnen.

Zu den Zeichen, die absolut nicht verwendet werden dürfen, gehören:

\ / : * ? “ < > |

Die Verwendung von allen weiteren Sonderzeichen ist zwar möglich, kann jedoch zu einem unerwarteten Verhalten des Systems führen. Daher wird von der Verwendung von Leerzeichen und Sonderzeichen, Bindestrich (-) und Unterstrich (_) ausgenommen, abgeraten.

Groß- und Kleinschreibung Die Groß- und Kleinschreibung in einem Dateinamen wird von unterschiedlichen Systemen verschieden gedeutet. In Windows beispielsweise kann, wenn es eine Datei mit dem Namen „*TestDatei*“ schon gibt, keine Datei mit dem Namen „*testdatei*“ angelegt werden. Andere Systeme könnten dies jedoch erlauben, was aber nicht bedeutet, dass man dies auch tun sollte.

Hat man sich in der Praxis einmal für eine Schreibweise entschieden, muss diese auch konsequent eingehalten werden und insbesondere bei der Arbeit auf verschiedenen Systemen darauf geachtet werden.

Länge Der Dateiname sollte so kurz wie möglich und so lang wie nötig sein. Eine aktuelle Obergrenze, die auf manchen Systemen nicht überschritten werden darf, sind 255 Zeichen. Kryptische oder untypische Kürzel sollten vermieden werden, da sie in der Regel in Vergessenheit geraten.

2.3.3 Versionskontrolle

Wenn unterschiedliche Personen an einer Datei arbeiten, ist es wichtig, die verschiedenen Änderungen und Entwicklungsstadien zu verfolgen und zu kennzeichnen. Nur so kann vermieden werden, dass an der falschen Dateiversion gearbeitet wird oder diese gar gelöscht wird. Dateiversionen, die nicht mehr benötigt werden, sollten bei Bedarf gelöscht werden.

Es gibt mehrere Strategien, um die Versionskontrolle durchzuführen, die im Folgenden erläutert werden.

Angabe im Dateinamen Eine einfache und übersichtliche Methode ist, die Versionsangabe in den Dateinamen zu integrieren. Das kann beispielsweise mit einer Datumsangabe oder Ziffern erfolgen. Durch ein vorangestelltes **v** werden die Ziffern als eine Versionsnummer gekennzeichnet, wie zum Beispiel **v001**. Führende Nullen stellen sicher, dass die Versionsnummern einheitlich und leichter lesbar sind und richtig sortiert werden.

Eine Kennzeichnung von Versionen durch Worte wie „*neu*“, „*neuer*“ und „*alt*“ ist unbedingt zu vermeiden. Eine Ausnahme können endgültige Dateiverisionen bilden, die der Übersicht halber etwa durch *FINAL* am Ende gekennzeichnet werden können. Es darf aber nur eine Datei mit dieser Kennzeichnung in einem Ordner und einem bestimmten Format vorliegen. Eine endgültige Version, die zum Beispiel sowohl als *docx* als auch als *pdf* vorliegt, ist also erlaubt.

Angabe in der Datei Angaben zum Erstellungsdatum und den verschiedenen Versionen und deren Änderungen können im Header der Datei oder in standardisierten Kopfzeilen in der Datei selbst angegeben werden. Bei Textdokumenten bietet sich die Möglichkeit einen Innentitel mit einer Versionshistorie zu verwenden. Ein Beispiel für solch einen Innentitel findet sich am Anfang der PDF-Version dieser Empfehlungen.

Änderungsprotokoll Statt einzelne Dateiversionen abzuspeichern, kann auch ein Änderungsprotokoll geführt werden. Dabei werden die einzelnen Änderungen in einer einfachen Textdatei protokolliert, die zusammen mit der eigentlichen Datei abgelegt wird. Im Englischen wird dafür der Begriff *ChangeLog* verwendet.

Software Die bisher beschriebenen Methoden sind hauptsächlich manuell anzuwendende Vorgänge. Es gibt jedoch auch Software zur Versionsverwaltung. Deren Einsatz lohnt sich vor allem in großen Projekten, die zentral auf einem Server abgelegt werden. Versionsverwaltungssoftware kann aus den Dateiänderungen automatisch *ChangeLogs* erstellen. Die am weitesten verbreiteten Systeme zur Versionsverwaltung sind Subversion (SVN), Mercurial und Git. Primär wurden sie für die Bedürfnisse von Softwareentwicklern konzipiert, jedoch eignen sie sich auch für allgemeinere Aufgaben. Für die einzelnen Computerarbeitsplätze werden Clients wie TortoiseSVN für SVN oder GitHub für Git benötigt.

Eine einfache Versionsverwaltung für Dateien bietet ownCloud, wobei die Menge der unterschiedlichen gespeicherten Versionen abhängig von dem persönlichen Speicherplatz ist.

Git:

<https://git-scm.com/>

Clients für Git:

<http://git-scm.com/downloads/guis>

Mercurial:

<http://mercurial.selenic.com/>

ownCloud:

<https://owncloud.org/>

Subversion (SVN):

<http://subversion.apache.org/>

TortoiseSVN:

<http://tortoisessvn.net/>

Vergleich von Versionsverwaltungssystemen auf Wikipedia:

http://en.wikipedia.org/wiki/Comparison_of_revision_control_software

2.4 Dateispeicherung und -sicherung

R. Förtsch, R. Komp, M. Trognitz

Um Datenverlust vorzubeugen, ist es unerlässlich eine geeignete Sicherungsstrategie zu verwenden. Dabei muss zwischen einer kurzfristigen Speicherung und einer mittelfristigen Sicherung, unterschieden werden. Ersteres meint, wie Daten während der Erstellung und Bearbeitung gespeichert werden. Letzteres bezieht sich auf einen längeren Speicherzeitraum, der durchaus auch mehrere Monate betragen kann, stellt also ein klassisches Backup der Daten dar.

Die Sicherungsstrategie legt fest wie die Datensicherung erfolgen soll und berücksichtigt folgende Fragen:

- Wer ist für die Datensicherung verantwortlich?
- Wer hat Zugriff auf die gesicherten Daten?
- Wann und wie oft soll die Datensicherung durchgeführt werden?
- Auf welche Weise soll gesichert werden?
- Welche Daten sollen gesichert werden?
- Welche Speichermedien sollen verwendet werden?
- Wie viele Sicherungskopien sollen angelegt werden?
- Wo sollen die Sicherungen aufbewahrt und wie sollen sie geschützt werden?
- Wie soll der Transport der Sicherungskopien erfolgen?
- Wie lange soll eine Sicherung aufbewahrt werden?

Die Sicherungsstrategie kann mit den zu verwendenden Richtlinien zur Datensicherung eng verzahnt sein, um etwa einen nahezu automatisierten Sicherungsvorgang zu ermöglichen.

Die richtige Speicherstrategie beugt zwar möglichen Datenverlusten durch Hardware-, Software- oder menschliche Fehler vor, stellt aber noch *keine* Archivierung der Daten dar. Eine Archivierung ist auf Langfristigkeit ausgelegt und impliziert immer eine bewusste Auswahl und umfassende Dokumentation der Daten, da eine Nachnutzung derselben das Ziel ist. Während der Projektlaufzeit kann bereits ein eigener Archivordner angelegt werden, worin beispielsweise finale Dateien oder unprozessierte Rohdaten abgelegt werden können, um die spätere Auswahl der zu archivierenden Daten vorzubereiten und zu vereinfachen.

Weiterführende Hinweise zur Datensicherung sind auf den Seiten des Bundesamtes für Sicherheit in der Informationstechnik zu finden, die sich sowohl an Einsteiger² als auch an Experten³ richten.

²https://www.bsi-fuer-buerger.de/BSIFB/DE/MeinPC/Datensicherung/Sicherungsmethoden/sicherungsmethoden_node.html

³https://www.bsi.bund.de/DE/Themen/ITGrundschutz/ITGrundschutzKataloge/Inhalt/_content/baust/b01/b01004.html

2.4.1 Kurzfristige Speicherung

Zur Bearbeitung geöffnete Dateien sollten in kürzeren Abständen gespeichert werden, um einem Verlust von stundenlanger Arbeit infolge etwa eines Stromausfalls vorzubeugen. Von diesen Arbeitsversionen sollten regelmäßig, beispielsweise nach jedem Arbeitstag, Kopien auf externen Speichermedien angefertigt werden, welche in die mittelfristige Sicherung übergehen.

Vorgaben zur Form der Dateiablage, der Dateibenennung und Versionierung müssen schon beim Anlegen der Dateien beachtet werden, um einen reibungslosen Ablauf der darauf abgestimmten Datensicherungsstrategie zu gewährleisten. Hinweise dazu sind im Abschnitt Dateiverwaltung ab Seite 43 zu finden.

2.4.2 Mittelfristige Sicherung

Digitale Daten sind nur dann über einen längeren Zeitraum sicher, wenn sie mehr als einmal gespeichert werden. Es müssen also mindestens zwei Sicherungskopien auf physisch getrennten Speichermedien vorliegen, um etwa bei einem Hardwa-redefekt auf eine alternative Sicherungskopie zugreifen zu können.

Die Sicherung von Daten auf unterschiedlichen Arbeitsgeräten wird erleichtert, wenn diese zunächst auf einem Datenträger gesammelt werden und dieser dann gesichert wird.

Speichermedium Welches Speichermedium gewählt wird, hängt vor allem von der Speichergröße der zu sichernden Dateien und der vorhandenen Infrastruktur ab. So eignen sich eine CD oder eine DVD nur für Daten bis maximal 900 MB bzw. 17 GB Speichervolumen. Für größere Datenmengen werden entsprechend größere Datenträger benötigt.

Steht ein lokaler oder institutioneller Server zur Verfügung, sollte dieser zur Datensicherung verwendet werden, da diese üblicherweise gewartet werden und meist selbst in eigene Sicherungsstrategien eingebunden sind. In Umgebungen mit Zugriff auf das Internet, kann eine Sicherung auch auf einen entfernten Server, beispielsweise in einem Rechenzentrum, erfolgen. In allen anderen Fällen können externe Festplatten verwendet werden. Von externen kommerziellen Diensten, wie beispielsweise Dropbox, ist für eine mittelfristige Sicherung abzuraten, weil es sich oft um ausländische Anbieter handelt, die nicht dem deutschen Recht unterliegen und den Nutzer im Unklaren darüber lassen, wie mit den Daten umgegangen wird.

Da die Speichermedien eine begrenzte Lebenszeit von etwa zehn Jahren haben, sollten die Sicherungen von Zeit zu Zeit auf neue Speichermedien migriert werden. Außerdem sollten sie zwischenzeitlich an verschiedenen Orten, mindestens in verschiedenen Räumen, gelagert werden, um etwa im Brandfall den Schaden zu begrenzen. Auch der Transport von Original- und Sicherungsdatenbeständen sollte getrennt durch unterschiedliche Personen erfolgen, um zu vermeiden, dass etwa durch Diebstahl alle Sicherungskopien auf einmal abhanden kommen.

Für das Kopieren der Daten auf ein Speichermedium gibt es verschiedene Methoden, wie die im Folgenden beschriebene Volldatensicherung, inkrementelle und differentielle Datensicherung.

Volldatensicherung Bei der Volldatensicherung werden die zu sichernden Dateien komplett und eins zu eins auf ein Speichermedium kopiert. Wird zum

ersten Mal eine Sicherungskopie angelegt, so muss dies als Volldatensicherung geschehen. Dabei ist zu beachten, dass der Kopiervorgang viel Zeit in Anspruch nehmen kann, wobei dies auch über Nacht geschehen kann. Vor der Ausführung der Volldatensicherung muss geprüft werden, dass auf dem Speichermedium ausreichend freier Speicherplatz vorhanden ist.

Die Volldatensicherung ist die einzige Methode, die manuell, ohne weitere Hilfsprogramme, durchgeführt werden kann. Wird eine Volldatensicherung angelegt, sollten, beispielsweise in einer Textdatei, Informationen gespeichert werden, die Auskunft über den Zeitpunkt der Sicherung, den Ansprechpartner und den Umfang der Daten geben.

Liegt eine erste Volldatensicherung vor, können neue oder geänderte Daten mittels inkrementeller oder differentieller Datensicherung gesichert werden, was die Möglichkeit bietet, verschiedene Versionen der Daten zu sichern. Für beide Methoden werden jedoch eigene Programme benötigt.

Inkrementelle Sicherung Die inkrementelle Sicherung speichert nur jene Daten, die sich im Vergleich zu der vorliegenden Volldatensicherung oder vergangenen inkrementellen Sicherungsvorgängen geändert haben. Dies spart Zeit und Speicherplatz, erfordert jedoch im Fall einer Wiederherstellung der Daten, dass zunächst die Volldatensicherung übertragen wird und anschließend alle erfolgten inkrementellen Sicherungen nacheinander eingespielt werden.

Differentielle Sicherung Bei der differentiellen Sicherung werden immer jene Daten gespeichert, die sich im Vergleich zu der vorliegenden letzten Volldatensicherung geändert haben. Im Unterschied zu der inkrementellen Sicherung entfällt also der Vergleich mit Daten aus zwischenliegenden Sicherungsvorgängen seit Erstellung der Volldatensicherung. Diese Methode erfordert zwar etwas mehr Zeit und Speicherplatz als die inkrementelle Sicherung, erleichtert aber die Wiederherstellung der Daten, da nur die Volldatensicherung und die letzte differentielle Sicherung auf das System übertragen werden muss. Die vorherigen differentiellen Sicherungen werden nicht benötigt.

Automatisierung Die Datensicherung kann mit Hilfe von Programmen oder Skripten automatisiert erfolgen. Wichtig ist dabei zu kontrollieren, dass die Datensicherung auch tatsächlich in der gewünschten Form ausgeführt wird. Geeignete Programme gibt es viele, wie beispielsweise Bacula, DirSync Pro, TrayBackup, PersonalBackup oder rsync. Eine ausführliche Liste ist außerdem auf der englischen Wikipedia unter List of backup software zu finden. Oft können die Programme auch zur Wiederherstellung von gesicherten Daten verwendet werden.

Bacula:

<http://bacula.org/>

DirSync Pro:

<http://www.dirsyncpro.org/>

TrayBackup:

<http://www.traybackup.de>

PersonalBackup:

<http://personal-backup.rathlev-home.de>

rsync:

<http://rsync.samba.org/>

Liste von Backup-Programmen auf Wikipedia:

http://en.wikipedia.org/wiki/List_of_backup_software

Sicherungszeitpunkt Der Zeitpunkt der Sicherung spielt eine nicht zu unterschätzende Rolle. Beispielsweise kann eine Sicherung von Datenbanken im laufenden Betrieb zu Inkonsistenzen führen, weshalb in solchen Fällen ein Administrator diese Aufgabe übernehmen sollte.

Regelmäßige Sicherungsvorgänge, die zu bestimmten Zeitpunkten täglich, wöchentlich oder monatlich automatisiert gestartet und durchlaufen werden, gewährleisten, dass eine Sicherung auch tatsächlich erfolgt, zumal sie nicht vom individuellen Verhalten eines Nutzers abhängen.

Für Daten, die verschiedene Arbeitsschritte durchlaufen und dadurch große Änderungen erfahren, können die verschiedenen Zustände als Sicherungspunkte gespeichert werden. Beispielsweise können zunächst unbearbeitete digitale Bilder gespeichert werden, sowie ein daraus erzeugtes photogrammetrisches Modell oder andere nicht oder nur schwer reproduzierbare signifikante Verarbeitungsschritte. Solche Sicherungspunkte müssen allerdings manuell gemacht werden, indem beispielsweise ein Ordner mit den Rohdaten angelegt wird. Hinweise für solche Sicherungspunkte werden für bestimmte angewendete Methoden in dem Kapitel Forschungsmethoden ab Seite 194 gegeben.

Sicherungsversionen Es empfiehlt sich, eine vorhandene Volldatensicherung nicht direkt mit einer neuen Volldatensicherung zu überschreiben. Tritt nämlich während des Sicherungsvorganges ein Fehler auf, könnte auch die bereits vorhandene Sicherungsversion beschädigt werden.

Bei der Anwendung von inkrementellen oder differentiellen Sicherungsmethoden sollte trotzdem in geregelten Zeitabständen eine neue Volldatensicherung angelegt werden, welche mindestens bis zur nächsten Volldatensicherung aufgehoben wird. Dies ist im Sinne des Generationenprinzips oder Großvater-Vater-Sohn-Prinzips, einer Strategie zur Datensicherung, bei der mehrere Sicherungen für verschiedenen Zeitpunkte und Zwecke angelegt werden. Hierfür können beispielsweise wöchentlich und monatlich Volldatensicherungen angelegt werden (Vater und Großvater), die primär den Datenverlust durch Hardware- oder Software-Fehler reduzieren sollen. Tägliche inkrementelle oder differentielle Sicherungen (Sohn) erlauben es beispielsweise auf Arbeitsversionen des Vortages zuzugreifen und können mit dem Beginn einer neuen Woche nach und nach überschrieben werden.

Disaster Recovery Eine gute Sicherungsstrategie umfasst zusätzlich die Erprobung der Wiederherstellung der Daten, um im Ernstfall sicher zu sein, dass die verwendete Strategie auch wirklich funktioniert und die Daten schnell wieder einsatzfähig sind. Somit können vorhandene Sicherungsversionen ebenfalls auf ihre Funktionsfähigkeit überprüft werden.

3 | Dateiformate

F. Schäfer

Im Folgenden werden konkrete technische Hinweise und Hintergrundinformationen zu den Dateiformaten gegeben, die nach heutigem Wissen für die Langzeitarchivierung geeignet sind. Des Weiteren werden die minimalen Anforderungen an die technische wie inhaltliche Dokumentation von einzelnen Dateiformaten genannt, damit diese nicht nur heute, sondern auch in Zukunft von Dritten verstanden und nachgenutzt werden können.

Es geht hier nur um die Dateiformate an sich, ihre Entstehungsweise wird dabei nicht berücksichtigt. Abhängig von der jeweiligen Generierung und Verwendung sind gegebenenfalls zusätzliche methodische und fachspezifische Metadaten erforderlich. Diese werden in dem Kapitel Forschungsmethoden ab Seite 194 beschrieben.

Bereits die Auswahl der verwendeten Software beeinflusst das Dateiformat und dessen Nachhaltigkeit. Daher sollte bei der Entscheidung für oder gegen eine bestimmte Software darauf geachtet werden, dass Dateiformate unterstützt werden, die möglichst folgende Eigenschaften aufweisen:

- Weit verbreitet und standardisiert
- Nicht proprietär, also nicht von einer Anwendung oder einem Hersteller abhängig und mit unterschiedlichen Programmen verwendbar
- Offen dokumentiert mit frei verfügbaren technischen Spezifikationen
- Verlustfreie oder keine Kompression
- Einfach dekodierbar oder sogar unmittelbar lesbar, also nicht durch Kodierung versteckt

Werden während eines Arbeitsprozesses andere als die hier empfohlenen Dateiformate erzeugt oder verwendet, sollte darauf geachtet werden, dass diese leicht und mit möglichst geringem Verlust an Informationen und Funktion in eines der in den folgenden Kapiteln aufgeführten Langzeitformate überführt werden können.

In den Übersichtstabellen mit den Formaten für bestimmte Dateitypen sind die präferierten Formate mit einem grünen Haken, die noch akzeptablen mit einer gelben Tilde und für die Langzeitarchivierung ungeeignete Formate mit einem roten Kreuz gekennzeichnet.

3.1 PDF-Dokumente

S. Jahn

Das PDF (Portable Document Format) wurde 1993 von Adobe Systems entwickelt, um den Datenaustausch zu erleichtern. Es ist ein plattformunabhängiges, offenes Dateiformat, das 2008 mit der Version 1.7 als ISO-Standard zertifiziert wurde und seitdem von der ISO weiter gepflegt wird.

Der große Vorteil von PDFs liegt darin, dass Dateiinhalte unabhängig vom Betriebssystem, dem ursprünglichen Anwendungsprogramm und der Hardwareplattform unverändert dargestellt werden. Das Aussehen von Dokumenten wird, wie bei einem analogen Ausdruck, eingefroren und kann somit angezeigt werden, wie es ursprünglich vom Autor intendiert war. Gleichzeitig sind die Möglichkeiten zur nachträglichen Bearbeitung begrenzt, wodurch eine große Authentizität gewährleistet wird.

Für das Öffnen einer PDF-Datei gibt es verschiedene freie Anwendungen, worin ein Grund für die große Verbreitung und Akzeptanz von PDFs liegt. Viele Programme können Dateien direkt im PDF-Format speichern oder exportieren. Darüber hinaus lassen sich mit Hilfe von zusätzlich installierten Druckertreibern PDF-Dateien aus allen Programmen heraus erzeugen.

Langzeitformate Nicht jede Datei mit der Dateierweiterung .pdf ist gleichermaßen für die Langzeitarchivierung geeignet. Für diesen Zweck wurde das PDF/A-Format entwickelt und als ISO-Standard zertifiziert. Es beschreibt in welcher Form bestimmte Elemente in einer PDF-Datei enthalten sein müssen und welche nicht erlaubt sind. Die erste Version der PDF/A-Norm wurde nachträglich um zwei weitere, aufeinander aufbauende Normteile ergänzt. Wird das PDF/A-Format für die Langzeitarchivierung einer Datei verwendet, die nicht nur textuelle Informationen enthält, sollten nach Möglichkeit auch die ursprünglichen Ausgangsdateien in archivierungstauglichen Formaten entweder als separate Dateien oder integriert in das Container-Format PDF/A-3 mitarchiviert werden.

Format	Begründung
✓ PDF/A-1 ✓ PDF/A-2	PDF/A ist gezielt als stabiles, offenes und standardisiertes Format für die Langzeitarchivierung unterschiedlicher Ausgangsdateien entwickelt worden.
~ PDF/A-3	PDF/A-3 ist nur dann für die Langzeitarchivierung geeignet, wenn alle eingebetteten Dateien in einem anerkannten Archivformat vorliegen.
✗ andere PDF-Varianten	Viele gängige PDF-Varianten sind nicht für die Langzeitarchivierung geeignet. Stattdessen sollten entweder die Ausgangsdateien in einem passenden Format archiviert oder eine Migration in ein PDF/A-Format vorgenommen werden.

Dokumentation Welche Metadaten für ein PDF-Dokument relevant sind, hängt von dessen Inhalt ab. Bei PDF-Dateien, die nur Text enthalten, sind

meist weniger Informationen erforderlich, als bei PDFs, die Bilder, GIS-Karten oder 3D-Modelle enthalten. Die Metadaten können unmittelbar in einem PDF-Dokument im XMP-Format gespeichert werden.

Die hier angegebenen Metadaten sind als minimale Angabe zu betrachten und ergänzen die angegebenen Metadaten für Projekte und Einzeldateien in dem Abschnitt Metadaten in der Anwendung ab Seite 36.

Metadatum	Beschreibung
Titel	Titel des Dokuments, nicht der Dateiname
Autor	Name des Verfassers oder Erstellers der Datei; gegebenenfalls Name der Einrichtung
Stichwörter	Schlagworte, wie z.B. Periode, Fundstelle oder charakteristische Merkmale. Wenn vorhanden, angemessene Thesauri verwenden
Beziehungen	Dateien oder Ressourcen, die mit dem PDF zusammenhängen, insbesondere der Name der Originaldatei, aus der heraus ein PDF erstellt wurde
Anwendung	Programm, mit dem der Inhalt ursprünglich erzeugt wurde
Datum	Datum der Erstellung oder letzten Änderung der Datei
Copyright-Angaben	Angaben zur Person oder Einrichtung, die das Copyright oder die Lizenzrechte an der Datei oder deren Inhalt besitzt
Zusätzliche Metadaten	
Kurzbeschreibung	Kurzbeschreibung über den Inhalt des Dokuments
Sprache	Sofern schriftliche Inhalte vorhanden sind, die Sprache
Verfasser der Metadaten	Name der Person, welche die Metadaten ausgefüllt hat

Weitere Metadaten sind abhängig vom Inhalt und der Methoden und können in den jeweiligen Abschnitten im Kapitel Forschungsmethoden ab Seite 194 nachgelesen werden.

Vertiefung

Eine wesentliche Eigenschaft und Stärke von PDF-Dateien ist, dass sie ganz unterschiedliche Inhalte enthalten können. So lassen sich neben Texten und Navigationshilfen, wie etwa Inhaltsverzeichnisse mit Textankern, auch andere Informationen wie Bilder, Verktorgrafiken, 3D-Visualisierungen, Tabellen, Präsentationen oder GIS-Karten einbinden, die zusammen mit allen ursprünglich verwendeten Layoutmerkmalen gespeichert werden. Daher werden die meisten Publikationen wie elektronische Zeitschriften oder Informationsmaterialien auf Webseiten üblicherweise als PDF-Dateien zur Verfügung gestellt. Bei der Langzeitarchivierung von Dateien spielt das PDF/A-Format eine besondere Rolle. In

dessen Spezifikation wird unter anderem vorgeschrieben, Schriften in das Dokument einzubetten und Farben geräteneutral zu definieren. Auf diese Weise wird gewährleistet, dass das originale Aussehen eines Dokuments wie in der ursprünglichen Anwendung und wie vom Autor intendiert erhalten bleibt. Die Intention von PDF-Dateien ist insofern die Erhaltung der ursprünglichen Darstellung, nicht aber der ursprünglichen Bearbeitungsfunktionalitäten.

Um eine möglichst große Authentizität einer PDF-Datei zu erlangen, sind die Möglichkeiten zur nachträglichen Bearbeitung sehr begrenzt. Es können Kommentare hinzugefügt, Seiten eingebunden oder gelöscht, oder Textstellen geschwärzt werden. Eingeschränkt sind Veränderungen der eigentlichen Inhalte oder des Layouts möglich, wie z. B. Änderungen von Farben in einem Foto oder Umformulierungen von Texten. Als Nachteil ergibt sich daraus, dass die ursprüngliche Funktionalität der nach PDF/A konvertierten Datei verloren geht und eine Nachnutzung der Inhalte wenn überhaupt nur mit Qualitätsverlusten möglich ist. So können beispielsweise tabellarische Daten aus einer PDF-Datei nicht ohne einen zusätzlichen Mehraufwand in eine neue Tabelle übernommen werden, wobei Bearbeitungsfehler auftreten können.

Da aber zunehmend Softwarelösungen existieren, die diese Einschränkungen der nachträglichen Bearbeitung auflösen oder umgehen, wächst der Bedarf an Schutzmechanismen gegen ungewollte Änderungen durch Dritte. Dazu zählen beispielsweise die Festlegung von einfachen Nutzungseinschränkungen, wie etwa die Erlaubnis zum Drucken, zum Kopieren von Inhalten oder zur Bearbeitung der im PDF gespeicherten Metadaten, oder die Vergabe von Passwörtern. Für Dateien, bei denen eine Langzeitarchivierung vorgesehen ist, dürfen allerdings keine Schutzfunktionen verwendet werden, damit sie ohne Hindernisse jederzeit nutzbar sind. Falls notwendig können Einschränkungen als Lizenzhinweise formuliert werden; die Authentizität eines Dokumentes kann durch Signaturen, Zeitstempel und/oder digitale Prüfsummen gewährleistet werden.

Inhalte Dateien im PDF-Format können durch unterschiedliche Anwendungen erzeugt werden und daher ganz unterschiedliche Inhalte enthalten und darstellen. Für jede Dateiart gibt es spezifische Aspekte, die es zu beachten gilt, insbesondere dass einige Inhalte nicht mit den Vorgaben von PDF/A vereinbar sind.

Texte: Standardmäßig sind in PDF-Readern bereits 14 Schriftarten verfügbar. Es lassen sich aber weitere in ein PDF-Dokument einbetten, um eine korrekte Darstellung zu gewährleisten. Mithilfe von OCR-Software kann in PDF-Dokumenten eine nachträgliche Texterkennung durchgeführt werden.

Bilder: Wie bei normalen Rastergrafiken können auch in PDF-Dateien eingebettete Bilder durch spezielle Kompressionsverfahren in ihrer Speichergröße reduziert und in ihrer Auflösung heruntergerechnet werden. Bei fast allen Verfahren ist dies mit einem Informationsverlust verbunden, der vor allem die Genauigkeit und Qualität der Anzeige betrifft.

Vektorgrafiken: Vektorgrafiken in PDF-Dateien können wie die Ausgangsdateien frei skaliert werden.

3D: Es gibt zwei PDF-Formate, in die 3D-Dateien eingebettet und mit Adobe Reader geöffnet und betrachtet werden können. Auf diese Weise lassen sich auch 3D-Objekte austauschen und mit einfachen Funktionen bedienen, z. B. in der Ansicht drehen, Lichteinstellungen verändern oder Aufsichten und Schnitt-

flächen erzeugen.

PDF-Varianten PDF-Dateien sind inzwischen ein sehr weit verbreitetes Austauschformat, das in ganz unterschiedlichen Kontexten Verwendung findet. Um eine gleichbleibende Qualität zu gewährleisten, wurden für die häufigsten Anwendungsfälle mehrere Normen und Standards festgelegt:

- PDF/X – Format für die Übermittlung von Druckvorlagen (seit 2001)
- PDF/A – Format für die elektronische Archivierung (ISO 19005-1 seit 2005)
- PDF/E – Format für technische Dokumente aus den Bereichen Ingenieurwesen, Architektur und Geo-Informationssysteme. Dieses Format besitzt die Fähigkeit zur interaktiven Darstellung von 3D-Objekten. Einzelteile können gedreht, auseinandergezogen, ein- oder ausgeblendet werden (ISO 24517 seit 2008)
- PDF/UA – Richtlinien zum Aufbau eines barrierefreien Dokuments im Format PDF (seit 2008)
- PDF/VT – Format für den Einsatz im hochvolumigen und Transaktionsdruck (seit 2010)

Mehrere PDF-Standards lassen sich miteinander verbinden. Genügt eine Datei dem PDF/X-Standard, eignet sie sich als digitale Druckunterlage. Erfüllt diese Datei zusätzlich den PDF/UA-Standard, kann sie als barrierefreies PDF von Screenreader-Programmen verarbeitet werden.

Das Archivierungsformat PDF/A Von den verschiedenen existierenden PDF-Formaten ist nur das PDF/A-Format ein sicheres und stabiles Archivformat, das speziell für die elektronische Langzeitarchivierung von Dokumenten definiert und publiziert wurde. Die klar festgelegten und als Standard allgemein anerkannten Regeln schreiben vor, welche Bestandteile in einer PDF-Datei enthalten sein müssen und welche nicht zulässig sind, um mit dem PDF/A-Standard konform zu gehen. Während einige Inhaltstypen wie Texte, Bilder, Zeichnungen, Tabellen und Schriften standardkonform in PDF/A-Dateien eingebunden werden können, sind andere Inhalte wie 3D-Objekte, Audio- und Video-Sequenzen, Flash-Animationen oder Scripte nicht erlaubt. Ebenso darf eine PDF/A-Datei keine externen oder dynamischen Inhalte enthalten, die in einem Viewer geladen werden müssen. Beispielsweise werden URL-Links nur als Zeichenkette, nicht aber als aktive Verknüpfung gespeichert. Enthält eine PDF-Datei nicht zulässige Inhalte, kann sie nicht nach PDF/A konvertiert werden. Eine Validierung, ob eine vorliegende Datei dem gültigen PDF/A-Format entspricht, ist über entsprechende Prüfwerkzeuge möglich und wird im Praxisteil auf Seite 59 beschrieben.

Insgesamt sind aktuell drei, aufeinander aufbauende Versionen von PDF/A-Formaten zu unterscheiden, von denen PDF/A-1 und PDF/A-2 echte Archivformate sind, da sie das Dokument in einem stabilen Zustand konservieren. Bei PDF/A-3 handelt es sich hingegen um einen Container, in den beliebige Dateiformate eingebettet werden können. Der große Vorteil liegt also darin, dass

editierbare Originaldateien in das Dokument integriert werden können. Es gibt allerdings keine Garantie, dass die angehängten Dateien auch künftig alle angezeigt werden können. Für die automatische Weiterverarbeitung besteht die Möglichkeit einer Integration von Daten im XML-Format.

Da es bislang keine Vorgabe über die Datenformate gibt, die in eine PDF/A-3-Datei eingebettet werden dürfen – erlaubt sind beispielsweise auch nicht offene, nicht standardisierte und nicht zukunftssichere Formate – hängt die Frage der Archivierbarkeit einer PDF/A-3-Datei vor allem von ihrem Inhalt ab. Es wird daher aktuell diskutiert, ob PDF/A-3 als Format für das Archival Information Package (AIP) nach ISO 14721:2012 OAIS geeignet ist.

Für alle Varianten des PDF/A-Formates gilt, dass Schriften und Zeichen direkt in die Datei eingebettet werden. Ebenso werden Sonderzeichen und Formeln verlässlich dargestellt. Auch Schriftsysteme wie Chinesisch oder Arabisch werden auf allen Geräten richtig wiedergegeben. Für die korrekte Anzeige von Farben werden festgelegte ICC-Profiles verwendet. In PDF/A Dokumenten ist grundsätzlich eine Volltextsuche möglich, auch wenn es sich um ein gescanntes Dokument handelt. Zusätzlich kann auch eine digitale Texterkennung (OCR) integriert werden. Über die vergebenen XMP-Metadaten sind erweiterte Suchoptionen verfügbar.

PDF/A-1 (seit 2005):

- Bilder, Grafiken, verwendete Schriftzeichen etc. müssen im PDF/A-Dokument selbst eingebettet sein
- präzise, plattformunabhängig kodierte Farbangaben mittels ICC-Profilen
- Verwendung von XMP für Dokument-Metadaten
- *keine* transparenten Elemente
- *keine* Kompression
- *keine* PDF-Ebenen, Aktionen und JavaScript
- *kein* Passwortschutz
- Einbettung von Signaturen und Hyperlinks möglich
- Konformitätsstufen a und b

PDF/A-2 (seit 2011):

- identisch zu PDF/A-1 mit zusätzlichen Möglichkeiten:
- erlaubt Kompression mit JPEG2000, transparente Elemente und PDF-Ebenen
- ermöglicht Einbettung von OpenType-Fonts
- unterstützt digitale Signaturen in Übereinstimmung mit den PAdES (PDF Advanced Electronic Signatures)
- Dateianhänge: Einbettung von PDF/A-1 und PDF/A-2 Dateien möglich

- Konformitätsstufen a, b und u

PDF/A-3 (seit 2012)

- identisch zu PDF/A-2 mit zusätzlichen Möglichkeiten:
- Einbettung von beliebigen Dateien möglich (z. B. die Ursprungsdatei aus der ein PDF/A-3 erstellt wurde)
- keine Gewährleistung, dass die Anhänge zu einem späteren Zeitpunkt verarbeitet werden können

Konformitätsstufen bei PDF/A Die Qualität der archivierten Dokumente wird durch verschiedene Konformitätsstufen bezeichnet. Die Stufen sind abhängig von Eingangsmaterial und Verwendungszweck und ihre Bezeichnung wird an das Ende des PDF/A-Formates angehängt, z. B. „PDF/A-1a“.

Stufe b (basic): Das Dokument sieht bei späterer Verarbeitung unverändert aus.

Stufe u (unicode): Das Dokument sieht bei späterer Verarbeitung unverändert aus. Zusätzlich wird der gesamte Text in Unicode abgebildet. Es besteht die Möglichkeit, Text zu durchsuchen und zu extrahieren.

Stufe a (accessible): Das Dokument sieht bei späterer Verarbeitung unverändert aus. Der gesamte Text wird in Unicode abgebildet. Es besteht die Möglichkeit, Text zu durchsuchen und zu extrahieren. Zusätzlich bleiben die Struktur des Dokuments und die natürliche Lesereihenfolge erhalten. Das Dokument ist somit barrierefrei und kann mithilfe von Screenreader-Programmen vorgelesen werden.

Praxis

Für die Erstellung von PDF-Dokumenten gibt es eine Vielzahl an Programmen, die in manchen Fällen auch für die Bearbeitung der Metadaten verwendet werden können. Die Validierung klärt die Frage, ob es sich bei einem vorhandenen PDF-Dokument um eine Datei handelt, die auch PDF/A-konform ist.

Erstellung von PDF-Dokumenten Eine PDF-Datei lässt sich auf vielfältige Art und Weise erzeugen. Beim Betriebssystem Mac OS X ist die Möglichkeit zur Erstellung einer PDF-Datei über den Druckdialog voreingestellt. Windows bietet diese Möglichkeit nicht immer von Haus aus. Deshalb gibt es ein großes Angebot an kommerzieller und freier Software. Diese wird häufig über die

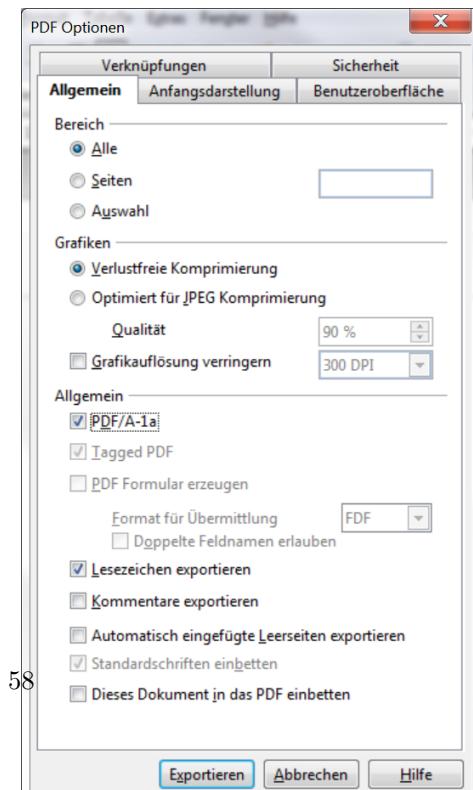


Abb. 3.1: Speichern im PDF/A-Format in OpenOffice

Druckfunktion in die Anwendungen eingebunden, so dass aus beliebigen Programmen heraus über den Druckdialog ein PDF erstellt werden kann. Gescannte Vorlagen lassen sich in der Regel über die Scansoftware direkt als PDF speichern.

Eine weitere Möglichkeit bieten viele Programme über eine Exportschnittstelle. Die Funktionen dazu finden sich in der Regel im Menü unter „Datei > Speichern unter“ oder „Datei > Exportieren“. Als Beispiele von Anwendungssoftware bzw. Produktsuiten mit der Möglichkeit zum direkten PDF-Export sind zu nennen: OpenOffice, Microsoft Office (ab 2010), Adobe Creative Suite, Framemaker und CorelDraw.

Nur in manchen Fällen lassen sich gezielt Dateien im PDF/A-Format erstellen. Diese Option beim Exportieren von Dateien bieten z. B. OpenOffice (Abbildung 3.1) und Microsoft Office oder Systeme, bei denen entsprechende Module zur Erstellung von PDF/A-Dateien installiert wurden. Werden PDF-Dateien über einen Druckdialog erzeugt, sollte individuell überprüft werden, ob die jeweiligen Druckeinstellungen entsprechende Optionen anbieten.

PDF-Dateien, die nicht als PDF/A vorliegen oder sich in diesem Format erzeugen lassen, können auch mithilfe von Programmen umgewandelt werden. Als Teil der Adobe Suite ist das Programm *Adobe Acrobat Professional* weit verbreitet. Es bietet vielfältige Werkzeuge zur Bearbeitung von PDF-Dateien. Dazu gehört auch die Möglichkeit, existierende PDF-Dateien in andere Varianten umzuwandeln. Zusätzlich gibt es ein großes Angebot an Plug-Ins, um die Funktionalitäten zu erweitern, z. B. zur Erstellung von PDFs mit 3D-Inhalten.

PDFCreator:

<http://de.pdfforge.org/>

FreePDF:

http://freepdfxp.de/index_de.html

PDF-XChange Editor:

www.pdf-xchange.de/pdf-xchange-viewer

PDF Split and Merge (einzelne PDFs zusammenfügen und trennen):

www.pdfsam.org

Validierung: Liegt eine PDF-Datei im PDF/A-Format vor? Häufig zeigen Viewer von PDF-Dateien unmittelbar an, ob eine Datei im PDF/A-Standard vorliegt oder diese Angabe ist in den Dateieigenschaften einsehbar. Will man aber sicher gehen, ob eine vorliegende Datei mit der Erweiterung .pdf dem PDF/A-Format entspricht, ist eine Validierung erforderlich. Bei diesem Vorgang werden alle relevanten Bestandteile des Dokuments untersucht und auf PDF/A-Vorgaben getestet. Nur so kann die uneingeschränkte ISO-Konformität gewährleistet werden und entschieden werden, ob eine Datei für die Langzeitarchivierung geeignet ist.

Da PDF/A-Dateien grundsätzlich nicht gegen Bearbeitung geschützt werden können, weil dies gegen die Vorschrift der Zugänglichkeit verstößen würde,

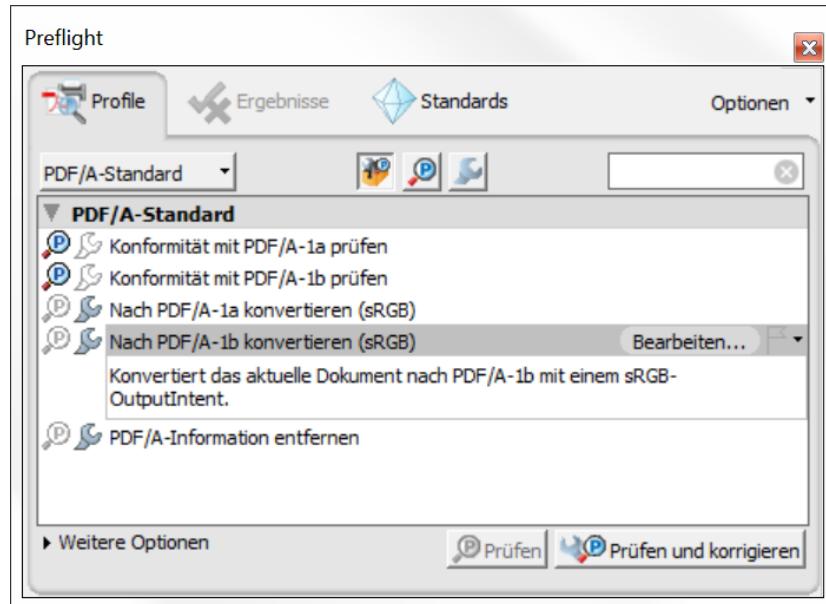


Abb. 3.2: PDF/A Validierung mit Preflight in Adobe Acrobat Professional

muss eine solche Validierung nach jeder Bearbeitung durchgeführt werden, um nachträgliche, nicht-standardkonforme Veränderungen ausschließen zu können.

Viele Firmen haben sich mittlerweile auf die Archivierung von PDF-Dateien spezialisiert und bieten verschiedene Tools zur Überprüfung oder zur Reparatur von nicht validen Dateien an. *Adobe Acrobat Professional* bietet mit dem *Preflight*-Werkzeug (im Menü „Erweitert > Preflight > Profile > PDF/A-Standard prüfen“) ein Tool zur Validierung an. Es ist in Abbildung 3.2 dargestellt. Auch mit kostenfreier Software lassen sich Validierungen durchführen. Mit PDF/A Live! 6.0 kann man ohne Lizenzschlüssel bis zu 50 Dokumente pro Tag überprüfen und bis zu 15 Dokumente konvertieren. Online besteht die Möglichkeit über den Free PDF/A Validator.

Isartor Test Suite:

www.pdfa.org/2011/08/download-isartor-test-suite

PDF/A Live! 6.0:

www.intarsys.de/pdf-produkte/pdfa-live

Free PDF/A Validator:

www.validatepdfa.com/online.htm

Bearbeitung von Metadaten Metadaten lassen sich im kostenlosen *Adobe Reader* unter dem Menü „Datei > Eigenschaften“ anzeigen und bearbeiten, wie Abbildung 3.3 verdeutlicht. Zu den hier aufgeführten Angaben für die inhaltliche Kurzbeschreibung gehören vier Felder: Titel, Verfasser, Thema und Stichwörter. Einige technische Informationen wie z. B. Dateigröße, Zeitstempel,

Programmversion oder Seitenformat werden bei der Speicherung automatisch erzeugt und können hier ebenfalls eingesehen werden. Nur in *Acrobat Professional* werden die vollständigen XMP Metadaten angezeigt, die in einem weiteren Reiter um zusätzliche Informationen, wie Beschreibung, Verfasser der Beschreibung oder Copyright-Angaben, ergänzt werden können. Diese zusätzlichen Metadaten sind in Abbildung 3.4 zu sehen.

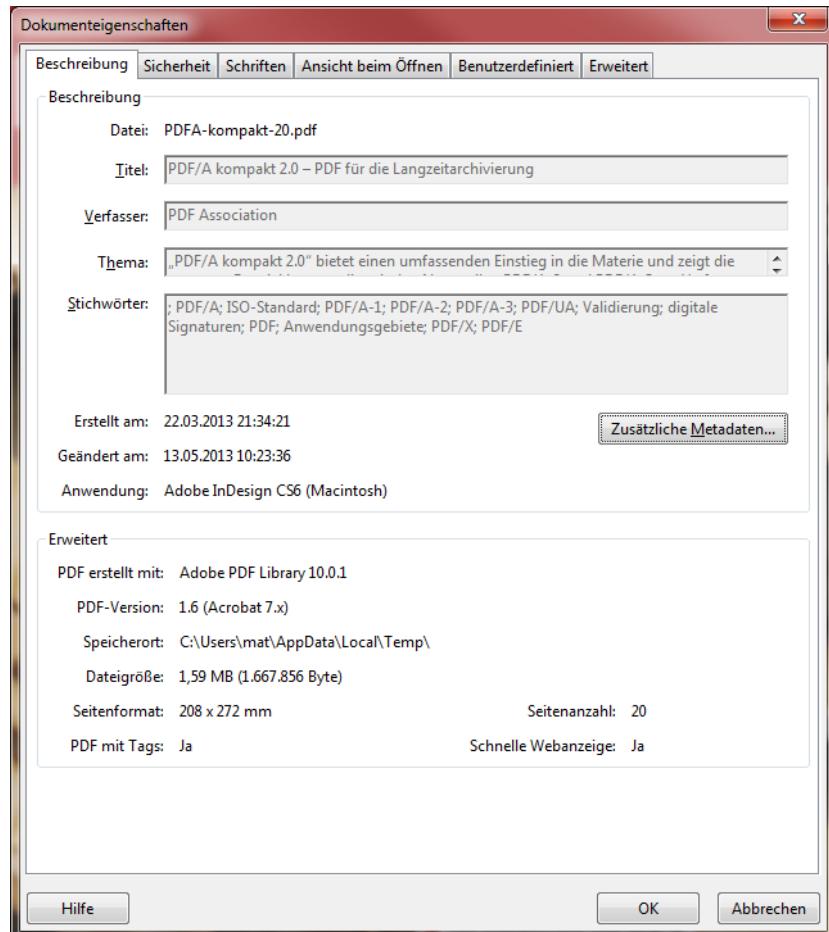


Abb. 3.3: Metadaten unter „Dokumenteneigenschaften“ in Adobe Acrobat Professional

Neben Adobe Acrobat gibt es weitere Editoren, mit denen sich Metadaten für ein PDF vergeben oder nachträglich ändern lassen. Als uneingeschränkte kostenlose Anwendungen sind für Windows BeCyPDFMetaEdit und für Mac PDFInfo erhältlich. Mit Einschränkungen gibt es für Windows den Hexonic PDF Metadata Editor, der kostenfrei eine Verarbeitung von maximal fünf PDF-Dateien gleichzeitig erlaubt. Das Metadata Extraction Tool kann mit Metadaten von vielen verschiedenen Formaten umgehen.

BeCyPDFMetaEdit:

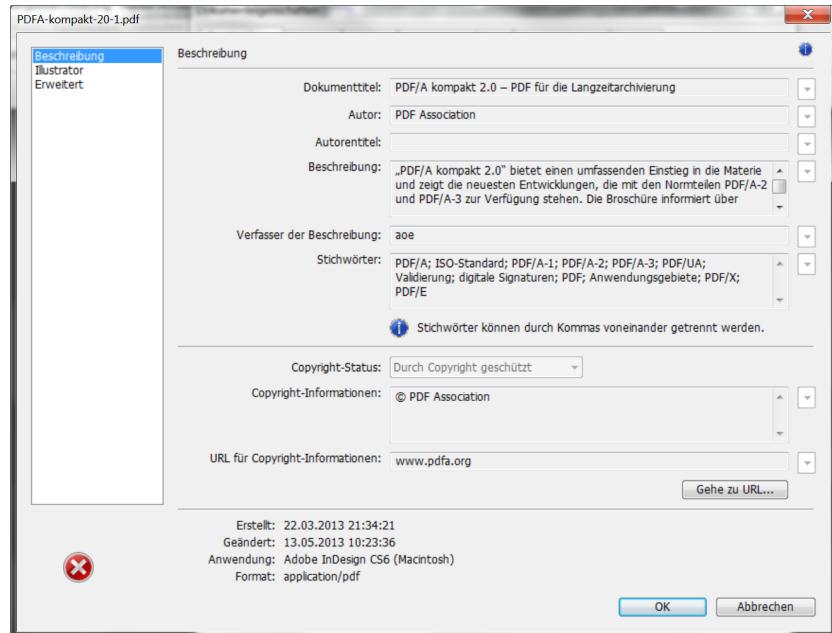


Abb. 3.4: Zusätzliche Metadaten unter „Dokumenteigenschaften“ in Adobe Acrobat Professional

www.becyhome.de/becypdfmetaedit/description_ger.htm

PDFInfo:

www.macupdate.com/app/mac/23356/pdfinfo

Hexonic PDF Metadata Editor:

www.hexonic.de/index.php/hexonic-pdf-metadata-editor

Metadata Extraction Tool:

<http://meta-extractor.sourceforge.net/>

Quellen

H. Bärfuss – M. Winkler, Pflicht und Kür in der Langzeitarchivierung, DOK Technologien, Strategien & Services für das digitale Dokument 5, 2013, 48-51
www.pdfa.org/wp-content/uploads/2013/11/DOK.magazin-2013-11-Compliance-DE.pdf

O. Drümmer – A. Oettler – D. von Seggern, PDF/A kompakt. Digitale Langzeitarchivierung mit PDF (Berlin 2007)
www.pdfa.org/wp-content/uploads/2011/08/PDFA_kompakt_pdfa1b.pdf

H.-J. Hübner, PDF/A – Update und Dienstleistungen für die Langzeitarchivierung, Präsentation vom 18.06.2013 auf dem Nestor-Praktikertag in Berlin
<http://files.dnb.de/nestor/veranstaltungen/Praktikertag2013/2013-06-pdfa-huebner.pdf>

K. Krieg, PDF/A und Metadaten: XML, RDF und XMP: Was, wie und

- woher? Präsentation vom 18.11.2008 beim Nestor-Seminar in Koblenz
http://nestor.sub.uni-goettingen.de/seminar_bundesarchiv/docs/04-Metadaten-KK.pdf
- A. Mantke, OpenOffice.org. Export in das PDF-Dateiformat (2010)
www.openoffice.org/de/doc/howto_2_0/office/exportpdfformat.pdf
- A. Oettler, PDF/A kompakt 2.0. PDF für die Langzeitarchivierung. Der ISO-Standard – von PDF/A-1 bis PDF/A-3 (Berlin 2013)
www.pdfa.org/wp-content/uploads/2013/05/PDFA-kompakt-20.pdf
- PDF/A Competence Center (Hrsg.) TechNote 0008: Predefined XMP Properties in PDF/A-1 (2008)
www.pdfa.org/wp-content/uploads/2011/08/tn0008_predefined_xmp_properties_in_pdfa-1_2008-03-20.pdf
- PDF/A Competence Center (Hrsg.) PDF/A Metadaten XMP, RDF & Dublin Core (2011)
<http://www.pdfa.org/2011/09/pdfa-metadaten-xmp-rdf-dublin-core/>
- D. von Seggern, PDF/A und XMP Metadaten: Nutzen und Anforderungen, Präsentation vom 8. Juli 2009 zum Webinar
www.pdfa.org/wp-content/uploads/2011/09/callas_webinar_xmp_deu.pdf
- Adobe Acrobat DC. Adobe PDF
www.adobe.com/de/products/acrobat/adobepdf.html
- Open Format. Archiving of 3D Documents
www.newformat.se/openformat/eng/odf-and-pdf-details/3d-archiving-pdfa-pdf-e-english-popup.html
- PDF/A Flyer (2014)
www.pdfa.org/wp-content/uploads/2014/01/Flyer-PDFA-DEU.pdf
- PDF Association: PDF Competence Center
www.pdfa.org/pdf-assocation
- PDFlib: Die PDF/A-Archivierungsstandards
www.pdflib.com/de/knowledge-base/pdfa

Formatspezifikationen

PDF/A-1 (ISO 19005-1:2005):

http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

PDF/A-2 (ISO 19005-2:2011):

http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50655

PDF/A-3 (ISO 19005-3:2012):

http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57229

Tools und Programme

Adobe Reader:

<http://get.adobe.com/de/reader>

Foxit Reader:

http://www.foxitsoftware.com/Secure_PDF_Reader

Ghostscript, Ghostview and GSview:

http://pages.cs.wisc.edu/~ghost

PDFCreator:
http://de.pdfforge.org

FreePDF:
http://freepdfxp.de/index_de.html

PDF-XChange Editor:
www.pdf-xchange.de/pdf-xchange-viewer

PDF Split and Merge:
www.pdfsam.org

Isartor Test Suite:
www.pdfa.org/2011/08/download-isartor-test-suite

Free PDF/A Validator:
www.validatepdfa.com/online.htm

PDF/A Live! 6.0:
www.intarsys.de/pdf-produkte/pdfa-live

BeCyPDFMetaEdit:
www.becyhome.de/becypdfmetaedit/description_ger.htm

PDFInfo:
www.macupdate.com/app/mac/23356/pdfinfo

Hexonic PDF Metadata Editor:
www.hexonic.de/index.php/hexonic-pdf-metadata-editor

Metadata Extraction Tool:
http://meta-extractor.sourceforge.net/

3.2 Textdokumente

M. Trognitz

Textdokumente stellen in der altertumswissenschaftlichen Forschung einen häufig vertretenen Dateityp dar. In Artikeln, Berichten, Anträgen, Tagebüchern, Notizen oder Dokumentationen sind wichtige Informationen enthalten, deren Fortbestehen und Lesbarkeit gewährleistet sein müssen. Auch Beschreibungen von anderen Dateien und Datensätzen oder des gesamten Projektes können als Textdokumente vorliegen.

Die Mehrzahl der Dokumente besteht aus strukturiertem Text, nämlich Sätzen, Absätzen, Seiten, Fußnoten und Kapiteln, und kann Formatierungsangaben, wie verschiedene Schriftgrößen, Fett- oder Kursivschreibung enthalten. Zusätzlich können Medien, wie Bilder, Tabellen oder Videos in die Dokumente integriert sein.

Da dasselbe Dokument auf verschiedenen Systemen unterschiedlich dargestellt werden kann, kann die Speicherung von Textdokumenten problematisch sein. Insbesondere wenn bestimmte Formatierungen von Textelementen mit einer Bedeutung verbunden sind und die Authentizität des Erscheinungsbildes, also das Aussehen des Dokumentes wichtig ist, ist bei der Speicherung besondere Aufmerksamkeit erforderlich.

Langzeitformate Finalisierte Dokumente mit Formatierungsangaben können im Format PDF/A gespeichert werden. Dieses Format erlaubt eine konsistente Darstellung des Dokumentes auf verschiedenen Systemen, verhindert aber auch eine nachträgliche Bearbeitung. Nähere Informationen sind im Abschnitt über PDF-Dokumente ab Seite 53 zu finden.

Textdokumente mit Formatierungsangaben, bei denen auch weiterhin eine Bearbeitung möglich sein soll, sollten in einem offenen auf XML basierenden Format gespeichert werden, wie beispielsweise DOCX oder ODT. Ersteres ist das Standardformat, das in Microsoft Word seit 2007 verwendet wird und auch von Microsoft entwickelt wurde. Letzteres ist das Format für Textdokumente, welches in OpenOffice oder LibreOffice verwendet wird. ODT ist ein Teil vom OpenDocument Format (ODF) und wurde von einem technischen Komitee unter der Leitung der *Organization for the Advancement of Structured Information Standards* (OASIS) entwickelt. Die Darstellung von DOCX- oder ODT-Dokumenten kann jedoch von System zu System unterschiedlich ausfallen, wenn beispielsweise bestimmte Schriftarten fehlen. Gegebenfalls kann das Dokument parallel im Format PDF/A gespeichert werden.

Bearbeitbare Textdokumente ohne Formatierungsangaben werden am besten als TXT-Datei gespeichert. Neben diesem einfachen, reinen Textformat (*plain text*) gibt es weitere textbasierte Formate, die auf eine bestimmte Weise strukturiert sind oder eine Auszeichnungssprache verwenden. Es handelt sich dabei um sogenannte Textdateien, die im Gegensatz zu binären Formaten darstellbare Zeichen enthalten und in Abhängigkeit ihrer Strukturierung unterschiedliche Dateiformate beschreiben. Beispielsweise werden mit Hilfe von CSV-Dateien Tabellen oder mit PLY-Dateien 3D-Inhalte gespeichert. Diese Formate werden in den Abschnitten „Tabellen“ und „3D und Virtual Reality“ ab Seite 97 bzw. 159 behandelt. Für textuelle Inhalte gibt es spezialisierte Formate, wie beispielsweise SGML-, XML- oder HTML-Dateien. Die Archivierung dieser Dateien, die

weit verbreiteten Konventionen folgen, ist unproblematisch, bedarf jedoch zusätzlicher Dateien, die die verwendete Struktur beschreiben, wie beispielsweise die sogenannte Dokumenttypdefinition (DTD, Document Type Definition) oder ein XML Schema (XSD, XML Schema Definition). Auch andere Textdateien mit spezieller Strukturierung können archiviert werden, wenn die Struktur im Dokument oder in einer separaten Datei erläutert und mitarchiviert wird.

Alle Textdokumente sollten Unicode für die Zeichenkodierung verwenden, wobei UTF-8 ohne BOM besonders empfohlen wird, falls keine speziellen Anforderungen dagegen sprechen. Wenn die verwendete Zeichenmenge es erlaubt, ist ASCII ebenfalls geeignet.

Hinweis: Eingebettete Bilder oder andere Medien sollten zusätzlich separat gespeichert werden. Außerdem muss beachtet werden, dass Links oder dynamische Inhalte, nicht immer dauerhaft erhalten bleiben.

Format	Begründung
PDF/A ✓	Wenn neben dem Inhalt auch das Aussehen des Dokumentes erhalten bleiben soll und die Bearbeitung des Dokuments abgeschlossen ist, eignet sich PDF/A am besten. Nähere Informationen sind in dem Abschnitt über PDF-Dokumente ab Seite 53 zu finden.
ODT	ODT basiert auf XML und ist Teil vom Open-Document Format. Damit können bearbeitbare Dokumente mit Formatierungssangaben gespeichert werden. ODF verwendet standardmäßig UTF-8 und erlaubt das Einbetten von Fonts.
DOCX	DOCX ist das auf XML basierende Format von Microsoft, das ebenfalls bearbeitbare Dokumente mit Formatierungssangaben speichern kann. DOCX verwendet standardmäßig UTF-8 und erlaubt das Einbetten von TrueType-Fonts.
TXT und <i>plain text</i>	Das Format eignet sich für reinen Text ohne Formatierungssangaben, wie Kursiv-, Fettschreibung oder Schriftgrößen. Die Zeichen sollten in UTF-8 ohne BOM kodiert sein.
strukturierter Text	Alle anderen textbasierten Formate, wie beispielsweise valide SGML-, XML- oder HTML-Dateien können ebenfalls archiviert werden. Für SGML und XML ist zusätzlich die DTD-Datei oder ein XML Schema erforderlich. Anders strukturierte textbasierte Dateien benötigen eine Erläuterung der Struktur innerhalb der Datei oder als zusätzliche separate Datei. Die Zeichen sollten in UTF-8 ohne BOM kodiert sein.
RTF ~	RTF ist ein proprietäres Format von Microsoft für den Datenaustausch, das von vielen Programmen unterstützt wird. Wegen möglichen Kompatibilitätsproblemen sollte DOCX oder ODT bevorzugt werden.
SXW	SXW ist ein Vorgängerformat von ODT, weshalb letzteres auch bevorzugt werden sollte.
DOC ✗	Das DOC-Format von Microsoft eignet sich nicht zur Archivierung, da es proprietär ist und die Inhalte nicht textbasiert gespeichert werden.
PDF	Für die Archivierung wurde speziell das Format PDF/A entwickelt, weshalb dieses verwendet werden sollte.

Dokumentation Metadaten für Textdokumente können in vielen Fällen direkt in das Dokument eingetragen werden. Beispielsweise als Deckblatt oder

in dafür vorgesehenen Teilen von strukturierten Dokumenten. Zusätzlich können einige Informationen als Dokumenteigenschaften in der Datei gespeichert werden.

Neben den allgemeinen Angaben zu Einzeldateien, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 gelistet sind, benötigen Textdokumente insbesondere Angaben zur verwendeten Zeichenkodierung und eine Auflistung der Sprachen.

Falls das Dokument publiziert wurde und eine ISBN oder einen anderen persistenten Identifikator erhalten hat, müssen diese neben den allgemeinen Angaben zur Publikation ebenfalls angegeben werden. Eingebettete Medien, wie Bilder oder Tabellen mit Formeln, sollten separat gespeichert und archiviert werden und in einer Liste weiterer Dateien aufgeführt werden.

Wenn das Aussehen wichtig ist und ein Format verwendet wird, welches das Einbetten von Schriftarten nicht ermöglicht, müssen die verwendeten Schriftarten explizit genannt werden.

Die hier angegebenen Metadaten sind als minimale Angabe zu betrachten und ergänzen die angegebenen Metadaten für Projekte und Einzeldateien in dem Abschnitt Metadaten in der Anwendung ab Seite 36.

Metadatum	Beschreibung
Zeichenkodierung	Welches Zeichenkodierung wird verwendet?
Sprache	In welchen Sprachen ist das Dokument verfasst? Sprachkennungen nach ISO 639 angeben.
Identifikator	Wenn das Dokument bereits veröffentlicht wurde und eine ISBN oder einen anderen persistenten Identifikator erhalten hat, sollte dieser angegeben werden.
weitere Dateien	Liste von eingebetteten Medien, die zusätzlich separat gespeichert wurden. Liegt eine Dokumentationsdatei für das Dokument vor, muss diese ebenfalls genannt werden.
Schriftarten	Angabe der verwendeten Schriftarten (Fonts), für Dokumente ohne eingebettete Fonts.

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Vertiefung

Textdokumente und Textdateien bestehen aus einer Folge von Zeichen, die Wörter, Sätze und Absätze bilden. Auf Maschinenebene werden diese Zeichen durch Zahlenwerte gespeichert und die Zeichenkodierung beschreibt, welcher Zahlenwert für welches Zeichen steht.

Wie ein Zeichen dargestellt wird, hängt von der verwendeten Schriftart, dem sogenannten Font ab, der einen Satz an Bildern für die verschiedenen Schriftzeichen bereitstellt.

Die Inhalte von Textdateien können durch die Verwendung einer Auszeichnungssprache strukturiert und beschrieben werden und somit auch eine maschinelle Verarbeitung ermöglichen.

Zeichenkodierung und Zeichensatz Zur korrekten Darstellung der Zeichen in einem Textdokument muss der Computer wissen, welche Zeichenkodierung (*encoding*) verwendet wird. Auf Maschinenebene wird ein Zeichen als eine Folge von Nullen und Einsen, in Form von Bytes gespeichert, die wiederum bestimmte Zahlenwerte angeben. Diese Zahlenwerte können in Abhängigkeit der Zeichenkodierung unterschiedlich interpretiert werden.

Eine Zeichenkodierung kann abstrakt als eine Tabelle verstanden werden, in der einer bestimmten Zeichenmenge, dem Zeichensatz, Zahlenwerte zugeordnet werden. Beispielsweise hat der Buchstabe *A* in dem *American Standard Code for Information Interchange* (ASCII) den dezimalen Zahlenwert von 65. Der ASCII-Zeichensatz besteht aus insgesamt 128 Zeichen die jeweils mit einem Byte gespeichert werden. Er enthält keine diakritischen Zeichen, wie etwa ä, oder gar andere Schriften, weshalb verschiedene Erweiterungen der ASCII-Kodierung entwickelt wurden, um insgesamt 256 verschiedene Zeichen zu kodieren.

Beispiele für diese Erweiterungen sind ISO 8859-1 für lateinische Schriften oder ISO 8859-7 für das griechische Alphabet. In beiden Zeichenkodierungen hat das Zeichen *A* jeweils den Wert 65. Jedoch stellt der Wert 228 in ISO 8859-1 das Zeichen ä und in ISO 8859-7 das Zeichen δ dar. Die Angabe der verwendeten Zeichenkodierung ist entscheidend dafür, ob auf dem Bildschirm öå÷íç oder τεχνη dargestellt wird.

In der Vergangenheit war es besonders schwierig, wenn in einem Text gleichzeitig Umlaute und griechische Buchstaben verwendet werden sollten, da jede ASCII-Erweiterung jeweils nur insgesamt 256 Zeichen kodiert und einem Dokument nicht mehr als eine Zeichenkodierung zugewiesen werden kann. Deshalb wurde Unicode entwickelt.

Unicode ist ein Zeichensatz, in dem aktuell für 113.021 Zeichen aus 123 Schriftsystemen eindeutige Codepunkte (*code points*) zugewiesen werden. Die Codepunkte werden mittels einer hexadezimalen Zahl und einem vorangestellten *U+* dargestellt, wie beispielsweise *U+00C4* für ä. Zugleich stellt dieser Zeichensatz die Umsetzung von dem in ISO 10646 beschriebenen universellen Zeichensatz *Universal Character Set* dar.

Um den Unicode-Zeichensatz in einem System anwenden zu können, wurden Zeichenkodierungen definiert, die unter dem Namen *Unicode Transformation Format* (UTF) subsumiert werden. Zu den häufigsten gehören dabei UTF-8 und UTF-16, die im Web und in verschiedenen Betriebssystemen eine große Verbreitung gefunden haben. Der Unterschied besteht dabei in der Zahl der pro Zeichen verwendeten Bytes. Eine Besonderheit von UTF-8 besteht darin, dass die Bytedarstellungen der ersten 128 Zeichen denen der 128 Zeichen des ASCII-Zeichensatzes entspricht.

Das Unicode-Zeichen *U+FEFF* gibt am Anfang des kodierten Dokumentes an, in welcher Reihenfolge die Bytes angeordnet sind. Diese Bytereihenfolge-Markierung (engl. *byte order mark*) wird als *BOM* abgekürzt und ist bei der Verwendung von UTF-16 und UTF-32 zwingend in der Datei erforderlich. Zusätzlich kann das BOM ein Hinweis auf die Verwendung von UTF-Kodierungen sein, jedoch wird von dessen Verwendung außer für UTF-16 und UTF-32 abgeraten.

Schriftart Das optische Erscheinungsbild eines Textdokumentes hängt vorwiegend von den verwendeten Schriftarten (Fonts) ab. Es handelt sich dabei um

die elektronische Form von Schriftarten, die für jedes Zeichen eine Raster- oder Vektorgrafik zur Verfügung stellt.

Nicht auf jedem Rechner sind die gleichen Schriftarten installiert. Wenn ein Textdokument auf einem anderen System geöffnet wird, wo die Schriftarten nicht verfügbar sind, werden diese automatisch durch andere ersetzt. Das kann zu Inkonsistenzen der Dokumentdarstellung auf unterschiedlichen Systemen führen, weil beispielsweise Wörter, Sätze oder Absätze von einer Seite auf die nächste oder vorhergehende wandern, was für die Referenzierung von Inhalten problematisch ist.

Daher muss für Dokumente, deren optischer Eindruck erhalten bleiben soll, zumindest der verwendete Font in den Metadaten angegeben werden. Wenn es das Format erlaubt, kann der Font auch in die Datei eingebettet werden, was im Praxisteil ab Seite 72 erläutert wird.

Auszeichnungssprachen Der Inhalt von reinen Textdateien kann durch die Verwendung von Auszeichnungssprachen (*Markup Languages*) näher beschrieben werden. Beispielsweise können verschiedene Gliederungsebenen mit Hilfe von bestimmten Auszeichnungselementen (auch *Tags*) annotiert werden. Wie diese Tags aussehen und wie sie angewendet und kombiniert werden können, beschreibt eine Dokumentgrammatik.

Abstrakt können Tags mit Etiketten verglichen werden, die einzelne Wörter, Wortgruppen oder ganze Textbereiche umschließen. Abbildung 3.5 veranschaulicht, wie mit einem Tag die Zeichenkette „24-28“ als Größenangabe etikettiert wird. Das Tag besteht aus einem öffnenden Teil vor und einem schließendem Teil nach der fraglichen Zeichenkette, wobei das schließende Element zusätzlich durch einen Schrägstrich gekennzeichnet ist.

Mit Hilfe von Auszeichnungssprachen wird das Aussehen eines Textdokumentes von dessen Struktur und Inhalt getrennt. Beispielsweise basieren Webseiten auf HTML-Dateien in denen Überschriften, Absätze, Links etc. mit Tags gekennzeichnet werden, die den Inhalt strukturieren. Wie dann beispielsweise die Überschriften formatiert werden, hängt von einer zusätzlichen Datei mit Formationsangaben ab, die austauschbar ist.

<groesse>24-28</groesse>

Abb. 3.5: Die Zeichenkette „24-28“ wird durch das Umschließen mit einem Tag als Größenangabe gekennzeichnet.

Die Grundlage vieler heute verwendeter Auszeichnungssprachen bildet die Standard Generalized Markup Language (SGML, Normierte Verallgemeinerte Auszeichnungssprache), die seit 1986 ein ISO-Standard (ISO 8879) ist. Die Regeln für die zu verwendenden Auszeichnungselemente und deren Kombinationsmöglichkeiten sind üblicherweise in einer externen Datei hinterlegt und werden zu Beginn der Datei in der Dokumenttypdeklaration angegeben. Bei SGML handelt es sich dabei um die sogenannte Dokumenttypdefinition (DTD).

Eine Anwendung von SGML ist die Hypertext Markup Language (HTML, Hypertext-Auszeichnungssprache), welche als Grundlage von Webseiten eine sehr große Verbreitung gefunden hat. HTML wird vom World Wide Web Consortium (W3C) und der Web Hypertext Application Technology Working Group (WHATWG) gepflegt und entwickelt. Die aktuellste Version ist HTML5.

Eine Teilmenge von SGML bildet die Extensible Markup Language (XML,

Erweiterbare Auszeichnungssprache) und erlaubt im Gegensatz zu HTML die Definition von eigenen Auszeichnungselementen, um beliebige Strukturen annotieren zu können. De facto wurde SGML von der einfacher anwendbaren XML verdrängt. Auch XML wird vom W3C gepflegt und entwickelt. XML bildet die Grundlage von vielen weiteren Dateiformaten wie ODT, DOCX, SVG etc. Für XML-Dateien gibt es als Alternative zu einer DTD die Möglichkeit der Verwendung eines XML Schemas (XSD, XML Schema Definition).

Auszeichnungssprachen kennzeichnen implizite Informationen, die nur für den menschlichen Leser verständlich sind, explizit. Dadurch wird ein Dokument maschinenlesbar und eine automatische Verarbeitung von semantisch annotierten Informationen in Texten möglich. Beispielsweise kann eine Münze mit Tags beschrieben werden, die das Material, das Gewicht, die Größe, den Avers und Revers kennzeichnen. So weiß auch ein Computerprogramm, welche Zeichenfolge in einer Datei sich auf das Material oder das Gewicht einer Münze bezieht.

Speziell für die Geistes-, Sozial- und die Sprachwissenschaften wird von der Text Encoding Initiative (TEI) ein auf XML basierendes Dokumentenformat entwickelt, das den Austausch von maschinenlesbaren Texten unterstützen und standardisieren soll. Die aktuelle Version ist P5.

Es gibt weitere Auszeichnungssprachen, die speziell die Darstellung der Dokumente beschreiben, also definieren, wie ein Dokument auf dem Bildschirm oder gedruckt aussehen soll. Beispiele hierfür sind das Textsatzsystem TeX mit dem Makropaket LATEX, PDF oder PostScript.

Es ist erforderlich, dass alle Dateien, die Auszeichnungssprachen verwenden, wohlgeformt und valide sind. Wohlgeformt meint das Einhalten der Regeln der jeweiligen Auszeichnungssprache. Die Validität bezieht sich auf die verwendete Grammatik und gilt insbesondere für SGML-, HTML- und XML-Dateien. Beispielsweise muss eine XML-Datei einen Verweis auf eine DTD oder ein XML Schema enthalten und auch die dadurch vorgegebene Struktur einhalten, um als valide zu gelten.

Praxis

Dieser Abschnitt liefert Hinweise zum Umgang mit Textdokumenten und Textdateien in der Praxis. Es wird erläutert, was bei der Speicherung von Text-



```
<objekt typ="muenze">
  <material>Silber</material>
  <gewicht einheit="gramm">16,96</gewicht>
  <groesse einheit="mm">24-28</groesse>
  <avers>Kopf der Athena</avers>
  <revers>AΘE. Eule</revers>
</objekt>
```

Abb. 3.6: Tetradrachme; Objektnummer 18214973 Münzkabinett – Staatliche Museen zu Berlin, Lizenz: CC-BY-NC-SA 3.0 mit einer Beschreibung in XML-Form. Das Material, das Gewicht, die Größe, der Avers und Revers sind mit Tags gekennzeichnet. Zusätzlich ist die Maßeinheit von Gewicht und Größe als Attribut angegeben.

dokumenten mit Formatierungsangaben zu beachten ist und wie Schriftarten eingebettet werden können. Speziell für Textdateien werden Texteditoren und das Einstellen der Zeichenkodierung thematisiert. Auch Hinweise zur Ergänzung und Extraktion von Metadaten werden gegeben. Für die Digitalisierung von Texten wurden die wichtigsten Informationen aus den DFG-Praxisregeln „Digitalisierung“ zusammengefasst.

Textdokumente mit Formatierungsangaben Textdokumente mit Formatierungsangaben, wie verschiedene Schriftgrößen, Fett- oder Kursivschreibung, oder in welche zusätzlich Medien, wie Bilder, Tabellen oder Videos integriert sind, erfordern eine besondere Aufmerksamkeit bei der Speicherung. Das gilt insbesondere wenn bestimmte Formatierungen von Textelementen mit einer Bedeutung verbunden sind und die Authentizität des Erscheinungsbildes, also das Aussehen des Dokumentes, wichtig ist, denn dasselbe Dokument könnte auf verschiedenen Systemen unterschiedlich dargestellt werden.

Für die Bearbeitung von Textdokumenten mit Formatierungsangaben und eingebetteten Medien gibt es dezidierte Textverarbeitungsprogramme, wie OpenOffice Writer, LibreOffice Writer oder Microsoft Word. OpenOffice und LibreOffice speichern Textdokumente standardmäßig im ODT-Format. Seit 2007 speichert Microsoft Word im DOCX-Format. Beide Formate sind offen dokumentiert, basieren auf XML und sind für die Langzeitarchivierung geeignet. In allen genannten Programmen ist die Zeichenkodierung bereits auf UTF-8 voreingestellt.

Eingebettete Bilder oder andere Medien sollten zusätzlich als separate Dateien in einem geeigneten Langzeitformat gespeichert werden. Dies stellt sicher, dass die Qualität der ursprünglichen Datei erhalten bleibt.

Die Darstellung von Textdokumenten kann auf verschiedenen Computern unterschiedlich ausfallen, was vor allem an unterschiedlichen Einstellungen liegt. Wenn bestimmte Schriftarten auf einem System fehlen, werden sie automatisch ersetzt, was ebenfalls zu unterschiedlichen Darstellungsweisen führt. Daher sollten nach Möglichkeit die verwendeten Schriftarten eingebettet werden, was im nächsten Unterabschnitt erläutert wird.

Eine stabile systemübergreifende Darstellung von Textdokumenten kann nur mittels Konvertierung in ein PDF-Dokument gewährleistet werden. Für die Langzeitspeicherung sollte PDF/A verwendet werden. Hinweise zum Erstellen von PDF- und PDF/A-Dokumenten sind im Praxisteil zu PDF-Dokumenten ab Seite 58 zu finden.

OpenOffice Writer:

<https://www.openoffice.org/>

LibreOffice Writer:

<http://www.libreoffice.org/>

Einbettung von Schriftarten Da das optische Erscheinungsbild eines Textdokumentes unter anderem von den verwendeten Schriftarten abhängt, kann die Einbettung derselben ratsam sein. Dabei muss darauf geachtet werden, dass die Lizenzen für die verwendeten Fonts vorhanden sind.

Ab Version 4.1 können in LibreOffice die benutzten Fonts in das ODT-Format eingebettet werden. Dazu im Menü auf „Datei > Eigenschaften“ ge-

hen, in dem Dialog den Reiter „Schriftart“ anwählen und dort den Haken bei „Schriftarten ins Dokument einbetten“ setzen. Dieser Vorgang muss für neue oder andere Dokumente wiederholt werden.

Auch in Microsoft Word ist diese Einstellung für das DOCX-Format möglich. Dazu auf „Datei > Optionen“ gehen, in dem Dialog den Punkt „Speichern“ auf der linken Seite auswählen und einen Haken bei „Schriftarten in der Datei einbetten“ setzen. Diese Einstellung ist ebenfalls nur für das aktuelle Dokument gültig und muss bei anderen Dokumenten bei Bedarf wiederholt werden.

Werden Textdokumente als PDF exportiert, so werden die verwendeten Schriftarten automatisch eingebettet. Aktuell funktioniert die Einbettung von Fonts in andere Dateiformate als PDF nicht völlig fehlerfrei.

Texteditoren und Editoren für Auszeichnungssprachen Für die Bearbeitung von Textdateien wie TXT, XML oder HTML sind einfache spezialisierte Texteditoren am besten geeignet. In den verschiedenen Betriebssystemen ist üblicherweise mindestens ein Texteditor vorinstalliert, wie beispielsweise *Editor* oder *Notepad* bei Microsoft Windows. Im Vergleich zu Textverarbeitungsprogrammen ist der Funktionsumfang bei Texteditoren deutlich kleiner, was bei reinen Textdateien aber kein Nachteil ist.

Gerade für den täglichen Umgang mit Textdateien empfiehlt sich die Verwendung von leistungsfähigen Editoren, die neben ausgefeilten Suchfunktionen auch Auto vervollständigung oder für Auszeichnungssprachen Syntaxhervorhebung bieten. Für Mac OS X gibt es beispielsweise TextWrangler und für Windows Notepad++ als kostenlose Angebote. Eine umfangreiche vergleichende Liste von Texteditoren ist auf Wikipedia zu finden.

Für den regelmäßigen Umgang mit einem bestimmten Format, wie etwa HTML oder XML, können weiter spezialisierte Editoren praktisch sein.

Notepad++:

<http://www.notepad-plus-plus.org/>

TextWrangler:

<http://www.barebones.com/products/textwrangler/>

Vergleich von Texteditoren auf Wikipedia:

http://en.wikipedia.org/wiki/Comparison_of_text_editors

Einstellen der Zeichenkodierung Wenn keine besonderen Anforderungen dagegen sprechen, sollte Unicode für die Zeichenkodierung verwendet werden. Dabei sollte UTF-8 ohne BOM bevorzugt werden.

In modernen Textverarbeitungsprogrammen, die DOCX oder ODT speichern, ist dies für die genannten Formate voreingestellt und muss nicht explizit angepasst werden.

Bei der Bearbeitung von Textdateien mit Texteditoren muss auf die richtigen Einstellungen und Speicheroptionen geachtet werden. Insbesondere wenn eine Datei auf verschiedenen Geräten bearbeitet wird, ist es wichtig, dass die ursprünglichen Dateieinstellungen, wie eben die Zeichenkodierung, beibehalten werden.

In *Notepad++* kann für alle neuen Dateien eine Zeichenkodierung vorgegeben werden. Dazu im Menü auf „Einstellungen > Optionen“ klicken und unter „Neue Dateien“ die gewünschte Kodierung auswählen. Wird eine vorhandene

Textdatei mit *Notepad++* geöffnet und bearbeitet, werden beim Speichern die ursprünglichen Einstellungen der Datei üblicherweise beibehalten. Die Kodierung einer vorhandenen Datei kann über den Menüpunkt „Kodierung > Konvertiere zu...“ geändert werden.

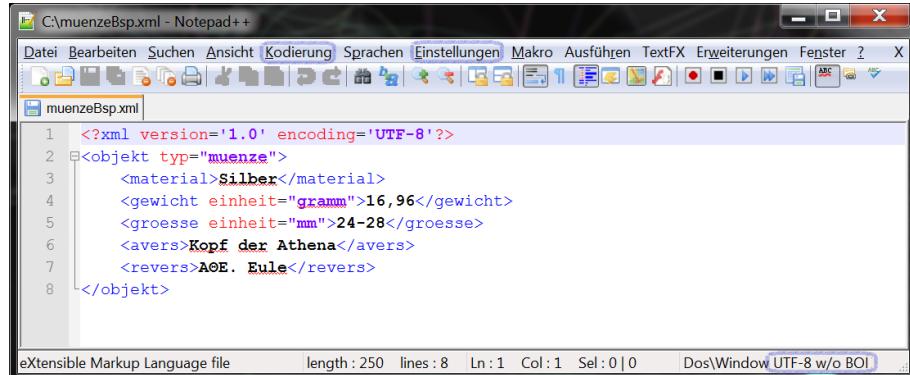


Abb. 3.7: Screenshot von Notepad++ mit einer geöffneten XML-Datei. Die Menüpunkte „Einstellungen“ und „Kodierung“ wurden hervorgehoben. Im unteren rechten Bereich ist die Anzeige der verwendeten Zeichenkodierung gekennzeichnet.

In *Text Wrangler* ist diese Option unter „TextWrangler > Preferences > Text Encoding“ zu finden. Auch hier werden die Einstellungen der Zeichenkodierung einer vorhandenen Datei beibehalten. Zusätzlich besteht die Möglichkeit die Zeichenkodierung zu ändern, indem eine Datei über „File > Reopen Using Encoding“ und der gewünschten Kodierung geöffnet wird.

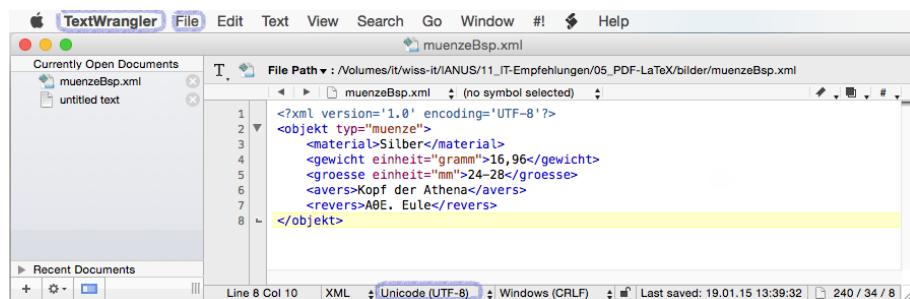


Abb. 3.8: Screenshot von TextWrangler mit einer geöffneten XML-Datei. Die Menüpunkte „TextWrangler“ und „File“ wurden hervorgehoben. Im unteren linken Bereich ist die Anzeige der verwendeten Zeichenkodierung gekennzeichnet.

Metadaten bearbeiten und ergänzen In der Regel werden nur wenige Metadaten automatisch in Textdokumenten von Textverarbeitungsprogrammen wie Microsoft Word, OpenOffice Writer oder LibreOffice Writer angelegt und

gespeichert. Dazu gehören vor allem technische Informationen, wie Dateigröße, Dateiname, Erstellungs- und Änderungsdatum. Auch eine Statistik mit der Anzahl der Zeichen, Wörter, Absätze etc. wird erstellt. Als Autor wird der für das jeweilige Programm angegebenen Nutzernamen gespeichert. Über die Menüpunkte „Datei > Informationen > Eigenschaften“ bzw. „Datei > Eigenschaften“ lassen sich die Angaben anpassen und ergänzen. Beispielsweise kann ein Titel, Schlagworte und ein Beschreibungstext eingefügt werden. Zusätzliche Angaben können unter „Anpassen“ bzw. „Benutzerdefinierte Eigenschaften“ aus einer Liste gewählt und ausgefüllt werden. Darüber hinausgehende Informationen wie beispielsweise ein Identifikator oder Angaben zur Lizenz, können in einer getrennten Text- oder XML-Datei hinterlegt werden. Ausführlichere Angaben sind in „Verfahren zur Produktion interoperabler Metadaten in digitalen Dokumentenverarbeitungsprozessen“ von Alexander Haffner (2011) zu finden.

Bei Textdokumenten bietet sich die Möglichkeit, neben einem Deckblatt auch einen Innentitel mit den relevanten Metadaten zu integrieren. Hier können zusätzlich ein Zitierhinweis und eine längere Versionshistorie untergebracht werden. Ein Beispiel für solch einen Innentitel findet sich am Anfang der PDF-Version dieser Empfehlungen.

In reinen Textdateien, wie TXT oder *plain text*, können keine Metadaten als Eigenschaften in das Dateiformat integriert werden. Es besteht jedoch die Möglichkeit, sie mit in das Dokument einzutragen oder eine separate Datei anzulegen. Auszeichnungssprachen bieten zu diesem Zweck meist einen eigens dafür vorgesehenen Bereich am Beginn der Datei, den sogenannten Kopfbereich oder *Header*.

Tools wie beispielsweise das Metadata Extraction Tool oder eines der Tools, die auf forensicswiki.org gelistet sind, können verwendet werden, um Metadaten zu extrahieren und in separaten Dateien zu speichern.

Metadata Extraction Tool:

<http://meta-extractor.sourceforge.net/>

Tools zur Extraktion von Metadaten:

http://www.forensicswiki.org/wiki/Document_Metadata_Extraction#Office_Files

Digitalisate Für die Digitalisierung von analogen Schriftstücken mittels eines Scanners gibt es ausführliche Hinweise in den *DFG-Praxisregeln "Digitalisierung"*.

Eine kurze Übersicht aus dem oben angegebenen Dokument ist in der folgenden Tabelle zu finden:

Größe des kleinsten signifikanten Zeichens	Auflösung
bis 1 mm	min. 400 dpi
ab 1,5 mm	min. 300 dpi

Die Speicherung erfolgt in Form unkomprimierter Baseline TIFF-Dateien.

Um zu verdeutlichen, dass von der Vorlage nichts abgeschnitten wurde, sollten Seiten immer vollständig mit einem umlaufenden Rand gesichert werden.

Der Scan eines Textdokumentes ist zunächst eine digitale Rastergrafik, die erst durch optische Zeichenerkennung (OCR, von engl. *Optical Character Recognition*) oder Transkription zu einem digitalen Textdokument wird. Mit OCR bearbeitete Texte benötigen eine Angabe zur Genauigkeit der Buchstaben in Prozent. Ab Seite 30 der Praxisregeln wird die Ermittlung der Buchstabengenauigkeit beschrieben.

Die DFG-Praxisregeln beziehen sich teilweise auf die Richtlinien der Federal Agencies Digitization Guidelines Initiative (FADGI), die in englischer Sprache in dem Dokument „*Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*“ zu finden sind.

Bei der Neubeschaffung eines Scanners muss darauf geachtet werden, dass er die Mindestanforderungen für den jeweiligen Digitalisierungszweck erfüllt.

Quellen

Archaeology Data Service, Documents and Digital Texts: A Guide to Good Practice

http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_Toc

A. Haffner, Verfahren zur Produktion interoperabler Metadaten in digitalen Dokumentenverarbeitungsprozessen (Frankfurt am Main 2011)

http://www.kim-forum.org/Subsites/kim/DE/Materialien/Dokumente/dokumente_node.html#doc42066bodyText7

R. Ishida, Zeichencodierung für Anfänger

<http://www.w3.org/International/questions/qa-what-is-encoding>

R. Ishida, Zeichencodierungen: grundlegende Konzepte

<http://www.w3.org/International/articles/definitions-characters/>

A. Morrison – M. Popham – K. Wikander, Creating and Documenting Electronic Texts: A Guide to Good Practice

<http://ota.ahds.ac.uk/documents/creating/cdet/index.html>

A. Morrison – M. Wynne, AHDS Preservation Handbook: Marked-up Textual Data (2005)

http://ota.ahds.ac.uk/documents/preservation/preservation_markup.pdf

nestor (Hrsg.) Nicht von Dauer: Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen (2009) 22-28

H. Neuroth – A. Oßwald – R. Scheffel – S. Strathmann – M. Juhn (Hrsg.) nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchiverung. Version 2.0 (2009) Kap. 17.2

G. Rehm, Texttechnologische Grundlagen, in: K.-U. Carstensen – Ch. Ebert – C. Endriss – S. Jekat – R. Klabunde – H. Langer (Hrsg.) Computerlinguistik und Sprachtechnologie. Eine Einführung ²(München 2004) 138-147

TEI (Hrsg.) A Gentle Introduction to XML

<http://www.tei-c.org/release/doc/tei-p5-doc/de/html/SG.html>

DFG-Praxisregeln „Digitalisierung“

http://www.dfg.de/formulare/12_151/12_151_de.pdf

FAQ zu UTF und BOM

http://www.unicode.org/faq/utf_bom.html

Formatspezifikationen

ODT:

<https://www.oasis-open.org/standards#opendocumentv1.2>

DOCX: ECMA-376

<http://www.ecma-international.org/publications/standards/Ecma-376.htm>

SGML: ISO 8879

http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387

HTML: WHATWG – W3C (Hrsg.) HTML5 (2014)

<http://www.w3.org/TR/html5/>

XML: T. Bray – J. Paoli – C. M. Sperberg-McQueen – E. Maler, F. Yergeau, Extensible Markup Language (XML) 1.0⁵(2008)

<http://www.w3.org/TR/xml/>

XSD:

http://www.w3.org/TR/#tr_XML_Schema

TEI: Text Encoding Initiative (TEI)

<http://www.tei-c.org/index.xml>

TEI: TEI (Hrsg.) P5: Richtlinien für die Auszeichnung und den Austausch elektronischer Texte

<http://www.tei-c.org/release/doc/tei-p5-doc/de/html/index.html>

Unicode: Unicode Consortium (Hrsg.) Unicode 7.0.0

<http://www.unicode.org/versions/Unicode7.0.0/>

Tools und Programme

Metadata Extraction Tool:

<http://meta-extractor.sourceforge.net/>

Tools zur Extraktion von Metadaten:

http://www.forensicswiki.org/wiki/Document_Metadata_Extraction#Office_Files

Notepad++:

<http://www.notepad-plus-plus.org/>

TextWrangler:

<http://www.barebones.com/products/textwrangler/>

Vergleich von Texteditoren auf Wikipedia:

http://en.wikipedia.org/wiki/Comparison_of_text_editors

3.3 Bilder – Rastergrafiken

M. Trognitz, P. Grunwald

Bei Rastergrafiken, auch Pixelgrafiken, handelt es sich um digitale Bilder, die mittels rasterförmig angeordneter Bildpunkte, den Pixeln, beschrieben werden. Jedem Pixel ist dabei ein Farbwert zugeordnet. Rastergrafiken haben eine fixe Größe und sind im Gegensatz zu Vektorgrafiken nicht beliebig skalierbar.

Zu den Rastergrafiken gehören: Digitale Fotografien jeder Art, Satellitenbilder, digitalisierte Bilder (Scans), Screenshots sowie digitale Originalbilder und -grafiken.

Langzeitformate Alle als Rastergrafiken vorliegenden Roh- und Urfassungen (Master) von Bildern sind in angemessener Qualität und unkomprimiert im baseline TIFF- oder DNG-Format abzuspeichern. Für georeferenziertes digitales Bildmaterial ist zwecks Erhalt der Referenzdaten das Format GeoTIFF zu verwenden.

Nur für Grafiken, *nicht* für Fotos, eignet sich auch das PNG-Format. Allerdings ist jederzeit das TIFF-Format vorzuziehen. JPEG, bzw. JPG, eignet sich nicht zur Langzeitarchivierung, da es keine verlustfreie Komprimierung anbietet.

Es muss darauf geachtet werden, dass die Bildgröße und -auflösung der originalen Datei erhalten bleibt, wenn in andere Formate umgewandelt wird. Außerdem muss bei der Konvertierung darauf geachtet werden, dass eine verlustfreie Komprimierung verwendet wird. Auch Farbtiefe und Farbraum sollten nach der Konvertierung erhalten bleiben.

Bilder, die Ebenen enthalten, müssen vorher auf eine Ebene reduziert werden. Bei Bedarf, sollten die verschiedenen Ebenen und Komponenten als einzelne Dateien abgespeichert werden.

Hinweis: Wiederholtes Bearbeiten und Abspeichern, führt zu einer allmählichen Abnahme der Qualität. Dieser sogenannte Generationsverlust tritt insbesondere bei der verlustbehafteten Bildkomprimierung auf.

Format	Begründung
✓ Baseline TIFF, unkomprimiert	TIFF ist quasi ein Standardformat für die digitale Langzeitarchivierung von Bilddateien und unterstützt auch die Speicherung von Metadaten im Exif-Format. Obwohl das Format auch Kompression verwenden kann, kommt für die Langzeitarchivierung nur die unkomprimierte Form in Frage.
DNG	Das von Adobe entwickelte Digital Negative Format ist ein offenes Format, das für die Langzeitarchivierung geeignet ist. Damit können RAW-Dateien und deren Metadaten (Exif oder IPTC-NAA) gelesen und gespeichert werden. Außerdem können über XMP weitere Metadaten eingespeist werden.
GeoTIFF	GeoTIFF basiert auf TIFF und ist besonders für georeferenzierte Bilddaten geeignet, da somit die Referenzdaten erhalten bleiben.
~ PNG	PNG ist eine verlustfreie Alternative zu dem GIF-Format, welches eine verlustbehaftete Kompression verwendet. Es bietet eine Farbtiefe von 32 Bit, einen Alphakanal für Transparenz und verlustfreie Kompression. Allerdings wird nur RGB als Farbraum unterstützt und es können keine Exif-Daten gespeichert werden. Das Format eignet sich <i>nicht</i> für digitale Fotos.
✗ JPEG	Trotz einiger Vorteile, eignet sich das JPEG-Format <i>nicht</i> für die Langzeitarchivierung, da es <i>keine</i> verlustfreie Komprimierung bietet.
GIF	GIF kann sowohl statische, als auch animierte Bilder speichern. Da es aber verlustbehaftet komprimiert, wird PNG als Alternativformat empfohlen.

Dokumentation Eingebettete Metadaten, wie beispielsweise Exif, IPTC-NAA oder XMP, sollten behalten und archiviert werden. Am besten werden sie in eine eigene Text- oder XML-Datei transferiert und getrennt gespeichert.

Neben technischen Informationen, die sich hauptsächlich mit der Erstellung der Rastergrafik befassen, sollten vor allem auch beschreibende und administrative Metadaten über das Bild gespeichert werden.

Hinweis: Werden digitale Aufnahmen mit Programmen bearbeitet, welche die eingebetteten Metadaten ignorieren, gehen diese verloren.

Die hier angegebenen Metadaten sind als minimale Angabe zu betrachten und ergänzen die angegebenen Metadaten für Projekte und Einzeldateien in dem Abschnitt Metadaten in der Anwendung ab Seite 36.

Metadatum	Beschreibung
Identifikator	Name der Datei, z.B. grabung01.tif
Bildunterschrift	Der Titel oder eine passende Bildunterschrift des Bildes
Beschreibung	Beschreibung des Bildes
Urheber	Name des Fotografen oder Erstellers
Datum	Datum der Erstellung oder letzten Änderung des Bildes
Rechte	Details zum Urheberrecht
Schlagworte	Schlagworte, wie z.B. Periode, Fundstelle oder charakteristische Merkmale. Wenn vorhanden, angemessene Thesauri verwenden
Ort	Ortsinformationen zu dem Bild. Möglichst in einem standardisierten Format angeben, wie z.B. Lat/Long oder Schlagworte aus einem geeigneten Thesaurus, z.B. Getty Thesaurus of Geographic Names oder GeoNames
Dateiformat & Version	z.B. Baseline TIFF 6.0
Dateigröße	Größe der Datei in Bytes
Bildgröße	Maße des Bildes gemessen in Pixeln, z.B. 400px × 700px
Auflösung	Bildauflösung, gemessen in Punkten pro Zoll (dpi)
Farbraum	Der in dem Bild verwendete Farbraum, z.B. RGB oder Graustufen
Farbtiefe	z.B. 24 bit oder 8 bit
Aufnahmegerät	Beispielsweise Details zur Kamera oder dem Scanner
Software	Software mit der das Bild aufgenommen, erstellt oder bearbeitet wurde, wie z.B. Adobe Photoshop CS3

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Vertiefung

Da eine Rastergrafik mit in einem rechteckigen Raster angeordneten Bildpunkten beschrieben wird und jedem dieser Punkte (Pixel), ein Farbwert zugeordnet wird, gibt es zwei Hauptmerkmale: die Bildgröße und die Farbtiefe. Abhängig von der Bildgröße ist die Auflösung. Hinzu kommen noch weitere Eigenschaften, wie Farbmodell und Farbraum, Komprimierung, Transparenz, Ebenen und Metadaten die hier ausführlicher erklärt werden.

Die einzelnen Eigenschaften können in Abhängigkeit von Quelle und Verwendung der Grafik stark variieren. Es ist praktisch unmöglich, genaue Vorgaben für die einzelnen Einstellungsmöglichkeiten zu machen. Sie sollten im Kontext des

Projektes betrachtet werden und ihren Zweck erfüllen. Nichtsdestotrotz werden in diesem Abschnitt Anmerkungen über die Qualität gemacht.

Entscheidend für die Nachnutzung von Rastergrafiken ist die Dokumentation, worin auch die Entscheidung für ein bestimmtes Format und dessen Einstellungen begründet werden kann.

Bildgröße und Auflösung Die Bildgröße beschreibt den Detaillierungsgrad einer Grafik mittels der Pixelanzahl. Dabei gilt: Je mehr Bildpunkte, desto höher ist auch der Detaillierungsgrad und die Dateigröße. Die Pixelanzahl kann durch die Gesamtanzahl der Bildpunkte, wie beispielsweise in der Digitalfotografie mittels Megapixeln, oder mit der Anzahl der Bildpunkte je Zeile mal der Anzahl der Bildpunkte je Spalte (z.B. 1024×768) angegeben werden. Aus der zweiten Darstellungsvariante geht auch das Seitenverhältnis hervor.

Umgangssprachlich wird die Bildgröße auch als Bildauflösung bezeichnet. Allerdings hängt die Auflösung von einem physikalischen Wiedergabemedium (z.B. Bildschirm oder ein A4 Ausdruck) ab, wobei die Punktdichte maßgeblich für die Wiedergabequalität ist. Die Punktdichte wird üblicherweise in Punkten (*dots per inch; dpi*), Pixeln (*pixel per inch; ppi*) oder Linien (*lines per inch; lpi*) pro Zoll (*inch*) angegeben.

Soll das Bild nur digital verwendet werden, reicht eine minimale Auflösung von 72dpi aus. Wenn das Bild jedoch gedruckt werden soll, muss mit einer Mindestauflösung von 300dpi gearbeitet werden.

Abhängig von der Aufgabe einer Rastergrafik muss eine geeignete Bildgröße gewählt werden. Dabei sollte bedacht werden, dass die Dateigröße mit dem Detaillierungsgrad steigt, und somit eine Balance zwischen dem benötigtem Detaillierungsgrad und der Dateigröße gefunden werden muss.

Farbtiefe Mit der Farbtiefe (engl. auch *bit depth*) wird die Anzahl der Bits angegeben, die den Farbwert eines Pixels speichern. Ein Bit kann dabei 2^1 , also zwei Farbwerte (z.B. Schwarz und Weiß) speichern. Die Anzahl der darstellbaren Farbwerte steigt mit der Anzahl der Bits exponentiell. So können mit 8 Bits bereits 2^8 also 256 Farbwerte (üblicherweise Graustufen) und mit 24 Bits schon $2^{24} = 16.777.216$ Farbwerte (*True color*) dargestellt werden. Größere Farbtiefen von 30, 32, 36, 40 und 48 Bit werden hauptsächlich im Scan-, Kino-, TV- und Druckbereich verwendet.

Wie bei der Bildgröße steigt auch hier die Dateigröße mit der Farbtiefe, weshalb nur die minimal notwendige Farbtiefe gewählt werden sollte. Beispielsweise reicht es aus, eine Schwarz-weiß Grafik mit 8-Bit Graustufen zu speichern.

Ein Spezialfall sind indizierte Farben. Dabei wird für ein Pixel nicht direkt

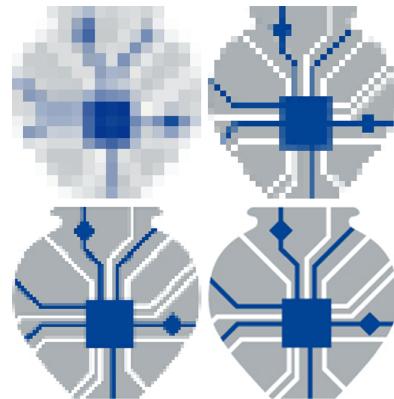


Abb. 3.9: Je mehr Bildpunkte eine Rastergrafik enthält, desto detaillierter wird die Abbildung. Die gleiche Grafik wurde hier mit 16x16, 32x32, 64x64 und 185x185 Pixeln dargestellt.

der Farbwert, sondern ein Index auf eine Farbe aus einer vorgegebenen Farbtabelle oder Farbpalette gespeichert. Somit können Bilder mit wenigen Farben Speicherplatz einsparen. Beispielsweise bietet das Dateiformat GIF eine Farbtiefe von 8 Bit.

Farbmodell und Farbraum Ein Farbmodell ist ein mathematisches Modell, das üblicherweise mit Hilfe von Zahlentupeln beschreibt wie Farben dargestellt werden können. Alle Farben eines Farbmodells können dreidimensional als Farbraum dargestellt werden.

Übliche Farbmodelle sind RGB und CMYK. RGB wird hauptsächlich für die Bildschirmanzeige verwendet, während CMYK im Druckbereich verwendet wird. Insgesamt enthält RGB mehr Farbkombinationen als CMYK, weshalb es sein kann, dass RGB-Grafiken nicht exakt farbtreu gedruckt werden können.

Ein Farbmodell kann auf mehrere Farbräume abgebildet werden, weshalb es für RGB unter anderem die Farbräume *sRGB* und *Adobe RGB* gibt. Die eben genannten Farbräume sind genormte, ausreichend große Farbräume, die für die meisten Anwendungen ausreichen.

Die Kombination von Farbtiefe und Farbmodell beschreibt wie viele Bits pro Farbwert zur Speicherung zur Verfügung stehen und wie viele Farben im Endeffekt dargestellt werden. Beispielsweise bietet *True Color* für RGB, mit einer Farbtiefe von 24 Bit, jeweils 8 Bit für Rot, Grün und Blau. Da man bei CMYK noch einen vierten Wert berücksichtigen muss, hat *True Color* für CMYK eine Farbtiefe von 32 Bit.

Wenn man im Zweifel ist, ob RGB oder CMYK verwendet werden soll, ist zu empfehlen das RGB-Farbmodell zu verwenden, da damit mehr Farben abgebildet werden können. Bei Bedarf lässt sich der Farbraum nachträglich in CMYK oder in einen indizierte Farbraum konvertieren.

Komprimierung Das Ziel einer Komprimierung ist, die Dateigröße für einen bestimmten Zweck zu reduzieren, wie etwa zur Darstellung im Internet. Die Komprimierung von Rastergrafiken kann entweder verlustfrei oder verlustbehaftet erfolgen. Grundsätzlich sollte bei der Speicherung einem Format den Vorzug gegeben werden, das entweder gar keine (z.B. TIFF oder PNG) oder verlustfreie (z.B. GIF, PNG oder TIFF mit LZW) Komprimierung verwendet. Verlustbehaftete Formate, wie etwa JPEG, sollten nur dann verwendet werden, wenn es nicht anders geht (z.B. weil die verwendete Digitalkamera nur dieses Speicherformat bietet) und möglichst bald in ein verlustfreies Format konvertiert werden.

Im Umgang mit Rastergrafiken ist es wichtig, sich bewusst zu machen, wann

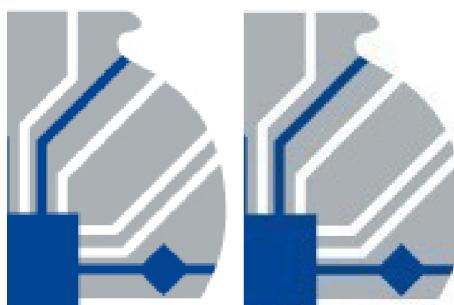


Abb. 3.10: Visualisierung des Generationsverlustes. Ein Ausschnitt einer JPG-Datei links im originalen Zustand und rechts nach 2000-maliger Speicherung.

eine Komprimierung erfolgt und in welchem Grad dies passiert. So führt etwa häufiges Bearbeiten und Abspeichern von JPEGs zum sogenannten Generationsverlust, der mit der Verwendung von TIFF-Dateien vermieden werden kann.

TIFF bietet die Möglichkeit, eine verlustfreie Komprimierung mit LZW anzuwenden. Allerdings ist dieses komprimierte TIFF-Format noch nicht für die Langzeitarchivierung erprobt.

Transparenz Transparente Elemente in Grafiken werden von vielen Vektorgrafikformaten unterstützt. Doch da Transparenz üblicherweise mit dem sogenannten Alphakanal gespeichert wird, benötigt man ein Rastergrafikformat, das diesen berücksichtigt, wie z.B. TIFF, PNG oder GIF.

Bei der Konvertierung von einem Format in ein anderes, sollte darauf geachtet werden, ob Transparenz auch unterstützt wird.

Ebenen Eine weit verbreitete Funktion von Grafikprogrammen ist die Möglichkeit, verschiedene Bildelemente auf verschiedenen Schichten, den sogenannten Ebenen, zu verteilen. Diese Eigenschaft wird von den meisten Rastergrafikformaten nicht unterstützt.

Wenn die Datei als Rastergrafik gespeichert wird, werden die Ebenen von oben nach unten verschmolzen. Sollen einzelne Ebenen auch getrennt zugänglich sein, kann jede Ebene für sich als eigene Bilddatei zu gespeichert werden.

Metadaten Einige Bildformate unterstützen die Speicherung der Metadaten direkt in der Datei. Dazu gehören auch TIFF, DNG und JPEG. Es gibt drei gebräuchliche Metadatenformate bzw. -standards, welche die Informationen jeweils in den Header der Bilddatei schreiben. Der erste für professionelle Arbeitsabläufe gestaltete Standard IIM wurde ab 1995 von Adobe teilweise in Photoshop übernommen. Daraus entwickelte sich dann der heute gebräuchliche IPTC-NAA-Standard. Gleichzeitig führte Adobe aber auch den XMP-Standard ein. Ein weiterer Standard ist Exif, der vor allem von Digitalkameras zur Speicherung der Aufnahmeinformationen verwendet wird.

Etwas problematisch ist, dass jeder Standard die Informationen in den Header der Bilddatei schreibt, weshalb die Gefahr besteht, dass sie sich teilweise überschreiben. Deshalb wird empfohlen, die Metadaten zu extrahieren und in einer gesonderten Textdatei zu speichern. Was für diese gesonderte Datei beachtet werden muss, wird im allgemeinen Abschnitt über Metadaten, 2.2, auf Seite 28 erläutert.

Die Informationen im Header sollten nicht komplett gelöscht werden, da im Dateiformat gespeicherte Metadaten das Aufkommen von verwaisten Werken verhindern, die keinerlei Rückschlüsse auf den Urheber zulassen.

Die Problematik der drei unterschiedlichen Standards wurde auch von namhaften Herstellern erkannt, weshalb die Metadata Working Group gegründet wurde, welche die Verwendung von Metadaten in Bilddateien vereinheitlichen will. Es soll aber nicht ein neuer Standard geschaffen werden, sondern die Nutzung der bestehenden Standards durch einen übergreifenden Rahmen geregelt werden.

Anmerkungen zur Qualität Gerade weil Rastergrafiken ein breites Anwendungsspektrum bieten, können hier keine spezifischen Aussagen zur Qua-

lität gemacht werden. Im Prinzip muss also der Ersteller entscheiden, was für die jeweilige Aufgabe angemessen ist. Dabei sollte aber nicht nur auf die aktuellen konkreten Anforderungen, sondern gerade im Hinblick auf die spätere Verwendung (z.B. Publikation) oder Arbeitsschritte (z.B. Konvertierung) auch auf zukünftige Anforderungen geachtet werden.

Eine große Datei ist nicht gleichbedeutend mit guter Qualität. Andererseits sollte auch nicht zugunsten von Speicherplatz auf die Qualität verzichtet werden.

Praxis

In diesem Abschnitt sind Hinweise zum Umgang mit Rastergrafiken gesammelt. Neben einem ausführlichen Abschnitt über Digitalfotografie gibt es kürzere Erläuterungen mit Literatur- und Programmhinweisen über das Rastern von Vektorgrafiken, die Konvertierung und Überprüfung von Dateiformaten, das Ergänzen und Extrahieren von Metadaten, die Stapelverarbeitung und Digitalisate.

Außerdem werden einige Angaben über Anforderungen an die benötigte Hardware gemacht.

Digitalfotografie Die digitale Fotografie hat bis auf wenige Ausnahmen und spezielle Anwendungen die analoge Fotografie weitgehend abgelöst. Um aber auch hier mit der digitalen Fotografie einen möglichst hohen Standard in Bildqualität und für die Archivierung zu gewährleisten, so wie er aus der analogen Fotografie bekannt ist und sich bewährt hat, bleibt es unerlässlich, Mindestanforderungen an Hard- und Software zu stellen und diese auch einzuhalten.

In diesem Falle ist unter Hardware die Kameraausrüstung und Speichermedium zu verstehen und unter Software die Programme, mit der die Fotos nach der Aufnahme kameraintern bzw. extern zum Bearbeiten bzw. Speichern und zur Langzeitarchivierung aufbereitet werden.

Während der Aufnahme speichert die Kamera Metadaten im standardisierten Exif-Format undbettet sie in die Bilddatei ein. Diese enthalten u.a. Datum, Uhrzeit, aufnahmespezifische Parameter wie Blende, Verschlusszeit, Brennweite, laufende Nummerierung und evtl. auch GPS-Koordinaten. Um korrekte Exif-Daten zu speichern, müssen Datum und Uhrzeit kameraseitig richtig eingestellt werden.

Je nach Kameramodell und Einstellung können die Bilddaten als unkomprimierte Rohdaten (RAW), bereits interpoliert als unkomprimiertes TIFF oder komprimiert als JPEG auf der Speicherkarte abgelegt werden. Verschiedene Kameramodelle lassen sich so einstellen, dass auf der Speicherkarte gleichzeitig ein RAW und JPEG bzw. TIFF und JPEG abgelegt werden. Diese Methode belegt allerdings sehr viel mehr an Speicherplatz. Einige wenige Kameramodelle gestatten auch das Speichern der Aufnahmen im unkomprimierten JPEG-Format.

Die kameraabhängigen Rohdaten sind um 30% kleiner als TIFF-Dateien, da ihnen der Datenzuwachs durch die Interpolation fehlt; diese findet erst beim Öffnen der Bilddatei statt. JPEG-Dateien können in Abhängigkeit vom gewählten Kompressionsfaktor bis auf unter 1/10 der Rohdaten-Dateigröße bei entsprechendem Qualitätsverlust komprimiert werden.

Bei der Neubeschaffung einer Kamera ist zu bedenken, dass die Bildqualität einer digitalen Kamera nicht nur von der Auflösung, also der Pixelzahl, abhängig

ist. Weitere wichtige Aspekte sind Objektivqualität, Sensorgröße, Rauschverhalten, Dynamikbereich, Farbdarstellung etc. Weiterhin sind auch die Handhabung und die Kompaktheit der Kamera wichtige Auswahlkriterien. Die Kamera sollte die Aufnahmen idealerweise als RAW, DNG oder TIFF speichern.

Sollte nur eine Kamera zur Verfügung stehen, die nur im JPEG-Format speichern kann, so sollte man mit der höchsten Bildqualität fotografieren. Die Bilder müssen zum nächstmöglichen Zeitpunkt in das TIFF-Format konvertiert werden, um dem Generationsverlust vorzubeugen.

Digitale Kameras gibt es in drei prinzipiellen Grundbauarten:

- Kompaktkamera mit optischem Sucher, seitlich versetzt vom Aufnahmeobjektiv angeordnet und mit Display, je nach Modell schwenkbares Display. Das Objektiv ist fest eingebaut. Kein Eindringen von Staub und Fremdkörpern auf den Sensor.
- Systemkamera mit elektronischem Sucher und Okular, fest eingebautes oder auswechselbares Objektiv und rückseitigem fest eingebauten oder schwenkbaren Display. Kann sowohl automatisch, als auch manuell betrieben werden.
- DSLR-Kamera (Digitale Spiegelreflex-Kamera) mit Wechselobjektiven und rückseitigem Display, fest eingebaut oder schwenkbar.

Unabhängig davon, ob es sich um eine digitale Kompakt-, System- oder Spiegelreflexkamera handelt, sind gewisse Ausstattungsmerkmale erforderlich, um geforderte Standards in Bezug auf Bildqualität und Dateiformat einzuhalten. Sowohl die Qualität des Objektives als auch eine evtl. ab Werk eingestellte kamerainterne Datenkompression nehmen direkten Einfluss auf die Bildqualität.

Derzeit gebräuchliche Bildwandlerformate (kurz als Chip bezeichnet) sind:

- Kleiner und bis $\frac{1}{4}$ -Format-Chip für Mini-, Kompakt- und Sucherkameras
- $\frac{1}{2}$ -Format-Chip für DSLR- und Monitor-Sucher-Kameras mit und ohne Wechselobjektiven
- Vollformat-Chip für DSLR Kamerassen mit Wechselobjektiven
- Großformat-Chips für Mittelformat-Rückteile, Kamera-Scanner etc.

Das Chip-Format und die Anzahl der Pixel beeinflussen direkt die Schärfeleistung, Kontrastwiedergabe und Farbtrennung.

Die „Normal-Brennweite“ des Objektives berechnet sich nach der Diagonale des verwendeten Chips; sie entspricht der Diagonale des Bildwandlers, angegeben in Millimetern. Die Qualität des Objektives beeinflusst ebenfalls direkt die Bildqualität und die effektive Auflösung des Bildwandlers.

In der Regel hat der Bildwandler das Format 3:4, die (theoretische) Auflösung berechnet sich aus der Anzahl der Pixel in Breite \times Höhe (z.B.: 1500×2000 Pixel entsprechen einer Auflösung von 3 Megapixeln und einer unkomprimierten Bilddatengröße von ca. 9 Megapixel).

Die Angabe des rechnerischen Pixelmaßes kann durchaus um mehr als 10% über der effektiv nutzbaren Pixelzahl liegen. Bei Neubeschaffung einer Kamera sollte derzeit eine effektive Auflösung von mindestens 10 Millionen Pixel nicht unterschritten werden.

Weitere Ausstattungsmerkmale, wie USB-Anschluss zur direkten Datenübertragung von Kamera zum Notebook, Macro-Bereich, schwenkbares Display, externer Blitzanschluss, elektrischer Draht- oder Fernauslöser sind zusätzliche Funktionen, welche die Arbeit mit der digitalen Kamera erheblich erleichtern können.

Rastern aus Vektorgrafiken Tritt der Fall ein, dass Grafiken, die ursprünglich als Vektordatei vorlagen, in eine Rastergrafik konvertiert, also geraستert werden sollen, so muss eine geeignete Bildgröße ausgewählt werden, die den gewünschten Anforderungen genügt. Die orginale Vektordatei sollte zu Archivierungszwecken ebenfalls aufbewahrt werden.

Das Rastern kann am besten in dem Programm gemacht werden, in dem die Grafik erstellt wurde. Dazu wählt man entweder die Option *Speichern unter* oder *Export*. Die weiteren Einstellungen werden dann üblicherweise von dem Programm abgefragt.

Stapelverarbeitung Oft tritt der Fall ein, dass eine ganze Reihe von Bildern mit gleichförmigen Abläufen, wie etwa Umbenennen oder Beschneiden, bearbeitet werden muss. Mittels Stapelverarbeitung (auch Batchverarbeitung) kann dies automatisch und zügig gemacht werden.

Für sogenannte Batch-Jobs bieten die verschiedenen Betriebssysteme eigene Skriptsprachen an, wie etwa Microsoft Batch. Speziell für die Bildverarbeitung gibt es eigene spezialisierte und bedienfreundliche Programme. Beispielsweise ist in der Creative Suite von Adobe das Programm Bridge enthalten, dessen Funktionsumfang beachtlich ist. Eine gern verwendete kostenlose und nicht ganz so umfangreiche Alternative ist IrfanView.

IrfanView:

<http://www.irfanview.de/>

Dateiformate konvertieren und überprüfen Möchte man Bilder von einem Format in ein anderes überführen, so kann man dies am besten mit den üblichen Grafikprogrammen machen. Mittels *Speichern unter* oder der Exportfunktion können die meisten Formate konvertiert werden. Allerdings muss man dabei beachten, dass eingebettete Metadaten aus dem Quellformat auch ins Zielformat überführt werden.

Generell müssen die verschiedenen Bildeinstellungen beachtet werden, weshalb geprüft werden sollte, ob das Zielformat die gewünschten Anforderungen erfüllt.

Neben den Grafikprogrammen findet man im Internet auch zahlreiche Online-Dienste, die Dateikonvertierungen anbieten. Ein Beispiel für so einen Online-Dienst ist Zamzar.

Speziell für digitale Fotos im RAW-Format eignet sich der frei verfügbare DNG Converter von Adobe. Weitere DNG Converter werden von Kameraherstellern zur Verfügung gestellt und sind oft speziell für ein bestimmtes Kameramodell gemacht.

Dateiformate können mit speziellen Programmen überprüft werden. Auch wenn das Dateiformat unbekannt ist, kann ein solches Programm helfen. Ein Programm, das speziell für die Langzeitarchivierung geeignete Datenformate

überprüft ist JHOVE. Das Programm bildet die Grundlage für JHOVE2, einem Nachfolger von JHOVE.

Zamzar:

<http://www.zamzar.com/url/>

Adobe DNG Converter:

<http://www.adobe.com/products/photoshop/extend.displayTab2.html#downloads>

JHOVE:

<http://jhoove.sourceforge.net/>

JHOVE2:

<https://bitbucket.org/jhove2/main/wiki/Home>

Ergänzen und extrahieren von Metadaten Technische Metadaten von Bildern werden in vielen Fällen schon bei der Erstellung einer Rastergrafik erzeugt und mit im Dateiformat abgespeichert (z.B. digitale Fotografie oder Scan). Weitere Metadaten können nachträglich hinzugefügt werden. Das kann man beispielsweise für einzelne Bilder in einem Grafikprogramm machen, welches die Funktionalität bietet, wie z.B. Photoshop oder Gimp (für Gimp muss allerdings noch ein Plugin installiert werden). Wenn es um eine sehr große Menge Fotos geht, empfiehlt sich ein eigenes Bildverwaltungsprogramm, wie z.B. Adobe Bridge, FotoWare oder XnView. Eine ausführliche Liste ist auf Wikipedia zu finden.

Für die Archivierung von Bildern ist es empfehlenswert, wenn man die Metadaten extrahiert und in einer eigenen Textdatei oder XML-Struktur unterbringt. Metadaten aus Bildern können aus den Bildverwaltungsprogrammen, wie Adobe Bridge oder FotoWare, exportiert oder mit eigenen Programmen extrahiert werden. Beispielsweise kann man das Metadata Extraction Tool, das ExifTool oder eines der Tools, die auf forensicswiki.org gelistet sind, verwenden. Es gibt auch die Möglichkeit mit dem ExifViewer die Metadaten online zu extrahieren.

Werden Bilder und deren Metadaten in einer eigenen Datenbank verwaltet, so muss der Abschnitt über Datenbanken 3.6 auf Seite 109 für die Langzeitarchivierung berücksichtigt werden.

Bildverwaltung mit XnView:

<http://www.xnview.com/>

Liste von Bildverwaltungsprogrammen auf Wikipedia:

<http://de.wikipedia.org/wiki/Bilderverwaltung#Software>

Metadata Extraction Tool:

<http://meta-extractor.sourceforge.net/>

ExifTool:

<http://www.sno.phy.queensu.ca/~phil/exiftool/>

Tools zur Extraktion von Metadaten:

http://www.forensicswiki.org/wiki/Document_Metadata_Extraction#Images

Metadaten online ansehen und extrahieren mit ExifViewer:

<http://www.exifviewer.org/>

Digitalisate Für die Digitalisierung von analogen Vorlagen mittels eines Scanners, gibt es ausführliche Hinweise in den *DFG-Praxisregeln "Digitalisierung"*.

Eine kurze Übersicht aus dem oben angegebenen Dokument ist in der folgenden Tabelle zu finden:

Vorlage	Auflösung	Farbtiefe
Aufsichtsvorlagen (z.B. Fotos) von Farb- und Graustufenabbildungen	min. 300 dpi	Farbe: 24 Bit Grau: 8 Bit
Durchsichtsvorlagen (z.B. Dias) im Kleinbildformat (24 × 36cm)	3000 dpi	Farbe: 48 Bit Grau: 16 Bit

Die Speicherung erfolgt in Form unkomprimierter Baseline TIFF-Dateien.

Die DFG-Praxisregeln beziehen sich teilweise auf die Richtlinien der Federal Agencies Digitization Guidelines Initiative (FADGI), die in englischer Sprache in dem Dokument "*Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*" zu finden sind.

Bei der Neubeschaffung eines Scanners muss darauf geachtet werden, dass er die Mindestanforderungen für den jeweiligen Digitalisierungszweck erfüllt.

Quellen

- Archaeology Data Service, Raster Images: A Guide to Good Practice
http://guides.archaeologydataservice.ac.uk/g2gp/RasterImg_Toc
- Significant Properties Testing Report: Raster Images
<http://www.significantproperties.org.uk/rasterimages-testingreport.html>
- nestor (2009) Nicht von Dauer: Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen
http://files.d-nb.de/nestor/ratgeber/ratg01_2_de.pdf
- DFG-Praxisregeln „Digitalisierung“
http://www.dfg.de/formulare/12_151/12_151_de.pdf
- Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files
http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf
- (2006) Metadaten in Fotos:
<http://www.drweb.de/magazin/metadaten-in-fotos/>
- Metadata Working Group – Guidelines for Handling Image Metadata:
http://www.metadataworkinggroup.com/pdf/mwg_guidance.pdf

Formatspezifikationen

- TIFF:
<http://www.fileformat.info/format/tiff/egff.htm>
- DNG:
http://wwwimages.adobe.com/content/dam/Adobe/en/products/photoshop/pdfs/dng_spec_1.4.0.0.pdf

GeoTiff:

<http://www.remotesensing.org/geotiff/spec/geotiffhome.html>

PNG:

<http://www.w3.org/TR/PNG/>

Tools und Programme

Irfanview:

<http://www.irfanview.de/>

Adobe DNG Converter:

<http://www.adobe.com/products/photoshop/extend.displayTab2.html#downloads>

Zamzar:

<http://www.zamzar.com/url/>

JHOVE:

<http://sourceforge.net/projects/jhove/>

JHOVE:

<http://jhove.sourceforge.net/>

JHOVE2:

<https://bitbucket.org/jhove2/main/wiki/Home>

Bildverwaltung mit XnView:

<http://www.xnview.com/>

Liste von Bildverwaltungsprogrammen auf Wikipedia:

<http://de.wikipedia.org/wiki/Bilderverwaltung#Software>

Tools zur Extraktion von Metadaten:

http://www.forensicswiki.org/wiki/Document_Metadata_Extraction#Images

Metadata Extraction Tool:

<http://meta-extractor.sourceforge.net/>

ExifTool:

<http://www.sno.phy.queensu.ca/~phil/exiftool/>

Metadaten online ansehen und extrahieren mit Exifviewer:

<http://www.exifviewer.org/>

3.4 Bilder – Vektorgrafiken und CAD-Daten

M. Trognitz, F. Martin, I. Schwarzbach

Vektorgrafiken sind Grafiken, die mittels grafischer Primitiven, wie beispielsweise Linien, Punkte, Polygone, Kreise und Kurven beschrieben werden. Sie sind daher im Gegensatz zu Rastergrafiken ohne Qualitätsverlust beliebig skalierbar. Zeichnungen von Scherben, Isometrische Darstellungen von Objekten, oder aus CAD- oder GIS-Anwendungen exportierte Pläne werden häufig als Vektorgrafik angelegt oder gespeichert.

Auch Daten, die mittels CAD (computer-aided design, dt. rechnerunterstützte Konstruktion) entstehen, handelt es sich um Vektordaten mit grafischen Primitiven, die in den entsprechenden Programmen zusätzliche Funktionalitäten bieten, die speziell auf architektonische und technische Zeichnungen zugeschnitten sind. CAD-Daten können sowohl in 2D als auch in 3D vorliegen und werden in den Altertumswissenschaften meist für Befundpläne von Ausgrabungen, archäologische Karten oder architektonische Pläne verwendet.

Dieser Abschnitt beschäftigt sich mit der Archivierung von zweidimensionalen Vektorgrafiken und CAD-Daten. Die Archivierung von GIS-Daten wird in dem eigenen Abschnitt GIS ab Seite 126 beschrieben. Hinweise zur Archivierung von dreidimensionalen Objekten sind in dem Abschnitt 3D und Virtual Reality ab Seite 159 zu finden.

Langzeitformate Für zweidimensionale Vektorgrafiken ist das vom W3C entwickelte und seit 2001 auch empfohlene Format Scalable Vector Graphics (SVG) für die Langzeitarchivierung geeignet, da es ein XML-basiertes, offen dokumentiertes Format ist, eine weite Verbreitung gefunden hat und als MIME-Type `image/svg+xml` registriert ist. Die aktuelle Version ist 1.1. Neben grafischen Primitiven, geometrischen Formen, Text und eingebetteten Rastergrafiken können SVG-Dateien auch Skripte und Animation enthalten. Mit Hilfe des Gruppierungselementes `< d`

`>` können Ebenen ausgezeichnet werden, wobei dieses Element in der Praxis auch für andere Zwecke genutzt wird. Umfassende Metadaten können in einem eigens dafür vorgesehenen Bereich gespeichert werden. Speziell für mobile Geräte gibt es die Variante SVG Tiny 1.2. Das W3C arbeitet seit 2012 an SVG 2, das in absehbarer Zeit die Version 1.1. ablösen soll. SVG-Dateien können auch komprimiert mittels gzip als SVGZ gespeichert werden, wobei dies nicht für die Archivierung zu empfehlen ist.

Computer Graphics Metafile (CGM) ist unter ISO/IEC 8632 standardisiert, wurde 1986 entwickelt und zuletzt 1999 aktualisiert. CGM kann umfangreiche grafische Informationen und geometrische Primitiven speichern, ist inzwischen jedoch für viele Bereiche veraltet, weshalb SVG für die Archivierung verwendet werden sollte. Das W3C hat 2001 die Variante WebCGM veröffentlicht, die zusätzliche Funktionalitäten für die Verwendung im Internet enthält. Die aktuellste und empfohlene Fassung von WebCGM ist Version 2.1.

Speziell für CAD-Daten gibt es kein Format, das uneingeschränkt für die Langzeitarchivierung zu empfehlen ist. Ein Grund dafür ist die immerwährende und schnelle Entwicklung von CAD-Software, die sich auch auf die Ausgabeformate auswirkt und zu einer hohen Heterogenität und mangelnden Interoperabilität der Datenformate führt. Durch die weite Verbreitung des von der Firma

Autodesk vertriebenen Programms AutoCAD haben die von Autodesk entwickelten Dateiformate DXF und DWG eine weite Verbreitung in der Archäologie und Bauforschung gefunden und sich de facto als Standard durchgesetzt.

Das Drawing Interchange Format (DXF, aktuelle Versionsnummer v.u.28.1.01) wurde von Autodesk entwickelt, um den Austausch von CAD-Dateien zwischen AutoCAD und anderen Programmen zu ermöglichen. Dabei wurde das Format so gestaltet, dass es möglichst alle in einer Datei enthaltenen Informationen speichern kann, ohne dass dafür die Optimierungen des proprietären DWG-Formates von Autodesk enthüllt werden müssen. Neue Versionen von DXF werden fast jedes Jahr veröffentlicht und durch neue Funktionalitäten angereichert. Für die jüngsten DXF-Versionen stellt Autodesk die Spezifikationen frei zur Verfügung, wobei allerdings ab Release 13 (v.u13.1.01) manche Bereiche ungenügend spezifiziert sind, so dass unter Umständen nicht alle Funktionen in anderen Programmen umgesetzt und unterstützt werden. DXF gibt es sowohl in einer textbasierten ASCII-Variante, als auch als Binärformat, wobei für die Archivierung erstere bevorzugt werden sollte. Maße werden von DXF nicht unterstützt, weshalb die Angabe des Maßstabes und der Zeicheneinheiten erforderlich ist.

DWG (abgeleitet vom englischen *drawing*) ist das native Speicherformat von AutoCAD und weder offen dokumentiert noch standardisiert. Mit jeder neuen Version von AutoCAD wird auch eine neue Version des DWG-Formates veröffentlicht. Da AutoCAD aber eine so weite Verbreitung gefunden hat, wird dieses Format auch von anderen CAD-Programmen unterstützt. Eine frei verfügbare nachkonstruierte Spezifikation wird von der Open Design Alliance zur Verfügung gestellt.

Für die Archivierung von CAD-Dateien sollten bis zur Etablierung von geeigneten Archivformaten neben dem Originalformat auch Versionen der Datei als DXF und als DWG vorgehalten werden. Um eine möglichst hohe Nachnutzbarkeit zu gewährleisten, sollte eine ältere verbreitete Version der Formate verwendet werden, wie beispielsweise DWG 2010 (AC1024) und DXF 2010 (AC1024).

Für alle Vektorgrafiken und insbesondere für CAD-Daten gilt die zusätzliche Empfehlung, verschiedene Druckansichten in geeigneten Formaten, wie etwa PDF/A, zu speichern. Dadurch bleibt das intendierte Aussehen der Ursprungsgrafik erhalten auch wenn die direkte Bearbeitbarkeit verloren geht. Jedoch können Vektorgrafiken größtenteils auch wieder aus einer PDF/A-Datei extrahiert werden. Nähere Informationen zu dem Format sind im Abschnitt über PDF-Dokumente ab Seite 53 zu finden.

Externe zugehörige Dateien müssen ebenfalls in einem geeigneten Archivformat gespeichert werden.

PostScript (PS) ist eine Sprache zur Beschreibung von Seiten, die von Adobe entwickelt wurde, und einen Grundstein des PDF-Formates bildet. Darauf aufbauend wurde das Format Encapsulated PostScript (EPS) entwickelt, das neben der Beschreibung in PostScript auch eine eingebettete Datei enthält, die eine Vorschau auf den Inhalt der Datei ermöglicht. Die Vorschaudatei wird jedoch nicht einheitlich gespeichert, was zu einer eingeschränkten Interoperabilität von EPS führt. Ferner kann EPS keine Ebenen und Transparenz speichern. Für die Langzeitarchivierung sollte daher SVG oder PDF/A verwendet werden.

Proprietäre Formate für Vektorgrafiken, wie etwa die Formate AI von Adobe Illustrator und INDD von Adobe InDesign, sind nicht für die Langzeitarchivierung geeignet. Dies trifft ebenfalls auf das von Autodesk entwickelte

Format DWF zu, da es primär für Visualisierungen gedacht ist und nicht für den vollständigen Datenaustausch. In der Industrie werden als alternative offene Austauschformate auch IGES und STEP für CAD-Daten und IFC für Gebäudedatenmodellierung verwendet.

Format	Begründung
✓ SVG	Ein XML-basierter offener Standard vom W3C. Für die Archivierung sollte unkomprimiertes SVG 1.1 verwendet werden.
~ WebCGM	Vom W3C 2001 als Variante von CGM veröffentlicht, um die Verwendung im Internet zu ermöglichen. Wenn es verwendet wird, dann in Version 2.1, wobei SVG für die Langzeitarchivierung besser geeignet ist.
DXF	Das Drawing Interchange Format wurde von Autodesk entwickelt und ist offen dokumentiert. Da Maße nicht gespeichert werden, ist die Angabe des Maßstabes und der Zeicheneinheiten erforderlich. Um eine hohe Nachnutzbarkeit zu erlauben, sollte eine ältere Version, wie etwa DXF 2010 (AC1024), verwendet werden.
DWG	Ist ein proprietäres Format von Autodesk und sollte nur verwendet werden, wenn eine Speicherung in DXF einen erheblichen Funktionsverlust darstellen würde. Um eine hohe Nachnutzbarkeit zu erlauben, sollte eine ältere Version, wie etwa DWG 2010 (AC1024), verwendet werden.
PDF/A	PDF/A ist für die langfristige Speicherung von zweidimensionalen Vektorgrafiken und Druckansichten von CAD-Daten geeignet. Jedoch bleibt nur das Aussehen erhalten. Nähere Informationen zu dem Format sind im Abschnitt über PDF-Dokumente ab Seite 53 zu finden.
✗ CGM	Ist unter ISO/IEC 8632 standardisiert, jedoch veraltet. Stattdessen sollte SVG verwendet werden.
✗ PS, EPS	Encapsulated PostScript beruht auf PostScript, einer Sprache zur Beschreibung von Seiten für den Druck, die von Adobe entwickelt wurden. Für die Archivierung sollte SVG oder PDF/A verwendet werden.
AI, INDD	Proprietäre Formate von Adobe Illustrator und Adobe InDesign sind nicht für die Langzeitarchivierung geeignet.
DWF	Wurde von Autodesk für die Visualisierung von CAD-Daten entwickelt und ist nicht für die Langzeitarchivierung geeignet.

Dokumentation Neben den allgemeinen minimalen Angaben zu Einzeldateien, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 aufgelistet sind, werden für Vektorgrafiken und insbesondere für CAD-Daten weitere Angaben benötigt.

Neben technischen Informationen sollten vor allem auch beschreibende und administrative Metadaten über die Datei erfasst werden. Dabei können manche Metadaten auch Bestandteil der Grafik selbst sein, wie beispielsweise eine Legende, die über Zeichnungskonventionen für verwendete Strichstärken oder Farben auskunft gibt.

Zur Dokumentation von CAD-Daten gehören unter anderem die Angabe der Quellen oder Messdaten, die zur Erstellung der Zeichnung herangezogen wurden, die Dokumentation der wesentlichen Arbeitsschritte, die im Verlauf der Erstellung durchlaufen wurden, die Dokumentation der angewendeten Methoden und eine Beschreibung der Abhängigkeitsverhältnisse zwischen den unterschiedlichen Bestandteilen, wie etwa den Ebenen.

Metadatum	Beschreibung
Identifikator	Name der Datei, z.B. grabung01.dxf
Dateiformat & Version	z.B. DXF Release 14
Farbraum	Der in dem Bild verwendete Farbraum, z.B. RGB oder Graustufen
Farbtiefe	z.B. 24 bit oder 8 bit
Bildunterschrift	Der Titel oder eine passende Unterschrift des Bildes
Beschreibung	Beschreibung der Datei
Urheber	Name des Erstellers oder der Bearbeiter
Datum	Datum der Erstellung oder letzten Änderung der Datei
Rechte	Details zum Urheberrecht
Schlagworte	Schlagworte, wie z.B. Periode, Fundstelle oder charakteristische Merkmale. Wenn vorhanden, angemessene Thesauri verwenden
Ort	Ortsinformationen zu der Datei. Möglichst in einem standardisierten Format angeben, wie z.B. Lat/Long oder Schlagworte aus einem geeigneten Thesaurus, z.B. Getty Thesaurus of Geographic Names oder GeoNames
Software	Angaben zu Software (inklusive Plugins) und Version mit der die Datei erstellt oder bearbeitet wurde, wie z.B. Autodesk AutoCAD 2012
Externe Dateien	Eine Liste aller externen zugehöriger Dateien, wie etwa Messdaten, Rastergrafiken oder eingebundener Datenbanken
Maßstab	Angabe zum verwendeten Maßstab, bzw. was eine Zeicheneinheit darstellt
Koordinatensysteme	Falls vorhanden, verwendetes Koordinatensystem oder Kartenprojektion angeben
Konventionen	Dokumentation der Bedeutung von Farben, Liniestärken, Linientypen, Füllarten etc.
Ebenen	Angabe der Ebenen, deren Inhalt und der Konventionen für deren Benennung und Befüllung
Aufnahmegerät	Beispielsweise Details zum Laserscanner
Messgenauigkeit	Angaben zur Messgenauigkeit des Aufnahmegerätes
Aufnahmemethode	Angabe zur verwendeten Methode zur Datengewinnung

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Quellen

- Archaeology Data Service, Vector Images: A Guide to Good Practice
http://guides.archaeologydataservice.ac.uk/g2gp/VectorImg_Toc
- A. Ball, Preserving Computer-Aided Design (CAD), DPC Technology Watch Report 13-02 April 2013 (2013)
<http://dx.doi.org/10.7207/twr13-02>
- D. Bibby – R. Göldner, CAD-Daten, in: AG Archivierung (Hrsg.), Ratgeber zur Archivierung digitaler Daten, Version 1.0 (2011)
http://www.landesarchaeologen.de/fileadmin/Dokumente/Dokumente_Kommissionen/Dokumente_Archaeologie-Informationssysteme/Dokumente_AIS_Archivierung/Ratgeber_Archivierung_V1.0.pdf
- CADexchange (Hrsg.), Empfehlung CAD-Datenformate (2011)
<http://www.cadexchange.ch/de/index.php?section=egov&cmd=detail&id=20>
- A. Cocciole, Digitally Archiving Architectural Models and Exhibition Designs: The Case of an Art Museum, Practical Technology for Archives 4, 2015
https://practicaltechnologyforarchives.org/issue4_cocciole/
- M. Coyne – D. Duce – B. Hopgood – G. Mallen – M. Stapleton, The Significant Properties of Vector Images (2007)
www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf
- D. Duce – B. Hopgood – M. Coyne – M. Stapleton – G. Mallen, SVG and the Preservation of Vector Images (2008)
http://www.svgopen.org/2008/papers/40-SVG_and_the_Preservation_of_Vector_Images/
- H. Eiteljorg, II, The CSA CAD Guide for Archaeologists and Architectural Historians (08. 2010)
<http://www.csanet.org/inftech/cadgd/cadgd.html>
- D. Evans, PostScript vs. PDF
<https://web.archive.org/web/20160413212438/https://www.adobe.com/print/features/psvspdf/>
- K. Green – K. Niven – G. Field, Migrating 2 and 3D Datasets: Preserving AutoCAD at the Archaeology Data Service, ISPRS International Journal of Geo-Information 5, 44, 2016
<http://dx.doi.org/10.3390/ijgi5040044>
- S. Kirch, Archäoinformatik – Digitale Archäologie. Informationstechnologien und Visualisierungstechniken in der Archäologie (2010)
https://www.uni-due.de/imperia/md/content/computerlinguistik/digitale_arch__ologie___berarbeitet___-sebastian_kirch.pdf
- Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (Hrsg.) Katalog archivischer Dateiformate: Vektorgrafiken
<http://www.kost-ceco.ch/wiki/welph/KaD/pages/Vektorgrafiken.html>
- Vereinigung der Landesdenkmalpfleger in der Bundesrepublik Deutschland (Hrsg.), Empfehlungen zum Umgang mit digitalen Baudokumentationen für eine Langzeitarchivierung, Arbeitsblatt Nr. 30 (2009)
<http://www.denkmalpflege-forum.de/Download/Nr30.pdf>

- W3C (Hrsg.), About SVG. 2d Graphics in XML (2004)
<https://www.w3.org/Graphics/SVG/About.html>
- H. Wright, Archaeological Vector Graphics and SVG: A case study from Cricklade, Internet Archaeology 20, 2006
<http://dx.doi.org/10.11141/ia.20.1>
- Wikipedia Computer Graphics Metafile (englisch, 10. 2016)
https://en.wikipedia.org/wiki/Computer_Graphics_Metafile
- Wikipedia Scalable Vector Graphics (deutsch, 10. 2016)
https://de.wikipedia.org/wiki/Scalable_Vector_Graphics

Formatspezifikationen

- SVG 1.1:
<https://www.w3.org/TR/SVG/>
- SVG 1.1: Library of Congress
<http://www.digitalpreservation.gov/formats/fdd/fdd000020.shtml>
- SVG 2.0:
<https://www.w3.org/TR/SVG2/>
- SVG Tiny 1.2:
<https://www.w3.org/TR/SVGTiny12/>
- CGM:
<http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html>
- WebCGM 2.1:
<https://www.w3.org/TR/webcgm21/>
- DXF:
<http://usa.autodesk.com/adsk/servlet/index?siteID=123112&id=24240325>
- DXF 2010:
http://images.autodesk.com/adsk/files/acad_dxf1.pdf
- DWG: Open Design Alliance
https://www.opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf
- EPS:
http://partners.adobe.com/public/developer/en/ps/5002.EPSF_Spec.pdf

3.5 Tabellen

M. Trognitz

Tabellen werden verwendet, um Informationen strukturiert in Zellen zu speichern, die in Zeilen und Spalten angeordnet sind. Im Gegensatz zu analogen Tabellen bieten digitale Tabellen, die mit entsprechenden Programmen erstellt werden, eine Vielzahl an weiteren Funktionalitäten. Beispielsweise können Zellinhalte nach bestimmten Kriterien sortiert, dynamisch mittels Formeln erzeugt oder Grafiken aus den Daten generiert werden. Um solche und andere Funktionen zu erhalten, erfordert die Speicherung besondere Aufmerksamkeit.

Langzeitformate Für einfache Tabellen ohne interaktive Elemente wie Formeln oder für Tabellenkalkulationen bei denen es ausreicht die Ergebnisse der Formeln zu speichern, wird ein textbasiertes Format mit Trennzeichen für die Archivierung empfohlen.

Dazu eignet sich beispielsweise das CSV-Format, wobei als Trennzeichen ein Komma (,) und als Textbegrenzungszeichen das Anführungszeichen (") verwendet werden sollte, um den Vorgaben von RFC 4180 gerecht zu werden. Andere Trennzeichen und Textbegrenzungszeichen können in begründeten Ausnahmefällen ebenfalls eingesetzt werden und müssen entsprechend dokumentiert werden. Als Zeichenkodierung sollte UTF-8 ohne BOM verwendet werden.

Ein alternatives textbasiertes Format für Tabellen ist das TSV-Format, das als Trennzeichen das Tabulator-Zeichen (U+0009) verwendet. Auch TSV-Dateien sollten UTF-8 ohne BOM als Zeichenkodierung verwenden.

Bei Dateien mit mehr als einem Arbeitsblatt (Tabellenkalkulationen), die als CSV- oder TSV-Datei gespeichert werden sollen, muss jedes Arbeitsblatt gesondert gespeichert werden. Dabei gilt für die Dateinamen der Arbeitsblätter, dass der Name des Arbeitsblattes an den Namen der Tabellenkalkulation, am besten durch ein Unterstrich (_) getrennt, angefügt wird (z. B. Tabellennname_Blatt1.csv, Tabellennname_Blatt2.csv usw.).

Tabellenkalkulationen, deren zusätzlichen Funktionalitäten erhalten bleiben sollen, werden am besten in einem offenen auf XML basierenden Format gespeichert, wie beispielsweise XLSX oder ODS. Ersteres ist das Standardformat, das in Microsoft Excel seit 2007 verwendet wird und auch von Microsoft entwickelt wurde. Letzteres ist das Format für Tabellenkalkulationen, welches in OpenOffice oder LibreOffice verwendet wird. ODS ist ein Teil vom OpenDocument Format (ODF) und wurde von einem technischen Komitee unter der Leitung der *Organization for the Advancement of Structured Information Standards* (OASIS) entwickelt.

Grafiken, die in Tabellenkalkulationen anhand der Daten erstellt wurden, müssen zusätzlich exportiert und gesondert in einem geeigneten Format gespeichert werden. Dies gilt ebenfalls für eingebettete Bilder oder andere Medien. Passende Formate sind beispielsweise in dem Abschnitt Rastergrafiken ab Seite 78 oder in dem Abschnitt Vektorgrafiken ab Seite 90 zu finden.

Tabellen können auch im XML-Format gespeichert werden. Es gibt eine ganze Reihe an DTDs oder XSDs, die hier als Grundlage dienen können, wie beispielsweise das Schema der TEI, das OASIS Exchange Table Model oder ISO 12083. Auch das Speichern von Tabellen im HTML-Format ist möglich. In jedem Fall muss die zugrundeliegende DTD oder XSD angegeben und ge-

gebenenfalls mit archiviert werden. Die Dateien sollten UTF-8 ohne BOM als Zeichenkodierung verwenden.

Wenn neben den eigentlichen Daten in den einzelnen Zellen auch das Aussehen der Tabelle archiviert werden soll, kann neben einer CSV-, TSV-, XLSX- oder ODS-Datei zusätzlich eine Version der Tabelle im PDF/A-Format gespeichert werden. Bei der Erstellung von Tabellen sollte aber darauf geachtet werden, dass Informationen nicht nur durch Formatierungsangaben, wie beispielsweise die Farbe von Zellen, vermittelt werden, da je nach gewähltem Format die Formatierungsangaben verloren gehen können.

Hinweis: Obwohl Textverarbeitungsprogramme entsprechende Funktionen bieten, sollten Tabellen auch tatsächlich als Tabellen in einem der hier gelisteten Formate gespeichert werden.

Format	Begründung
CSV ✓	Das textbasierte CSV-Format sollte mit einem Komma als Trennzeichen und mit Anführungszeichen als Textbegrenzungszeichen verwendet werden. Ausnahmen müssen dokumentiert werden. Die Zeichen sollten in UTF-8 ohne BOM kodiert sein.
TSV	TSV (MIME-Type <code>text/tab-separated-values</code>) ist ein textbasiertes Format, welches das Tabulator-Zeichen (<code>U+0009</code>) als Trennzeichen verwendet. Die Zeichen sollten in UTF-8 ohne BOM kodiert sein.
ODS	ODS basiert auf XML und ist Teil vom Open-Document Format. Eingebettete Bilder und Medien müssen gesondert gespeichert werden.
XLSX	XLSX ist das auf XML basierende Format von Microsoft. Eingebettete Bilder und Medien müssen gesondert gespeichert werden.
XML, HTML	Tabellen im textbasierten XML- oder HTML-Format können ebenfalls archiviert werden. XML-Dateien benötigen zusätzlich eine DTD-Datei oder das XML Schema. Die Zeichen sollten in UTF-8 ohne BOM kodiert sein.
PDF/A ~	Wenn neben der Daten auch das Aussehen der Tabelle erhalten bleiben soll, eignet sich PDF/A am besten. Zusätzlich sollten die tabellarischen Daten in einem nachnutzbaren Format, wie etwa CSV, gespeichert werden.
SXC	SXC ist ein Vorgängerformat von ODS, weshalb letzteres auch bevorzugt werden sollte.
XLS ✗	Das XLS-Format von Microsoft eignet sich nicht zur Archivierung, da es proprietär ist und die Inhalte nicht textbasiert gespeichert werden.

Dokumentation Neben den allgemeinen minimalen Angaben zu Einzeldateien, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 gelistet sind, werden für Tabellen und Tabellenkalkulationen weitere Angaben benötigt.

Um die Verständlichkeit einer Tabelle auch für Dritte zu gewährleisten, müssen Name und Zweck der jeweiligen Tabelle und der einzelnen Arbeitsblätter bekannt sein. Jede Spalte benötigt eine Überschrift, und zusätzlich müssen die verwendeten Formatvorgaben, Abkürzungen, Codes, Wertelisten und sonstige Terminologien dokumentiert werden. Um leere Zellen auch explizit als solche zu kennzeichnen, sollte ein vorher festgelegtes Zeichen (z.B. -) eingetragen und dokumentiert werden. Wenn Maßeinheiten nicht direkt aus der Tabelle ersichtlich sind, müssen diese ebenfalls gesondert dokumentiert werden.

Um sicher zu gehen, dass die Tabelle auch vollständig vorliegt, sollten die Anzahl der Spalten, Zeilen und der Arbeitsblätter angegeben werden.

Tabellen in textbasierten Formaten brauchen Angaben zu den verwendeten Trennzeichen, Textbegrenzungszeichen und der Zeichenkodierung.

Für Tabellenkalkulationen müssen bei Bedarf weitere Informationen zu Relationen, Formeln und Makros dokumentiert werden. Eingebettete Medien, wie etwa Bilder, sollten separat gespeichert und archiviert werden und in einer Liste zugehöriger Dateien aufgeführt werden. Dies gilt ebenfalls für Grafiken, die aus den Daten in der Tabellenkalkulation erzeugt wurden.

Metadatum	Beschreibung
Beschreibung der Tabelle oder des Arbeitsblattes	Welchen Zweck verfolgt die Tabelle oder das Arbeitsblatt?
Bezeichnung der Arbeitsblätter	Auflistung der Bezeichnungen der Arbeitsblätter.
Spaltenüberschrift	Jede Spalte einer Tabelle muss einen Namen haben.
Spaltenbeschreibung	Beschreibung und Auflistung der in der jeweiligen Spalte verwendeten Formatvorgaben, Abkürzungen, Codes, Wertelisten, Eingabekonventionen, Fachvokabulare, Zeichen für leere Zellen oder Maßeinheiten.
Anzahl Spalten	Wie viele Spalten enthält die Tabelle?
Anzahl Zeilen	Wie viele Zeilen enthält die Tabelle?
Anzahl Arbeitsblätter	Anzahl der Arbeitsblätter in einer Tabellenkalkulation.
Trennzeichen	Angabe des verwendeten Trennzeichens bei textbasierten Speicherformaten wie CSV.
Textbegrenzungszeichen	Angabe des verwendeten Textbegrenzungszeichens bei textbasierten Speicherformaten wie CSV.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Metadatum	Beschreibung
Zeichenkodierung	Angabe der verwendeten Zeichenkodierung bei textbasierten Speicherformaten wie CSV oder TSV.
Relationen	Welche Querverweise gibt es innerhalb der Tabellenkalkulation?
Formeln	Welche Formeln werden in der Tabellenkalkulation verwendet?
Makros	Welche Makros gibt es in der Tabellenkalkulation?
Abgeleitete Grafiken	Aus den Daten erzeugte Grafiken müssen zusätzlich separat gespeichert werden und in der Liste zugehöriger Dateien aufgenommen werden.
Sprache	In welchen Sprachen ist das Dokument verfasst? Sprachkennungen nach ISO 639 angeben.
Identifikator	Wenn das Dokument bereits veröffentlicht wurde und eine ISBN oder einen anderen persistenten Identifikator erhalten hat, sollte dieser angegeben werden.
Weitere Dateien	Liste abgeleiteter Grafiken und eingebetteter Medien, wie Bilder, die zusätzlich separat gespeichert wurden. Liegt eine Dokumentationsdatei für das Dokument vor, muss diese ebenfalls genannt werden.

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Vertiefung

Tabellen bestehen in ihrer einfachsten Form aus Zeilen und Spalten, deren Überschneidung eine Zelle bilden, in der ein Wert eingetragen wird. Die Namen der Spalten werden in die erste Zeile, der Kopfzeile (*header*), einge tragen. In der ersten Spalte, der Vorspalte, können die Zeilenbezeichnungen stehen.

Digitale Tabellen und Tabellenkalkulationen (*spreadsheets*) bieten erweiterte Funktionalitäten.

Für die Langzeitarchivierung von Tabellen sind strukturierte Textdateien mit Trennzeichen (*delimiter separated*

A	B	C	D	
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4

Abb. 3.11: Die Spalte *B* einer Tabelle ist blau und die Zeile 3 gelb markiert. Die Überschneidung aus Zeile und Spalte in grün ist die Zelle *B3*. Die erste Spalte stellt die Vorspalte und die erste Zeile die Kopfzeile dar.

values), wie CSV- oder TSV-Dateien, gut geeignet. Um Tabellen als XML-Dateien zu speichern, gibt es verschiedene XML Schemata und Dokumenttypdefinitionen.

Tabellenkalkulationen Als Tabellenkalkulation wird sowohl das Programm, als auch die damit erstellte Datei mit einer oder mehreren Tabellen bezeichnet. Im Folgenden ist die resultierende Datei gemeint.

Eine Tabellenkalkulation kann mehrere Tabellen, sogenannte Arbeitsblätter (*worksheets* oder *spreadsheets*) enthalten. Die Daten selbst können innerhalb der Tabellenkalkulation verwendet werden, um mittels Formeln neue Daten zu erzeugen oder Grafiken zu erstellen. Dazu erlauben die Programme eine Referenzierung auf Werte in anderen Zellen, die sich auch in einem anderen Arbeitsblatt befinden können. Es handelt sich dabei um Relationen.

Die gängigsten Tabellenkalkulationsprogramme ermöglichen die Erstellung und Verwendung von Makros, mit denen typische Befehlsfolgen und Bedienungsschritte aufgezeichnet und automatisiert wiederholt werden können.

Formatierungsangaben können in digitalen Tabellen und Tabellenkalkulationen ebenfalls vorgenommen werden. Dabei können wie in Textdokumenten nicht nur die Schriftart, Schriftgröße und ähnliches angepasst werden, sondern auch das Aussehen von Zellen, Zeilen und Spalten mittels Angaben zu Rahmenlinien und Hintergrundfarben. Soll der durch die Formatierungsangaben erzeugte optische Eindruck ebenfalls archiviert werden, weil er für das Verständnis der Tabelle zentrale Informationen transportiert, empfiehlt sich eine zusätzliche Speicherung der Tabellenkalkulation als PDF/A-Dokument.

Strukturierte Textdateien mit Trennzeichen Tabellen können in einem textbasierten Format gespeichert werden. Dabei handelt es sich um eine Textdatei, die auf eine bestimmte Weise strukturiert ist. Für Tabellen gibt es insbesondere die Formate CSV (*comma-separated values*) und TSV (*tab-separated values*), die sich nur geringfügig voneinander unterscheiden. Textbasierte Formate berücksichtigen keine Formatierungsangaben, sondern speichern nur die reinen Werte aus jeder Zelle ab.

Jede Zeile einer CSV- oder TSV-Datei entspricht der Zeile einer Tabelle. Die einzelnen Zellen werden durch sogenannte Trennzeichen voneinander getrennt. In TSV-Dateien ist dies das Tabulator-Zeichen (U+0009), welches nach dem Standard der *Internet Assigned Numbers Authority* (IANA) nicht als Inhalt der Zellen erlaubt ist.

	A	B	C	D
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4

Abb. 3.12: Die Tabelle aus Abbildung 3.11 im TSV-Format.

Für CSV-Dateien gibt es mit RFC 4180 bisher nur einen De-facto-Standard, der als Trennzeichen ein Komma (,) vorsieht. Da Kommas innerhalb von Zellen auch als Wert erlaubt werden, sind für diesen Fall Anführungszeichen (")

als Textbegrenzungszeichen vorgesehen, die vor und nach der jeweiligen Zelle eingefügt werden. In Abbildung 3.13 ist links ein Beispiel einer CSV-Datei zu sehen und rechts daneben wie diese in einem Tabellenkalkulationsprogramm dargestellt wird.

Objektnr., Material, Gew. (g), Dm. (mm)	Objektnr.	Material	Gew. (g)	Dm. (mm)
18214973, Silber, "16,96", 24-28	18214973	Silber	16,96	24-28
18200244, Gold, "7,80", 20	18200244	Gold	7,80	20
18202951, Gold, "17,19", 22	18202951	Gold	17,19	22
18204771, Bronze, "3,67", 17	18204771	Bronze	3,67	17

(a)

(b)

Abb. 3.13: (a) Eine CSV-Datei. In der Spalte 'Gew. (g)' sind die Werte jeweils mit Textbegrenzungszeichen (") versehen, da innerhalb des Wertes ein Komma verwendet wird. (b) Die CSV-Datei aus a, wie sie in einem Tabellenkalkulationsprogramm dargestellt werden könnte. Die Tabelle stellt Informationen zu Münzen aus dem Münzkabinett der Staatlichen Museen zu Berlin dar.

Für den Fall, dass das Textbegrenzungszeichen auch innerhalb einer Zelle verwendet werden soll, sieht das RFC 4180 vor, dass dieses gedoppelt wird und die Zelle zusätzlich von den Textbegrenzungszeichen umschlossen wird. Die Abbildung 3.14 verdeutlicht die Verwendung der Textbegrenzungszeichen in den Zellwerten.

Gold, """Au""", "3,67", "0,67"""	Gold	"""Au"""	3,67	0,67"""
----------------------------------	------	----------	------	---------

(a)

(b)

Abb. 3.14: (a) Eine CSV-Datei, in der Textbegrenzungszeichen (") innerhalb der Zellen verwendet werden. (b) Die CSV-Datei aus a, wie sie in einem Tabellenkalkulationsprogramm dargestellt werden könnte.

In der Praxis werden auch andere Trennzeichen und Textbegrenzungszeichen in CSV-Dateien verwendet. Beispielsweise ist das Trennzeichen in den von Microsoft Excel gespeicherten CSV-Dateien ein Semikolon (;) statt eines Kommas. Diese Abweichungen von der Empfehlung des RFC 4180 müssen in den Metadaten angegeben werden. Außerdem ist zu beachten, dass die Verwendung von mehreren verschiedenen Trennzeichen oder Textbegrenzungszeichen in einer Datei nicht erlaubt ist.

Das Speichern von Tabellen in strukturierten Textdateien bringt einige Einschränkungen mit sich. Es können keine Formatierungsangaben oder Makros gespeichert werden. Im Gegensatz zu einer Tabellenkalkulation, wo bei Formeln sowohl die Formel als auch deren Ergebnis gespeichert ist, kann in textbasierten Tabellenformaten entweder nur die Formel selbst oder das Ergebnis gespeichert werden.

Weiterhin ist es nicht möglich verbundene Zellen zu speichern. Bei verbundenen Zellen handelt es sich im Prinzip um einen visuellen Effekt, bei dem eine Zelle die anderen Zellen überdeckt. Dementsprechend wird in einem textbasierten Format der Wert aus den verbundenen Zellen in die erste Zelle aus der

Gruppe geschrieben, während die übrigen Zellen leer bleiben. Wird die so gespeicherte Tabelle später wieder geöffnet, ist der visuelle Effekt der verbundenen Zellen nicht mehr sichtbar, was Abbildung 3.15 verdeutlicht.

	A	B	C	D
1	A1-C1		D1	
2	A2	B2	C2	D2
3	A3	B3	C3	D3

(a)

	A,B,C,D
1	A1-C1,,,D1
2	A2,B2,C2,D2
3	A3,B3,C3,D3

(b)

	A	B	C	D
1	A1-C1			D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

(c)

Abb. 3.15: (a) Eine Tabelle mit verbundenen Zellen in einem Tabellenkalkulationsprogramm. (b) Die Tabelle aus (a) im CSV-Format. Der Wert aus den verbundenen Zellen steht in der ersten der vorher verbundenen Zellen. (c) Wird die CSV-Tabelle aus (b) wieder in einem Tabellenkalkulationsprogramm geöffnet, werden die Zellen auch nicht mehr als verbunden angezeigt.

Tabellen mit Auszeichnungssprachen Tabellen können auch als Textdatei mit Auszeichnungssprachen wie XML oder HTML gespeichert werden. Allgemeine Eigenschaften von Auszeichnungssprachen werden in dem Kapitel „Textdokumente“ in dem Abschnitt „Auszeichnungssprachen“ ab Seite 70 beschrieben.

Vom Prinzip her sind Tabellen mit Auszeichnungssprachen ebenfalls strukturierte Textdateien, die aber weitere Angabemöglichkeiten beispielsweise für Metadaten oder verbundene Zellen bieten.

Für das XML-Format gibt es das OASIS Exchange Table Model (DTD) oder das von der TEI speziell für die Geistes-, Sozial- und Sprachwissenschaften entwickelte XSD-Schema. Ersteres stammt von dem Tabellenmodell CALS ab, welches von dem US-Verteidigungsministerium entwickelt wurde.

Speziell für SGML gibt es den Standard ISO 12083, der entwickelt wurde, um Publikationen auszuzeichnen.

Auch HTML und TeX stellen für die Eingabe von Tabellen jeweils eine spezielle Syntax zur Verfügung, was Abbildung 3.16 veranschaulicht.

Wichtig für die Archivierung von Tabellen mit Auszeichnungssprachen ist, dass alle Dateien wohlgeformt und valide sind, also die Regeln der jeweiligen Auszeichnungssprache und deren Grammatik einhalten. Die jeweils verwendeten DTD- oder XSD-Dateien müssen in jedem Fall angegeben werden und gegebenfalls mit archiviert werden. Die Dateien selbst sollten UTF-8 ohne BOM als Zeichenkodierung verwenden.

Praxis

Dieser Abschnitt liefert Hinweise zum Umgang mit Tabellen und Tabellenkalkulationen in der Praxis und stellt Tabellenkalkulationsprogramme vor. Es wird erläutert, wie Tabellen als CSV- oder TSV-Datei gespeichert und wie Grafiken aus Tabellenkalkulationen exportiert werden können. Für die Bereinigung von Tabellen wird eine automatisierte Lösung vorgeschlagen.

Tabellenkalkulationsprogramme Für die Bearbeitung von Tabellen und Tabellenkalkulationen gibt es dezidierte Tabellenkalkulationsprogramme, wie

Material	Gewicht	Objektnr.
Bronze	3,67	18204771
Silber	16,96	18214973
Gold	7,80	18200244
	17,19	18202951

(a)

```
\begin{tabular}{| r | c | c | c |}
\hline Material & Gewicht & Objektnr. \\
\cline{2-4}
Bronze & 3,67 & 18204771 \\
\cline{2-4}
Silber & 16,96 & 18214973 \\
\cline{2-4}
\multirow{2}{*}{Gold} & 7,80 & 18200244 \\
\cline{2-3}
& 17,19 & 18202951 \\
\cline{2-4}
\hline
\end{tabular}
```

(b)

```
<table rows="4" cols="3">
<head>Muenzgewichte</head>
<row role="label">
<cell>Material</cell>
<cell>Gewicht</cell>
<cell>Objektnr.</cell>
</row>
<row>
<cell><material>Bronze</material></cell>
<cell>
<gewicht einheit="gramm">3,67</gewicht>
</cell>
<cell>18204771</cell>
</row>
<row>
<cell><material>Silber</material></cell>
<cell>
<gewicht einheit="gramm">16,96</gewicht>
</cell>
<cell>18214973</cell>
</row>
<row>
<cell><material>Gold</material></cell>
<cell>
<gewicht einheit="gramm">7,80</gewicht>
</cell>
<cell>18200244</cell>
</row>
<row>
<cell><material>Gold</material></cell>
<cell>
<gewicht einheit="gramm">17,19</gewicht>
</cell>
<cell>18202951</cell>
</row>
</table>
```

(c)

```
<table>
<tr>
<th>Material</th>
<th>Gewicht</th>
<th>Objektnr.</th>
</tr>
<tr>
<td>Bronze</td>
<td>3,67</td>
<td>18204771</td>
</tr>
<tr>
<td>Silber</td>
<td>16,96</td>
<td>18214973</td>
</tr>
<tr>
<td rowspan="2">Gold</td>
<td>7,80</td>
<td>18200244</td>
</tr>
<tr>
<td>17,19</td>
<td>18202951</td>
</tr>
</table>
```

(d)

Abb. 3.16: Eine Tabelle (a), wie sie in TeX (b), mit XML (c) und in HTML (d) dargestellt werden könnte.

OpenOffice Calc, LibreOffice Calc oder Microsoft Excel. OpenOffice und LibreOffice speichern Tabellenkalkulationen standardmäßig im ODS-Format. Seit 2007 speichert Microsoft Word im XLSX-Format. Beide Formate sind offen dokumentiert, basieren auf XML und sind für die Langzeitarchivierung geeignet.

Aus den Daten erzeugte Grafiken, eingebettete Bilder oder andere Medien sollten zusätzlich als separate Dateien in einem geeigneten Langzeitformat ge-

speichert werden. Dies stellt sicher, dass die Qualität der ursprünglichen Datei erhalten bleibt, automatisch erzeugte Grafiken nicht verloren gehen und wie ursprünglich intendiert aussehen.

Die Darstellung von Tabellenkalkulationen mit umfangreichen Formatierungsangaben kann auf verschiedenen Computern unterschiedlich ausfallen, was vor allem an unterschiedlichen Einstellungen liegt. Wenn bestimmte Schriftarten auf einem System fehlen, werden sie automatisch ersetzt, was ebenfalls zu unterschiedlichen Darstellungsweisen führt. Wenn eine stabile systemübergreifende Darstellung von Tabellenkalkulationen gewünscht ist, kann dies nur mittels Konvertierung in ein PDF-Dokument gewährleistet werden. Für die Langzeitspeicherung sollte PDF/A verwendet werden. Hinweise zum Erstellen von PDF- und PDF/A-Dokumenten sind im Praxisteil zu PDF-Dokumenten ab Seite 58 zu finden.

OpenOffice Calc:

<https://www.openoffice.org/>

LibreOffice Calc:

<http://www.libreoffice.org/>

Tabellen als CSV- oder TSV-Dateien speichern Alle gängigen Tabellenkalkulationsprogramme bieten Exportfunktionen zur Speicherung von Tabellen als CSV- oder TSV-Dateien. Dabei ist zu beachten, dass jeweils nur das aktuell gewählte Arbeitsblatt gespeichert wird und Formatierungsangaben und Makros nicht gespeichert werden. Bei den Formeln wird üblicherweise nur der Ergebniswert gespeichert, wenn keine andere Einstellung vorgenommen wird.

Eine TSV-Datei kann in Microsoft Excel über „Datei > Speichern unter“ mit dem Dateityp „Text (Tabstopp-getrennt) (*.txt)“ erzeugt werden. Die Datei wird als TXT-Datei gespeichert, erfüllt jedoch alle Anforderungen an eine TSV-Datei.

Für das CSV-Format bietet Microsoft Excel über „Datei > Speichern unter“ gleich mehrere Optionen, von denen „CSV (Trennzeichen-getrennt) (*.csv)“ verwendet werden sollte, da „CSV (Macintosh) (*.csv)“ und „CSV (MS-DOS) (*.csv)“ andere problembehaftete Zeichenkodierungen verwenden, die für ältere Systeme gedacht waren. Die von Microsoft Excel erzeugten CSV-Dateien werden standardmäßig mit einem Semikolon (;) als Trennzeichen und Anführungszeichen (") als Textbegrenzungszeichen gespeichert. Da abweichend von den Vorgaben des RFC 4180 als Trennzeichen ein Komma und kein Semikolon verwendet wird, ist in diesem Fall die Angabe des Trennzeichens in den Metadaten notwendig.

Microsoft Excel verwendet ANSI für die Zeichenkodierung der Dateien, weshalb dies nachträglich auf UTF-8 ohne BOM angepasst werden muss. Im Kapitel „Textdokumente“ sind im Abschnitt „Einstellen der Zeichenkodierung“ ab Seite 73 nähere Erläuterung dazu zu finden. Dieser Schritt sollte allerdings erst dann erfolgen, wenn die Tabellen nicht mehr mit Excel bearbeitet werden sollen, weil das Programm nicht mit UTF-8 kodierten Dateien umgehen kann und daher die Zeichen nicht richtig darstellt.

Wesentlich komfortabler und flexibler ist die Erstellung von CSV-Dateien mit LibreOffice Calc oder OpenOffice Calc. Über „Datei > Speichern unter“ wird der Dateityp „Text CSV (*.csv)“ angeboten. Wird bei der Opti-

on „Filtereinstellungen bearbeiten“ ein Haken gesetzt, erscheint bei dem Klick auf „Speichern“ der Dialog „Textexport“ (Abbildung 3.17), in dem die Zeichenkodierung bei „Zeichensatz“, das Trennzeichen bei „Feldtrenner“ und das Textbegrenzungszeichen bei „Texttrenner“ ausgewählt werden können.

Mit der Option „Zellinhalt wie angezeigt speichern“ wird der Inhalt der Zellen in der CSV-Datei so gespeichert, wie sie zu sehen sind. Wird die Option nicht ausgewählt, werden beispielsweise die Währungssymbole, die mittels Angaben für die Zellformatierung automatisch eingefügt wurden, nicht gespeichert. Die Option „Formeln anstatt berechneter Werten speichern“ ermöglicht die Speicherung der eingegebenen Formeln anstatt der Werte.

TSV-Dateien werden in LibreOffice Calc oder OpenOffice Calc zunächst als CSV-Datei angelegt, wobei als Feldtrenner {Tabulator} gewählt wird. Um zu verdeutlichen, dass es sich um eine TSV-Datei und nicht um eine CSV-Datei handelt, sollte die Dateiendung nachträglich manuell geändert werden.

Grafiken exportieren Grafiken, die in Tabellenkalkulationsprogrammen mittels der eingetragenen Daten erzeugt wurden, müssen für die Archivierung der Tabelle zusätzlich als gesonderte Bilddatei in einem geeigneten Format gespeichert werden. Passende Formate sind in den Kapiteln „Rastergrafiken“ ab Seite 78 und „Vektorgrafiken“ ab Seite 90 beschrieben.

Der Export von Grafiken funktioniert bei LibreOffice Calc mit einem Klick mit der rechten Maustaste auf die Grafik. In dem erscheinenden Menü gibt es den Eintrag „Als Bild exportieren“. Es erscheint ein Speicherdialog für die Grafik in dem verschiedene Formate zur Auswahl stehen. Wird die Grafik als Rastergrafik gespeichert, muss vorher auf eine geeignete Größe der Grafik geachtet werden, indem sie beispielsweise vorher in der Tabellenkalkulation vergrößert wird.

Eine alternative Exportmethode, die auch in OpenOffice Calc und Microsoft Excel funktioniert, ist das Zwischenspeichern der Grafik und anschließende Speichern in einem dezidierten Grafikprogramm. Die Grafik kann in dem Grafikprogramm bearbeitet und angepasst werden, bevor sie gespeichert wird. Auch hier muss beim Speichern auf eine geeignete Größe des Bildes geachtet werden.

Die oben beschriebenen Schritte müssen für alle Grafiken in der Tabellenkalkulation wiederholt werden, was bei Tabellenkalkulationen mit vielen Grafiken ein langwieriger Vorgang sein kann. Für diesen Fall bietet sich die Option an, die gesamte Tabellenkalkulation mittels „Datei > Speichern unter“ als HTML-Datei („Webseite“ oder „HTML-Dokument“) in einem gesonderten Ordner zu

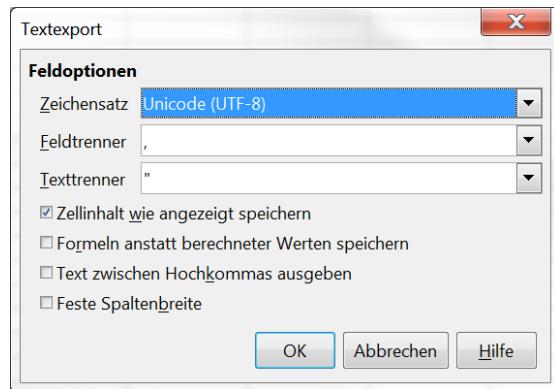


Abb. 3.17: Der Dialog „Textexport“ in LibreOffice Calc zur Speicherung von CSV-Dateien.

speichern. In dem Ordner werden alle Grafiken als PNG- oder JPG-Datei abgelegt. Bei dieser Exportvariante ist zu beachten, dass die erzeugten Einzelbilder eventuell Mängel in der Qualität aufweisen.

Bereinigung von Tabellen Bei Tabellen, die über einen längeren Zeitraum von mehreren Bearbeitern gepflegt werden, können trotz festgelegter Vorgaben Inkonsistenzen auftreten. Solche Inkonsistenzen können beispielsweise darin bestehen, dass für die Materialbeschreibung mal „Silber“ und mal „Ag“ verwendet wird. Bei einer automatisierten statistischen Auswertung könnte das zu Verzerrungen führen. Für solche Fälle kann das frei verfügbare Programm OpenRefine verwendet werden. Es ermöglicht die Aufdeckung und Bereinigung von Inkonsistenzen in Tabellen. Darauf hinaus bietet es Funktionen, um die Daten zusätzlich über Transformationen oder durch externe Webinhalte anzureichern.

OpenRefine:
<http://openrefine.org/>

Quellen

Archaeology Data Service, Databases and Spreadsheets: A Guide to Good Practice

http://guides.archaeologydataservice.ac.uk/g2gp/DbSht_Toc

T. Hicks, Should We Be Using ISO 12083?, The Journal of Electronic Publishing 3, 1998, 4

<http://quod.lib.umich.edu/j/jep/3336451.0003.407?view=text;rgn=main>

Research Data Management Service Group, Preparing tabular data for description and archiving

<http://data.research.cornell.edu/content/tabular-data>

Formatspezifikationen

ODT:

<https://www.oasis-open.org/standards#opendocumentv1.2>

XSLX: ECMA-376

<http://www.ecma-international.org/publications/standards/Ecma-376.htm>

CSV: RFC 4180

<https://tools.ietf.org/html/rfc4180>

TSV: IANA:

<http://www.iana.org/assignments/media-types/text/tab-separated-values>

XML: OASIS (Hrsg.) Standards/Specs: Table Models

<https://www.oasis-open.org/specs/tablemodels.php>

XML: OASIS (Hrsg.) XML Exchange Table Model Document Type Definition

<https://www.oasis-open.org/specs/tm9901.html>

XML: TEI (Hrsg.) P5: Guidelines for Electronic Text Encoding and Interchange

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FT.html>

HTML: W3C:

<http://www.w3.org/TR/html401/struct/tables.html>

ISO 12083:

http://www.iso.org/iso/catalogue_detail.htm?csnumber=20866

TeX:

<https://en.wikibooks.org/wiki/TeX/Tables>

Tools und Programme

OpenOffice Calc:

<https://www.openoffice.org/>

LibreOffice Calc:

<http://www.libreoffice.org/>

OpenRefine:

<http://openrefine.org/>

3.6 Datenbanken

P. Gerth, F. Schäfer, K. Hofmann

Übersicht

Datenbanksysteme (DBS) werden zur Strukturierung, Pflege und Verwaltung von Daten verwendet. Sie bestehen aus zwei Komponenten - einerseits aus der Datenbank, welche den logisch zusammengehörigen Datenbestand umfasst, sowie aus dem Datenbankmanagementsystem (DBMS), der Datenbanksoftware, welche u.a. die Zugriffsrechte regelt und Vorkehrungen zur Datensicherheit gewährleistet. Es gibt verschiedene Datenbankmodelle, von denen relationale Datenbanksysteme am häufigsten verwendet und im Rahmen der IT Empfehlungen betrachtet werden. Dabei werden die Daten in Tabellen (Relationen) strukturiert, wobei Verweise von einer Tabelle auf eine andere möglich sind. Näheres zu Tabellen ist in dem Kapitel Tabellen ab Seite 97 zu finden.

Langzeitformate Für relationale Datenbanken wird eine Archivierung in Form eines Exports der Datenbankinhalte, der Strukturinformationen und weiterer Funktionalitäten in textbasierte Formate empfohlen, welche unabhängig vom verwendeten Datenbankmanagementsystem sind. Neben den Daten in den Tabellen müssen die Datenbankstrukturdefinitionen wie Attributdatentypen, Relationen zwischen den Tabellen und Formeln zwingend mit archiviert werden. Unabhängig vom gewählten Archivformat ist es nötig, die grafische Benutzeroberfläche zu dokumentieren, zum Beispiel in Form von Bildschirmfotos der Benutzeroberfläche oder eines eventuell vorhandenen Benutzerhandbuchs. Eine darüberhinaus gehende Erhaltung der Funktionalität ist zumeist nur mit Hilfe der technisch sehr aufwendigen Emulation möglich, bei der das ursprüngliche Datenbanksystem durch ein anderes System mit ähnlicher Funktionsweise ersetzt wird. Unabhängig von den vorgestellten Archivformaten ist der Umgang mit sogenannten Binary Large Objects (BLOBs). Dabei handelt es sich um intern gespeicherte Mediendateien, meistens Fotos, Zeichnungen, z. B. von Funden, oder PDF-Dokumente. Diese Medien müssen in jedem Fall exportiert, auf die dann externen Dateien muss referenziert und je nach Dateityp in für die Archivierung geeigneten Formaten gespeichert werden (siehe entsprechendes Dateiformatkapitel). Grundsätzlich wird von der Verwendung von BLOBs abgeraten und eine externe Speicherung von Medieninhalten empfohlen.

Ein geeignetes Archivformat ist SQL (Structured Query Language), welches seit 1986 bei der ANSI standardisiert ist und seit 1987 ebenfalls als ISO/IEC 9075 zertifiziert ist. Seitdem wurde der SQL-Standard in sieben Revisionen weiter ausgebaut, wobei SQL:2011 die neueste Version ist. Jedoch wird dieser Standard in den verschiedenen Datenbanksystemen wie MySQL, PostgreSQL oder OracleDB unterschiedlich umgesetzt und erweitert. Daher gibt es je nach Datenbanksystem mehrere spezifische Befehle, welche über die Spezifikation des Standards hinausgehen. In Folge dessen sind die verschiedenen DBMS nicht vollständig kompatibel, weswegen eine exportierte Datenbank im SQL-Format nur dann problemlos und ohne zusätzlichen Aufwand importiert werden kann, wenn das ursprüngliche DBMS verwendet wird. Daher kann SQL für die Archivierung nur dann empfohlen werden, wenn ein einheitlicher ANSI-/ISO-Standard, z. B.

SQL:2008, verwendet wird, welcher von DBMS-spezifischen Befehlen bereinigt ist. Der verwendete Standard muss dokumentiert werden. Der Vorteil in der Speicherung in SQL liegt in der gleichzeitigen Erhaltung der kompletten Daten, der Tabellen und ihrer Relationen zueinander, der Attributsspezifikationen sowie der kompletten Metadaten. Nach einem Export, erhält man eine singuläre textbasierte Datei, welche sich mit relativ geringem Aufwand wiederherstellen lässt. Die Speicherung einer Datenbank in einem SQL-Skript erfolgt unter Angabe der SQL-Befehle, welche zum Import der Datenbankinhalte benötigt und ggf. für jeden Datensatz wiederholt werden müssen. Die damit einhergehende Redundanz erhöht die Datenmenge. Datenbanken aus geisteswissenschaftlichen Forschungsprojekten erreichen jedoch nur selten eine Größe, bei welcher dieser Nachteil zum Tragen kommt.

XML eignet sich als Archivformat für Datenbanken aus Gründen der Systemunabhängigkeit, Standardisierung, Lesbarkeit und Erweiterbarkeit. Für die Speicherung der Datenbankinhalte kann XML uneingeschränkt empfohlen werden. Die Beschreibung der Datenbankstruktur sollte mithilfe einer entsprechenden XML-Schema-Definition (XSD) wie der Database Markup Language erfolgen. Alternativ kann die Datenbankstruktur im SQL-Format und als Entitäten-Relationen-Diagramm (ERD) gespeichert werden.

Das 2008 initial vom schweizerischen Bundesarchiv entwickelte und im Rahmen des e-ARK-Projekts weiterentwickelte Format SIARD (Software Independent Archiving of Relational Databases) in der Version 2.0 beruht im Wesentlichen auf den beiden ISO-Standards XML gemäß ISO/IEC 19503:2005 und SQL:2008. Es vereinigt damit die Vorteile einer Datenhaltung in XML mit einer Dokumentation der Datenbankstruktur in SQL. Es ist als eCH-Standard offen und wird unter einer freien Lizenz zur Verfügung gestellt.

Ein alternativer bewährter Ansatz zur Archivierung der Tabellen einer Datenbank ist der Export in das CSV-Format, welches besonders langlebig und weit verbreitet ist. Diese Methode ist technisch mit relativ geringem Aufwand verbunden, besitzt aber einen starken Fokus auf die Datenbankinhalte und geht daher mit einem vollständigen Verlust der Funktionalitäten und Datenbankstruktur einher. Infolgedessen ist es zwingend erforderlich, bei der Speicherung im CSV-Format zusätzlich die Datenbankstruktur in Form von strukturierten Diagrammen, wie dem Entitäten-Relationen-Diagramm (ERD) festzuhalten. Selbst dann ist die Datenbankstruktur nur teilweise und sehr aufwendig rekonstruierbar, weswegen das CSV-Format nur bei einer hybriden Lösung mit der Archivierung der Datenbankstruktur in einem geeigneten Format (z. B. SIARD, SQL, XML/DML) bei gleichzeitiger Speicherung der Daten im CSV-Format, empfohlen werden kann. Als Zeichenkodierung sollte UTF-8 verwendet werden, wobei RFC-4180-konform als Trennzeichen ein Komma (,) und als Textbegrenzungszeichen das Anführungszeichen (") verwendet werden sollte. Mehr dazu ist in dem Kapitel Tabellen ab Seite 97 zu finden.

Bei der Javascript Object Notation (JSON) handelt es sich um ein einfaches textbasiertes Datenformat, welches für einen schnellen Datentransfer im Internet entwickelt wurde und als RFC 7159 und ECMA-404 standardisiert ist. Es besitzt bei der Speicherung von Daten vergleichbare Vorteile wie XML, ist dabei aber wesentlich kompakter. Zur Zeit gibt es allerdings keinen Standard zur Beschreibung von Datenbankstrukturen in JSON, weswegen auch hier auf eine hybride Archivierungslösung mit der Speicherung der Datenbankinhalte in JSON und gleichzeitiger Speicherung der Datenbankstruktur in einem anderen

Format (z. B. SIARD, SQL, XML/DML) zurückgegriffen werden muss.

Die proprietären Dateiformate von Desktop-DBMS, wie Microsoft Access, dBase oder FileMaker eignen sich nicht für die Archivierung, da die Formatspezifikationen nicht offen vorliegen und verschiedene Versionen oft nicht untereinander kompatibel sind. Ebenso verhält es sich mit den binären Exportformaten serverbasierter, relationaler DBMS, wie Oracles DMP-Format oder das komprimierte BAK-Format von PostgreSQL. Auch diese folgen keinem allgemeingültigen Standard, liegen nur als Binärdatei vor und sind teilweise proprietär. Eine spätere Nachnutzung ist eng an das jeweilige für alle Nutzer kostenpflichtige Produkt sowie spezielle Tools gebunden.

ODB ist ein Container-Dateiformat des Programms, welche in LibreOffice und OpenOffice enthalten ist. ODB ist ein Teil vom OpenDocument Format (ODF) und wurde von einem technischen Komitee unter der Leitung der Organization for the Advancement of Structured Information Standards (OASIS) entwickelt und unter ISO/IEC 26300 standardisiert. Der Zip-Container enthält verschiedene Dateien, die zum Betrieb der Datenbank erforderlich sind, wie persönliche Konfigurationsdateien, eine binäre Datei des Datenbankmanagementsystems HSQLDB mit den Datenbankinhalten sowie der Datenbankstruktur und die XML-basierten Formulare und Reports. Aufgrund der binären Datenstruktur kann ODB nicht für die Langzeitarchivierung empfohlen werden.

Keines der genannten und empfohlenen Formate unterstützt und enthält alle notwendigen Informationen und signifikanten Eigenschaften, die für die Archivierung und Nachnutzung einer Datenbank relevant sind. Und einige Informationen, etwa ein Handbuch zur Benutzung, müssen ohnehin gesondert gespeichert werden. Daher ist eine hybride Archivierungsstrategie erforderlich, welche verschiedene Dateien mit unterschiedlichen Dateiformaten umfasst. In der folgenden Tabelle werden die Kategorien von Informationen, welche von den für die Archivierung geeigneten Datenbankformate prinzipiell gespeichert werden können, durch Abkürzungen kenntlich gemacht: I = Datenbankinhalt, S = Datenbankstruktur, F = Funktionalitäten, B = Benutzung. Eine zusätzliche Übersicht ist in dem Abschnitt Vertiefung zu finden.

Format	Begründung
✓ SIARD, SIARD2	SIARD ist offen, frei verfügbar und setzt auf die Standards SQL und XML. Der wesentliche Vorteil von SIARD liegt darin, dass sowohl die Datenbankstruktur als auch die Datenbankinhalte gemeinsam beschrieben und archiviert werden. Für die Archivierung sollte Version 2.0 verwendet werden. Speichert: I, S, F.
SQL	Geeignet für die Langzeitarchivierung bei Verwendung eines offiziellen ISO/IEC 9075 Standards, z.B. SQL:2008. Daten, Datenbankstruktur und ein Großteil der Funktionalität bleiben erhalten. Der verwendete Standard muss dokumentiert sein. Speichert: I, S, F.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Format	Begründung
XML	XML bietet sich insbesondere zur Speicherung der Daten einer Datenbank an. Die Datenbankstruktur sollte mit Hilfe eines XSD-Schemas, wie DBML, beschrieben werden. Alternativ ist dies auch mithilfe einer SQL-Datei oder SIARD möglich. Speichert I und teilweise S.
~ CSV	Das CSV Format ist ein langlebiges, textbasiertes und weit verbreitetes Format. Bei dem Export von Datenbankinhalten in das CSV-Format gehen allerdings Informationen zur Datenbankstruktur und zu Funktionalitäten verloren, weshalb es nur bei einer gleichzeitigen Speicherung der Datenbankstruktur in einem anderen Format (z. B. SIARD, SQL, XML/DBML) zu empfehlen ist. Speichert: I.
JSON	JSON bietet sich als Format zur strukturierten Speicherung von Datenbankinhalten an. Da es keine Standards zur Beschreibung der Datenbankstruktur im JSON Format gibt, muss diese gesondert in einem geeigneten Format (z. B. SIARD, SQL, XML/DBML) gespeichert werden. JSON bietet sich insbesondere für die Archivierung von Daten aus NOSQL Datenbanken an. Speichert: I.
MDB, ACCDB	Binäre, proprietäre Formate von Microsoft Access, die nicht zur Archivierung geeignet sind.
✗ FP5, FP7, FMP12	Binäre, proprietäre Formate von FileMaker, die nicht zur Archivierung geeignet sind.
ODB	Containerformat von OpenOffice, welches verschiedene XML-basierte Dateien, wie z.B. Formulare und Abfragen, und die binären Dateien mit den Inhalten und der Struktur der Datenbank enthält. ODB ist nicht zur Archivierung geeignet. Speichert: teilweise B.
DBF	Binäres, proprietäres Format von dBase, das nicht zur Archivierung geeignet ist.
BAK, DB, DMP	Die verschiedenen, binären Exportformate von relationalen Datenbanksystemen sind für die Archivierung nicht geeignet.

Dokumentation Die Empfehlungen zur Dokumentation von Datenbanken orientieren sich an verschiedenen Standards, darunter SIARD Metadata Schema, Archival Data Description Mark-up Language (ADDML) und Database Markup Language (DBML). Allen Schemata gemein ist eine differenzierte Beschreibung der verschiedenen Ebenen eines Datenbanksystems. Dazu gehören Metadaten zur allgemeinen Beschreibung der Datenbank als auch spezifische Metadaten zur Beschreibung der hierarchisch aufgebauten Ebenen einer Datenbank: den in SQL Datenbanken vorkommenden Schemen, die mehrere Tabellen beinhalten, welche wiederum mit Hilfe von mehreren Attributen definiert werden können.

Weitergehende Dokumente, die zum Verständnis einer Datenbank, ihrer spezifischen Nutzung oder zur Eigenart der Inhalte notwendig sind, wie beispielsweise Benutzerhandbuch, Bildschirmfotos, ERD-Diagramme, Beschreibungen semantischer Konventionen (wie Wertelisten), Anforderungsanalysen oder Rechte- und Rollendokumentation sollten auf alle Fälle mit archiviert werden.

Ergänzend zu den Metadaten, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 gelistet sind, sollten folgende minimale Metadaten zur nachhaltigen Beschreibung von Datenbanken aufgenommen werden.

Metadatum	Beschreibung
Datenbankname	Ansprache/Name der Datenbank
Beschreibung der Datenbank	Zusammenfassung über Zweck, Bedeutung und Inhalt der Datenbank
Sprache	Liste der Sprachen, in welchen die Daten eingegeben wurden sowie Sprachkennungen nach ISO 639 angeben.
Identifikator	Falls die Datenbank online öffentlich zugänglich ist, sollte ein persistenter Identifikator (z.B. DOI, URI) oder eine eindeutige Adresse (z.B. URL) angegeben werden.
Rechteinhaber	Personen, Institutionen oder Unternehmen, die Rechte an der Datenbankstruktur und/oder den Datenbankinhalten besitzen.
DBMS	Name des Datenbankmanagementsystems mit Versionsnummer mit der die Datenbank betrieben wurde.
Schemata Liste	Liste der Schemata innerhalb der Datenbank sofern vorhanden.
Benutzer	Liste der Benutzer und ihrer Rolle innerhalb der Datenbank.
Rolle	Liste der verschiedenen Rollen und Gruppen mit ihren definierten Zugriffsrechten.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Metadatum	Beschreibung
Standard	Name und Version verwendeter Standards, etwa zur Definition des Datentyps eines Attributs, zur Dokumentation von Abfragen oder zu dem Datenbankformat, z.B. SQL:2008 oder SIARD 2.0.
Abgeleitete Dateien	Aus den Datenbankinhalten erzeugte Grafiken, Abbildungen, Diagramme müssen zusätzlich separat archiviert und gelistet werden.
Weitere Dateien	Liste weiterer Dokumente, die für das Verständnis und die Dokumentation einer Datenbank notwendig sind, wie beispielsweise ein ERD oder ein Benutzerhandbuch.
Schema Metadatum	Beschreibung
Schema Name	Name des Schemas.
Schema Beschreibung	Bedeutung und Inhalt des Schemas.
Tabellenliste	Liste der Tabellen im Schema.
Funktion	Name und Art der Funktion (Gespeicherte Funktionen, Sichten).
Funktion Beschreibung	Beschreibung der Funktion hinsichtlich Sinn und Zweck.
Funktion Befehl	Befehle der Funktion nach einem einheitlich festgelegten Standard (Syntax Standard).
Tabelle Metadatum	Beschreibung
Tabelle Name	Name der Tabelle.
Tabelle Beschreibung	Bedeutung und Inhalt der Tabelle.
Anzahl Datensätze	Anzahl der Datensätze in der Tabelle.
Attributliste	Liste der Attribute in der Tabelle.
Attribut Metadatum	Beschreibung
Attribut Name	Name des Attributs.
Attribut Beschreibung	Bedeutung und Inhalt des Attributs. Wird das Attribut mit Hilfe von Einheiten, z. B. metrische Maßeinheiten, ausgedrückt, ist die Angabe der Einheit erforderlich.
Attribut Typ	Datentyp des Attributs nach einem einheitlich festgelegtem Standard (Syntax Standard) z. B. „Integer“ nach SQL:2008.
Kontrolliertes Vokabular	Sofern für das Attribut eine Werteliste vorliegt oder ein Thesaurus verwendet wurde, ist dieses zu dokumentieren.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Metadatum	Beschreibung
Attribut Schlüssel	Sofern es sich um ein Schlüsselattribut handelt, Benennung der Schlüsselart z. B. Primär- oder Fremdschlüssel.
Fremdschlüssel Referenz	Im Falle eines Fremdschlüssels Angabe des referenzierten Attributs mit Angabe der Tabelle und des Schemas.

Vertiefung

Ein Datenbanksystem (DBS) ist ein System zur elektronischen Verwaltung großer Datenmengen, um diese effizient, widerspruchsfrei und dauerhaft zu speichern und in den benötigten Teilmengen in unterschiedlichen Darstellungsformen für Benutzer und Anwendungsprogramme bereitzustellen. Gegenüber einer Speicherung in strukturierten Einzeldateien, wie Tabellen, werden unter anderem folgende Probleme effizienter gelöst:

- **Redundanzen:** Bei einer Speicherung in Einzeldateien werden viele Daten mehrfach gespeichert. Änderungen werden dadurch sehr aufwendig, da sie an verschiedenen Stellen durchgeführt werden müssen.
- **Inkonsistenzen:** Sowohl bei der gleichzeitigen Änderung von Daten durch mehrere Benutzer als auch bei Änderungen von redundanten Daten kann es leicht zu Widersprüchen und Inkonsistenzen in den Daten kommen, was mit der Verwendung von DBs vermieden wird.
- **Integritätsbedingungen:** Sofern einzelne oder globale Informationen direkt von anderen Informationen abhängen, lassen sich entsprechende Bedingungen in der Datenbank abbilden, so dass diese bei der Eingabe neuer Informationen auch eine Kontrollfunktion übernehmen und bestimmte Fehler in den Daten verhindern bzw. minimieren kann.
- **Datenschutzprobleme:** Einzeldateien besitzen kein dezentriertes User-Management, weswegen ein Zugriff auf den gesamten Datenbestand vorliegt, wohingegen DBS differenzierte Zugriffsregelungen ermöglichen.

Ein DBS besteht aus der Verwaltungssoftware, genannt Datenbankmanagementsystem (DBMS), und der Menge der zu verwaltenden Informationen, der eigentlichen Datenbank (DB). Die Datenbank enthält zusätzlich zu den eigentlichen Inhalten noch die Beschreibung der Datenstruktur, die in einem Datenkatalog und in einem Datenbankschema mittels Metadaten festgehalten wird.

Bei einem DBMS handelt es sich um ein Programm (kommerzielle Produkte sind u. a. FileMakerPro, Access und Oracle, OpenSource sind u. a. MySQL und PostgreSQL), das als Steuerungsprogramm intern die strukturierte Speicherung der Daten organisiert und alle lesenden und schreibenden Zugriffe auf die Datenbank ausführt. Der Benutzer greift also nicht direkt auf die Datenbank zu, sondern auf das DBMS als Kontrollinstanz. Die Aufgaben eines DBMS umfassen u. a. folgende Bereiche:

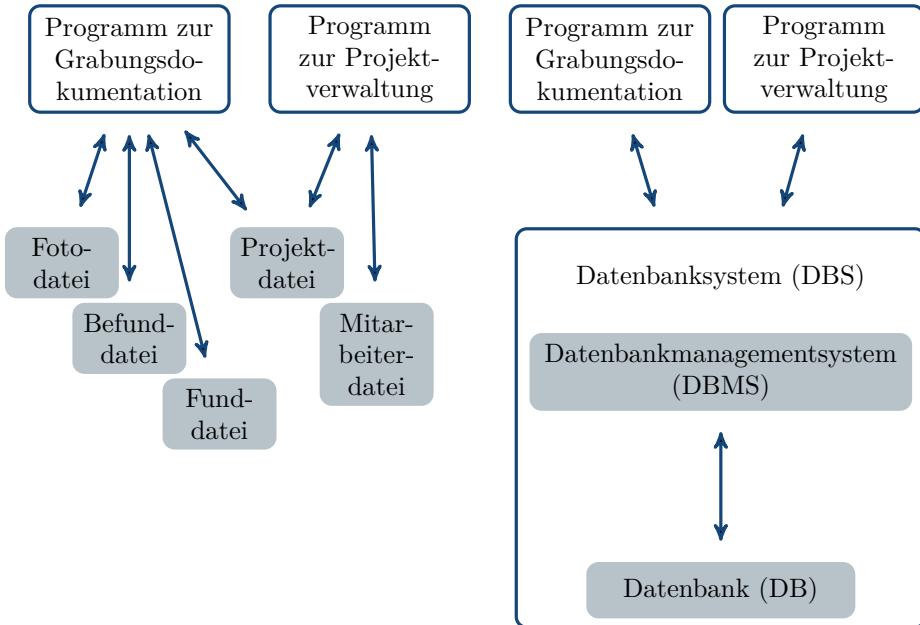


Abb. 3.18: Vergleich von Datenhaltung in Einzeldateien (links) gegenüber Datenhaltung mit Hilfe eines Datenbanksystems (rechts).

- Eingabe, Änderung und Löschen von Daten
- Erstellen von Datenbanken inklusive Umsetzung des Datenmodells
- Suchen in den Datenbankinhalten mittels Abfragen
- generelle Verwaltung von Benutzern, Zugriffen und Zugriffsrechten

Der Begriff Datenbank (DB) bezieht sich auf die eigentlichen Sachdaten, die entweder von Benutzern manuell eingegeben oder automatisch generiert werden. Die Datenbankinhalte können in unterschiedlichen Formaten vorliegen und als Texte, Zahlen, Links oder Medien (Fotos, Zeichnungen, Filme etc.) in die Datenbank einfließen. Dieser Datenbestand wird von einem DBMS verwaltet, auf Speichermedien abgelegt und je nach Vorgabe kontrolliert (damit z. B. ein Funddatum immer auch einem Fund zugeordnet ist).

Datenbankmodelle Zur formalen Beschreibung aller in der Datenbank enthaltenen Daten, ihrer Speicherung und ihrer Beziehungen untereinander verwendet man auf der konzeptionellen Ebene ein Datenmodell. Es ist die konzeptionelle Grundlage für ein Datenbanksystem und legt fest, auf welche Art und Weise der jeweils interessierende Ausschnitt aus der realen Welt in einer Datenbank abgebildet wird.

Es gibt verschiedene Datenmodelle, von denen das relationale das am weitesten verbreitete ist.

Relationales Datenmodell: Bei Relationalen Datenbanken (Abbildung 3.19) werden die Daten sowie deren Beziehungen untereinander in Form von

Tabellen (Relationen) beschrieben. Die einzelnen Einträge in unterschiedlichen Tabellen können mittels Verweise miteinander in Beziehung gesetzt werden.

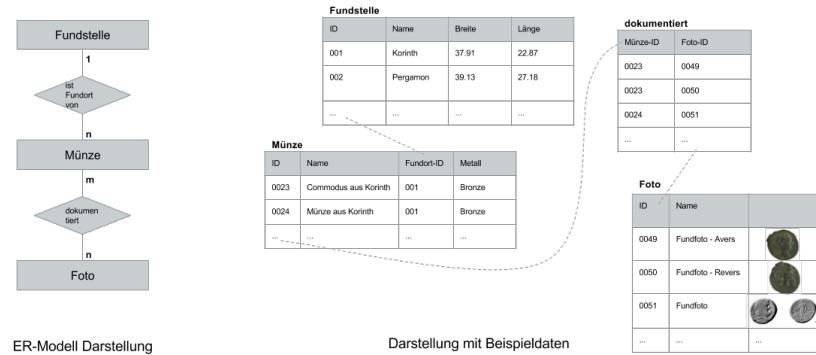


Abb. 3.19: Beispiel einer relationalen Datenbank anhand von Münzdaten

Hierarchisches Datenmodell: Bei hierarchischen Datenbanken (Abbildung 3.20) werden die Informationen in Datensätze gruppiert und in einer baumartigen Datenstruktur gespeichert, die einer Mutter-Kind-Beziehung entspricht: Jede Mutter kann mehrere Kinder haben, ein Kind aber nur eine Mutter besitzen.



Abb. 3.20: Beispiel einer hierarchischen Datenbank anhand von Münzdaten

Netzwerk Datenmodell: Netzwerkdatenbanken stellen eine Verallgemeinerung und Weiterentwicklung des hierarchischen Datenmodells dar, wobei die Daten in logischen Graphen strukturiert werden. Da die einzelnen Knoten im Graph beliebig miteinander über Beziehungen verknüpft werden können, entfallen die Beschränkungen, die mit Mutter-Kind-Beziehungen einhergehen.

Objektorientiertes Datenmodell: Analog zu den objektorientierten Programmiersprachen versuchen objektorientierte Datenbanken Objekte aus der realen Welt mit ihren Eigenschaften und ihrem Verhalten nachzubilden. Die Definition der Objekte erfolgt über Klassen, welche Attribut- und Verhaltensdefinitionen enthalten und als Bauplan für daraus erzeugte Objekte fungieren.

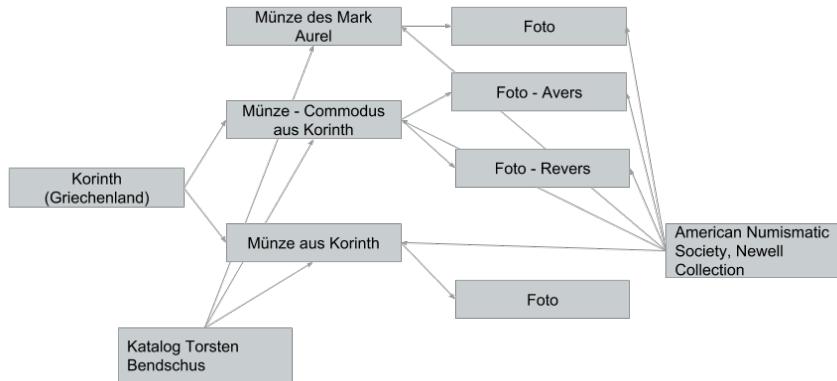


Abb. 3.21: Beispiel einer Netzwerkdatenbank anhand von Münzdaten

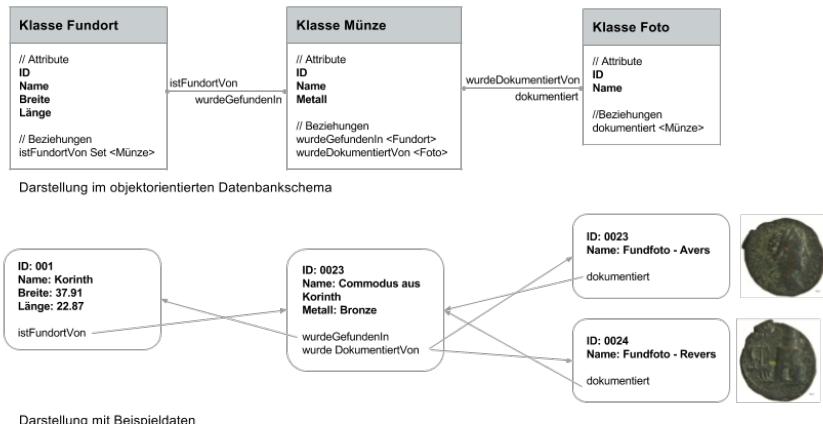


Abb. 3.22: Beispiel einer objektorientierten Datenbank anhand von Münzdaten

NoSQL Datenmodell: Der Begriff NoSQL steht für „Nicht nur SQL“ (Not only SQL) und bezeichnet Datenbanken, welche keinem relationalen Datenschema, sondern andere Paradigmen verfolgen. Sie können in zwei Typen unterschieden werden:

- Bei **dokumentenorientierten Datenbanken** wird die zusammengehörige Informationsmenge in einem einzelnen Dokument gespeichert. Die Daten werden in Feldname-Wert-Paaren (z. B. Datierung: Römisch) z.B. im JSON Format gespeichert, wobei nicht jedes Paar vorhanden sein muss, was eine gewisse Flexibilität erlaubt.
- **Graphen-Datenbanken** können als moderne Nachfolger von Netzwerkdatenbanken verstanden werden, die vor allem bei sehr stark vernetzten Daten verwendet werden. Auch hier werden mit Hilfe von Graphen die Daten als Knoten dargestellt und die Beziehungen zwischen Ihnen als Kanten,

wobei sowohl die Knoten als auch die Kanten Eigenschaften besitzen.

```
{  
    "id":1,  
    "Fundstellen_Name":"Korinth",  
    "Breite":37.91,  
    "Länge":22.87,  
    "Münzen":[  
        {  
            "id":23,  
            "Münzen_Name":"Commodus aus Korinth",  
            "Metall":"Bronze",  
            "Fotos": [  
                {  
                    "id":49,  
                    "Foto_Name":"Fundfoto -- Avers"  
                },  
                {  
                    "id":50,  
                    "Foto_Name":"Fundfoto -- Revers"  
                }  
            ]  
        },  
        {  
            "id":24,  
            "Münzen_Name":"Münze aus Korinth",  
            "Metall":"Bronze",  
            "Fotos": [  
                {  
                    "id":51,  
                    "Foto_Name":"Fundfoto"  
                }  
            ]  
        }  
    ]  
}
```

Relationale Datenbanken Das relationale Datenmodell hat sich als Standard von Datenbankanwendungen etabliert, da sich damit die Realwelt mit recht einfachen und realitätsnahen Strukturen abbilden lässt. In der archäologischen Forschung wird dieses Datenmodell fast ausschließlich eingesetzt.

Vereinfacht kann man sich eine relationale Datenbank als eine Sammlung von Tabellen (Relationen) vorstellen, in denen jeder Datensatz in einer Zeile (Tupel) einer Tabelle gespeichert wird (Abbildung 3.23). Jede Zeile besteht aus einer Anzahl von Attributen (Eigenschaften), den Spalten der Tabelle. Jedes Attribut hat einen eindeutigen zur Referenzierung dienenden Namen und einen bestimmten Attributtyp, welcher definiert welche Art von Daten – beispielsweise Zeichensätze (String) oder Fließkommazahlen (Float) – in dem Attribut gespeichert werden können. Jedes Attribut kann Werte entsprechend ihres Wertebereiches aufnehmen, der sich aus den Attributtyp ergibt. Zum Beispiel kann eine Münze mit Hilfe folgender Attribute (mitsamt Attributtypen) beschrieben

werden: ID (Integer), Nominal (String), Datierung (String), Material (String), Durchmesser (Integer), Gewicht (Float), Funddatum (Date).

Attribute (Spalten)						
ID	Nominal	Datierung	Material	Dm.	Gew.	Funddatum
1	Sesterz	römisch	Bronze	36	28,68	25.02.1993
2	Tetradrachme	römisch	Gold	25	13	13.08.2008
3	Tessera	römisch	Silber	18	2,5	12.02.1932
4	Drachme	griechisch	Silber	27	5,68	12.03.2012

Abb. 3.23: Aufbau der Relation „Münze“ am Beispiel von Informationen zu Münzen.

Pro Tabelle werden immer nur gleichartige, reale Objekte (Entitäten) erfasst, z. B. nur Befunde, nur Münzen, nur Fotos, nur Bauwerke etc. Die eindeutige Referenzierung eines Datensatzes erfolgt mit Hilfe eines oder mehrerer Schlüsselattribute, dem sogenannten Primärschlüssel. Dieser Schlüssel ist einzigartig und darf sich niemals ändern, da damit die Zeile in der Tabelle referenziert wird. Der Aufbau aller Tabellen, die zu einer Datenbank gehören, also die Datenbankstruktur, muss in dem Datenkatalog dokumentiert werden. Dieser enthält somit alle Festlegungen und Detailinformationen, die die Struktur der Datenbankinhalte in einem konkreten DBMS beschreiben, z.B. den Typ verschiedener Attribute (Kurztext, Langtext, Zahlen, Datum, Medien, etc.) und die Abhängigkeiten der Attribute zueinander. Der wichtigste Teilespekt des Datenkataloges ist das Datenbankschema, das in abstrakter Form die Beziehung und Art der Speicherung von Einzelinformationen festlegt.

Beziehungen Einzelne Tabellen können anhand von Schlüsseln miteinander in Beziehung gesetzt werden, um Beziehungen zwischen einzelnen Datensätzen aus den unterschiedlichen explizit auszudrücken. Beispielsweise um einer Münze, welche in der Tabelle „Münze“ gespeichert ist, einem Fundort, welcher in der Tabelle „Fundort“ gespeichert ist, zuzuweisen (Abbildung 3.24). Mit Hilfe von Primärschlüsseln (Hauptschlüssel, engl. Primary Key) wird der Datensatz einer Relation eindeutig definiert, wie im vorliegenden Beispiel mit Hilfe des Attributs „Muenze-ID“. Der Primärschlüssel einer Relation darf dabei nur einmal vorkommen, muss also einzigartig sein, um eine Referenzierung zu gewährleisten.

Ein Fremdschlüssel (engl. Foreign Key) dient zur Referenzierung eines Primärschlüssels einer anderen Tabelle. In dem Münzenbeispiel wurde die Relation „Muenze“ um eine Attribut „FS_FundortID“ erweitert. Es ist ein Fremdschlüssel, welcher auf den Primärschlüssel „PS_FundortID“ der Relation Fundort verweist.

Die Beziehungen zwischen zwei verschiedenen Relationen können dabei mit den folgenden gebräuchlichen Beziehungstypen beschrieben werden. Der Beziehungstyp wird dabei mit Hilfe einer Mengenangabe, der Kardinalität, angesprochen, der angibt wie viele Datensätze einer Relation mit wie vielen Datensätzen einer weiteren Relation in Beziehung stehen:

- In 1 : 1 - Beziehungen ist jedem Datensatz einer Tabelle A nur ein einziger passender Datensatz in Tabelle B zugeordnet. Diesen Beziehungstyp ver-

ID	Nominal	...	Funddatum	FS_FundortID	↔	PS_FundortID	Ortsname
1	Sesterz	...	25.02.1993	2		1	Paris
2	Tetradrachme	...	13.08.2008	1		2	Haifa
3	Tessera	...	12.02.1932	1		3	Alexandria
4	Drachme	...	12.03.2012	4		4	Athen

Abb. 3.24: Beispielhafte Beziehung zwischen der Relation „Muenze“ und der Relation „Fundort“.

wendet man unter anderem, um eine unübersichtliche Tabelle mit vielen Attributen in mehrere übersichtliche Tabellen aufzuteilen. Beispielsweise werden in einer Tabelle für Münzen allgemeine Eigenschaften abgebildet und die Einzelheiten für die Vorder- und Rückseite einer Münze werden in gesonderten verknüpften Tabellen beschrieben.

- In 1 : N - Beziehungen, können einem Datensatz der Tabelle A mehrere passende Datensätze aus Tabelle B zugeordnet sein, wobei jedoch einem Datensatz aus Tabelle B nie mehr als ein Datensatz aus Tabelle A zugeordnet werden kann. Beispielsweise können an einem Fundort mehrere Münzen gefunden worden sein, aber eine Münze kann nur einen Fundort haben, wie in der Abbildung oben dargestellt ist.
- Bei M : N - Beziehungen können jedem Datensatz aus einer Tabelle A mehrere passende Datensätze aus einer Tabelle B zugeordnet sein und umgekehrt. Auf einem Foto können beispielsweise mehrere Münzen zu sehen sein, während es für eine Münze auch mehrere Fotoansichten geben kann.

Attributtypen Es gibt verschiedene Attributtypen, die spezifische Informationen und Wertebereiche speichern:

- Ganze Zahlen (integer, long) sind zu verwenden für Schlüssel, ordinalskalierte Variablen, wie etwa Schulnoten, oder sonstige numerische Angaben, bei denen die Genauigkeit ganzer Zahlen ausreichend ist.
- Fließkommazahlen (float) kommen zum Beispiel bei metrischen Maßen zum Einsatz. Bei allen Maßeinheiten ist zu beachten, dass in der Attributdefinition die verwendete Maßeinheit notiert und diese während der Benutzung auch einheitlich verwendet wird.
- Bool'sche Felder lassen nur zwei Werte zu: „ja“ (1) oder „nein“ (0) bzw. „wahr“ oder „falsch“.
- Textfelder (varchar) können mit einer maximalen Länge definiert werden oder aber ohne weitere Angaben standardmäßig bis maximal 255 Zeichen verwendet werden.
- Größere Textfelder (text) erlauben einen beliebig langen Freitexteintrag. Sie sollten so wenig wie möglich verwendet werden, da sie verhindern Informationen strukturiert einzutragen und somit ein gezieltes Auffinden erschweren. Sie können bei einer späteren Konvertierung einer DB in ein neues System Schwierigkeiten verursachen.

- Medienfelder/Binary Large Objects (BLOBs) ist ein binärer Datentyp, der zur datenbankinternen Speicherung von externen Medieninhalten (Fotos, Zeichnungen, Filme, Texte etc.) verwendet werden kann. Jedoch empfiehlt es sich diese Medieninhalte außerhalb der eigentlichen Datenbank zu speichern und nur die Referenz auf den Speicherort der externen Ressource in einem Attribut zu erfassen. Dadurch werden u. a. die Datenbankgröße überschaubar gehalten, eine konsistente Datenhaltung der Einzeldaten gewährleistet und Probleme beim Export der Datenbankinhalte reduziert. Ist im Ausnahmefall dennoch das Importieren von externen Medieninhalten notwendig, sollten diese für den jeweiligen Zweck optimiert sein (z. B. sind für die Bildvorschau geringe Bildauflösungen ausreichend).

Structured Query Language (SQL) Mit Hilfe von speziellen Datenbanksprachen kann ein Benutzer oder ein externes Programm über das DMBS mit einer Datenbank kommunizieren. Die am weitesten verbreitete Datenbanksprache im Bereich der relationalen Datenbanken ist SQL (structured query language). Sie wird rechnerunabhängig und produktübergreifend seit 1986 im ISO Standard 9075 definiert. Seitdem erfuhr SQL sieben Erweiterungen, von denen die letzte Aktualisierung 2011 als ISO/IEC 9075:2011 erschienen ist.

SQL-Befehle lassen sich in vier verschiedene Befehlsguppen unterteilen:

Befehlsguppe	Beschreibung	Beispiel
Data Definition Logic (DDL)	Befehle zur Definition des Datenbankschemas: Erstellen von Datenbanken, Tabellen und Indizes.	<ul style="list-style-type: none"> – Anlegen der Tabelle „muenzen“ mit beispielhaften Attributen unter Angabe des Datentyps CREATE TABLE muenzen (PS_MuenzID serial NOT NULL, Nominal varchar, Gewicht float);
Data Query Language (DQL)	Befehle zum Abfragen von Daten.	<ul style="list-style-type: none"> – Abfrage nach allen Münzen mit Fundort Paris. SELECT * FROM muenzen WHERE FS_FundortID = '1'
Data Manipulation Logic (DML)	Befehle zur Datenmanipulation, also dem Ändern, Bearbeiten und Löschen von Datensätzen.	<ul style="list-style-type: none"> – Ändern des Materials in Gold für den dritten Datensatz UPDATE muenzen SET Material = „Gold“ WHERE PS_MuenzID = „3“
Data Control Language (DCL)	Befehle zum Anlegen von Benutzern und zur Definition von Zugriffsrechten	<ul style="list-style-type: none"> – Anlegen eines Nutzers mit Passwort CREATE USER mmustermann PASSWORD 'passwort'

Tabelle 3.4

Entity-Relationship-Modell (ERM) Ein wichtiges grafisches Hilfsmittel beim Entwurf und zur Dokumentation einer Datenbank ist das Entity-Relationship-Modell (ERM). Neben der grafischen Visualisierung mittels eines Entity-Relationship-Diagramm (ERD) ist auch die tabellarische Erfassung und Beschreibung aller Entitätstypen, Beziehungstypen und Attribute für eine vollständige Dokumentation einer Datenbank notwendig. Das ERD ist ein beschriftetes Diagramm mit festgelegten Symbolen, das dazu dient, den gewünschten Ausschnitt der realen Welt formal zu beschreiben. Grundlage der ER-Modelle

ist die Abstraktion von Objekten und deren Beziehungen untereinander. Folgende grundlegende Begriffe in einem ERM und deren grafische Wiedergabe in einem ERD werden in diesem Zusammenhang unterschieden:

Element	Beschreibung	Grafische Darstellung
Entität (Datensatz)	Objekt der Wirklichkeit, materiell oder abstrakt, z. B. ein Fund „Münze XY“, ein Befund „Pfostenloch 123“ oder eine Abteilung „DAI Rom“.	keine grafische Darstellung
Entitätstyp (Tabelle)	Abstraktion gleichartiger Entitäten, z. B. Personen, Funde, Befunde, Adressen.	
Beziehung	semantische Beziehung zwischen verschiedenen Datensätzen aus unterschiedlichen Tabellen, z. B. „Münze“ in „Pfostenloch“	
Beziehungs- typ	Abstraktion gleichartiger Beziehungen, z. B. „gefunden in“.	
Kardinalität	mögliche Anzahl der an einer Beziehung beteiligten Entitäten (s. u. Beziehungsarten); Symbol: 1, N oder M über Strich	
Attribut	Eigenschaften eines Entitätstyps oder eines Beziehungstyps, z. B. Geburtsdatum von Personen.	
Schlüssel	Attribut/e, welche/s zur eindeutigen Identifizierung einer Entität von einem Entitätstypen dient/en.	

Tabelle 3.5

Bestandteile einer Datenbank Um eine Datenbank auch zukünftig nutzen und die in ihr erfassten Inhalte verstehen zu können, ist eine umfassende und vollständige Dokumentation erforderlich, zu der verschiedene Komponenten gehören:

- Die schriftliche und/oder tabellarische Anforderungsanalyse
- Die eigentlichen Datenbankinhalte
- Die Beschreibung der Datenstruktur
- Die Definition der implementierten Funktionalitäten
- Semantisch-inhaltliche Aspekte wie vorgegebene Wertelisten, Thesauri oder individuelle Definitionen von Begriffen und Datensätzen

- Dokumentation der Benutzeransichten (z.B. Eingabelayouts, Suchmasken, Ausgabeformate), technischer Schnittstellen, Regelwerke

Da mit keinem der empfohlenen Dateiformate alle Komponenten erfasst und archiviert werden können, ist eine hybride Archivierungsstrategie erforderlich. Beispielsweise können der Inhalt, die Struktur und die Funktionalitäten einer Datenbank im SIARD-Format archiviert werden. Die Anforderungen werden schriftlich beschrieben und um ein graphisches Entity-Relationship-Diagramm ergänzt. Individuelle Vorgaben zur Benutzung (insbesondere der Dateneingabe) werden mittels eines Benutzerhandbuchs im Format PDF/A dokumentiert und die einzelnen Layoutmasken werden mittels Screenshots als unkomprimierte TIFF-Dateien gespeichert.

Die folgenden Tabelle gibt einen Überblick über die verschiedenen Formate für Datenbanken und welche davon für die langfristige Speicherung der genannten Dokumentationskomponenten von Datenbanken geeignet sind.

Format, Dokument	I	S	F	Benutzung
SIARD	✓	✓	✓	
SQL	✓	✓	✓	
XML	✓	~		
CSV	✓			
JSON	✓			
ERD (z. B. TIFF)			✓	
Benutzerhandbuch (z. B. PDF/A)				✓
beschriebene Bildschirmfotos (z. B. TIFF)				~
MDB, ACCDB				
FP5, FP7, FMP12				
ODB				~
DMP, BAK, DB				

Tabelle 3.6: Formate für Datenbanken und welche Elemente sie langfristig Speichern können. I = Datenbankinhalt, S = Datenbankstruktur, F = Funktionalitäten. Leere Zellen bedeuten fehlende Speichereigenschaften.

Quellen

- H. Eiteljorg, Archaeological Computing ²(Bryn Mawr 2008)
(insbes. Kap. III ; download: <http://archcomp.csanet.org/>)
- P. Kandzia - H. J. Klein, Theoretische Grundlagen relationaler Datenbanksysteme. Reihe Informatik 79 (1993).
- R. Elmasri - S. B. Navathe. Fundamentals of Database Systems ⁴(2004).
- A. Kemper - A. Eickler, Datenbanksysteme. Eine Einführung ⁷(2009).
- R. Steiner, Grundkurs relationale Datenbanken. Einführung in die Praxis für Ausbildung, Studium und IT-Beruf ⁷(Wiesbaden 2009).
- A. G. Taylor, Datenbanken für Dummies (Bonn 2001)
- P. Kleinschmidt - Chr. Rank, Relationale Datenbanksysteme. Eine praktische Einführung ³(Berlin 2005).
- G. Lock, Using Computers in Archaeology (London 2003) 85-98.

D. Smith, Database fundamentals for archaeologists, in: S. Ross et alii (hrsg.),
Computing for Archaeologists (Oxford 1991) 111-125.

3.7 GIS

K. Rassman

Für die Langzeitarchivierung von GIS-Daten ist derzeit noch kein allgemeingültiger Standard etabliert, weswegen auf die Vorgaben zu den in das GIS importierten Daten (z.B. Luftbilder, Geländemodell etc.) verwiesen wird. Es sollen jedoch möglichst keine programminternen Formate, die nicht von anderen GIS Systemen importiert werden können, genutzt werden. Bevorzugt werden sollen:

- Vektordaten als Esri Shapefile (shp + shx + dbf)
- Rasterdaten als GeoTiff

Die Haltung von Geodaten sollte vorzugsweise in Geodatenbanken erfolgen. Falls dies nicht möglich ist, müssen die Metadaten ausgefüllt werden. Folgende Informationen sind in der Geodatenbank / den Metadaten bereitzustellen:

- Datenquelle (z.B. SRTM, Tachymeter)
- Verarbeitung: Programm, Algorithmus, Parameter (z.B. QGIS , r.shaded.relief, altitude, azimuth etc.)
- Projektion (WGS 1984, EPSG: 4326)
- Datenqualität (Auflösung 90m, Ungenauigkeit)
- Information, was mit den Layern ausgedrückt wird, da es nicht immer aus dem Dateinamen hervorgeht (z.B. Höhenmodell)

Vertiefung

GIS in der Archäologie Geographische Informationssysteme (GIS) dienen der Aufnahme, Verwaltung, Darstellung und Auswertung raumbezogener (Sach- und Geometrie-) Daten. In der Regel beschreibt der Terminus GIS die Gesamtheit eines GIS-Projektes, d.h. ein Gesamtsystem, das alle für die digitale, raumbezogene Arbeit erforderlichen Werkzeuge (Hardware, Software) und Daten (archäologische Basisdaten, geographisch - naturräumliche Daten) umfasst. Seltener beschränkt sich der Terminus GIS lediglich auf die Software mit der Aufnahme, Verwaltung, Darstellung und Auswertung raumbezogener Daten.

Den Datenbestand eines GIS könnte im Kern bereits eine digitale Liste mit Keramikfunden einer Ausgrabung bilden, unverzichtbar sind dabei Informationen zur Lage der Funde. Daraus lassen sich mit Hilfe von GIS-Software Karten erstellen, die verschiedene Informationen zu den Keramikfunden wiedergeben (Verbreitung der Verzierungsmotive, Typen, Menge der Scherben usw.). Viele Nutzer beschränken sich beim Gebrauch eines GIS auf die computergestützte Kartographie. Der entscheidende Vorteil eines GIS besteht aber in den darüber hinausgehenden vielfältigen Möglichkeiten der Datenanalyse. Derartige Analysen setzen in der Regel Berechnungen zum Raumbezug voraus, wie die Abstände zwischen den Funden, die Größe der ausgegrabenen Flächen oder der entsprechenden Ausgrabungsbefunde. Solche Daten lassen sich mit geringem Aufwand mit Hilfe eines GIS automatisch erheben und in Bezug zueinander setzen. Damit eröffnen sich neue Wege, die Verteilungen des Fundmaterials mit statistischen Methoden zu vergleichen, häufig werden hier Funddichteberechnungen durchgeführt, wie die Verteilung von Scherbenanzahl oder -gewicht.

Maßstabsebenen Die Aufnahme, Verwaltung, Darstellung und Analyse raumbezogener Daten mit einem GIS ist auf verschiedenen Maßstabsebenen archäologischer Arbeit sinnvoll. In diesem kurzen Ratgeber wird vor allem der Einsatz innerhalb einer Ausgrabung thematisiert. GIS-Anwendungen sind jedoch auch für die Auswertung größerer Landschaftsausschnitte attraktiv. Der Fundplatz, den man ausgräbt, ist eingebettet in eine Umgebung, die archäologisch teils durch frühere Arbeiten bekannt ist, teils durch eigene Oberflächenbefahrungen weiter untersucht wird. Dies ist das Niveau von „Siedlungskammer“ oder „Schlüsselgebiet“. Die naturräumlich - geographischen Daten stehen auf diesem Niveau meist im Maßstab 1:25.000 oder 1:50.000 zur Verfügung. Die Größe solcher Mikro-Regionen beträgt typischerweise zwischen wenigen $10km^2$ und wenigen $100km^2$. Auf dieser Maßstabsebene ist es möglich, alle Nachbarplätze inklusive Fundstoff selbst zu überschauen und man kann Vollständigkeit anstreben.

Auf der nächst höheren Maßstabsebene, der der Regionalstudien, bewegt man sich in einer Größenordnung von wenigen $1.000km^2$. Wenn man die archäologischen Daten einer solchen Region selbst erfassen will, muss mit einem Arbeitsaufwand in der Größenordnung eines Dissertationsprojektes gerechnet werden. Bei Maßstäben ab 1:500.000 und einigen $10.000km^2$ oder gar größer muss man sich auf die Richtigkeit von Angaben aus der Literatur verlassen und es ist unmöglich, die Vollständigkeit des archäologischen Datensatzes zu gewährleisten. Trotzdem ist auch die Betrachtung dieser Maßstabsebenen notwendig, denn Fragen zu kulturellen Diffusionsprozessen, zu Migrationen, zu Tausch und Handel, zur Mensch-Umwelt-Beziehung und auch zur Lage des selbst ausgegrabenen Fundplatzes in Bezug zu kulturellen Zentren sowie peripheren Regionen usw. sind nicht in Maßstabsebenen zu beantworten, die nur die Darstellung von wenigen $100km^2$ erlauben. GIS-Anwendungen erlauben theoretisch einen gleitenden Übergang von einer Skalenebene zur anderen. Die archäologischen Konsequenzen der eben beschriebenen, scheinbar trivialen Zusammenhänge steht in der Archäologie erst ganz am Anfang. Lösungen sind nur vom Einsatz derjenigen Rechenverfahren zu erwarten, die GIS-Programme mit anbieten und für die sich Archäologen bisher nur am Rande interessiert haben. Da man sich nicht darauf verlassen kann, dass auf einer Karte, die ein Gebiet von mehreren $10.000km^2$ Größe darstellt, alle archäologischen Fundpunkte vorhanden sind, muss man an dieser Stelle eben von der Punkt- zur Schraffurdarstellung übergehen. Allerdings sollte der schraffierte Bereich, in dem die meisten Fundpunkte liegen, nach nachvollziehbaren Kriterien bestimmt sein. Es liegen bereits Vorschläge vor, welche Methoden man zu diesem Zweck verwenden sollte.

Aber natürlich gewinnt auch der Einsatz von GIS auf Ausgrabungen eine zunehmende Bedeutung. Umso mehr, da mit höherer Komplexität der erhobenen Daten (Funde, Architekturbefunde, naturwissenschaftliche Daten) die Schwierigkeit wächst, diese Informationseinheiten in ihrer Gesamtheit zu erfassen und die vielfältigen Beziehungen zwischen ihnen auswerten zu können.

Datenmodell eines GIS Man unterscheidet allgemein zwischen Sach- und Geometriedaten. Um beim Beispiel der Liste der Keramikfunde zu bleiben, handelt es sich bei diesen Informationen um Sachdaten. Aus Geometriedaten besteht der ins GIS eingebundene digitalisierte Grabungsplan. Er könnte beliebig viele geometrische Objekte enthalten, wie die Flächen von Häusern oder

Pfostenlöcher.

Den Kern des GIS bilden die Sachdaten, die in unterschiedlich komplexen Datenbanken organisiert werden. Diese Sachdaten sind in Abhängigkeit von der archäologischen Fragestellung zu strukturieren. Unverzichtbar sind Informationen zur Lage der Funde. Das sind zumeist x,y-Koordinaten und Tiefenangaben, häufig aber auch Angaben des entsprechenden Quadranten mit der horizontalen Auflösung von 1x1m bei Nennung des jeweiligen Planums.

Geometriedaten stehen in zwei verschiedenen Klassen zur Verfügung. Man unterscheidet zwischen Vektor- und Rasterdaten.

Vektordaten ermöglichen eine exakte Abbildung räumlicher Objekte, z. B. der Grenze einer Hausgrube oder einen Flusslauf. Jedes Objekt besteht dabei aus einer variablen Menge von Koordinatenpaaren, die gewissermaßen Knotenpunkte in diesem System darstellen. Die Knotenpunkte können zu beliebigen Strukturen kombiniert werden. Auf diese Weise lassen sich Punkte, Linienzüge oder Flächen abbilden. Die Kombination der Knotenpunkte zu Flächen oder zu Linien ergibt sich aus Zusatzinformationen. Ein Pfostenloch kann in einem Grabungsplan als Objekt dargestellt werden, das aus fünf oder mehr Knotenpunkten besteht. Die Zusatzinformationen beschreiben die Zusammengehörigkeit der Knotenpunkte und die Eigenschaft des Objektes als Fläche. Die Geometriedaten werden im GIS dann mit Sachdaten verknüpft. Für das Objekt Pfostenloch könnten dies Daten zur Höhe bzw. Tiefe der Unterkante des Pfostenloches sein oder Informationen zu dem hier vorgefundenen Keramikmaterial.

Rasterdaten bestehen aus nur einem geometrischen Grundelement - der Rasterzelle. Die in der Regel quadratischen Rasterzellen werden in Zeilen und Spalten angeordnet. Die Zellgröße bestimmt die Auflösung der so entstehenden Karte. Jeder Zelle wird dann ein Attributwert zugewiesen. Das häufigste Anwendungsbeispiel für eine Rasterdatei ist ein Digitales Geländemodell (DGM), in der jede Zelle die Information des Höhenwertes enthält. Am bekanntesten dürfte das weltweit kostenlos zur Verfügung stehende SRTM-Höhenmodell (Shuttle Radar Topography Mission) aus dem Jahr 2000 mit einer Rastergröße von ca. 90x90m sein. In einer Rasterdatei lässt sich auch die Verteilung der Scherbenanzahl oder des Scherbengewichts in einer Grabungsfläche mit Rasterzellen von 1x1m Quadraten veranschaulichen.

In einem GIS-Projekt lassen sich je nach Fragestellung Raster- und Vektordaten integrieren. Für die Umwandlung von Rasterdaten in Vektordaten oder umgekehrt stehen dabei in den meisten GIS Standardwerkzeuge zur Verfügung. Dabei gilt es zu bedenken, dass Rasterdateien in der Regel größer sind als Vektordateien vergleichbarer Auflösung, sie bieten aber beim Verschneiden mit Hilfe von Basisrechenverfahren einige Vorteile.

Praxis

Die Software eines GIS setzt sich aus zwei Hauptbestandteilen zusammen. Erster Bestandteil ist ein Datenbankprogramm für die Verwaltung der Sachdaten. Häufig wird für diese Arbeit auf spezielle Datenbanksoftware zurückgegriffen (MS Access, FileMaker, MySQL). Den zweiten Teil bilden spezielle GIS-Programme wie ArcView, Manifold, MapInfo, QuantumGIS, SAGA, gvSIG oder GRASS. Derartige Programme ermöglichen es, die Daten in ihrem räumlichen Bezug zu analysieren und deren Strukturen zu visualisieren. Die graphische Funktionalität der GIS-Programme, eröffnet für die Erstellung thematischer Karten vielfältige

Möglichkeiten. Der große Vorteil von GIS-Programmen sind deren Werkzeuge zur Analyse raumbezogener Daten. Durch die Möglichkeiten der Visualisierung, der Herstellung von Karten, lassen sich viele Fragen auf eine empirische Weise verfolgen. Es ist es jedoch unverzichtbar, dass der Visualisierung eine Analyse folgt, mit der geprüft wird, inwieweit die herausgearbeiteten Unterschiede oder Gemeinsamkeiten bedeutsam sind. Die analytischen Werkzeuge ermöglichen einfache Abfragen, die Berechnung von Vektor- und Rasterflächen bis zu komplexen statistischen Operationen. Spezielle GIS-Programme enthalten daneben auch Bausteine, mit denen Datenbanken erstellt und verwaltet werden können. Meist wird man aber auf spezielle Datenbanksoftware zurückgreifen.

In der Praxis werden häufig Grabungszeichnungen in AutoCAD erzeugt und dann als DXF-Datei in das GIS importiert und mit den vorhandenen Sachdaten zusammengefügt. Ein weiteres klassisches Anwendungsbeispiel für ein vektorbasiertes GIS ist ein digitales Landschaftsmodell, in dem Straßen, Waldflächen, überbaute Flächen u.a. enthalten sind. Der Vorteil liegt darin, dass jedem dieser Objekte Zusatzinformationen zugewiesen werden können (Breite einer Straße, Belag, Steigung usw.).

Besondere Aufmerksamkeit ist bei der Planung von GIS-Vorhaben dem Einsatz der entsprechenden Software zu widmen. In den zurückliegenden Jahren sind leistungsfähige OpenSource-Programme entwickelt worden, die nahezu alle Anwendungsgebiete abdecken. Dabei ist zu bedenken, dass die hohen Kosten kommerzieller GIS-Software ihrem Einsatz bislang Grenzen gesetzt haben. An Universitäten stehen die Programme häufig nur in den PC-Pools der Rechenzentren oder aber auf einigen Rechner zur Verfügung. Für weit vernetzte, internationale Forschungsvorhaben ist dieser Umstand von erheblichen Nachteil.

Eine besondere Empfehlung unter den neu aufkommenden leistungsfähigen OpenSource-Programme verdient die spanische GIS-Software gvSIG, die plattformübergreifend auf Windows-, Linux- und Mac-Betriebssystemen eingesetzt werden kann.

Die eingesparten Kosten lassen sich zukünftig effektiver durch die gezielte Anpassung der Software an die spezifischen Projektanforderungen einsetzen oder aber für Schulungen.

Ausgrabungen und GIS Die ersten Anwendungen zur Digitalisierung von Grabungsdaten basierten auf der Konstruktionssoftware AutoCAD. Diese Versuche reichen bis in die frühen 90er Jahre zurück.

Die Vorzüge dieser Werkzeuge kamen bei der Herstellung und Vorlage digitaler Planzeichnungen zum Tragen. Nachteilig ist die umständliche Einbindung von Datenbanken und das weitgehende Fehlen von Werkzeugen für die Analyse von Raumstrukturen.

Die Vorzüge eines GIS bei der Durchführung und Auswertung einer Ausgrabung liegen auf der Hand. Nahezu alle auf einer Ausgrabung gewonnenen Informationen haben einen Raumbezug. Viele der Sachdaten sind in Datenbanken zu verwalten, wie Funddaten, Vermessungsdaten, Schichtbeschreibungen usw. Die Form, Ausdehnung und Lage der Befunde und Architekturreste lassen sich durch geometrische Daten eines GIS abbilden. Die reine graphische Dokumentation ist auch mit einer Software wie AutoCAD möglich, nur bieten derartige Programme standardmäßig keine Werkzeuge, mit denen sich bspw. Flächengrößen und Mittelpunkte von Flächenobjekten berechnen lassen. Das GIS er-

öffnet darüber hinaus die Möglichkeit Verteilungen zu vergleichen, Flächen zu verschneiden etc. Bliebe man bei der digitalen Umsetzung der konventionellen Grabungspläne, wäre ein Programm wie AutoCAD ausreichend. Folgerichtig wird AutoCAD bei Ausgrabungen da eingesetzt, wo das besondere Augenmerk auf der Dokumentation liegt. Bewährt haben sich speziell für AutoCAD entwickelte Werkzeuge, wie das Programm TachyCAD von der Firma Kubit oder das Programm ArchäoCAD der Firma ArcTron. Mit diesen Lösungen lassen sich mittels einer Totalstation auf Ausgrabungen Befunde einmessen. Computer und Totalstation sind auf der Grabung miteinander verbunden. Die Daten werden direkt ins AutoCAD eingelesen. Später lassen sich die Daten problemlos in ein GIS exportieren. Das gängige Austauschformat ist DXF/DWG.

Ein weiteres AutoCAD-Werkzeug der Firma Kubit ist Photoplan, mit dem sich Grabungsbefunde photogrammetrisch dokumentieren lassen. Diese Arbeitsschritte ließen sich auch mit einer GIS-Lösung ohne AutoCAD realisieren. Nach unseren Erfahrungen sind diese Programmbestandteile jedoch nicht so anwendergerecht. In der Regel wird man bei der Wahl der Software nicht allein das nach den theoretischen Leistungsparametern beste Programm wählen, sondern diejenigen Programme bevorzugen, für die ein guter Ausbildungsstand vorhanden ist und kompetente Ansprechpartner für die Diskussion von Problemen zur Verfügung stehen. Zudem hat sich herausgestellt, dass keines der angebotenen GIS-Programme alle benötigten Rechenverfahren gleich oder in gleicher Anwendergerechtigkeit zur Verfügung stellt. Häufig ist es daher - je nach Fragestellung - sinnvoll oder unumgänglich, dass verschiedene GIS-Programme zum Einsatz kommen.

Den Vorzug sollten zukünftig mehr und mehr OpenSource-Lösungen erhalten.

Quellen

- Bill, R. , Grundlagen der Geo-Informationssysteme (Heidelberg 1999).
Blankholm, H. P., Intrasite Spatial Analysis in Theory and Practice (Aarhus 1991).
Burrough, P. A. / R. A. McDonnel, Principles of Geographical Information Systems (Oxford 1999).
J. Conolly/M. Lake, Geographical Information Systems in Archaeology. Cambridge Manuals in Archaeology (Cambridge 2006).
Gaffney, V. / Z. Stancic GIS approaches to regional analysis: A case study of the island of Hvar (Ljubljana 1991).
Hietala, H. J. , Intrasite spatial analysis: a brief overview (Cambridge 1984).
Hodder, I. / C. Orton (1976). Spatial Analysis in Archaeology. Cambridge, Cambridge University Press.
Kunow, J. / J. Müller Archäoprognose Brandenburg 1. Symposium Landschaftsarchäologie und geographische Informationssysteme.
Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnung (Wünsdorf 2003).
G. Lock/B. Leigh Molyneaux (Hrsg.), Confronting scale in archaeology (Berlin 2006).
Neteler, M., GRASS-Handbuch, Der praktische Leitfaden zum Geographischen Informationssystem GRASS. Hannover, Abteilung Physische Geographie und

- Landschaftsökologie am Geographischen Institut der Universität Hannover (Hannover 2000).
- C. Orton, Sampling in Archaeology (Cambridge 2000).
- Schier, W. (1992). Zum Einsatz multivariater Verfahren bei der Analyse des Lage- und Umweltbezugs prähistorischer Siedlungen. *Archaeologische Information* 15/1&2: 117-122.
- Wheatley, D. and M. Gillings. Spatial Technology and Archaeology. The archaeological applications of GIS. (London / New York 2002).
- Zimmermann, A., J. Richter, et al., Landschaftsarchäologie II - Überlegungen zu Prinzipien einer Landschaftsarchäologie. Bericht der Römisch-Germanischen Kommission 2004, 85: 37-95.
- Preserving Geospatial Data:
http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-greg-janee
;
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:DE:PDF
;
http://www.gdi-de.org/
- Weiteres:
http://www-sul.stanford.edu/depts/gis/Archaeology.htm
- Weiteres:
http://oadigital.net/software/gvsigoade
- Weiteres:
http://www.fuerstensitze.de/1140_Publikationen.html

3.8 Video

M. Trognitz; Mit Beiträgen von: H. Lewetz

Videos sind in der altertumswissenschaftlichen Forschung vergleichsweise wenig verbreitet und dienen vor allem der Dokumentation von Arbeitsabläufen, Visualisierung von Ergebnissen oder publikumswirksamen Medienauftritten.

Eine digitale Videodatei enthält sowohl visuelle als auch auditive Inhalte. Diese sind jeweils in einem eigenen Format, dem sogenannten Codec, gespeichert und werden in einem Containerformat zusammengeführt. Je nachdem welches Containerformat verwendet wird, können weitere Informationen, wie Metadaten oder Untertitel gespeichert werden. Zu beachten ist, dass es sich bei manchen Containerformaten gleichzeitig um Codecs handeln kann. Die Digitalisierung von analogem Film- und Videomaterial ist nicht Gegenstand dieser Empfehlungen.

Langzeitformate Bei der Auswahl für die Langzeitarchivierung von digitalen Videos muss nicht nur ein passendes Format, sondern auch ein geeigneter Codec gefunden werden. Dabei gelten für die Auswahl des Codecs und des Containerformats die Kriterien, die für die Wahl von Dateiformaten ab Seite 52 beschrieben werden: Es sollte sich um einen weit verbreiteten, möglichst nicht proprietären und offen dokumentierten Standard handeln, der verlustfreie oder gar keine Kompression anwendet. Außerdem sollte darauf geachtet werden, dass das gewählte Format die Speicherung aller relevanten Elemente des Videos, wie beispielsweise Untertitel, unterstützt.

Das Containerformat Matroska (MKV) erfüllt alle Anforderungen an ein Archivformat. Es wird seit 2003 explizit als offenes Containerformat entwickelt, das modernen Ansprüchen genügt und viele verschiedene Codecs enthalten sowie zusätzliche Elemente, wie Untertitel, speichern kann. Das Format basiert auf einer binären Variante von XML, nämlich EBML (Extensible Binary Meta Language), was eine zukünftige flexible Erweiterung erlaubt, jedoch auch sicherstellt, dass ältere Programme weiterhin damit umgehen können. Zusätzlich ist das Format fehlertolerant und kann bis zu einem gewissen Grad auch beschädigte Dateien wiedergeben. Zusammen mit den Codecs FFV1 für Video und FLAC für Audio wird eine Standardisierung bei der IETF angestrebt, weshalb diese Kombination empfohlen wird.

Motion JPEG 2000 (MJ2) ist ein speziell für die Archivierung entwickeltes Containerformat, das unter ISO/IEC 15444-3:2007 zertifiziert und als MIME-Type video/mj2 registriert ist. Es verwendet einen eigenen Codec, JPEG 2000, der jedes Einzelbild verlustfrei mit JPEG 2000 komprimiert. Zugehörige auditive Inhalte werden von Motion JPEG 2000 ebenfalls unterstützt. Das Format ist rechenintensiv und nicht vollständig lizenzenfrei, weshalb es besser sein kann den JPEG 2000 Codec in einem anderen Containerformat, wie beispielsweise MXF zu verwenden. Außerdem ist zu Beachten, dass MJ2 auch verlustbehaftete Varianten kennt, die für die Archivierung nicht verwendet werden sollten. Motion JPEG 2000 darf nicht mit dem MPEG-Containerformat oder dem verlustbehafteten Motion JPEG (M-JPEG) verwechselt werden.

Die Familie der MPEG-Containerformate wird von der Moving Picture Experts Group (MPEG) entwickelt und ISO/IEC zertifiziert. Es gibt mehrere Generationen von Standards, darunter MPEG-1 (MPG), MPEG-2 (MPG) oder

MPEG-4 (MP4), die entsprechend unterschiedliche Eigenschaften aufweisen. Die Standards beschreiben nicht nur Containerformate, sondern auch Codecs für Video und Audio. MPEG-1 ist seit 1991 unter ISO/IEC 11172 zertifiziert und beschreibt einen Standard für verlustbehaftete Komprimierung von Audio- und Videodaten, der eine gute Qualität und weite Verbreitung hat. MPEG-1 ist insbesondere für das Audioformat MP3 bekannt, eignet sich aber nicht für die Langzeitarchivierung. Der ebenfalls verlustbehaftete Nachfolger, MPEG-2, ist unter ISO/IEC 13818 zertifiziert und hat eine weite Verbreitung in dem Containerformat MPEG-TS (TS) für Rundfunkübertragungen oder auf DVDs gefunden. MPEG-2 kann nur als Langzeitformat für Dateien verwendet werden, die ursprünglich in diesem Format entstanden sind und nicht mehr bearbeitet werden.

Der unter ISO/IEC 14496 zertifizierte MPEG-4-Standard weist eine höhere Effizienz in der Komprimierung und eine höhere Qualität der Videodaten als seine Vorgänger auf. Er verwendet den Codec H.264/MPEG-4 AVC, der meistens verlustbehaftet verwendet wird, jedoch auch eine kaum verwendete verlustfreie Variante hat. Dieser Codec bietet eine hohe Bildqualität, weshalb er eine weite Verbreitung auf Blu-ray Discs, als Aufnahmeformat oder für das Streaming über das Internet gefunden hat. Teil 14 der MPEG-4-Spezifikation definiert das Containerformat MP4, das auch als MIME-Type video/mp4 registriert ist. Wird dieses Format mit dem H.264-Codec verwendet, kann es zur Langzeitarchivierung verwendet werden, wenn dies entweder dem Ursprungsformat der Videodatei entspricht oder verlustfreie Kompression verwendet wird.

Das Containerformat Material eXchange Format (MXF) ist ein offenes Format, das für den Austausch von auditiven und visuellen Inhalten entwickelt wurde und durch eine Reihe von Standards der Society of Motion Picture and Television Engineers (SMPTE) beschrieben wird. MXF wird vor allem im Bereich Kino und Fernsehen verwendet, wo jeweils spezialisierte Varianten verwendet werden, die nicht immer vollständig untereinander kompatibel sind. MXF ist als MIME-Type application/mxf registriert. Zusammen mit der verlustfreien Variante des Codecs JPEG 2000 stellt MXF eine für die Langzeitarchivierung geeignete Variante dar.

Es gibt eine große Zahl an weiteren Containerformaten, wie etwa MOV und ASF/WMV. Sie eignen sich aber nicht für die langfristige Videoarchivierung, da es sich dabei um proprietäre Formate handelt. Bei Audio Video Interleave (AVI) handelt es sich ebenfalls um ein proprietäres Format, es wird jedoch von einigen Archiven als Archivformat verwendet, da es sich um ein einfaches und robustes Format mit einer großen Verbreitung handelt. Die Formate Ogg und Flash wurden vor allem für das Streaming von Videos entwickelt und sind aufgrund der verwendeten verlustbehafteten Codecs nicht als Langzeitformate geeignet.

Die Wahl des Containerformates hängt von dem zu verwendenden Codec ab, da nicht jeder Codec in jedem Container verwendet werden kann. Bekannte und verbreitete Codecs werden nachfolgend beschrieben.

Der Nachteil bei unkomprimierten Videodateien ist, dass sie sehr viel Speicherplatz (teilweise mehrere GB pro Filminminute) beanspruchen und auch für die Wiedergabe entsprechend leistungsfähige Hardware benötigen. Eine Alternative ist die Verwendung von Codecs mit verlustfreier Kompression, wie etwa FFV1, oder HuffYUV und Lagarith. Der FFmpeg Video Codec 1 (FFV1) wurde im Rahmen des FFmpeg-Projektes entwickelt, ist offen dokumentiert, hat

eine weite Verbreitung gefunden und komprimiert sehr gut und schnell. Diese Eigenschaften machen ihn zu einem empfehlenswerten Codec für die Langzeitarchivierung. HuffYUV und dessen Abspaltung Lagarith wurden eigens für Windows-Systeme entwickelt, weshalb diese nicht für die langfristige Archivierung empfohlen werden können.

Einige Codecs können sowohl verlustfrei als auch verlustbehaftet komprimieren. Dazu gehören: Dirac/Schroedinger, JPEG 2000, H.264/MPEG-4 AVC und H.265/MPEG-H (HEVC). Dirac/Schroedinger wurde beim BBC entwickelt, ist jedoch nicht sehr performant und wird von nur wenigen Programmen unterstützt, weshalb er aktuell nicht für die Archivierung empfohlen werden kann. Bei der Verwendung von JPEG 2000 muss darauf geachtet werden, dass die verlustfrei komprimierende Variante verwendet wird, da dieser Standard auch verlustbehaftet komprimieren kann. H.264/MPEG-4 AVC kann in einer verlustfreien Variante verwendet werden, wenn das entsprechende Programm dies unterstützt. Die verlustbehaftete Variante von H.264/MPEG-4 AVC kann nur für die Archivierung von Dateien verwendet werden, die in diesem Codec entstanden sind und nicht mehr bearbeitet werden. Seit 2013 gibt es den Codec High Efficiency Video Coding (HEVC, auch H.265/MPEG-H), der als Nachfolger für H.264 gedacht ist, jedoch noch nicht vollständig spezifiziert ist. Daher und aus patentrechtlichen Gründen ist dieser Codec nicht für die Archivierung zu empfehlen.

Neben den oben erwähnten Codecs gibt es weitere, jedoch verlustbehaftete Codecs, wie M-JPEG, Theora und DV. Im Gegensatz zu JPEG 2000 ist der Codec Motion JPEG (M-JPEG) nicht für die langfristige Archivierung von Videodateien geeignet, da dieser keine verlustfreie Komprimierung unterstützt. Theora ist der für Ogg-Dateien entwickelte Codec. DV ist ein Codec für digitale Videos auf Videokassetten, der mit der Ablösung von Kassetten durch andere Speichermedien obsolet wird. Theora und DV sind keine geeigneten Codecs für die Langzeitarchivierung.

Hinweise auf geeignete Audioformate und dazugehöriger Metadaten sind in dem Kapitel Audio ab Seite 149 zu finden.

Format	Begründung
✓ Matroska (MKV)	Ein offenes Containerformat, das eine große Bandbreite von Codecs und ergänzenden Inhalten unterstützt. Für die Archivierung können die Codecs FFV1 für Video und FLAC für Audio empfohlen werden. Weitere geeignete Codecs für Matroska sind H.264/MPEG-4 AVC und MPEG-2.
~ Motion JPEG 2000 (MJ2)	Motion JPEG 2000 verwendet den Codec JPEG 2000 und ist unter ISO/IEC 15444-3:2007 zertifiziert.
MP4	Der unter ISO/IEC 14496 zertifizierte MPEG-4-Standard verwendet den Codec H.264/MPEG-4 AVC, der meistens verlustbehaftet verwendet wird, jedoch auch verlustfrei komprimieren kann. Das Containerformat MP4 ist als MIME-Type video/mp4 registriert. Wird dieses Format mit dem H.264-Codec verwendet, kann es zur Langzeitarchivierung verwendet werden, wenn dies entweder dem Ursprungsformat der Videodatei entspricht oder verlustfreie Kompression verwendet wird.
MXF	Das Containerformat Material eXchange Format (MXF) ist ein offenes Format, das durch eine Reihe von Standards der SMPTE beschrieben wird. MXF ist als MIME-Type application/mxf registriert. Zusammen mit der verlustfreien Variante des Codecs JPEG 2000 stellt MXF eine für die Langzeitarchivierung geeignete Variante dar.
MPEG-2	MPEG-2 ist unter ISO/IEC 13818 zertifiziert und komprimiert verlustbehaftet. MPEG-2 kann nur als Langzeitformat für Dateien verwendet werden, die ursprünglich in diesem Format entstanden sind und nicht mehr bearbeitet werden.
MPEG (weitere Varianten)	Mit Ausnahme von MPEG-4 und MPEG-2 können andere MPEG-Varianten nicht für die Langzeitarchivierung empfohlen werden.
✗ AVI	Audio Video Interleave ist ein von Microsoft entwickeltes proprietäres Format mit einer weiten Verbreitung. Es eignet sich nicht für die Langzeitarchivierung.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Format	Begründung
MOV	Ein verbreitetes proprietäres Format von Apple, das in QuickTime verwendet wird und nicht für die Langzeitarchivierung geeignet ist.
ASF/WMV	Ein proprietäres Format von Microsoft, das nicht für die Langzeitarchivierung geeignet ist.
Ogg	Ein von Xiph entwickeltes und offenes Format für das Streaming von Videos, das jedoch nicht für die Archivierung geeignet ist.
Flash	Ein von Macromedia und dann von Adobe entwickeltes Format für das Streaming von Videos. Es ist nicht für die Langzeitarchivierung geeignet.

Dokumentation Neben den allgemeinen minimalen Angaben zu Einzeldateien, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 gelistet sind, werden für Videos weitere Angaben benötigt, die insbesondere technische Details dokumentieren.

Die technischen Angaben zu Bildformat, Seitenverhältnis, Bildfrequenz, Bitrate und Codec werden zur korrekten Wiedergabe der Datei benötigt. Angaben zu Länge, Tonkanälen, Profilen und weiteren Inhalten sind zur Prüfung auf Vollständigkeit der Datei erforderlich.

Bereits eingebettete Metadaten, wie beispielsweise Exif, oder Bestandteile im Containerformat sollten behalten und archiviert werden. Am besten werden sie in eine eigene Text- oder XML-Datei transferiert und getrennt gespeichert.

Metadatum	Beschreibung
Länge	Dauer des Videos. Diese Angaben sollten konform zu ISO 8601 erfolgen. Beispiel: P3Y6M4DT12H30M5S (3 Jahre, 6 Monate, 4 Tage, 12 Stunden, 30 Minuten und 5 Sekunden) oder T2H2M (2 Stunden und 2 Minuten)
Beteiligte	Angabe der Beteiligten, wie etwa Autor, Regisseur, Darsteller, Interviewpartner etc.
Bildgröße und Seitenverhältnis	Maße des Bildes gemessen in Pixeln, z.B. 1280px × 720px, und Angabe des Seitenverhältnisses für die korrekte Darstellung im Verhältnis Breite zu Höhe, z.B.: 16:9
Bildfrequenz	Angabe der Bildfrequenz in Bildern pro Sekunde, z.B. 25

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Metadatum	Beschreibung
Bitrate Video	Angabe der Datenrate in Bits pro Sekunde, z.B. 863 kbps
Farbraum	Angabe des verwendeten Farbraums, z.B. YUV
Farbtiefe	z.B. 8 bit oder 10 bit
Farbunterabtastung	Angabe der verwendeten Farbunterabtastung, z.B. 4:2:0
Videocodec	Name und Version des verwendeten Videocodecs und welche Kompression verwendet wird
Audiostreams	Anzahl der enthaltenen Audiostreams
Audiocodec	Angaben zu den verwendeten Audio codecs je enthaltenen Audiostreams
Bitrate Audio	Angabe der Datenrate in Bits pro Sekunde je enthaltenen Audiostreams, z.B. 666 kbps
Abtastrate	Angabe der Abtastrate in Hertz je enthaltenen Audiostreams, z.B. 44.1 kHz
Tonkanäle	Angabe über Anzahl der Tonkanäle je enthaltenen Audiostreams
Profil	Wenn ein bestimmtes Videoprofil für ein Containerformat verwendet wurde, bitte angeben.
Weitere Inhalte	Angabe über weitere Inhalte, die in dem Containerformat enthalten sind oder als zusätzliche Datei vorliegen, wie beispielsweise Untertitel
Aufnahmegerät	Herstellername und Modell des Aufnahmegeräts (z.B. eine Kamera)
Software	Name und Versionsnummer der Software, mit der das Video aufgenommen, erstellt oder bearbeitet wurde, wie z.B. Adobe Premiere Pro CC (2015.1)

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Vertiefung

Die Informationen für Bild und Ton von Videos werden jeweils in einem eigenen Format gespeichert, dem sogenannten Codec, welche in einem Containerformat zusammengeführt werden.

Das Bildformat von Videodateien ist nicht nur von der Anzahl der Pixel, sondern auch von dem Seitenverhältnis abhängig. Analog zu Rastergrafiken werden Videos in einem Farbraum mit einer bestimmten Farbtiefe gespeichert. Eine Besonderheit dabei ist die Farbunterabtastung, die zur verlustbehafteten Reduzierung der Datenmenge angewendet werden kann. Die Datenmenge ist außerdem von der Bitrate abhängig, die variabel oder konstant sein kann.

Containerformat In der Regel handelt es sich bei Multimediaformaten, zu denen Videodateien zählen, um Containerformate, die mehrere verschiedene Inhalte zusammenfassen. Zu den Inhalten gehört neben dem eigentlichen Bildinhalt auch begleitender Ton. Zusätzlich können Untertitel, strukturierende Angaben, wie etwa Kapitaleinteilungen, oder allgemeine Metadaten enthalten sein. Dabei wird jeder dieser Inhalte in einem eigenen Format gespeichert. Formate für die visuellen und auditiven Inhalte werden Codecs genannt. In Abbildung 3.25 ist ein Videocontainer mit einem Video- und zwei Audiostreams veranschaulicht.

Es gibt eine große Zahl an Containerformaten, die für unterschiedliche Anforderungen entwickelt wurden und daher auch jeweils einen unterschiedlichen Funktionsumfang aufweisen. Beispielsweise können manche Formate, wie beispielsweise Motion JPEG 2000, nur mit einem oder wenigen Codecs umgehen, während andere Formate, wie etwa Matroska (MKV), so gut wie jeden existierenden Video- und Audiocodec speichern können. Das bedeutet auch, dass eine MKV-Datei nicht gleich einer anderen MKV-Datei ist, da sie unterschiedliche Codecs enthalten kann.

Ob eine Mediendatei geöffnet und wiedergegeben werden kann hängt nicht nur davon ab, ob das verwendete Wiedergabeprogramm mit dem Containerformat umgehen kann, sondern auch, ob die verwendeten Codecs unterstützt werden. Die Dateiendung gibt dabei nur über das Containerformat Auskunft, während die darin enthaltenen Codecs erst durch Öffnen der Datei oder mittels Analyseprogrammen ermittelt werden können.

Das Zusammenführen der einzelnen Komponenten in ein Containerformat wird als Multiplexing oder Muxing bezeichnet, während das Aufsplitten der Inhalte eines Containers Demultiplexing oder Demuxing genannt wird.

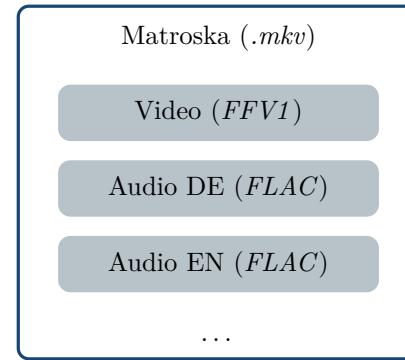


Abb. 3.25: Das Containerformat Matroska enthält einen Video- und zwei Audiostreams. Die Punkte deuten an, dass weitere Inhalte möglich sind. Der verwendete Codec für den Film ist FFV1, der für beide Tonspuren FLAC.

Codec und Kompression Für die Speicherung der eigentlichen Bild- und Toninhalte von Videos in den Containerformaten gibt es eigene Codecs, die definieren wie die Datenströme gespeichert und gelesen werden.

Der Begriff Codec setzt sich aus den englischen Wörtern *coder* (Codierer) und *decoder* (Decodierer) zusammen. Die Konvertierung von einem Codec in einen anderen wird Transcodierung genannt und wird schematisch in Abbildung 3.26 dargestellt.

Codecs bieten immer eine Kompression der Daten, die entweder verlustbehaftet oder verlustfrei erfolgen kann. Dabei verwenden die Codecs unterschiedliche Verfahren an, die grob in Intraframe- und Interframe-Kompression unterteilt werden können. Intraframe-Kompression beschränkt sich auf die Komprimierung der Daten eines Einzelbildes, während bei der Interframe-Kompression über eine ganze Videosequenz hinweg komprimiert wird. Letzteres hat den Nachteil, dass eventuell auftretende Fehler nicht nur ein Einzelbild, sondern eine ganze Sequenz an Bildern korrumpern.

Die einzelnen Codecs werden durch einen FourCC (Four Character Code) im Header der Datei oder des Dateiteils identifiziert. Es handelt sich dabei um einen vier Byte langen Bezeichner, der aus ASCII-Zeichen besteht und somit auch für Menschen lesbar und erkennbar ist, wie beispielsweise *FFV1* für FFV1 oder *MJ2C* für Motion JPEG 2000.

Videodaten können auch unkomprimiert in einem Container gespeichert werden, jedoch muss dabei beachtet werden, dass es verschiedene Varianten gibt, die nicht unbedingt untereinander kompatibel sind und im schlechtesten Fall sogar nur von spezifischen Programmen verarbeitet werden können. Daher sind zumindest weitere Angaben über Farbtiefe, Auflösung und Bildfrequenz erforderlich. Der Nachteil von unkomprimierten Videodateien ist, dass sie sehr viel Speicherplatz (teilweise mehrere GB pro Filmmminute) beanspruchen.

Da nicht jeder Codec in jedem Container verwendet werden kann, hängt die Wahl des Codecs von dem zu verwendenden Containerformat ab.

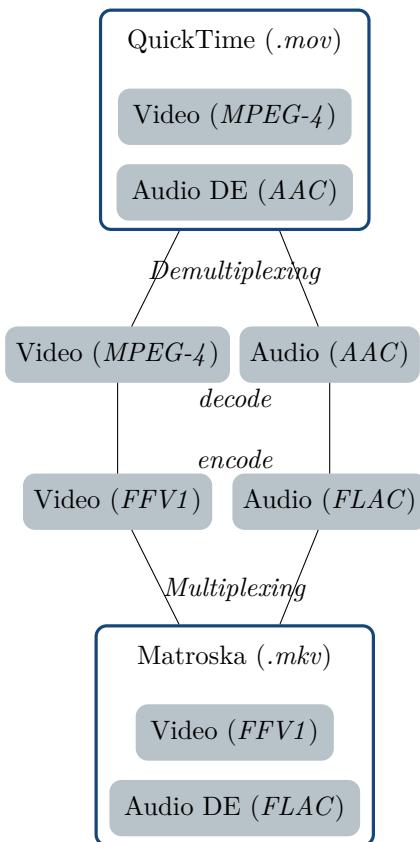


Abb. 3.26: Ablauf einer Transcodierung. Die Quelldatei wird entpackt (Demultiplexing). Die einzelnen Video- und Audioinhalte werden zunächst decodiert und anschließend im Zielcodec codiert. Schließlich werden die Einzelinhalte in der Zielformat verpackt (Multiplexing).

Bildgröße und Seitenverhältnis Für die Bildgröße von digitalen Videos gilt im Prinzip das Gleiche, wie für die Bildgröße von Rastergrafiken, die in dem Abschnitt Bildgröße und Auflösung ab Seite 81 beschrieben wird.

Zusammen mit der Bildgröße sollte immer das Seitenverhältnis angegeben werden, da aufgrund des analogen Ursprungs von einigen Videoformaten die Pixel nicht in jedem Fall quadratisch sind und für die Darstellung entsprechend umgerechnet werden müssen. Das Seitenverhältnis bezieht sich auf die Darstellung des Bildes und wird als Verhältnis von Breite zu Höhe angegeben, wie beispielsweise 4:3 für SD PAL.

Wenn ein analoges Video digitalisiert wird, muss darauf geachtet werden, dass das ursprüngliche Bild nicht verzerrt oder beschnitten wird, also das Seitenverhältnis gleich bleibt. Bei der Wiedergabe kann das Bild in das Format des Wiedergabemediums durch sogenanntes Pillarboxing eingefügt werden, wobei schwarze Ränder in Kauf zu nehmen sind. Ein typisches Beispiel hierfür ist die Wiedergabe von Videokassetten mit dem Verhältnis 4:3 auf Breitbildmonitoren mit einem Seitenverhältnis von 16:9.



Abb. 3.27: Das Bild links im Seitenverhältnis 4:3 wurde rechts mittels Pillarboxing in das Seitenverhältnis 16:9 übertragen, wodurch links und rechts davon schwarze Balken zu sehen sind. (Blender Foundation)

Farbinformationen Was Farbtiefe und Farbraum bedeutet, wird in dem Kapitel über Rastergrafiken in dem Abschnitt Farbtiefe ab Seite 81 und in dem Abschnitt Farbmodell und Farbraum ab Seite 82 erläutert.

Speziell für die Farbdarstellung auf selbstleuchtenden Geräten wurden eigene Farbmodelle entwickelt, zu denen YUV, YCbCr und YPbPr mit jeweils drei Kanälen gehören. Dabei beschreibt der erste Kanal, das Y, die Helligkeit der einzelnen Bildpunkte (Luma oder auch Luminanz). Der Y-Kanal wird somit für die Übertragung eines Schwarz-Weiß-Bildes verwendet. Der Farbanteil (Chroma oder Chrominanz) wird jeweils mit den zwei übrigen Kanälen (U & V, Cb & Cr oder Pb & Pr) repräsentiert.

Da das menschliche Auge Farben im Vergleich zu Helligkeiten schlechter auflösen kann, wird eine Reduzierung in der Auflösung von Farbanteilen schlechter wahrgenommen. Wenn also der Y-Kanal vollständig übertragen wird, während die beiden Farbkanäle jeweils reduziert übertragen werden, ist optisch für den Menschen kein Unterschied in der Schärfe feststellbar. Diesen Umstand macht sich die Farbunterabtastung (engl. *chroma subsampling*) zunutze, die im Folgenden und in Abbildung 3.29 auf Seite 142 kurz erläutert wird.

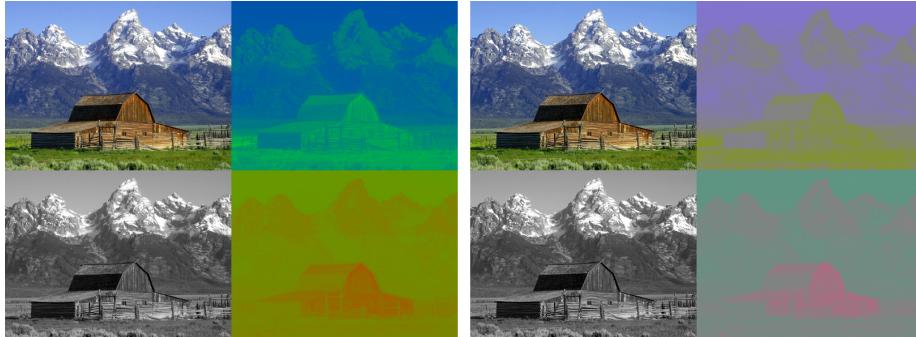


Abb. 3.28: Links die Aufteilung eines Bildes in die drei Kanäle Y, U und V. Auf der rechten Seite wurde das Bild in die drei Kanäle Y, Cb und Cr aufgeteilt. (Wikimedia Commons)

Farbunterabtastung wird meist durch drei Zahlen dargestellt, die jeweils mit einem Doppelpunkt voneinander getrennt werden, wie beispielsweise 4:2:0. Die erste Ziffer bezieht sich auf die Luminanz, die als Referenz für die weiteren Ziffern verwendet wird. Es handelt sich dabei üblicherweise um eine Vier.

Die zweite Ziffer beschreibt die horizontale Abtastung beider Farbkanäle im Verhältnis zur Luminanz. Handelt es sich dabei um eine Vier, wird die Chrominanz vollständig abgetastet, während eine Zwei bedeutet, dass die Chrominanz in der Horizontalen nur halb so oft, wie die Luminanz abgetastet wird.

Die dritte Ziffer bezieht sich auf die vertikale Abtastung. Wenn die dritte Ziffer gleich der zweiten Ziffer ist, wird keine vertikale Farbunterabtastung angewendet. Wenn die dritte Ziffer eine Null ist, findet eine vertikale Farbunterabtastung im Verhältnis 2:1 beider Farbkanäle statt.

Eine vierte Ziffer, die immer den gleichen Wert der ersten Ziffer hat, wird verwendet, wenn zusätzlich Transparenz kodiert werden soll.

Datenrate Die Datenrate, gibt an wie viele Daten in einer bestimmten Zeitspanne übertragen oder gelesen werden. Bei digitalen Videos wird die Datenrate in Bits pro Sekunde gemessen, weshalb hier auch von Bitrate gesprochen wird.

Wenn eine konstante Bitrate (*constant bitrate*, CBR) angewendet wird, wird pro Sekunde immer die gleiche Datenmenge übertragen, was zur Folge haben kann, dass komplexe Bildinhalte an Qualität verlieren. Bei der Anwendung einer variablen Bitrate (*variable bitrate*, VBR) variiert die Menge der übertragenen Daten je nach Komplexität des Bildinhaltes, was eine höhere Bildqualität zur Folge hat. Da bei der VBR ruhige und einfache Bildübergänge und -inhalte mit einer niedrigen Datenrate gespeichert werden, kann der Speicherplatzverbrauch verringert werden.

In den meisten Codecs wird eine variable Bitrate verwendet.

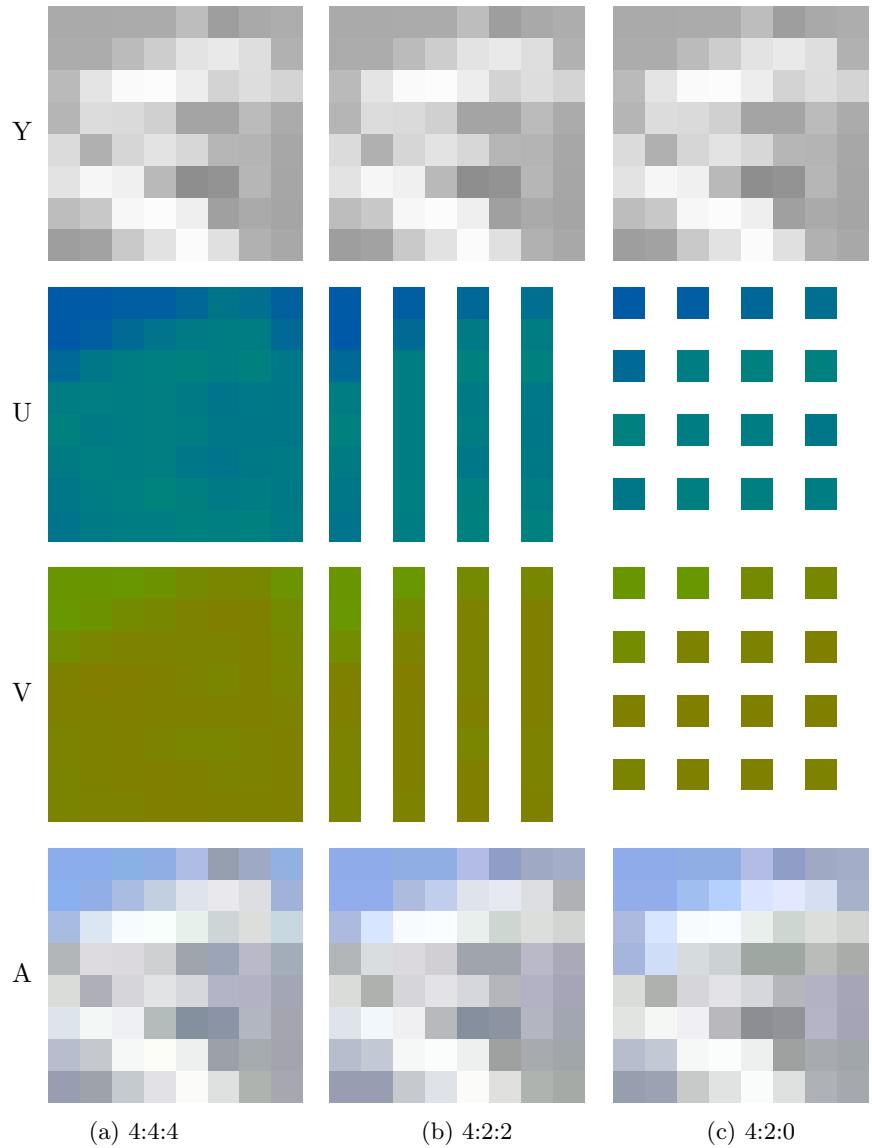


Abb. 3.29: Schematische Darstellung der Farbunterabtastung. Bei 4:4:4 werden die Informationen aller drei Kanäle vollständig abgetastet. Bei 4:2:2 wird der Y-Kanal vollständig und im U- und V-Kanal jeweils nur jeder zweite Pixel jeder Zeile gelesen. Bei 4:2:0 wird der Y-Kanal ebenfalls vollständig gelesen, während die Farbkanäle U und V jeweils nur zur Hälfte in der Horizontalen und in der Vertikalen gelesen werden, also nur jeder zweite Pixel in jeder zweiten Zeile. Werden die drei Kanäle zusammengerechnet, weisen die Gesamtbilder optisch kaum sichtbare Unterschiede auf.

Praxis

In diesem Abschnitt sind Hinweise zum Umgang mit Videodateien gesammelt. Es gibt Erläuterungen mit Literatur- und Programmhinweisen über die Ansicht und Extraktion von Metadaten, die Wiedergabe und die Transcodierung von Videos. Zur Digitalisierung von Filmmaterial können nur allgemeine Hinweise gegeben werden.

Metadaten Die technischen Metadaten der Containerdatei sind in der Regel in dessen Header gespeichert. Weitere Metadaten können in den Video- und Audiostreams gespeichert sein. Diese Daten können von verschiedenen Programmen ausgelesen werden. Ein Programm, das mit vielen Formaten umgehen kann und auch den Export von Metadaten in eine gesonderte Datei erlaubt ist MediaInfo.

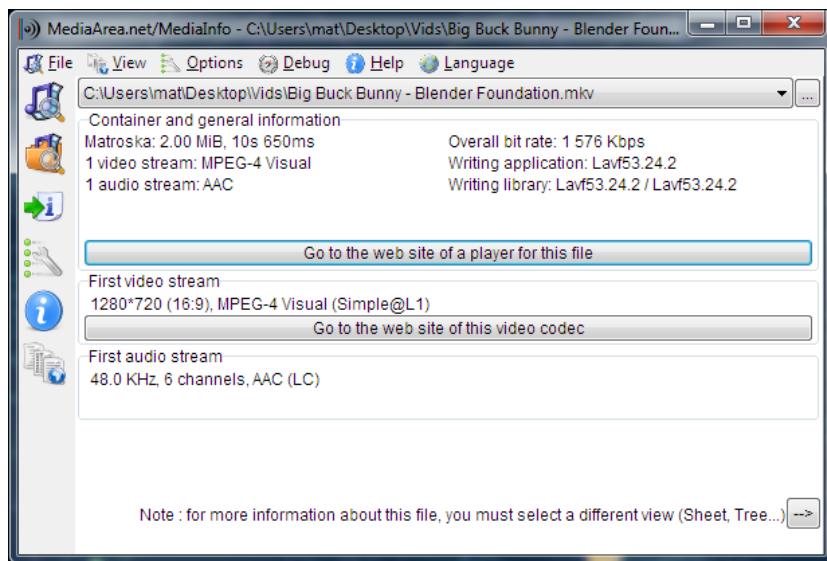


Abb. 3.30: Die Übersichtsseite von MediaInfo für eine analysierte MKV-Datei.

Zu den Exportformaten gehören PBCore und EBUCore. Es handelt sich dabei um Metadatenschemata, die speziell für Film und Ton von Public Broadcasting in den USA, beziehungsweise von der European Broadcasting Union (EBU) entwickelt wurden. Die Moving Pictures Experts Group (MPEG) hat MPEG-7 für die Dokumentation von Multimediateilen entwickelt, der insbesondere zur Erweiterung der anderen MPEG-Standards gedacht ist.

MediaInfo:

<https://mediaarea.net/MediaInfo>

PBCore:

<http://pbcore.org/>

EBUCore:

<https://tech.ebu.ch/MetadataEbuCore>

MPEG-7:

<http://mpeg.chiariglione.org/standards/mpeg-7>

Wiedergabeprogramme Jedes Betriebssystem hat meist schon ein Programm für die Wiedergabe von Videodateien vorinstalliert. Jedoch können diese Programme nur mit einer begrenzten Auswahl von Containerformaten und Codecs umgehen. Codecs können nachträglich installiert werden, wobei aber auf Kompatibilität mit dem Wiedergabeprogramm geachtet werden muss.

Ein Programm, das mit allen hier vorgestellten Formaten und Codecs umgehen kann, ist der VLC media player. Unter der einfach gehaltenen Oberfläche verbirgt sich eine Vielzahl an Funktionalitäten und das Programm gibt es für alle gängigen Betriebssysteme.

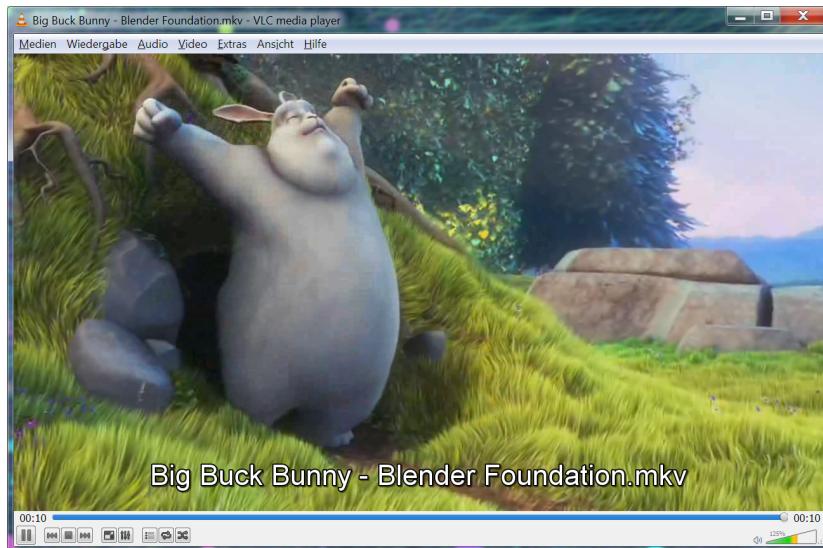


Abb. 3.31: VLC media player

VLC media player von VideoLAN:

<http://www.videolan.org/vlc/index.html>

Transcodierung Für die Transcodierung von einem Codec in einen weiteren gibt es einige frei verfügbare Programme, wie beispielsweise der bereits vorgestellt VLC media player, der eine leicht verständliche grafische Oberfläche anbietet. Auch das frei verfügbare Programm Handbrake ist für mehrere Betriebssysteme verfügbar und bietet eine grafische Oberfläche, ist jedoch im Gegensatz zu VLC nur auf die Transcodierung spezialisiert.

Für fortgeschrittene Anwender ist das ebenfalls frei verfügbare FFmpeg zu empfehlen, das komplexere und detailliertere Transcodierungsoptionen erlaubt und auch von der Kommandozeile aus gesteuert werden kann.

VLC media player von VideoLAN:

<http://www.videolan.org/vlc/index.html>

Handbrake:

<https://handbrake.fr/>

FFmpeg:

<https://www.ffmpeg.org/>

Digitalisierung und Aufnahme Die Digitalisierung von Film- und Videomaterial ist angesichts der beinahe unüberschaubaren Menge an analogen Film- und Videoformaten nicht Gegenstand dieser Empfehlungen.

Im Allgemeinen sollte darauf geachtet werden, dass das originale analoge Material weiterhin erhalten bleibt. Bei der Digitalisierung muss auf größtmögliche Qualität und Originaltreue geachtet werden, wobei vor allem Farbwiedergabe und Seitenverhältnis unverändert bleiben sollten. Die Konsultierung eines darauf spezialisierten Archivs oder Dienstleisters ist hierbei ratsam.

Die Aufnahmen neuer Videos sollten digital erfolgen, da diese die Analogtechnik inzwischen an Qualität meist übertrifft. Dabei sollten die Audiodaten in linearem PCM oder FLAC aufgenommen werden und die Videodaten entweder unkomprimiert, als FLAC oder in möglichst hochqualitativem H.264/MPEG-4 AVC gespeichert werden. Von der DFG gibt es weitere Hinweise in der Handreichung *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*.

Quellen

Archaeology Data Service, Digital Video: A Guide to Good Practice
http://guides.archaeologydataservice.ac.uk/g2gp/Video_Toc

W. Bergmeyer, Multimedia/Komplexe Applikationen, in: H. Neuroth – A. Oßwald – R. Scheffel – S. Strathmann – K. Huth (Hrsg.) nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchiverung. Version 2.3 (2010) Kap. 17.4
<http://www.nestor.sub.uni-goettingen.de/handbuch>

G. Blood, Refining Conversion Contract Specifications: Determining Suitable Digital Video Formats for Medium-term Storage. (2011)
http://www.digitizationguidelines.gov/audio-visual/documents/IntrmMastVidFormatRecs_2011001.pdf

P. Bubestinger – H. Lewetz – M. Jaks, The archivist's video codec and container FAQ (2013)
http://download.das-werkstatt.com/pb/mthk/info/video/FAQ-digital-video_archiving.html

P. Bubestinger – H. Lewetz – M. Jaks, Comparing video codecs and containers for archives (2015)
http://download.das-werkstatt.com/pb/mthk/info/video/comparison_video_codecs_containers.html

B. Devlin, MXF – The Material Exchange Format, in: EBU Technical Review (2002)
http://www.ebu.ch/en/technical/trev/trev_291-devlin.pdf

B. Devlin, MXF. What is it, how does it work, and why hasn't it solved the

- world's problems yet? (2012)
<http://www.tvtechnology.com/research-E-standards/0114/mxf/262167>
- DFG Handreichung Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora
http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf
- equasys (Hrsg.), Color Formats (2015)
<http://www.equasys.de/colorformat.html>
- Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (Hrsg.) Katalog archivischer Dateiformate: Videodaten
<http://kost-ceco.ch/wiki/whelp/KaD/pages/Video.html>
- C. Lacinak, A Primer on Codecs for Moving Image and Sound Archives. 10 Recommendations for Codec Selection and Management (2010)
http://www.avpreserve.com/wp-content/uploads/2010/04/AVPS_Codec_Primer.pdf
- E. Lorrain, A short guide to choosing a digital format for video archiving masters (2014)
<https://www.scart.be/?q=en/content/short-guide-choosing-digital-format-video-archiving-masters>
- Memoriav (Hrsg.), Digitale Archivierung von Film und Video: Grundlagen und Orientierung (2015)
http://memoriav.ch/wp-content/uploads/2015/04/Empfehlungen_Digitale_Archivierung_Version1.0.pdf
- G. Pearson – M. Gill, An Evaluation of Motion JPEG 2000 for Video Archiving, in: Proceedings IS&T Archiving 2005 vom 26. bis 29. April in Washington, D.C. (Washington, D.C. 2005) 237-243
http://dancearchivalproject.wikispaces.asu.edu/file/view/MJ2_video_archiving.pdf
- C. Poynton, Chroma subsampling notation (2008)
http://www.poynton.com/PDFs/Chroma_subsampling_notation.pdf
- D. Sauter, Video, In: H. Neuroth – A. Oßwald – R. Scheffel – S. Strathmann – K. Huth (Hrsg.) nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchiverung. Version 2.3 (2010) Kap. 17.5
<http://www.nestor.sub.uni-goettingen.de/handbuch>
- K. van Malssen, Digital Video Preservation and Oral History, in: D. Boyd – S. Cohen – B. Rakerd – D. Rehberger (Hrsg.) Oral History in the Digital Age (Washington, D.C. 2012)
<http://ohda.matrix.msu.edu/2012/06/digital-video-preservation-and-oral-history/>
- FFV1 vs other formats for preservation, Diskussion in Google Gruppe
<https://groups.google.com/forum/#!topic/archivematica/HuLV96gJ0go>
- What is a FOURCC? (englisch, 04. 2016)
<http://www.fourcc.org/fourcc.php>
- Wikipedia Bitrate (deutsch, 04. 2016)
<https://de.wikipedia.org/wiki/Bitrate>

Wikipedia MPEG-4 (englisch, 04. 2016)
<https://en.wikipedia.org/wiki/MPEG-4>

Filmausschnitte aus Big Buck Bunny von der Blender Foundation (2008)
<http://www.bigbuckbunny.org/>

Formatspezifikationen

Matroska:

<https://www.matroska.org/>

Matroska: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000342.shtml>

Matroska: IETF CELLAR Charter

<https://datatracker.ietf.org/doc/charter-ietf-cellar/>

Motion JPEG 2000:

<https://jpeg.org/jpeg2000/>

Motion JPEG 2000: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000127.shtml>

Motion JPEG 2000: RFC 3745

<http://www.ietf.org/rfc/rfc3745.txt>

Motion JPEG 2000: R. Kromer, Glossareintrag JPEG 2000

<http://avpres.net/Glossar/JPEG2000.html>

MPEG:

<http://mpeg.chiariglione.org/>

MPEG-1:

<http://mpeg.chiariglione.org/standards/mpeg-1>

MPEG-1: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000035.shtml>

MPEG-2:

<http://mpeg.chiariglione.org/standards/mpeg-2>

MPEG-2: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000335.shtml>

MPEG-4:

<http://mpeg.chiariglione.org/standards/mpeg-4>

MPEG-4: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000037.shtml>

MPEG-4: RFC 4337

<http://www.ietf.org/rfc/rfc4337.txt>

MXF: SMPTE 377-1

<http://dx.doi.org/10.5594/SMPTE.ST377-1.2011>

MXF: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml>

MXF: RFC 4539

<http://www.ietf.org/rfc/rfc4539.txt>

FFV1:

<http://www.ffmpeg.org/~michael/ffv1.html>

Tools und Programme

MediaInfo:

<https://mediaarea.net/MediaInfo>

PBCore:

<http://pbcore.org/>

EBUCore:

<https://tech.ebu.ch/MetadataEbuCore>

MPEG-7:

<http://mpeg.chiariglione.org/standards/mpeg-7>

VLC media player von VideoLAN:

<http://www.videolan.org/vlc/index.html>

Handbrake:

<https://handbrake.fr/>

FFmpeg:

<https://www.ffmpeg.org/>

3.9 Audio

M. Trognitz; Mit Beiträgen von: S. Rohde-Enslin, A. Romeyke

Tonaufnahmen werden in der altertumswissenschaftlichen Forschung eher selten erzeugt. Sie werden für die Aufnahme von Interviews, rekonstruierten Musikinstrumenten im Rahmen der Musikarchäologie oder akustischen Eigenschaften von archäologischen Orten erstellt. Sogar Hörbücher altertumswissenschaftlichen Inhalts existieren mittlerweile.

Eine Audiodatei enthält auditive Inhalte, die durch zusätzliche Komponenten wie Kapiteleinteilungen angereichert sein können. Die zusammenfassende Speicherung aller Inhalte erfolgt in einem Containerformat. Die auditiven Inhalte selbst werden nochmals in einem eigenen Format, dem sogenannten Codec, gespeichert. Je nachdem welches Containerformat verwendet wird, können weitere Informationen, wie Metadaten oder Transkriptionen gespeichert werden.

Langzeitformate Da es sich bei den Dateiformaten für digitale Audiodateien um Containerformate handelt, muss bei der Auswahl für die Langzeitarchivierung nicht nur ein passendes Format, sondern auch ein geeigneter Codec gefunden werden. Dabei gelten für die Auswahl des Codecs und des Containerformats die Kriterien, die für die Wahl von Dateiformaten ab Seite 52 beschrieben werden: Es sollte sich um einen weit verbreiteten, möglichst nicht proprietären und offen dokumentierten Standard handeln, der verlustfreie oder gar keine Kompression anwendet. Außerdem sollte darauf geachtet werden, dass das gewählte Format die Speicherung aller relevanten Elemente der Audiodatei, wie etwa Kapiteleinteilungen, unterstützt.

Da viele Containerformate für Audio nur ein Codec-Format speichern können, werden diese in der Regel unter einem Namen zusammengefasst.

Für die Speicherung von unkomprimierten Audiodaten hat sich lineares PCM (Lineare Puls-Code-Modulation, auch LPCM) als Standard durchgesetzt, das von verschiedenen Containerformaten unterstützt und auch von der IASA (International Association of Sound and Audiovisual Archives) empfohlen wird. Der einzige Nachteil ist der große Speicherplatzbedarf mit etwa 10 MB pro Minute.

Eine verlustfrei komprimierende Alternative für die Langzeitarchivierung stellt der Free Lossless Audio Codec (FLAC) dar, der offen dokumentiert und frei verfügbar ist. FLAC wird beispielsweise von dem Containerformat Matroska unterstützt oder kann als eigenständiges Format verwendet werden.

Das Waveform Audio File Format (WAVE) wurde von Microsoft und IBM als Teil des Resource Interchange File Format (RIFF) entwickelt. Es ist ein offen dokumentiertes aber proprietäres Format, das mehrere Audio-Codecs unterstützt. Aufgrund der weiten Verbreitung wird dieses Format jedoch zusammen mit linearem PCM von der IASA empfohlen und hat sich als Standard de facto durchgesetzt.

Die European Broadcast Union (EBU) hat auf WAVE aufbauend das Format BWF entwickelt, das zusätzlich die Einbettung von Metadaten unterstützt. Für die Langzeitarchivierung sollte dieses Format mit linearem PCM verwendet werden. Für Dateien, die größer als 4 GB sind, hat die EBU das Format RF64/MBWF entwickelt, das zusätzlich die Speicherung von mehr als zwei Tonkanälen erlaubt.

Das Containerformat Matroska (MKA) wird seit 2003 explizit als offenes

Containerformat entwickelt, das modernen Ansprüchen genügt und viele verschiedene Codecs enthalten sowie zusätzliche Elemente, wie Kapiteleinteilungen, speichern kann. Das Format basiert auf einer binären Variante von XML, nämlich EBML (Extensible Binary Meta Language), was eine zukünftige flexible Erweiterung erlaubt, jedoch auch sicher stellt, dass ältere Programme weiterhin damit umgehen können. Zusätzlich ist das Format fehlertolerant und kann bis zu einem gewissen Grad auch beschädigte Dateien wiedergeben. Wenn die Audiodaten in linearem PCM oder FLAC gespeichert werden, kann Matroska für die Langzeitarchivierung empfohlen werden.

Aus der Familie der MPEG-Containerformate, die von der Moving Picture Experts Group (MPEG) entwickelt werden und ISO/IEC zertifiziert sind, stammen das MP3- und das AAC-Format. MP3 (MPEG-1 Audio Layer III) wurde schon in MPEG-1 spezifiziert, das seit 1991 unter ISO/IEC 11172 zertifiziert ist. Es handelt sich dabei um einen Audiocodec mit verlustbehafteter Komprimierung, der jedoch eine weite Verbreitung gefunden hat und in einem gleichnamigen Containerformat gespeichert wird.

Advanced Audio Coding (AAC) ist ein Audiocodec, der als Nachfolger von MP3 im Rahmen von MPEG-2 und MPEG-4 spezifiziert wurde und unter ISO/IEC 13818-7 und 14496-3 standardisiert ist. Auch AAC komprimiert verlustbehaftet und kann in einem gleichnamigen Container gespeichert werden oder beispielsweise auch in dem Containerformat MP4. MP3 und AAC sollten nur dann als Langzeitformat für Dateien verwendet werden, wenn diese ursprünglich in dem Format entstanden sind.

Das von Apple entwickelte Audio Interchange File Format (AIFF) ist nicht für die Langzeitarchivierung geeignet, weil es proprietär ist und hauptsächlich nur auf Apple-Systemen Verbreitung gefunden hat. Das von Microsoft entwickelte Windows Media Audio (WMA) ist ebenfalls ein proprietäres Format, das sich wegen der verwendeten verlustbehafteten Kompression nicht für die Langzeitarchivierung eignet. Das Format Ogg wurde vor allem für das Streaming entwickelt und ist aufgrund der verwendeten verlustbehafteten Codecs Opus und Vorbis nicht als Langzeitformat geeignet.

Format	Begründung
FLAC ✓	Free Lossless Audio Codec ist ein verlustfrei komprimierender Codec, der offen dokumentiert und frei verfügbar ist.
WAVE	Waveform Audio File Format wurde von Microsoft und IBM entwickelt und ist offen dokumentiert aber proprietär. Die Audiodaten sollten als lineares PCM gespeichert werden.
BWF	Das Format BWF baut auf WAVE auf und wurde von der EBU entwickelt, um zusätzlich die Einbettung von Metadaten zu unterstützen. Die Audiodaten sollten als lineares PCM gespeichert werden.
Matroska ~	Ein offenes Containerformat, das eine große Bandbreite von Codecs und ergänzenden Inhalten unterstützt. Die Audiodaten sollten als lineares PCM oder FLAC gespeichert werden.
RF64/MBWF	Wurde von der EBU aus BWF entwickelt, um Dateien zu speichern, die größer als 4 GB sind oder mehr als zwei Tonkanäle beinhalten. Die Audiodaten sollten als lineares PCM gespeichert werden.
AAC/MP4	Advanced Audio Coding ist der Nachfolger von MP3 und unter ISO/IEC 13818-7 und 14496-3 standardisiert. Er kann unter anderem in den Formaten AAC oder MP4 gespeichert werden. Die Daten werden verlustbehaftet komprimiert, weshalb er nur für Dateien verwendet werden darf, die ursprünglich in diesem Format entstanden sind.
MP3	MPEG-1 Audio Layer III ist unter ISO/IEC 11172 zertifiziert und verwendet verlustbehaftete Kompression. MP3 kann nur als Langzeitformat für Dateien verwendet werden, die ursprünglich in diesem Format entstanden sind.
AIFF ✗	Das Audio Interchange File Format (AIFF) von Apple ist nicht für die Langzeitarchivierung geeignet, weil es proprietär ist.
WMA	Windows Media Audio ist ein von Microsoft entwickeltes Format mit verlustbehafteter Kompression.
Ogg	Ein von Xiph entwickeltes und offenes Format, das jedoch nicht für die Archivierung geeignet ist, da die Codecs Opus und Vorbis verlustbehaftet komprimieren.

Dokumentation Neben den allgemeinen minimalen Angaben zu Einzeldateien, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 gelistet sind, werden für Audiodateien weitere Angaben benötigt, die insbesondere technische Details dokumentieren.

Die technischen Angaben zu Codec, Bitrate, Abtastrate und Abtasttiefe werden zur korrekten Wiedergabe der Datei benötigt und vermitteln einen ersten Eindruck über die Qualität der Datei. Angaben zu Länge, Tonkanälen und weiteren Inhalten sind zur Prüfung auf Vollständigkeit der Datei erforderlich.

Bereits eingebettete Metadaten, wie beispielsweise Exif, oder Bestandteile im Containerformat sollten behalten und archiviert werden. Am besten werden sie in eine eigene Text- oder XML-Datei transferiert und getrennt gespeichert.

Metadatum	Beschreibung
Beteiligte	Angabe der Beteiligten, wie etwa Autor, Komponist, Interpret, Interviewpartner etc.
Länge	Dauer der Audiodatei. Diese Angaben sollten konform zu ISO 8601 erfolgen. Beispiel: P3Y6M4DT12H30M5S (3 Jahre, 6 Monate, 4 Tage, 12 Stunden, 30 Minuten und 5 Sekunden) oder T2H2M (2 Stunden und 2 Minuten)
Audio codec	Angabe des verwendeten Audio codecs
Tonkanäle	Angabe der Anzahl der Tonspuren und Benennung des Systems, z.B. 5 (Dolby 5.1)
Bitrate	Angabe der Datenrate in Bits pro Sekunde, z.B. 666 kbps
Abtastrate	Angabe der Abtastrate in Hertz, z.B. 44.1 kHz
Abtasttiefe	Angabe der Anzahl der Quantisierungsstufen als Bittiefe, z.B. 16 bit
Weitere Inhalte	Angabe über weitere Inhalte die in dem Containerformat enthalten sind oder als zusätzliche Datei vorliegen, wie beispielsweise Transkriptionen
Aufnahmegerät	Herstellername und Modell des Aufnahmegeräts (z.B. eines Analog-Digital-Umsetzers oder einer Kamera)
Software	Name und Versionsnummer der Software mit der die Audiodatei aufgenommen, erstellt oder bearbeitet wurde, wie z.B. Audacity 2.1.2

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Vertiefung

Die eigentlichen Audiodaten werden in einem eigenen Format gespeichert, dem sogenannten Codec, welcher wiederum in einem Containerformat eingebettet wird. Nähere Erläuterungen dazu sind in dem Kapitel Video in den Unterab-

schnitten Containerformat ab Seite 138 sowie Codec und Kompression ab Seite 138 zu finden.

Die Qualität einer Audiodatei ist zu einem großen Teil von der Abtastrate und Abtasttiefe abhängig. Die Anzahl der Tonkanäle und deren Anordnung werden mittels der Angabe des Tonsystems beschrieben, welche notwendig ist, um die Datei korrekt abzuspielen.

Die Datenmenge von Audiodaten ist nicht nur von der Abtastrate und der Abtasttiefe, sondern auch von der Bitrate abhängig. In dem Kapitel Video wird dieser Begriff in dem Unterabschnitt Datenrate ab Seite 141 erläutert.

Abtastrate und Abtasttiefe Mit analoger Technik kann Ton so aufgenommen werden, wie er physikalisch auftritt, nämlich als durchgängige Schallwelle. Um dieses Signal zu digitalisieren muss es in bestimmten Zeitintervallen gemessen und gespeichert werden. Die Zeitintervalle, also die Häufigkeit, in der das Signal gemessen wird, wird als Abtastrate bezeichnet. Je kleiner das Zeitintervall, desto größer die Abtastrate und desto mehr entspricht das digitale Signal dem analogen Signal. Die Abtastrate wird meist in Kilohertz (kHz) angegeben.

Wie genau der gemessene und gespeicherte Wert dann dem ursprünglichen Signal entspricht, hängt von der Messgenauigkeit ab. Diese Messgenauigkeit wird durch die Anzahl der Bits bestimmt und wird als Abtasttiefe bezeichnet. Sie ist vergleichbar mit der Farbtiefe, die in dem Kapitel über Rastergrafiken ab Seite 81 beschrieben wird. Je größer die Abtasttiefe, desto feiner kann das Signal gemessen werden und die ursprüngliche Tonqualität reproduziert werden.

Eine höhere Abtastrate und eine größere Abtasttiefe führen zu einer originalgetreueren digitalen Repräsentation des Signals.

Der gemessene Wert des Signals kann zwischen zwei Werten der durch die Messgenauigkeit vorgegebenen Messskala liegen, weshalb dieser auf den nächstliegenden Wert gerundet werden muss. Dieser Vorgang wird als Quantisierung bezeichnet.

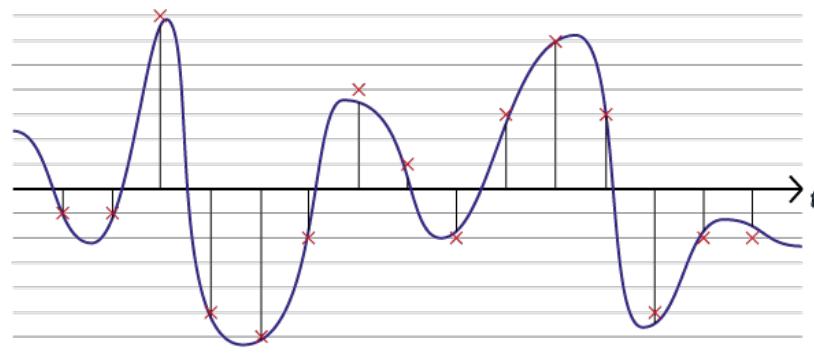


Abb. 3.32: Die Abtastung und Quantisierung eines Schallsignals. Das Schallsignal (blau) wird in einem vorgegebenen Intervall (senkrechte Linien) gemessen. Der gemessene Wert wird auf den nächstgelegenen Wert der durch die Abtasttiefe vorgegebenen Skala (horizontale Linien) quantisiert (Kreuze), weshalb die digitale Repräsentation des Signals leicht abweicht.

Puls-Code-Modulation (PCM) bezeichnet das Abtasten eines Signals in einer vorgegebenen Abtastrate und die anschließende Quantisierung des gemessenen Wertes. Die lineare Puls-Code-Modulation verwendet dabei einen immer gleich groß bleibenden Wertebereich, also immer die gleiche Messskala für die abgetasteten Werte.

Tonkanäle und Tonsysteme Bei der Wiedergabe von Audiodateien, beispielsweise über Kopfhörer, besteht die Möglichkeit Töne nur auf einem der beiden Hörer auszugeben, um einen räumlichen Schalleindruck zu erzeugen. Dafür werden technisch die in der Audiodatei gespeicherten Tonkanäle (Tonspuren) verwendet und an die unterschiedlichen Ausgabegeräte geleitet, wie etwa an den linken und den rechten Kopfhörer.

In einer Audiodatei können eine, zwei oder noch mehr Tonspuren enthalten sein. Dabei kann die intendierte Anordnung der Lautsprecher je nach Anzahl der Kanäle variieren. Bei einer Tonspur wird davon ausgegangen, dass sich das Ausgabegerät vor dem Hörer befindet und bei zwei Tonspuren von zwei Ausgabegeräten links und rechts vom Hörer. Ab drei Kanälen sind mehr als zwei Anordnungen möglich, weshalb die Benennung des Tonsystems oder eine Angabe zur Anordnung notwendig wird.

Praxis

Dieser Abschnitt vereint Hinweise zum Umgang mit Audiodaten. Es gibt Erläuterungen mit Literatur- und Programmhinweisen über die Ansicht und Extraktion von Metadaten, die Aufnahme, die Wiedergabe und die Transcodierung von Audiodateien. Für die Aufnahme und Digitalisierung von Ton werden Mindestanforderungen spezifiziert und Hinweise auf vertiefende Informationen gegeben.

Metadaten Die technischen Metadaten sind in der Regel in dem Header der Datei gespeichert und können von verschiedenen Programmen ausgelesen werden. Ein Programm, das mit vielen Formaten umgehen kann und auch den Export von Metadaten in eine gesonderte Datei erlaubt ist MediaInfo.

Zu den Exportformaten gehören PBCore und EBUCore. Es handelt sich dabei um Metadatenschemata, die speziell für Film und Ton von Public Broadcasting in den USA, beziehungsweise von der European Broadcasting Union (EBU) entwickelt wurden. Die Moving Pictures Experts Group (MPEG) hat MPEG-7 für die Dokumentation von Multimediateien entwickelt, der insbesondere zur Erweiterung der anderen MPEG-Standards gedacht ist.

Für das Editieren von Metadaten im BWF-Format gibt es das Programm BWF MetaEdit.

MediaInfo:

<https://mediaarea.net/MediaInfo>

PBCore:

<http://pbcore.org/>

EBUCore:

<https://tech.ebu.ch/MetadataEbuCore>

MPEG-7:

<http://mpeg.chiariglione.org/standards/mpeg-7>

BWF MetaEdit:

<https://sourceforge.net/projects/bwfmetaedit/>

Aufnahme und Bearbeitung Für die Aufnahme und Bearbeitung von Audiodaten kann das kostenlose Programm Audacity verwendet werden. Audacity gibt es für alle gängigen Betriebssysteme.

Bei der Aufnahme sollte auf eine angemessene Aufnahmegerätqualität geachtet werden. In der Handreichung *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora* der DFG werden eine Abtastrate von 22 kHz und eine Abtasttiefe von 16 bit als Mindestanforderungen spezifiziert. Jedoch ist eine höhere Abtastrate von 48 kHz empfehlenswert, da diese eher der Grenzfrequenz des menschlichen Gehörs entspricht, besser wären sogar 96 kHz. Die Abtastung sollte linear erfolgen.

Welche Abtastrate sich für die gewünschte Aufnahme am besten eignet, kann mittels der Nyquist-Frequenz ermittelt werden. Die Nyquist-Frequenz besagt, dass das originale Signal exakt reproduziert werden kann, wenn die Abtastrate doppelt so hoch wie die höchste Signalfrequenz ist.

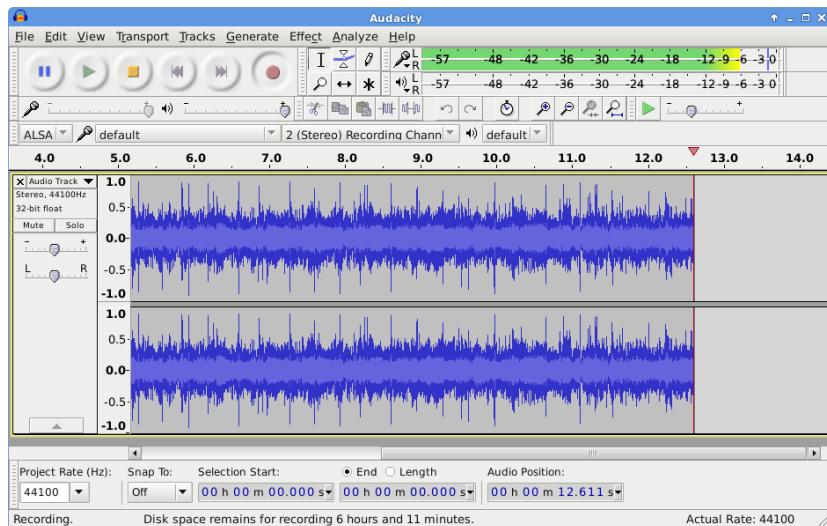


Abb. 3.33: Audacity (audacity.org)

Audacity:

<http://www.audacityteam.org/>

Handreichung der DFG:

http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

Wiedergabeprogramme Jedes Betriebssystem hat meist schon ein Programm für die Wiedergabe von Audiodateien vorinstalliert. Jedoch können diese

Programme nur mit einer begrenzten Auswahl von Containerformaten und Codecs umgehen. Codecs können nachträglich installiert werden, wobei aber auf Kompatibilität mit dem Wiedergabeprogramm geachtet werden muss.

Ein Programm, das mit allen hier vorgestellten Formaten und Codecs umgehen kann, ist der VLC media player. Unter der unscheinbaren Oberfläche verbirgt sich eine Vielzahl an Funktionalitäten und das Programm gibt es für alle gängigen Betriebssysteme.

VLC media player von VideoLAN:
<http://www.videolan.org/vlc/index.html>

Transcodierung Für die Transcodierung, also die Konvertierung von einem Codec in einen weiteren gibt es einige frei verfügbare Programme, wie beispielsweise der bereits genannte VLC media player, der eine leicht verständliche grafische Oberfläche anbietet. Für fortgeschrittene Anwender ist das ebenfalls frei verfügbare FFmpeg zu empfehlen, das komplexere und detailliertere Transcodierungsoptionen erlaubt und auch von der Kommandozeile aus gesteuert werden kann.

VLC media player von VideoLAN:
<http://www.videolan.org/vlc/index.html>

FFmpeg:
<https://www.ffmpeg.org/>

Digitalisierung Für die Digitalisierung von analogem Audiomaterial gibt es von der IASA konkrete Vorgaben, die eine minimale Abtastrate von 48 kHz und eine minimale Abtasttiefe von 24 Bits empfehlen, um eine möglichst hohe Qualität des Digitalisats zu gewährleisten. Als Formate werden WAVE oder BWF mit linearem PCM vorgegeben. Ausführliche Hinweise sind in den Guidelines on the Production and Preservation of Digital Audio Objects (IASA-TC 04) der IASA zu finden.

Es sollte darauf geachtet werden, dass das originale analoge Material weiterhin erhalten bleibt und die Konsultierung eines darauf spezialisierten Archivs oder Dienstleisters ist ratsam.

Quellen

Archaeology Data Service, Digital Audio: A Guide to Good Practice
http://guides.archaeologydataservice.ac.uk/g2gp/Audio_Toc

Hicks, Tony, Should We Be Using ISO 12083?, The Journal of Electronic Publishing 3, 1998, 4
<http://quod.lib.umich.edu/j/jep/3336451.0003.407?view=text;rgn=main>

W. Bergmeyer, Audio, In: H. Neuroth – A. Oßwald – R. Scheffel – S. Strathmann – K. Huth (Hrsg.) nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchiverung. Version 2.3 (2010) Kap. 17.6
<http://www.nestor.sub.uni-goettingen.de/handbuch>

IASA (Hrsg.), Guidelines on the Production and Preservation of Digital Audio Objects (IASA-TC 04)
<http://www.iasa-web.org/tc04/audio-preservation>

- JISC (Hrsg.), Infokit: Digital file formats. Audio
http://www.jiscdigitalmedia.ac.uk/infokit/file_formats/digitisation-audio
- JISC (Hrsg.), Infokit: Digital file formats. Digital audio formats and compression
http://www.jiscdigitalmedia.ac.uk/infokit/file_formats/digital-audio-formats-and-compression
- FADGI (Hrsg.), Guidelines: Embedded Metadata in Broadcast WAVE Files
<http://www.digitizationguidelines.gov/guidelines/digitize-embedding.html>
- Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (Hrsg.) Katalog archivischer Dateiformate: Audiodaten
<http://kost-ceco.ch/wiki/whelp/KaD/pages/Audio.html>
- R. Kromer, On Audio-Visual File Formats (Notizen zur Präsentation 2015)
<http://reto.ch/training/2015/20150424/audio.html>
- Memoriav (Hrsg.), Memoriav Empfehlungen Ton. Die Erhaltung von Tondokumenten (2014)
http://memoriav.ch/wp-content/uploads/2015/02/Empfehlungen_Ton_de.pdf
- Y. Trivedi, HTG Explains: What Are the Differences Between All Those Audio Formats? (2011)
<http://www.howtogeek.com/howto/40465/htg-explains-what-are-the-differences-between-all-those-audio-formats/>
- Wikipedia Mehrkanal-Tonsystem (deutsch 05. 2016)
<https://de.wikipedia.org/wiki/Mehrkanal-Tonsystem>

Formatspezifikationen

- FLAC:
<https://xiph.org/flac/format.html>
- WAVE: McGill University
<http://www-mmsp.ece.mcgill.ca/documents/audioformats/wave/wave.html>
- WAVE mit LCPM: Library of Congress
<http://www.digitalpreservation.gov/formats/fdd/fdd000002.shtml>
- WAVE als Teil von RIFF:
http://www.tactilemedia.com/info/MCI_Control_Info.html
- BWF:
<http://tech.ebu.ch/docs/tech/tech3285.pdf>
- MBWF/RF64:
<http://tech.ebu.ch/docs/tech/tech3306-2009.pdf>
- Matroska:
<https://www.matroska.org/>
- Matroska: Library of Congress
<http://www.digitalpreservation.gov/formats/fdd/fdd000342.shtml>
- Matroska: IETF CELLAR Charter
<https://datatracker.ietf.org/doc/charter-ietf-cellard/>
- MP3 (MPEG-1):

<http://mpeg.chiariglione.org/standards/mpeg-1>

MPEG-1: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000035.shtml>

AAC (MPEG-4):

<http://mpeg.chiariglione.org/standards/mpeg-4>

MP4 mit AAC: Library of Congress

<http://www.digitalpreservation.gov/formats/fdd/fdd000234.shtml>

Tools und Programme

MediaInfo:

<https://mediaarea.net/MediaInfo>

PBCore:

<http://pbcore.org/>

EBUCore:

<https://tech.ebu.ch/MetadataEbuCore>

MPEG-7:

<http://mpeg.chiariglione.org/standards/mpeg-7>

BWF MetaEdit:

<https://sourceforge.net/projects/bwfmetaedit/>

Audacity:

<http://www.audacityteam.org/>

VLC media player von VideoLAN:

<http://www.videolan.org/vlc/index.html>

FFmpeg:

<https://www.ffmpeg.org/>

3.10 3D und Virtual Reality

M. Trognitz; Mit Beiträgen von: D. Lengyel

Im Gegensatz zu statischen zweidimensionalen Bildern können dreidimensionale Repräsentationen von Objekten aus jeder Richtung betrachtet, skaliert und rotiert werden. Ein Punkt in einem 3D-Modell wird von seiner Lage auf der x-, der y- und der z-Achse eines kartesischen Koordinatensystems beschrieben, wobei die z-Achse in diesem Zusammenhang üblicherweise die Tiefe, seltener die Höhe, angibt.

Virtual Reality (Virtuelle Realität) bezeichnet digitale dreidimensionale Welten, mit welchen in Echtzeit interagiert werden kann.

3D-Inhalte können auf unterschiedliche Weise entstehen: durch manuelle Modellierung, wie Rekonstruktionen von Gebäuden, durch Aufnahme, wie etwa einem 3D-Scan von Objekten, oder durch automatisierte Berechnung aus Fotos, wie etwa Photogrammetrie oder Structure from Motion. Von der Entstehungsweise hängen weitere Angaben für die Dokumentation ab, die über die hier angegebenen hinausgehen. Zusätzlichen Angaben sind in den jeweiligen Abschnitten in dem Kapitel Forschungsmethoden ab Seite 194 zu finden, wobei vor allem die Abschnitte Bauforschung, Geodäsie, Geodatenanalyse und Materialaufnahme von Interesse sind. Außerdem bieten die Ergebnisse des Projektes 3D ICONS umfangreiche Informationen zur Dokumentation von 3D-Aufnahmemethoden und die anschließende Verarbeitung.

Langzeitformate 3D-Inhalte sollten in einem offen dokumentierten, textbasierten Format (ASCII) gespeichert werden. Dies ermöglicht bei Bedarf die Rückentwicklung unabhängig von einem bestimmten Programm.

Das zu verwendende Format hängt von der Entstehungsweise und dem Zweck des 3D-Inhaltes ab, da unterschiedliche Formate unterschiedliche Eigenschaften und Elemente speichern, wie beispielsweise Geometrie, Texturen, Lichtquellen oder Standpunkt und Bildausschnitt (auch Viewport genannt). Eine Übersicht über die Speichereigenschaften von den hier empfohlenen 3D-Formaten wird in Tabelle 3.11 auf Seite 168 gegeben.

Das vom Web3D Consortium entwickelte Format X3D (eXtensible 3D Graphics) ist seit 2006 unter ISO/IEC 19775/19776/19777 zertifiziert und eignet sich sowohl zur Speicherung von einzelnen 3D-Modellen, als auch komplexer 3D-Inhalte, wie etwa Virtual Reality. Es ist das Nachfolgeformat von dem seit 2007 unter ISO/IEC 14772-1 zertifizierten VRML-Format und sollte diesem daher vorgezogen werden.

Ein weiteres für komplexe 3D-Inhalte geeignetes Format ist das von der Khronos Group entwickelte COLLADA (collaborative design activity, DAE), das seit 2012 unter ISO 17506 standardisiert ist.

Das U3D-Format bietet einen ähnlichen Funktionsumfang wie X3D und COLLADA und ist insbesondere für das Teilen von 3D-Inhalten in PDF-Dokumenten gedacht. Es ist nicht für die Langzeitarchivierung geeignet.

Die Formate OBJ, PLY und STL eignen sich nicht für komplexe Szenen mit Lichtquellen oder gar Animationen, bieten aber alle Eigenschaften, um die visuellen Oberflächeneigenschaften eines 3D-Objektes zu speichern.

Aus dem CAD-Bereich stammt das Format DXF, welches neben 2D-Inhalten auch 3D-Inhalte speichern kann. Dieses Format sollte nur verwendet werden,

wenn die 3D-Inhalte mit CAD-Software erstellt wurden. In der Industrie werden als alternative Formate auch IGES und STEP verwendet.

Nicht für die Langzeitarchivierung geeignet sind programmspezifische oder binäre Formate, wie beispielsweise FBX, 3DS, MAX, SKP, BLEND, PRC oder NXS.

Um zukünftigen Nutzern einen schnellen Überblick über die 3D-Inhalte zu bieten, ist die zusätzliche Speicherung von Bild- oder Videodateien, die einen ersten Eindruck des Modells vermitteln, zu empfehlen.

Hinweis: Bei der Konvertierung von einem 3D-Format in ein anderes können bestimmte Informationen verloren gehen, wenn sie von dem Zielformat nicht unterstützt werden. Zusätzlich zu dem gewählten Archivierungsformat sollten die originalen Quelldateien, verwendete Texturdateien, Visualisierungen und alle weiteren Dateien, die für die Benutzung und das Verständnis des 3D-Modells relevant sind aufgehoben werden. In den entsprechenden Abschnitten in dem Kapitel Forschungsmethoden ab Seite 194 sind nähere Details zu finden.

Die Abkürzungen in der folgenden Tabelle geben die unterstützten Eigenschaften der 3D-Formate an: DG = Drahtgittermodell, P = Parametrisch, F = Farbe, X = Textur mittels Bild, B = Bumpmapping, M = Material, V = Viewport und Kamera, L = Lichtquellen, T = Transformationen, G = Gruppierung. Diese Begriffe werden in dem Abschnitt Vertiefung ab Seite 163 erläutert.

Format	Begründung
X3D ✓	Das X3D-Format wurde vom Web3D Consortium entwickelt und ist seit 2006 ISO-zertifiziert. Es darf nicht mit dem proprietären Format 3DXML verwechselt werden. Speichert: DG, P, F, X, B, M, V, L, T, G und Animationen.
COLLADA	COLLADA (.dae) wurde von der Khronos Group entwickelt und ist seit 2012 ISO-zertifiziert. Es basiert auf XML und speichert: DG, P, B-Rep, F, X, B, M, V, L, T, G und Animationen.
OBJ	Das offene OBJ-Format wurde von Wavefront Technologies entwickelt und hat eine weite Verbreitung gefunden. Material oder Texturen werden in gesonderten MTL- oder JPG-Dateien gespeichert, die auch archiviert werden müssen. Speichert: DG, P, F, X, B, M und G.
PLY	Das Polygon File Format (acuh Stanford Triangle Format) ist ein einfaches Format, das an der Universität Stanford für Daten von 3D-Scannern entwickelt wurde. Mittels Erweiterungen könnten weitere Eigenschaften gespeichert werden, die aber nicht von jedem Programm unterstützt werden. Für die Langzeitarchivierung sollte die ASCII-Version verwendet werden. Speichert: DG, F, X, B und M.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Format	Begründung
VRML ~	Die Virtual Reality Modeling Language ist das Vorgängerformat von X3D. Die jüngste Version wurde 1997 unter dem Namen VRML97 veröffentlicht. Speichert: DG, P, F, X, B, M, V, L, T, G und Animationen.
STL	Das Format STL wurde von 3D Systems entwickelt. Es steht für <i>stereolithography</i> oder <i>Standard Tessellation Language</i> und findet weite Verbreitung im Bereich von 3D-Druckern und digitaler Fabrikation. Die ASCII-Variante des STL-Formats kann nur die Geometrie eines 3D-Modells speichern. Die Binärvariante des Formats ist weniger speicherintensiv und kann mit einer entsprechenden Erweiterung auch Farben des 3D-Modells speichern, ist aber nicht für die Langzeitarchivierung geeignet.
DXF	Das von Autodesk entwickelte Drawing Interchange Format stammt aus dem CAD-Bereich und sollte nur für 3D-Inhalte verwendet werden, die mit CAD-Software erstellt wurden. Speichert: DG, P, CSG, B-Rep, F und G.
U3D ✗	Das Universal 3D Format ist ein 2005 von der Ecma standardisiertes 3D-Format, das vom 3D Industry Forum mit Mitgliedern wie Intel und Adobe Systems entwickelt wurde. Dieses Format ist nur für die Integration von 3D-Modellen in ein 3D-PDF relevant. Speichert: DG, F, X, B, V, L, T, G und Animationen.
FBX	Ein proprietäres Format von Autodesk, für den Datenaustausch mit anderen kommerziellen 3D-Programmen.
3DS, MAX	Proprietäre binäre Formate von AutoDesk.
SKP	Natives Format von Google SketchUp.
BLEND	Natives binäres Format von Blender.
PRC	PRC (Product Representation Compact) ist ein stark komprimiertes Format für den CAD-Bereich und für die Verwendung in 3D-PDFs, das unter ISO 14739-1 standardisiert ist.
PDF	3D-PDFs mit eingebetteten Modellen in U3D oder PRC eignen sich für einen unkomplizierten Datenaustausch, jedoch nicht für die Langzeitarchivierung.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Format	Begründung
NXS	Nexus ist ein von CNR-ISTI offen entwickeltes Format, für die Web-Visualisierung von 3D-Modellen.

Dokumentation Für die Dokumentation von 3D-Inhalten sollten die Leitsätze der *Londoner Charta* berücksichtigt und eingehalten werden. Dabei ist insbesondere „Leitsatz 4: Dokumentation“ wichtig, dessen erster Absatz lautet:
„Es sollen genügend Informationen dokumentiert und weitergegeben werden, um das Verstehen und Bewerten der computergestützten Visualisierungsmethoden und -ergebnisse in Bezug auf die Zusammenhänge und Absichten, für die sie eingesetzt werden, zu ermöglichen“

Dazu gehören unter anderem die Dokumentation der Forschungsquellen, die zur Erstellung des 3D-Modells herangezogen wurden, die Dokumentation der Prozesse, die im Verlauf der Modellerstellung durchlaufen wurden, die Dokumentation der angewendeten Methoden und eine Beschreibung der Abhängigkeitsverhältnisse zwischen den unterschiedlichen Bestandteilen.

Wenn das Dateiformat es erlaubt, sollte ein Teil der Metadaten zusätzlich dort integriert werden. Allerdings kann der Großteil der Metadaten nur extern abgelegt werden.

Die hier angegebenen Metadaten sind als minimale Angabe zu betrachten und ergänzen die angegebenen Metadaten für Projekte und Einzeldateien in dem Abschnitt Metadaten in der Anwendung ab Seite 36.

Metadatum	Beschreibung
Anzahl Eckpunkte	Aus wie vielen Eckpunkten besteht das 3D-Modell?
Anzahl Polygone	Wie viele Polygone hat das 3D-Modell?
Geometriertyp	Welcher Geometriertyp wird verwendet (Drahtgittermodell, parametrisch, CSG, B-Rep etc.)?
Maßstab	Welcher Maßstab liegt vor bzw. was stellt eine Einheit dar?
Koordinatensystem	Wird ein geographisches Koordinatensystem oder ein beliebiges verwendet?
Modellstatus	Handelt es sich bei dem Modell um das ursprünglich erzeugte und unverarbeitete Modell (den Master) oder ist es ein Modell, das mittels weiterverarbeitenden Schritten, wie Füllen von Löchern, Vereinfachung oder Glättung aus dem Master erzeugt wurde?

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Metadatum	Beschreibung
Detaillierungsgrad (LOD) oder Auflösung	Wie detailliert ist das Modell oder welche Auflösung wurde beim 3D-Scan verwendet?
Ebenen	Werden Ebenen verwendet? Wie viele?
Farbe und Textur	Werden Farben oder Texturen verwendet? Wie werden diese gespeichert? Müssen zusätzliche Texturdateien archiviert werden?
Material	Informationen über die Materialeigenschaften des Modells und inwieweit sie physikalisch korrekt sind.
Licht	Informationen über die Lichtquellen und inwieweit sie physikalisch korrekt sind.
Shader	Werden spezielle, erweiterte Shader verwendet?
Animation	Ist das Modell animiert? Welche Art von Animation wird verwendet (Keyframe, motion capture)?
Externe Dateien	Eine Liste aller externen Dateien, die für das Öffnen der Datei notwendig sind (z.B. Texturen)

Weitere Metadaten sind methodenabhängig und können in den jeweiligen Abschnitten nachgelesen werden.

Vertiefung

Ein übliches 3D-Modell besteht aus Eckpunkten (*vertices*), die Polygone beschreiben. Meistens handelt es sich dabei um Dreiecke. Sind die Eckpunkte und die zu den Polygone gehörenden Kanten erfasst, spricht man von einem Drahtgittermodell (*wire-frame model* oder *mesh*). Sind nur die Eckpunkte erfasst, handelt es sich um eine Punktwolke, welche beispielsweise von 3D-Scannern erzeugt wird. Um eine Punktwolke weiter zu verarbeiten, muss diese vorher meist in ein Drahtgittermodell umgewandelt werden.

Die Inhalte von 3D-Modellen können in drei Kategorien eingeteilt werden: Geometrie, Oberflächeneigenschaften sowie Belichtungs- und Kameraparameter. Zusätzlich kann noch eine Animation oder eine Interaktion hinterlegt sein. Aus diesen Angaben wird die Visualisierung berechnet.

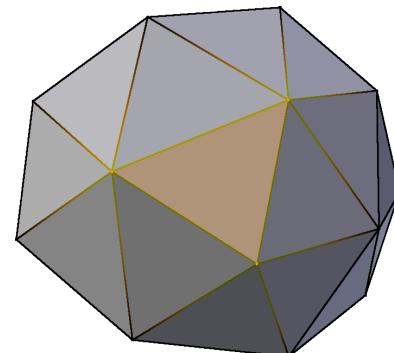


Abb. 3.34: Die gelb markierten Eckpunkte beschreiben das hervorgehobene Dreieck in dem 3D-Modell.

Dieser Vorgang wird als Bildsynthese bezeichnet (*rendering*) und resultiert in statischen Rastergrafiken, Videos oder interaktive Modelle.

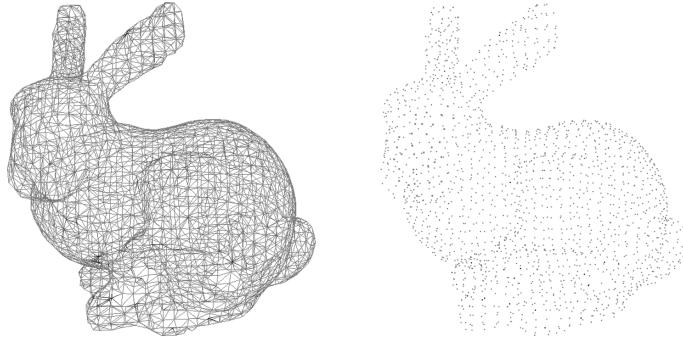


Abb. 3.35: Das Stanford Bunny links als Drahtgittermodell und rechts als Punkt-wolke.

Geometrie Die Form eines 3D-Modells kann mit vier verschiedenen Methoden beschrieben werden: durch ein Drahtgittermodell oder Facetten, parametrisch mit mathematisch beschriebenen Kurven und Flächen, mittels geometrischen Körpern, der sogenannten Konstruktiven Festkörpergeometrie, oder durch die begrenzenden Oberflächen, dem sogenannten Begrenzungsflächenmodell.

Drahtgittermodell: Hierbei wird die Form des 3D-Modells durch dessen Eckpunkte beschrieben, welche wiederum eine Reihe von Polygonen bilden, die zusammen die Oberfläche des Modells darstellen. Drahtgittermodelle sind einfach und schnell zu rendern, jedoch sind bei Detailansichten von konkaven oder konvexen Oberflächen immer die Flächen und scharfen Kanten der einzelnen Polygone erkennbar. Mittels sogenannten Shading-Algorithmen kann beim Rendern der Eindruck von gleichmäßigen Oberflächen generiert werden. Jedoch ist dann an der Kontur immer noch die polygonale Herkunft zu erkennen, wie in Abbildung 3.37 im Abschnitt Oberflächeneigenschaften ab Seite 165 dargestellt.

Eine Annäherung an eine gleichmäßige Oberfläche kann auch durch eine größere Anzahl an kleineren Polygonen erzielt werden, was aber zur Folge hat, dass die Dateigröße des Modells wächst.

Parametrische Darstellung: Viele Kurven und Flächen können mathematisch mit wenigen Parametern errechnet werden. Im Bereich der 3D-Grafik werden für die parametrische Darstellung häufig NURBS (Non-Uniform Rational B-Splines, nicht-uniforme rationale B-Splines) verwendet. Die Verwendung von mathematisch beschriebenen Kurven und Flächen erlaubt eine in der Detailliertheit der Oberfläche verlustfreie Skalierbarkeit. Wird ein 3D-Modell mit parametrischen Flächen in einem Format gespeichert, das diese nicht unterstützt, muss das Modell in Polygone umgerechnet werden, was dazu führt, dass Informationen der Oberflächenstruktur verloren gehen, was vergleichbar mit dem Informationsverlust bei der Konvertierung von einer Vektorgrafik in eine Rastergrafik ist.

Konstruktive Festkörpergeometrie: 3D-Modelle können, anstatt durch Eckpunkte, Flächen oder Kurven auch mit Hilfe von geschlossenen geometri-

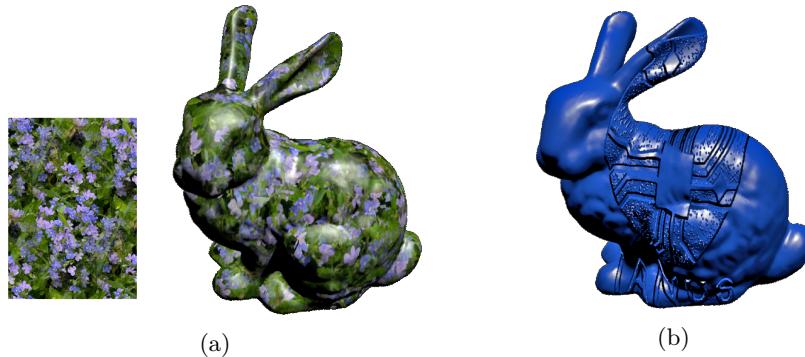


Abb. 3.36: (a) Eine Textur und deren Anwendung auf ein 3D-Modell. (b) Mittels Bumpmapping wurde das IANUS-Logo auf ein 3D-Modell aufgebracht. Es erweckt den Anschein einer dreidimensionalen Veränderung der Oberfläche, jedoch hat sich in der Geometrie des Modells nichts verändert, was an den Umrissen zu erkennen ist.

schen Körpern konstruiert werden. Dabei kann die Vereinigung, die Differenz oder die Schnittmenge der Körper gebildet werden. Beispielsweise kann ein Verkehrshütchen aus der Vereinigung eines Kegels und eines flachen Quaders konstruiert werden. Daher wird diese Vorgehensweise als Konstruktive Festkörpergeometrie (Constructive Solid Geometry, CSG) bezeichnet. Sie erfordert die Speicherung sowohl der einzelnen geometrischen Körper, als auch der angewendeten Methoden, um eine spätere Bearbeitung der Modelle einfach zu ermöglichen. Insbesondere Dateiformate aus dem CAD-Bereich berücksichtigen dies. Eine Konvertierung von Formaten, die CSG unterstützen, zu denen, die keine unterstützen, führt zu Einschränkungen bei der Bearbeitbarkeit, da nur noch Polygone und deren Eckpunkte verändert werden können. Eine Konvertierung von einem Polygonformat zu einem CSG-Format ist nur schwer möglich.

Begrenzungsflächenmodell: Eine weitere Form der Speicherung der Geometrien von 3D-Modellen aus dem CAD-Bereich stellt das Begrenzungsflächenmodell (Boundary Representation, B-Rep) dar. Dabei wird das Modell durch dessen begrenzende Oberflächen beschrieben.

Oberflächeneigenschaften Neben der Form müssen für ein 3D-Modell auch Oberflächeneigenschaften gespeichert werden, die das Aussehen des Modells bestimmen. Neben der Farbe gehören dazu Texturen und Materialeigenschaften, die kombiniert zu einem sehr realistisch wirkenden 3D-Modell führen können.

Jedem Punkt einer Punktwolke oder den Eckpunkten und Polygonen können Farbwerte zugeordnet werden. Texturen können abstrakt als eine auf das Modell aufgebrachte Tapete verstanden werden. Beispielsweise kann auf einen einfachen Zylinder ein Bild von Holzmaserung aufgebracht werden, um einen Baumstamm fotorealistisch darstellen zu können. Dabei wird jedem Punkt im 3D-Modell ein Punkt aus dem zweidimensionalen Bild zugewiesen.

Auch die Materialeigenschaften können in einem 3D-Modell modelliert werden, um dem Objekt die gewünschten Reflexionseigenschaften zuzuweisen, denn ein Holztisch weist bei Beleuchtung andere Eigenschaften auf als ein Glastisch. Hierfür werden in dem Modell verschiedene Parameter gespeichert, welche dif-

fuses Streulicht, spiegelndes Spekularlicht, Umgebungslicht, Transparenz, Lichtbrechung, Lichtemission etc. beschreiben.

Eine weitere Technik, um das Aussehen einer Oberfläche zu beeinflussen ist das Bumpmapping. Dabei werden mittels einer Textur die Normalenvektoren der einzelnen Polygone verändert, was dazu führt, dass die Schattierungen und Reflexionen auf der Oberfläche verändert werden und somit Oberflächenunebenheiten simuliert werden, die geometrisch eigentlich nicht vorhanden sind. Weitere Beispiele ähnliche Techniken sind Normal Maps oder Transparency Maps, welche den Schattenwurf, die Reflektion und die Transparenz eines Objektes beeinflussen können.

Die Oberflächeneigenschaften werden mit Shadern beim Rendern visuell umgesetzt. Shader (dt. Schattierer) sind Programme, die mittels verschiedener Algorithmen unter Berücksichtigung der Lichtquellen die Farbe für jeden darstellenden Pixel berechnen, um das Aussehen des 3D-Modells in der gerenderten Darstellung zu bestimmen. Ein Shader kann beispielsweise den Eindruck von gleichmäßigen Oberflächen erzeugen, wie in Abbildung 3.37 dargestellt.

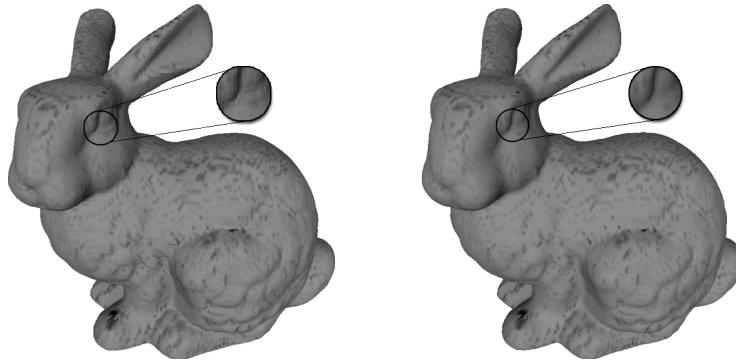


Abb. 3.37: Das linke Modell wurde ohne spezielle Shader gerendert, weshalb die einzelnen Polygone zu erkennen sind. Wird auf das gleiche Modell ein glättender Shader angewendet, entsteht der Eindruck von gleichmäßigen Oberflächen, wie im rechten Teil der Abbildung.

Belichtungs- und Kameraparameter Wie einem Nutzer ein 3D-Modell angezeigt wird, hängt von dem Szenenaufbau ab. Dazu gehören die Größe des Viewports, die Positionierung des 3D-Modells darin, die Position der Kamera und die Lichteinstellungen.

Der Viewport ist dabei vergleichbar mit einer Bühne, die den für die Darstellung zur Verfügung stehenden Bildausschnitt in seiner Höhe, Breite und Tiefe definiert. Für die Kamera muss nicht nur die Position, sondern auch die Blickrichtung gespeichert werden.

Werden 3D-Inhalte ohne Licht gerendert, wird ein schwarzes Bild erzeugt. Daher sind Lichtquellen notwendig, um die Szene auszuleuchten. Werden keine Angaben zu Anzahl, Position, Intensität und Art der Lichtquellen gespeichert, bleibt dies dem Nutzer überlassen oder sie werden gegebenenfalls automatisch vom Programm gesetzt.

Die Szene kann ein oder mehrere Modelle umfassen, welche auch gruppiert werden können. Gruppierungen sind insbesondere dann notwendig, wenn ein Modell aus mehreren Teilen besteht. Dabei muss außerdem die Lage der Teile zueinander gespeichert werden. Es handelt sich dabei um Transformationen, die Verschiebungen, Drehungen oder Skalierungen der Objekte betreffen.

Bei der Darstellung von umfangreichen Szenen können unterschiedliche Detailstufen für die einzelnen Objekte verwendet werden. Beispielsweise kann eine Mauer im Vordergrund, nahe der Kamera sehr detailliert dargestellt werden, während ein Baum weit im Hintergrund nur grob angedeutet werden muss, wobei der Detailgrad von 3D-Modellen dabei von der Anzahl der Polygone abhängt. Dieser Detailierungsgrad wird als „Level of Detail“ (LOD) bezeichnet und findet insbesondere dann Anwendung, wenn das Rendern von 3D-Modellen effizienter werden soll.

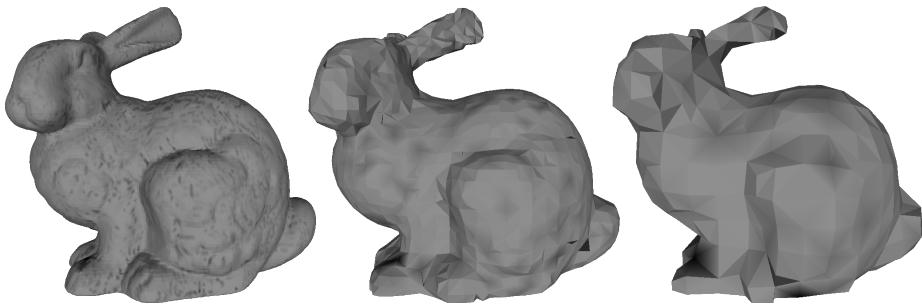


Abb. 3.38: Auswirkungen von verschiedenen Detaillierungsgraden (LOD) auf ein 3D-Modell. Das Stanford Bunny mit 69451, 3851 und 948 Polygonen.

Animation und Interaktion Für Animation und Interaktion ist die Speicherung von zusätzlichen Parametern erforderlich. Dies muss bei der Wahl eines Archivierungsformates berücksichtigt werden und erfordert zusätzliche Angaben in der Dokumentation. Werden Animationen beispielsweise als Videodateien exportiert, sollten die Hinweise in dem Abschnitt Video ab Seite 132 berücksichtigt werden.

Speicherung der verschiedenen Eigenschaften Nicht jedes 3D-Format unterstützt die Speicherung aller eben genannten Eigenschaften. Daher muss bei der Wahl des Speicherformates bedacht werden, welche Eigenschaften unbedingt erhalten bleiben sollen. Beispielsweise ist für ein gescanntes Objekt die Speicherung von Lichtquellen oder Animationen nicht unbedingt erforderlich, wohingegen bei einer Architekturrekonstruktion zur Darstellung von Lichtverhältnissen die Lichtquellen unerlässlich sind.

Eine Übersicht über die Speichereigenschaften der hier empfohlenen Formate zur Langzeitarchivierung gibt die folgende Tabelle.¹

Die Tabelle ist nicht erschöpfend, da es noch viele weitere spezielle Eigenschaften von 3D-Inhalten gibt, die aber oft nur in spezialisierten, häufig pro-

¹Nach K. McHenry – P. Bajcsy, An Overview of 3D Data Content, File Formats and Viewers (2008)

prietaryen Anwendungen und 3D-Formaten zum Tragen kommen. Daher ist außerdem zu empfehlen, dass die originalen Dateien ebenfalls archiviert werden, falls es sich dabei nicht schon um eines der hier angegebenen Formate handelt.

Format	Geometrie				Aussehen				Szene				Anim.
	DG	P	CSG	B-Rep	F	X	B	M	V	L	T	G	
X3D	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
COLL.	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
OBJ	✓	✓			✓	✓	✓	✓					✓
PLY	✓				✓	✓	✓	✓					
VRML	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
STL	✓				(✓)								
DXF	✓	✓	✓	✓	✓								✓
U3D	✓				✓	✓	✓		✓	✓	✓	✓	✓

Tabelle 3.11: Unterstützte Eigenschaften von verschiedenen 3D-Formaten. DG = Drahtgittermodell, P = Parametrisch, F = Farbe, X = Textur mittels Bild, B = Bumpmapping, M = Material, V = Viewport und Kamera, L = Lichtquellen, T = Transformationen, G = Gruppierung. Leere Zellen bedeuten entweder fehlende Speichereigenschaften oder es wurden keine Angaben in der jeweiligen Formatspezifikation gefunden.

Praxis

Im Umgang mit 3D-Inhalten müssen in der Praxis einige Details beachtet werden. Dazu gehören die Hinweise aus der Londoner Charta und dass bei der Konvertierung von einem Format in ein anderes Informationsverlust auftreten kann. Die Weitergabe von 3D-Modellen kann mittels verschiedenen Portalen oder auch mit 3D-PDF erfolgen, deren Erstellung in diesem Abschnitt erläutert wird.

London Charter Die Londoner Charta wurde 2006 ins Leben gerufen, um die digitale Visualisierung von Kulturgut so intellektuell und technisch riguros wie andere etablierte Forschungsmethoden zu machen. Die Londoner Charta strebt danach, strenge Regeln im Umgang mit computergestützten Visualisierungsmethoden zu definieren.

Es gibt sechs Leitsätze:

- Leitsatz 1 adressiert den Umgang mit der Londoner Charta.
- Leitsatz 2, Ziele und Methoden, regt dazu an, darüber nachzudenken, ob die computergestützte Visualisierungsmethode auch die geeignetste verfügbare Methode für den gewünschten Zweck ist.
- Leitsatz 3, Forschungsquellen, thematisiert die Auswahl und Dokumentation des Quellmaterials für eine Visualisierung.

- Leitsatz 4, Dokumentation, gibt wertvolle Hinweise darüber, was bei dem Entstehungsprozess des 3D-Inhaltes alles dokumentiert werden sollte, um das Verstehen und den Umgang mit den Inhalten zu erleichtern.
- Leitsatz 5, Nachhaltigkeit, behandelt die langfristige Zukunftsfähigkeit und vor allem die langfristige Benutzbarkeit von 3D-Inhalten.
- Leitsatz 6 thematisiert einen wichtigen Aspekt im Umgang mit Kulturgütern, nämlich den Zugang zum Kulturgut mittels computergestützten Visualisierungen. Auch der Zugang zu den Visualisierungen selbst wird besprochen.

Londoner Charta:

www.londoncharter.org

Deutsche Fassung der Londoner Charta:

www.londoncharter.org/fileadmin/templates/main/docs/london_charter_2_1_de.pdf

Konvertierung In dem Abschnitt Vertiefung wurden die Speichereigenschaften von verschiedenen 3D-Formaten besprochen und in der Tabelle auf Seite 168 zusammengefasst. Die Speichereigenschaften spielen insbesondere dann eine Rolle, wenn von einem Format in ein anderes konvertiert werden muss, um beispielsweise von einem proprietären Format verlustfrei zu einem langzeitarchivfähigen offenen Format zu gelangen.

Wenn die Software, mit der das 3D-Objekt erstellt wurde, einen Export in eines der hier empfohlenen Formate zulässt, so ist diese Variante vorzuziehen. In manchen Fällen kann es jedoch sein, dass das gewünschte Zielformat nicht unterstützt wird und zunächst ein Export in ein Zwischenformat vorgenommen werden muss, welches anschließend mit einem weiteren Programm in das Zielformat überführt werden kann. Soll beispielsweise ein 3D-Objekt aus *Blender* in ein 3D-PDF integriert werden, könnte dieses etwa im OBJ-Format gespeichert werden und anschließend mit *MeshLab* in U3D konvertiert werden. Mit der *ConversionSoftwareRegistry* können Programme gefunden werden, die von einem Format in ein anderes konvertieren können, wobei eventuelle Zwischenformate auch berücksichtigt werden.

Für die Konvertierung von umfangreichen 3D-Inhalten gibt es auch spezialisierte kommerzielle Programme, wie beispielsweise *Polytrans* und *NuGraph* der Firma Okino.

Blender:

www.blender.org

MeshLab:

<http://meshlab.sourceforge.net>

ConversionSoftwareRegistry:

<http://isda.ncsa.illinois.edu/NARA/CSR/php/search/conversions.php>

Okino (Polytrans/Nugraph):

www.okino.com

Weitergabe von 3D-Modellen Die Weitergabe von 3D-Modellen kann über das Internet erfolgen. Dazu gibt es eigene Dienste, wie beispielsweise Sketchfab, wo 3D-Modelle hochgeladen, veröffentlicht und visualisiert werden können. Zusätzlich kann der Inhalt von Sketchfab auch in andere Webseiten eingebettet werden.

Eine weniger kommerzielle Alternative stellt die Plattform 3DHOP (3D Heritage Online Presenter) dar, die von CNR-ISTI entwickelt wurde und frei zur Verfügung gestellt wird. Der Viewer verwendet ein eigenes Format (Nexus), wodurch auch große Modelle effizient dargestellt werden können.

Eine weitere Möglichkeit, um 3D-Modelle unkompliziert mit Dritten zu teilen, ist die Einbettung in ein 3D-PDF. Dabei können neben dem eigentlichen Modell zusätzliche Informationen in Form von Text, Bildern und Links eingefügt werden. Zusätzliche Programme zur Betrachtung des 3D-PDFs sind in der Regel nicht notwendig, da diese Funktion in Adobe Reader bereits integriert ist und dieser eine breite Anwenderbasis hat. Im Reader können Annotations- und Messwerkzeuge verwendet werden, um Strecken, Radien und Winkel an dem Modell auszumessen.

Ein 3D-PDF kann am komfortabelsten mit Adobe Acrobat Pro erstellt werden. Dabei muss das Modell bereits im U3D-Format vorliegen, welches über „Werkzeuge > Multimedia > 3D-Werkzeug“ in das zu erstellende PDF eingefügt werden kann. Die Firma tetra4D arbeitet eng mit Adobe zusammen und bietet neben dem Programm *3DPDF Converter* eine Reihe an Plugins für professionelle 3D-Software wie AutoCAD oder Maya an. Eine günstigere Alternative bieten die Produkte von *PDF3D*. Freie Alternativen gibt zurzeit es wenige und auch nur mit eingeschränktem Umfang. Ein Beispiel wäre *CutePDF*.

Problematisch ist dabei aber, dass 3D-PDFs aktuell nur mit dem Adobe Reader betrachtet werden können und das Programm aktuell nur für Windows, Mac und Android angeboten wird. Alternativen für Linux fehlen im Moment noch. Bedacht werden muss auch, dass große Modelle viel Arbeitsspeicher benötigen, der nicht auf jedem System vorhanden ist.

Sketchfab:

<https://sketchfab.com/>

3DHOP:

<http://vcg.isti.cnr.it/3dhop/>

3D-Modelle in Adobe Reader ansehen:

<https://helpx.adobe.com/de/acrobat/using/displaying-3d-models-pdfs.html>

3D-PDFs mit Acrobat erstellen:

<https://helpx.adobe.com/de/acrobat/using/adding-3d-models-pdfs-acrobat.html>

3D-PDFs für Europeana erstellen:

www.carare.eu/rum/Media/Files/3D-Training-Materials

tetra4D:

www.tetra4d.com

PDF3D:

www.pdf3d.com

CutePDF:

www.cutepdf.com

Quellen

- S. Boeykens – E. Bogani, Metadata for 3D models. How to Search in 3D Model Repositories? (2008)
<https://lirias.kuleuven.be/bitstream/123456789/202356/1/boeykens-bogani.pdf>
- K. McHenry – P. Bajcsy, An Overview of 3D Data Content, File Formats and Viewers (2008)
<http://isda.ncsa.illinois.edu/drupal/sites/default/files/NCSA-ISDA-2008-002.pdf>
- K. McHenry – P. Bajcsy, 3D+Time File Formats. Technical Report NCSA-ISDA10-001 (2010)
<http://isda.ncsa.uiuc.edu/peter/publications/techreports/2010/NCSA-ISDA-2010-001.pdf>
- D. Pletinckx – D. Haskiya, D5.1 Functional Specification of Requirements for Preparing 3D/VR for Europeana (2011)
http://www.carare.eu/eng/content/download/2026/15338/file/D5_1_Functional_specification_3DVR_final.pdf
- J. Slick, 3D Basics – Getting Started in 3D
<http://3d.about.com/od/3d-101-The-Basics/u/3d-Basics-Getting-Started-In-3d.htm>
- M. Trognitz – K. Niven – V. Gilissen, 3D Datasets in Archaeology: A Guide to Good Practice, in: ADS (Hrsg.) Guides to Good Practice (2016)
http://guides.archaeologydataservice.ac.uk/g2gp/3d_Toc
- 3D ICONS Projekt:
www.3dicons-project.eu
- Stanford Bunny aus dem Stanford Computer Graphics Laboratory:
<http://graphics.stanford.edu/data/3Dscanrep/>

Formatspezifikationen

X3D und VRML:

www.web3d.org

COLLADA:

www.khronos.org/collada

OBJ:

http://en.wikipedia.org/wiki/Wavefront_.obj_file

PLY:

<http://paulbourke.net/dataformats/ply/>

U3D:

www.ecma-international.org/publications/standards/Ecma-363.htm

STL:

www.ennex.com/~fabbers/StL.asp

DXF:

<http://usa.autodesk.com/adsk/servlet/index?siteID=123112&id=24240325>

DXF 2010:

http://images.autodesk.com/adsk/files/acad_dxif1.pdf

Tools und Programme

Londoner Charta:

www.londoncharter.org

Deutsche Fassung der Londoner Charta:

www.londoncharter.org/fileadmin/templates/main/docs/london_charter_2_1_de.pdf

Blender:

www.blender.org

MeshLab:

<http://meshlab.sourceforge.net>

ConversionSoftwareRegistry:

<http://isda.ncsa.illinois.edu/NARA/CSR/php/search/conversions.php>

Okino (Polytrans/Nugraph):

www.okino.com

Sketchfab:

<https://sketchfab.com/>

3DHOP:

<http://vcg.isti.cnr.it/3dhop/>

3D-Modelle in Adobe Reader:

<https://helpx.adobe.com/de/acrobat/using/displaying-3d-models-pdfs.html>

3D-Modelle erstellen mit Acrobat:

<https://helpx.adobe.com/de/acrobat/using/adding-3d-models-pdfs-acrobat.html>

3D-PDFs für Europeana erstellen:

www.carare.eu/rum/Media/Files/3D-Training-Materials

tetra4D:

www.tetra4d.com

PDF3D:

www.pdf3d.com

CutePDF:

www.cutepdf.com

3.11 Webseiten

D. Hagmann

Eine Webseite stellt eine Ressource aus strukturiertem Text im World Wide Web (WWW) dar und besteht in ihrer einfachsten Form aus einer HTML-Datei. Sie kann via Hyperlinks mit beliebig vielen weiteren Ressourcen vernetzt sein. Webseiten sind ein integraler Bestandteil des WWW im Internet. In der Regel ist eine Webseite Teil einer Website, bzw. eines Webauftrittes, also eines zusammengehörenden Paketes von miteinander vernetzten Webseiten und weiteren Ressourcen.

In der altertumswissenschaftlichen Forschung können Webseiten für die Öffentlichkeit zugängliche Informationen beinhalten, wie etwa Blogbeiträge oder ausführliche Projektbeschreibungen.

Der vorliegende Artikel beschäftigt sich vornehmlich mit der Archivierung einzelner Webseiten und nicht mit der Archivierung ganzer Websitesysteme. Um umfangreiche Websites mit mehreren Webseiten zu archivieren, empfehlen sich Online-Speicherdienste, spezialisierte Internetarchive oder dezidierte Softwarelösungen.

Langzeitformate Webseiten können archiviert werden, wenn die nötigen Nutzungsrechte der Inhalte vorliegen. Die Archivierung kann dabei auf verschiedene Arten erfolgen:

- Speicherung als statisches PDF mit zugehörigen Ressourcen
- lokale Speicherung als HTML oder MHTML mit zugehörigen Ressourcen
- Speicherung bei einem spezialisierten Internetarchiv

Als Optimum ist die Langzeitarchivierung einer Webseite in einer Form, die möglichst wenig Informationsverlust garantiert und einfach umzusetzen ist, anzustreben. Webseiten bestehen zum einen aus mindestens einer strukturierten HTML-Datei, zum anderen aus beliebig vielen via Hyperlinks mit der HTML-Datei verbundenen Ressourcen, die teilweise auf demselben Webserver gespeichert sind, aber auch von jedem anderen Ort im Internet bezogen werden können. Dies stellt jedoch nur den theoretisch sehr einfachen Aufbau dar, praktisch bestehen Webseiten aus einer Vielzahl an weiteren strukturierten Textdateien, die etwa das Design der Website regeln (CSS-Dateien) und können zudem über verschiedenste von anderen Websites bezogene und auf der Webseite eingebettete Inhalte (Videos, 3D-Modelle, interaktive Karten etc.) verfügen. Eine der Hauptintentionen jedweder Webseite ist es verschiedene Informationen nach einem vorgegebenen Design in einer bestimmten Abfolge und einem bestimmten Layout dem Nutzer zu vermitteln, vergleichbar zu einer gedruckten Seite in einem Buch. Bei der Archivierung muss beachtet werden, dass für Webseiten teilweise Dateiformate verwendet werden, die für die Langzeitarchivierung dezidiert ungeeignet sind, etwa JPEGs.

Generell empfiehlt es sich, die der Webseite zugrundeliegenden Daten (z.B. Text und Bilder) als Einzeldateien jeweils separat in einem geeigneten Archivformat zu archivieren. Nähere Informationen zu den Archivierungsformaten sind in den entsprechenden Kapiteln zu finden. Auch werden nicht alle multimedialen

Inhalte, Webanwendungen (z.B. Web-GIS) oder über externe Dienste eingebettete Inhalte mit jeder Archivierungsmethode gespeichert, weshalb in solchen Fällen besonderes Augenmerk auf die Auswahl der zu verwendenden Methode gelegt werden muss.

Eine Webseite kann als PDF mit Hilfe des Webbrowsers und eines PDF-Generators als PDF-Datei gespeichert und anschließend mit entsprechender Software in ein archivierbares PDF/A-Datei konvertiert werden. Informationen dazu finden sich im Abschnitt PDF-Dokumente. Diese Methode führt praktisch immer zu Änderungen des ursprünglichen Layouts. Jedoch können mittels Plug-ins im Webbrowser oder bestimmten Softwareprogrammen Webseiten unter großteiliger Wahrung des Layouts als PDF gespeichert werden. Abschließend muss auch hier eine Konvertierung in das PDF/A-Format vorgenommen werden. Multimediale Inhalte (Videos, 3D-Objekte etc.) werden mit dieser Methode nicht gespeichert.

Die lokale Speicherung einer Webseite aus dem WWW mittels eines Webbrowsers ist einfach möglich und wird durch alle aktuellen Webbrowser unterstützt. Hierbei gilt es jedoch, bestimmte Speicherformate zu beachten, da nicht alle in den Webbrowsers verfügbaren Formate für die Archivierung geeignet sind. Für die Strukturierung und Formatierung von Webseiten werden üblicherweise die Hypertext Markup Language (HTML) oder die Extensible Hypertext Markup Language (XHTML), sowie Cascading Style Sheets (CSS) verwendet. Es handelt sich dabei um Standards, die vom W3C entwickelt und empfohlen werden, weshalb diese in den Versionen HTML5, XHTML5 und CSS 3 für die Archivierung empfohlen werden können.

Es bietet sich hier also die Möglichkeit der Speicherung der Webseite als HTML- oder XHTML-Datei an. HTML-Dateien (und XHTML) archivieren den strukturierten Text und die Hyperlinks, jedoch nicht die verknüpften Ressourcen (z.B. Bilder, multimediale Inhalte oder externe Inhalte), zudem wird hierdurch nicht das Design der Webseite, welches durch CSS geregelt wird, übernommen, da die entsprechenden Dateien nicht gespeichert werden. Um auch die verknüpften und für das Design benötigte Ressourcen zu speichern, können diese automatisch in einen zusätzlichen lokalen Ordner geladen werden. In der Regel handelt es sich dabei um HTML/XHTML- und CSS-Dateien, Grafiken, JavaScript-Dateien sowie gegebenenfalls Java Applets und Multimedia-Dateien.

Die lokale Speicherung einer Webseite in einer einzigen Datei wird mittels MIME HTML (MHTML) ermöglicht. Es handelt sich um ein textbasiertes Format, das in RFC 2557 spezifiziert wird. In der Regel werden bei MHTML-Dateien das Layout und alle Hyperlinks vollständig übernommen. Auch hier muss das Speichern von eingebetteten Inhalten gegebenenfalls gesondert vorgenommen werden.

Das offen dokumentierte Mozilla Archive Format (MAFF) ermöglicht ebenfalls die Speicherung einer Webseite in Form einer einzelnen Datei. Dabei werden die einzelnen Bestandteile in einem ZIP-Container gespeichert. Da dieses Format aktuell nur von Mozilla Firefox unterstützt wird, sollte für die Archivierung jedoch ein anderes Format vorgezogen werden. Ähnlich verhält es sich mit dem Format Webarchive, das derzeit jedoch nur durch den Safari Webbrowser unterstützt wird und daher nicht empfohlen werden kann.

Auch HTML-Dateien mit Data-URIs ermöglichen die Speicherung einer gesamten Webseite meist unter Beibehaltung des Layouts in einer einzigen Datei. Data-URIs ermöglichen es, Ressourcen in HTML einzubetten und sind in RFC

2397 definiert. Es handelt sich dabei um eine spezielle Syntax, mit der binäre Daten als ASCII-Zeichenketten kodiert werden. Da Ressourcen als Data-URIs, wie beispielsweise Bilder, direkt und in menschenunlesbarer Form in die Datei integriert werden, können diese nicht nachgenutzt werden, weshalb von einer Speicherung als HTML mit Data-URIs für die Archivierung abgesehen werden sollte.

Eine andere häufig praktizierte, jedoch eindeutig nicht empfohlene Möglichkeit, stellt die Speicherung von Webseiten in der Form von Screenshots dar. Screenshots werden in der Regel im PNG- oder JPEG-Format gespeichert. Dies hat drei Nachteile: (1) Die Konvertierung erfolgt oft in das JPEG-Format, das zur Archivierung nicht geeignet ist. (2) Die Speicherung als Rastergrafik kann in manchen Fällen aufgrund einer zu niedrigen Auflösung zu Qualitätsverlusten führen. Außerdem wird Text nicht mehr als solcher erkannt und gespeichert. (3) Die Konvertierung der Webseite in eine Grafik führt dazu, dass sämtliche Hyperlinks desintegriert werden.

Es besteht zwar hinsichtlich der Punkte (1) und (2) die Möglichkeit, mit entsprechender Software eine Texterkennung und anschließende Speicherung als PDF/A durchzuführen, jedoch können hinsichtlich Punkt (3) dadurch keine Hyperlinks wiederhergestellt werden.

Ein anderer Ansatz ist die Archivierung einer Webseite über einen spezialisierten Archivierungsdienst. Solche werden etwa durch die Bayerische Staatsbibliothek (mit Anmeldung) oder die Organisation Internet Archive angeboten. Hier werden die Webseiten auf einem Server des Archivierungsdienstes gespeichert und können auf diesen Plattformen wiederum über das WWW abgerufen werden. Diese Dienste sind auch zur Archivierung ganzer Websites geeignet. Für die Archivierung ganzer Websites gibt es das Format Web ARChive (WARC), das seit 2009 als ISO 28500 standardisiert ist und von dem International Internet Preservation Consortium aufbauend auf dem Format ARC entwickelt wurde. In einer WARC-Datei werden alle Seiten, Ressourcen und weitere Komponenten einer Website gespeichert.

Hinweis: Angaben zur Archivierung von Programmen in JavaScript sowie Java (Java-Applets) finden sich im Kapitel Eigene Programme und Skripte, Ausführungen zu multimedialen Inhalten (z.B. 3D-Objekte, Audio oder Video) in den entsprechenden Kapiteln.

Format	Begründung
✓ PDF/A-1, PDF/A-2	PDF/A ist gezielt als stabiles, offenes und standardisiertes Format für die Langzeitarchivierung unterschiedlicher Ausgangsdateien entwickelt worden.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Format	Begründung
HTML und XHTML	HTML- und XHTML-Dateien können als streng strukturierte Textdokumente, die vom W3C standardisiert sind, problemlos archiviert werden. Die Datei sollte wohlgeformt und in UTF-8 ohne BOM kodiert sein. Es sollte möglichst HTML5 verwendet werden. Zusätzliche Dateien, wie CSS, JavaScript oder andere strukturierter Textformate müssen ebenfalls archiviert werden. Eingebettete Ressourcen müssen gesondert archiviert werden.
MHTML	MHTML-Dateien können als strukturierte Textdokumente mit genauen Spezifikationen für die Archivierung verwendet werden. Die Archivierung von eingebetteten Inhalten muss gegebenenfalls gesondert erfolgen.
WARC	Web ARCHive ist als ISO 28500 standardisiert und dient als Containerformat für mehrere Webseiten einer Website.
~ MAFF	Das Format ermöglicht die Speicherung einer ganzen Webseite samt aller zugehöriger Ressourcen komprimiert und verlustfrei in einem ZIP Container und eignet sich zur Langzeitarchivierung, solange die einzelnen Ressourcen selbst in archivfähigen Formaten vorliegen und Hyperlinks entsprechend aktualisiert werden.
HTML mit Data URIs	HTML-Dateien können als strukturierte Textdokumente, die weit verbreiteten Konventionen folgen und aufgrund der integrierten DTD, die die verwendete Struktur beschreibt, problemlos archiviert werden. Data URIs sind ebenso spezifiziert.
✗ andere PDF-Varianten	Viele gängige PDF-Varianten sind nicht für die Langzeitarchivierung geeignet. Stattdessen sollten entweder die Ausgangsdateien in einem passenden Format archiviert oder eine Migration in ein PDF/A-Format vorgenommen werden.
Screenshots	Screenshots eignen sich nur für die Dokumentation der Optik der Webseite, jedoch nicht für die Archivierung der Inhalte, da diese als Rastergrafik gespeichert werden und so kaum nutzbar sind.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Format	Begründung
Webarchive	Ist ein Format von Apple, das derzeit auch nur von Safari unterstützt wird. Es ist nicht für die Archivierung geeignet.

Dokumentation HTML, XHTML und MHTML verfügen über einen eigenen Dokumentenkopf, in dem verschiedene Metadaten eingebettet werden können. Es sollten Angaben zur verwendeten Zeichenkodierung, dem Titel des Dokumentes, dem/der AutorIn sowie Stichwörter gemacht werden. Ergänzende Metadaten können zusätzlich mit Hilfe eines Kommentars in den Kopfdaten der Datei eingefügt werden. Auch in CSS-Dateien können Metadaten als Kommentar eingetragen werden.

Die hier angegebenen Metadaten sind als minimale Angabe zu betrachten und ergänzen die angegebenen Metadaten für Projekte und Einzeldateien in dem Abschnitt Metadaten in der Anwendung.

Metadatum	Beschreibung
Titel	Titel der Webseite
Kurzbeschreibung	Kurze Beschreibung des Inhaltes.
Stichwörter	Schlagworte, die den Inhalt beschreiben.
Autor	Name des Verfassers oder Erstellers der Datei.
Erstellungsdatum	Datum der Erstellung der Datei, also der Archivierung der Webseite.
Bearbeitungsdatum	Datum der letzten Bearbeitung der Webseite.
Abschaltung Webserver	Datum an dem die Webseite zum letzten Mal online verfügbar war
URI	Internetadresse der archivierten Webseite
Identifikator	Wenn das Dokument bereits veröffentlicht wurde und einen Persistent Identifier erhalten hat, sollte dieser angegeben werden.
Sprache	Angabe der im Dokument verwendeten Sprachen. Sprachkennungen nach ISO639 angeben.
Rechte	Details zum Urheberrecht.
Standard	Name und Version des verwendeten Standards, z.B. HTML5 und CSS 3.
Zeichenkodierung	Angabe der verwendeten Zeichenkodierung, z.B. UTF-8 ohne BOM.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Metadatum	Beschreibung
Beziehungen	Dateien oder Ressourcen, die mit der Datei zusammenhängen, wozu auch frühere Versionen gehören. Bei der Archivierung einer Website mit mehreren Webseiten müssen die Beziehungen der einzelnen Seiten untereinander dokumentiert werden, beispielsweise mit einer Sitemap.
Versionsnummer	Angabe der Dateiversion, bezogen auf den Inhalt. z.B. 1.3.
Software	Name und Version der für die Archivierung der Seite verwendeten Programme
weitere Dateien	Liste von eingebetteten Medien, die zusätzlich separat gespeichert wurden. Liegt eine Dokumentationsdatei für das Dokument vor, muss diese ebenfalls genannt werden.

Vertiefung

Das Internet verbindet weltweit verschiedene Computer-Netzwerke. Systeme wie das WWW als Internetdienst ermöglichen dabei den Austausch der Daten. Das WWW selbst ist wiederum ein System aus Hypertexten und Hypermedia im Internet, das via HTTP als Protokoll kommuniziert. Eine Webseite ist in der Regel eine HTML-Datei mit Verweisen, die als Hyperlinks entweder auf andere Stellen im selben Dokument oder auf beliebig viele weitere Ressourcen im WWW verweist. Webseiten bilden die einzelnen Bestandteile einer Website und können sowohl als Homepage (resp. Startseite) als auch in Form jeder beliebigen anderen Seite innerhalb einer Website enthalten sein. Webseiten als HTML-Dokumente können nicht ausschließlich nur im WWW auftreten, sondern etwa auch lokal vorliegen.

Grundsätzliche Funktionsweise Das generelle Vorgehen zum Abrufen einer Webseite in Form einer HTML-Datei kann schematisiert wie folgt beschrieben werden: Nach dem Öffnen des Webbrowsers (1) erfolgt die Eingabe der URI (z.B. <http://www.ianus-fdz.de/>) bzw. der Klick auf einen Hyperlink (2). Dies bewirkt, dass der Webbrowser anhand der in der URI enthaltenen Domain (z.B. www.ianus-fdz.de) die Internet Protocol (IP) Adresse von einem Domain Name System (DNS) Server bezieht (3); den Datenaustausch regelt das Transmission Control Protocol (TCP). Jedes Gerät, mit dem eine Internetverbindung hergestellt werden kann, verfügt über eine IP-Adresse, die eine eindeutige Identifikation des Gerätes ermöglicht. Anhand der IP-Adresse fragt nun der Webbrowser als Web-Client um das gewünschte HTML-Dokument beim entsprechenden Web-Server, auf dem das Dokument gespeichert ist, an (4). Der Web-Server stellt nun das angeforderte HTML-Dokument zur Verfügung (5), welches vom Webbrowser empfangen (6) und dargestellt wird (7). Siehe dazu auch Abbildung 3.39.

Dies ist nur ein grundsätzlicher Überblick hinsichtlich der wichtigsten Elemente. Ausführlichere Informationen finden sich auf <https://wiki.selfhtml.org/wiki/Grundlagen>. Zur Geschichte des Internets und WWW existieren zahlreiche Abhandlungen, siehe hierfür etwa detailliert https://wiki.selfhtml.org/wiki/Grundlagen/Entstehung_des_Internet.

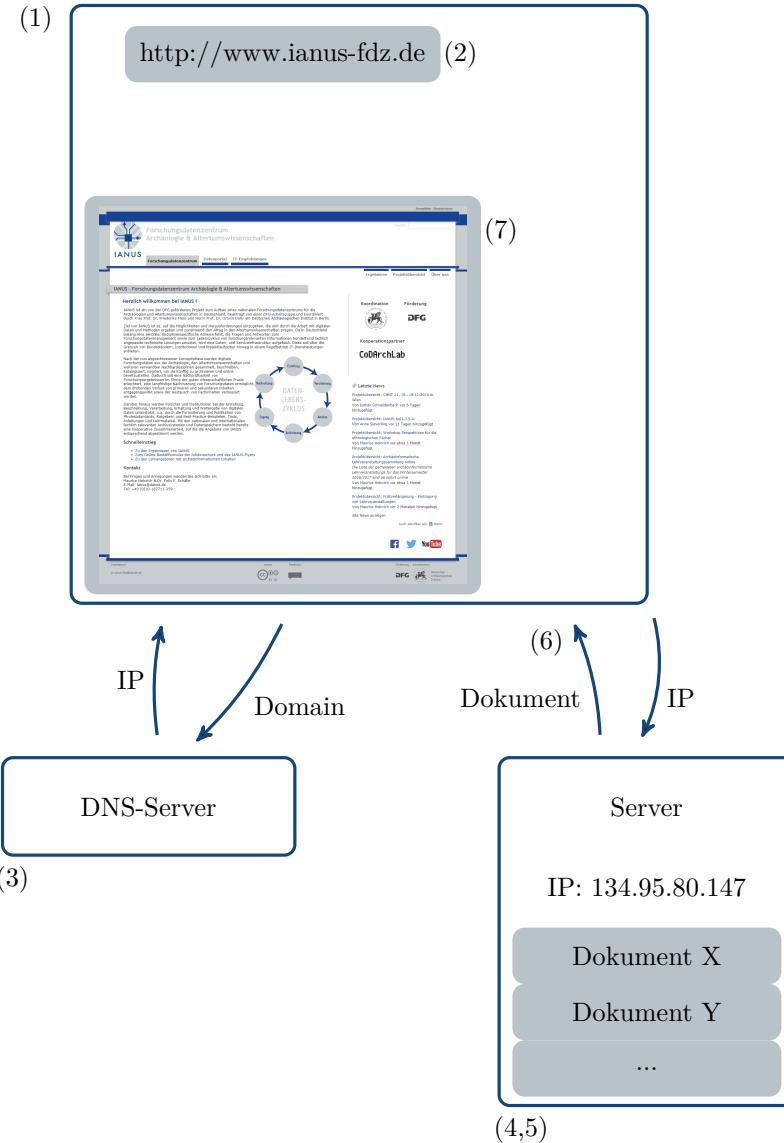


Abb. 3.39: In dem Webbrower (1) erfolgt die Eingabe der URI oder der Klick auf einen Hyperlink (2). Der Webbrower bezieht die IP-Adresse vom DNS-Server (3). Mit der IP-Adresse wird beim entsprechenden Web-Server das gewünschte Dokument angefragt und zur Verfügung gestellt (4,5). Das angeforderte Dokument wird vom Brower empfangen (6) und dargestellt (7).

Webbrowser Ein Webbrowser ist ein Computerprogramm zum Abrufen sowie Darstellen von Ressourcen (HTML-Dokumente, multimediale Inhalte, ganze Webanwendungen etc.) und ist die Schnittstelle zwischen dem Nutzer und dem WWW. Er ermöglicht das sequenzielle Abrufen und Betrachten von Webseiten unter Verwendung von Hyperlinks im WWW (surfen), wobei es keine Rolle spielt, ob eine anzuseigende Ressource über das WWW oder lokal zur Verfügung gestellt wird. Für die Darstellung der Ressourcen können Plugins herangezogen werden. Ein Webbrowser ermöglicht zudem die Speicherung von Dateien und Programmen aus dem Internet auf dem Computer. Heutige Webbrowser unterstützen die Anzeige mehrerer Fenster gleichzeitig in Form von Tabs (Registerkarten, Reiter). Beliebte aktuelle Webbrowser sind Google Chrome, Mozilla Firefox, Microsoft Internet Explorer, Apple Safari und Opera: <http://www.w3schools.com/browsers/>

URIs Uniform Resource Identifiers (URI, einheitlicher Bezeichner für Ressourcen) ermöglichen die Identifikation einer Ressource (z. B. einer Webseite oder eines PDF-Dokumentes) im Internet. Im Bereich des WWW treten URIs vor allem als Uniform Resource Locators (URL, einheitlicher Quellenanzeiger, also die eigentlichen „Internetadressen“, z. B. <http://www.ianus-fdz.de/>) und Uniform Resource Names (URN, einheitlicher Name für Ressourcen) auf: URLs definieren dabei im WWW den Ort einer Ressource, URNs benennen die Ressource selbst. Eine URN kann mit einer oder mehreren URLs verknüpft sein, etwa wenn dieselbe Ressource an verschiedenen Speicherorten verfügbar ist. Unter anderem verwenden Nationalbibliotheken URNs als Persistente Identifikatoren zur Kennzeichnung von Onlinepublikationen: Die Deutsche Nationalbibliothek hat beispielsweise das Dokument „Policy für die Vergabe von URNs im Namensraum urn:nbn:de“ mit der URN <urn:nbn:de:101-2012121200> ausgestattet. Man benötigt für gewöhnlich einen URN-Resolver, der die zum URN gehörige(n) URL(s) anzeigt, um an die gewünschte Ressource zu gelangen. Die Deutsche Nationalbibliothek bietet einen solchen unter der URI <http://nbn-resolving.org/> an, wo für die URN <urn:nbn:de:101-2012121200> (= „Name“) die URL <http://d-nb.info/1029114455/34> (= „Adresse“) angegeben wird. Detaillierte Informationen zu URIs, URLs, URNs und persistenten Identifikatoren sind in dem „Abschlussbericht Testbed ‘Persistent Identifiers’“² zu finden.

Webseiten Unter der Verwendung von HTML, CSS, JavaScript sowie weiteren Ressourcen, z. B. Rastergrafiken, Vektorgrafiken, Video- oder Audiodateien werden Webseiten erstellt. Die einzelnen Ressourcen können in verschiedenen Ordnern oder sogar an verschiedenen Orten liegen. Eine Webseite besteht hierbei aus drei grundsätzlichen Schichten: der Struktur, dem Layout und dem Verhalten. Die Struktur bzw. der Aufbau wird durch HTML organisiert, CSS definiert das Layout der Webseite und JavaScript bestimmt, wie sich die Webseite bei Interaktionen des Nutzers verhält. Weitere Informationen dazu finden sich auf https://wiki.selfhtml.org/wiki/HTML/Tutorials/Trennung_von_Inhalt,_Präsentation_und_Verhalten

²<http://www.ianus-fdz.de/attachments/download/560/Testbed-Persistent%20Identifiers.pdf>

HTML Bei Hypertext Markup Language (HTML) handelt es sich um eine Auszeichnungssprache, durch die Dokumente strukturiert beschrieben werden können. HTML kann zur Erstellung von Webseiten aber auch zur Erstellung von lokalen Dokumenten verwendet, ausgedruckt oder mit Hilfe synthetischer Stimmen barrierefrei für Menschen mit Sehbeeinträchtigungen auf Audio-Systemen ausgegeben werden. Webbrower visualisieren die Auszeichnungsbefehle unter eventueller Berücksichtigung von CSS-Dateien und machen so das Dokument menschenlesbar. Ein Kernelement von HTML ist die Verfügbarkeit von Verweisen in Form von Hyperlinks, durch die andere Stellen im selben Dokument, aber auch andere Ressourcen im WWW und im Internet aufgerufen werden können.

Der aktuelle Standard für HTML-Dateien ist HTML5 und wird mit einer Dokumenttypdeklaration am Beginn der HTML-Datei angegeben. Da HTML im ASCII-Zeichensatz verfasst wird, eignen sich Text-Editoren bzw. spezialisierte HTML-Editoren zur Bearbeitung. HTML5 verwendet als Standardzeichensatz UTF-8, was auch beibehalten werden sollte.

Das Grundgerüst einer HTML Datei besteht aus:

- Der Dokumenttypdeklaration in der Form <!DOCTYPE html> für HTML5.
- Dem HTML-Wurzelement <html>, welches den Inhalt der HTML-Datei umklammert.
- Dem Kopfelement <head>, welches die Kopfdaten (z.B. das verpflichtende Titellelement) beinhaltet. Auch ist hier der geeignete Ort für allgemeine Kommentare zum HTML Dokument (z.B. Metadaten). Die Kopfdaten werden im Browser nicht angezeigt.
- Dem Titellelement <title> als Teil des <head>, welches verpflichtend anzugeben ist.
- Dem Körperelement <body>, das den anzuseigenden Inhalt enthält.

Dieser Aufbau gestaltet sich in HTML folgendermaßen:

```
<!DOCTYPE html>
<html>
  <head>
    <title></title>
    <!-- Kommentare -->
  </head>
  <body>
    anzuseigender Inhalt
  </body>
</html>
```

Bei HTML-Dokumenten können zusätzliche Informationen wie z.B. Metadaten als Kommentare an beliebiger Stelle eingegeben werden, wie etwa der <head>-Bereich. Kommentare werden durch die Zeichenfolge <!-- eingeleitet und durch die Zeichenfolge --> abgeschlossen. Sie werden von einem Webbrower generell nicht angezeigt, können jedoch mittels eines Texteditors dargestellt werden.

Ein kurzes Beispieldokument mit rudimentären Metadaten in HTML, das in einem Webbrower den Text „Testüberschrift: Dies ist ein einfaches Beispiel“ anzeigt, könnte wie folgt aussehen:

```

<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <meta name="description" content="Ein Beispieldokument">
    <meta name="keywords" content="example, html, dai, ianus">
    <meta name="author" content ="Dominik Hagmann">
    <title>Beispieldokument</title>
    <!-- Weitere Metadaten: Erstellt: 11.2016, Lizenz: CC-BY -->
  </head>
  <body>
    <h1>Testüberschrift: </h1>
    <p>Dies ist ein einfaches Beispiel.</p>
  </body>
</html>

```

CSS Cascading Style Sheets (CSS) dienen zur Formatierung von HTML- und XML-Dokumenten. HTML-Dokumente besitzen einige vom jeweiligen Browser vorgegebene Formatierungen, etwa hinsichtlich der Gestaltung der Überschriften oder der Hyperlinks. Mittels CSS können deutlich umfangreichere Designs erzeugt werden; CSS-Dateien, die die Formatierung regeln, sind für professionelles Web-Design von höchster Bedeutung. CSS ermöglicht die Formatierung aller HTML-Elemente sowie zahlreicher weiterer Bestandteile, die nicht in HTML enthalten sind. Wahlweise kann mittels CSS global das gesamte Design auf einmal bestimmt oder aber auch die Formatierung einzelner HTML-Objekte individuell definiert werden. Für CSS eignen sich dieselben Editoren wie für HTML.

Durch einen Selektor wird in der CSS-Datei ein Element angewählt und der Wert der Eigenschaft definiert. Dies geschieht in der Form „Selektor {Eigenschaft : Wert; }“. Folgendes Beispiel färbt alle Buchstaben eines Absatzes blau ein:

```

|| p {color: blue; }

```

JavaScript Bei JavaScript (nicht zu verwechseln mit den Programmiersprachen Java und JScript) handelt es sich um eine Implementation der Skriptsprache ECMAScript und ist die am weitesten verbreitete Programmiersprache für Webseiten. Mittels JavaScript können Webseiten um Zusatzfunktionen ergänzt werden. Dabei kann der JavaScript Code abhängig von dessen Einsatzzweck direkt in der HTML-Datei im Head oder Body eingebettet oder in einer separaten Datei mit der Endung .js vorhanden sein. Allgemein empfiehlt es sich, JavaScript extern als js-Datei zu speichern und in dem HTML-Dokument mittels einer Skriptreferenz darauf zu referenzieren.

JavaScript stattet Webseiten mit Elementen aus, durch die der User mit der Webseite interagieren kann. Ein Beispiel für eine solche Interaktion auf einer Webseite ist, per Mausklick auf einen Button den Inhalt von Textabschnitten zu ändern: http://www.w3schools.com/js/tryit.asp?filename=tryjs_whereto_head. In diesem Beispiel ist das Skript in HTML eingebettet:

```

<!DOCTYPE html>
<html>
  <head>
    <script>
      function myFunction() {
        document.getElementById("Beispiel").innerHTML=
        "... der nun verändert wurde ;-)";
      }
    </script>
  </head>
  <body>
    <h1>Beispiel zu JavaScript</h1>
    <p id="Beispiel">Dies ist der ursprüngliche Text...</p>
    <button type="button" onclick="myFunction()">
      Bitte drücken
    </button>
  </body>
</html>

```

Ist der Code in einer separaten Datei gespeichert, benötigt man eine Skriptreferenz:

```
<script src="Beispielskript.js"></script>
```

Aufgrund des Umfangs wird nicht weiter auf JavaScript im Speziellen eingegangen. Weiterführende Informationen finden sich beispielsweise auf <https://wiki.selfhtml.org/wiki/JavaScript/Tutorials/Einf%C3%BChrung>

Dynamische Websites Wenn die angezeigten Inhalte einer Webseite aus einer zugrunde liegenden Datenbank stammen, handelt es sich um eine dynamische Webseite. Bei dem Besuch einer dynamischen Webseite wird also nicht ein fertiges statisches HTML-Dokument abgerufen, sondern ein anzuseigendes HTML-Dokument wird ad-hoc und individuell aus den Einträgen in der Datenbank generiert. Mit Hilfe von Content-Management-Systemen (CMS), wie beispielsweise Joomla, Drupal, Typo3 oder WordPress, können die Inhalte solcher Websites auch ohne HTML-Kenntnisse verwaltet und gepflegt werden.

Für die Archivierung von vollständigen CM-Systemen gelten umfangreichere Anforderungen, als bei der Archivierung einzelner, zum Zeitpunkt des Aufrufes, statischer Webseiten daraus. Es muss beispielsweise entschieden werden, ob der gesamte Funktionsumfang der Website oder lediglich die darin enthaltenen Informationen gesichert werden sollen.

Praxis

In diesem Abschnitt werden Programme und Editoren vorgestellt, um eine Webseite zu bearbeiten. Neben einer allgemeinen Übersicht der Archivierungsmöglichkeiten, wird auf die verschiedene Möglichkeiten eingegangen, um Webseiten als PDF, MHTML, MAFF oder HTML mit Data-URIs zu archivieren. Auch Archivierungsmöglichkeiten für gesamte Websites werden vorgestellt, sowie einige Hinweise zu Screenshots gegeben. Zahlreiche Hinweise finden sich auch im Praxisabschnitt des Kapitels Textdokumente ab Seite 65.

Editoren Um HTML, CSS und JavaScript Dateien von Webseiten zu bearbeiten, wird ein Texteditor benötigt. Eine ausführliche Übersicht bietet der Abschnitt Praxis im Kapitel Textdokumente. Zusätzlich zu den dort besprochenen Editoren existieren spezielle Editoren für das Webdesign, unter Umständen mit WYSIWYG-Modi, bzw. mächtige Code-Editoren, die umfangreiche Möglichkeiten zur Codeerstellung bieten, jedoch oftmals auch das nötige Hintergrundwissen hinsichtlich der verwendeten Sprache und deren Syntax verlangen. Ein Beispiel für einen WYSIWYG-Editor ist Adobe Muse CC, Beispiele für professionelle Code-Editoren sind Adobe Dreamweaver CC und Sublime Text von Jon Skinner wobei es sich jeweils um proprietäre Softwarelösungen handelt. Die Systeme von Adobe sind für Windows und Mac OS verfügbar, Sublime Text zusätzlich für Linux.

Frei verfügbare Alternativen stellen Aptana Studio, SeaMonkey und BlueGriffon dar, die alle für Windows, Mac OS und Linux verfügbar sind. Aptana Studio bietet Werkzeuge zur Erstellung von HTML, CSS und JavaScript. SeaMonkey und BlueGriffon haben einen WYSIWYG-Editor.

Adobe Muse CC:

<https://www.adobe.com/at/products/muse.html>

Adobe Dreamweaver CC:

<https://www.adobe.com/at/products/dreamweaver.html>

Sublime Text:

<https://www.sublimetext.com>

Aptana Studio:

<http://www.aptana.com>

Seamonkey:

<http://www.seamonkey-project.org>

BlueGriffon:

<http://www.bluegriffon.org>

Archivierungsmethoden Die Archivierung einer Webseite wird schnell und einfach durch ihre Konvertierung in eine PDF-Datei und anschließende Speicherung als PDF/A-Datei bewerkstelligt. Sie kann auf unterschiedliche Weise mittels des Webbrowsers, durch eigene Online-Konvertierungsdienste oder spezielle Softwareprogramme erfolgen.

Alternativ kann eine Speicherung und Archivierung der Webseite als HTML-Datei mit Data URIs, MHTML-Datei oder MAFF-Container vorgenommen werden. Hierbei ist zu beachten, dass nicht alle Webbrowers alle Formate unterstützen bzw. ein vorhergehendes Entpacken der komprimierten Webseite nötig ist. Eine Übersicht der unterstützten Formate für aktuelle Webbrower ist in Tabelle 3.14 und Tabelle 3.15 gegeben.

Eine weitere Methode stellt die Archivierung der Webseite durch spezielle Archivierungsdienste (z.B. Internet Archive <https://archive.org/web/>) dar.

Allen Methoden ist gemein, dass sie für gewöhnlich keine multimedialen (und extern von anderen Webseiten) eingebundene Inhalte (Video, 3D-Modelle) in die PDF-Datei integrieren. Derartige Inhalte müssen in der Regel separat archiviert werden.

Webbrowser	Data-URI	MAFF	MHTML	PDF
Chrome (54.0.x)	✓	~*	✓	✓
Edge	✓	~*	✗	✓
Firefox (49.0.x)	✓	✓	~**	✓
Internet Explorer (11.x)	✓	~*	✓	✓
Opera (40.0.x)	✓	~*	✓	✓
Safari (10.0.x)	✓	~*	~**	✓
Vivaldi (1.4.x)	✓	~*	✓	✓

* vorheriges Entpacken **PlugIn benötigt

Tabelle 3.14: Überblick über die näher behandelten Speichermöglichkeiten und deren Unterstützung durch aktuelle Webbrowser. Die Tabelle beschreibt, welche Dateiformate durch welchen Webbrowser zu öffnen sind.

Webbrowser	Data-URI	MAFF	MHTML	PDF
Chrome (54.0.x)	~*	~*	✓ **	✓
Firefox (49.0.x)	✗	✓	✓ *	✓
Internet Explorer (11.x)	✗	✗	✓	✓
Opera (40.0.x)	✗ *	✗	✓	✓
Safari (10.0.x)	✗	✗	✗	✓
Vivaldi (1.4.x)	✗	✗	✓ **	✓

* PlugIn benötigt ** MHTML muss aktiviert werden

Tabelle 3.15: Überblick über die näher behandelten Speichermöglichkeiten und deren Unterstützung durch aktuelle Webbrowser. Die Tabelle beschreibt, welche Dateiformate durch welchen Webbrowser generiert werden können.

Archivierung als PDF Jeder Webbrowser ermöglicht das Ausdrucken einer Webseite, durch Verwendung eines PDF-Druckers. Anschließend kann das PDF in PDF/A konvertiert werden. Nähere Informationen zur Generierung von PDF-Dateien findet sich im Praxisteil des Kapitels PDF-Dokumente.

Diese Methode stellt sicher, dass alle Informationen der angezeigten Webseite (exklusive multimedialer Inhalte, etwa Videos oder eingebettete 3D-Modelle) gespeichert und anschließend archiviert werden kann. Je nach Webseite, Webbrowser und PDF-Drucker variiert das Ergebnis jedoch hinsichtlich der Übernahme des Layouts. Meist können direkt im Browser oder in den Einstellungen des PDF-Generators für den Ausdruck typische Parameter konfiguriert werden: Druck der gesamten Webseite oder eines Auszugs, definiert durch die Seitenzahlen, Hoch- oder Querformat, Druck in Farbe oder Graustufen/Schwarz-Weiß, Papierformat, Seitenränder, Auflösung, Hintergrundgrafiken, Kopf- und Fußzeilen (diese enthalten in der Regel das Datum des Ausdrucks und die URI der ausgedruckten Webseite). Eingeschränkte Formatierungen können besonders durch die Anpassung der Ränder, des verwendeten Papierformats und dessen Orientierung vorgenommen werden.

Spezielle Plug-Ins für Webbrowser ermöglichen es, die Webseite auch unter Beibehaltung des Layouts als PDF-Datei zu speichern. Sie bieten oftmals umfangreiche Optionen hinsichtlich der gewünschten PDF-Datei, so z.B. die Spei-

cherung als ein langes, durchgehendes Dokument im Format der Webseite, was die Adaptierung der Webseite auf ein bestimmtes Format (z.B. A4) und die damit verbundene Aufsplittung der Webseite in ein mehrseitiges PDF-Dokument obsolet macht. Anschließend kann diese wieder in eine PDF/A konvertiert werden. Je nach verwendetem Webbrower und Plug-in sowie besuchter Webseite können die Ergebnisse zwischen Original und Kopie variieren. Den PDF-Dateien werden zudem meistens Angaben zum verwendeten Plug-in in Form eines Wasserzeichens hinzugefügt. Derzeit aktuelle Plug-ins sind etwa „Firefox Web2PDF Converter“, „Save as PDF“ sowie „FireShot“. Das Angebot an solchen Plugins ist sehr vielfältig, umfangreich und schnelllebig; diese und weitere Plugins können von den jeweiligen auf Webbrowerweiterungen spezialisierten Onlinestores der einzelnen Anbieter bezogen werden.

Eine weitere Möglichkeit bietet die Konvertierung einer Webseite durch einen Online-Konvertierungsdienst, etwa Web2PDF. Nach der Eingabe der URI der zu konvertierenden Webseite wird ein PDF-Dokument generiert und zum Download angeboten. Je nach Webseite und Konvertierungsdienst können die Ergebnisse zwischen Original und Kopie variieren. Den PDF-Dateien werden zudem meistens Angaben zum verwendeten Plug-in in Form eines Wasserzeichens hinzugefügt. Im Anschluss an den Download kann die PDF-Datei zu einer PDF/A-Datei konvertiert und archiviert werden.



Abb. 3.40: Das Glossar der IT-Empfehlungen als Screenshot (links), als mit dem Browser erzeugte PDF-Datei (mitte) und als mittels dem „Save as PDF“-Plug-in erzeugtes PDF (rechts).

Spezielle (kommerzielle) Programme wie die Literaturverwaltungssoftware Citavi beherrschen ebenso das Speichern von Webseiten unter Beibehaltung des Layouts als PDF. Auch bietet etwa Adobe Acrobat Pro DC eine Option zur Konvertierung von Webseiten in ein PDF-Dokument.

Mozilla Firefox „Firefox Web2PDF Converter“:

<https://addons.mozilla.org/en-US/firefox/addon/web2pdf-converter>

Mozilla Firefox „Save as PDF“:

<https://addons.mozilla.org/de/firefox/addon/save-as-pdf>

Mozilla Firefox „FireShot“:

<https://addons.mozilla.org/de/firefox/addon/fireshot>

Google Chrome „Save as PDF“:

<https://chrome.google.com/webstore/detail/save-as-pdf/kpdjmbiefanbdgnkcihllpmjnllbbc>

Google Chrome „FireShot“:

<https://chrome.google.com/webstore/detail/capture-webpage-screensho/mcbpblocmgfnpjjppndjkmgjaogfcieg>

Web2PDF:

<http://www.web2pdfconvert.com>

Swiss Academic Software Citavi:

<https://www.citavi.com/de/index.html>

Adobe Acrobat Pro DC:

<https://acrobat.adobe.com/at/de/acrobat.html>

Archivierung als MHTML MHTML-Dateien können mit Webbrowsersn erstellt und geöffnet werden. Auch mit Texteditoren können MHTML-Dateien angesehen werden. Jedoch benötigen Firefox und Safari noch ein Plug-in. Bei den Webbrowsersn Chrome und Vivaldi muss zuvor MHTML in den experimentellen Funktionen aktiviert werden: Bei Chrome wird hierfür `chrome://flags/`, bei Vivaldi `vivaldi://flags` in die Adresszeile eingegeben und die entsprechende Funktion aktiviert. Ein Neustart des Webbrowsers wird danach benötigt. Im Head der MHTML-Dateien können Metadaten mittels eines Texteditors einge tragen werden. Der Inhalt wie auch das Layout und alle Hyperlinks werden bei MHTML-Dateien in der Regel vollständig übernommen.

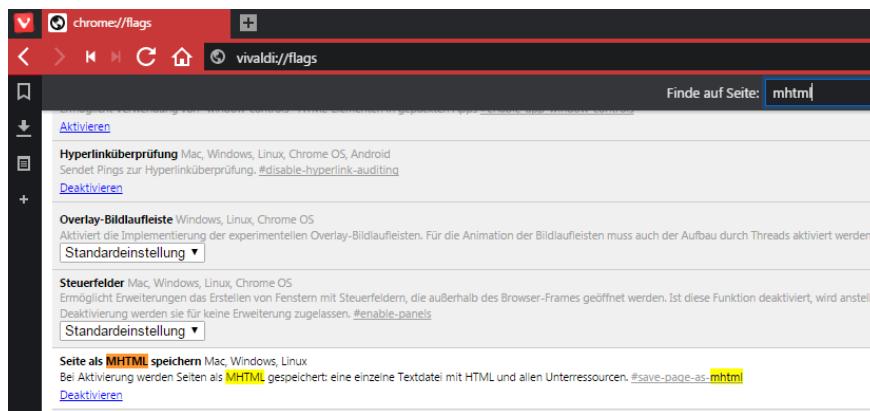


Abb. 3.41: Einstellungen im Webbrowser Vivaldi

Google Chrome „Archiveror“:

<https://chrome.google.com/webstore/detail/archiveror/cpjdnkhhgjdecpmjglkcegchhiijadpb>

Google Chrome „Save As MHTML“:

<https://chrome.google.com/webstore/detail/save-as-mhtml/eomfisclcdpkaghkehajpoldkdnkmegfa>

Google Chrome „In Google Drive speichern“:

<https://chrome.google.com/webstore/detail/save-to-google-drive/gmbmikajggmnabiglmoftipeabaddhgne>

Mozilla Firefox „Mozilla Archive Format, with MHT and Faithful Save“:

<https://addons.mozilla.org/de/firefox/addon/mozilla-archive-format>

Abb. 3.42: Der Quellcode einer MHTML-Datei. Rechts ist der Code einer eingebetteten Grafik zu sehen.

Archivierung als MAFF-Datei Die Speicherung einer Webseite als MAFF-Datei wird derzeit nur von Mozilla Firefox mittels des Plug-ins „Mozilla Archive Format, with MHT and Faithful Save“ unterstützt. MAFF-Dateien können nur von Mozilla Firefox mit diesem Plug-in geöffnet werden. Alle weiteren, aktuellen Webbrowsersn können MAFF-Dateien öffnen, indem sie mittels eines Datenkompressionsprogramms entpackt werden.

Mozilla Firefox „Mozilla Archive Format, with MHT and Faithful Save“:
<https://addons.mozilla.org/de/firefox/addon/mozilla-archive-format>

Archivierung als HTML mit Data-URI Alle aktuellen Webbrowser können HTML-Dateien mit Data-URIs öffnen. Derzeit beherrscht nur der Webbrowser Google Chrome die Speicherung von Webseiten als HTML-Dateien mit Data-URIs mittels des Plug-ins „SingleFile“. Eine alte Version des gleichen Plugins existiert auch für frühere Versionen des Webbrowsers Opera. Im Head der HTML-Dateien mit Data-URLs können Metadaten mittels eines Texteditors eingetragen werden. Der Inhalt wie auch das Layout und alle Hyperlinks werden bei HTML-Dateien mit Data-URIs in der Regel vollständig übernommen.

Google Chrome „SingleFile“:
<https://chrome.google.com/webstore/detail/mpiodijhokgodhhofbcjdecppffipk>

Archivierung von Websites Webseiten können durch Websitearchivierungsdienste archiviert werden, wie sie durch die Bayerische Landesbibliothek (auf Antrag) oder Internet Archive angeboten werden. Dabei erfolgt die Speicherung einer Webseite auf einem Server dieser Dienste und kann über das Internet abgerufen werden. Man gibt dazu die URI der zu archivierenden Seite bei dem Archivierungsdienst an und wird kurz darauf auf die archivierte Seite unter einer neuen URI weitergeleitet. Es kann sowohl die einzelne Webseite als auch die gesamte (oder ein Großteil) der gesamten Website archiviert werden. Metadaten müssen separat z.B. in Form einer XML-Datei, die auch den Link zur

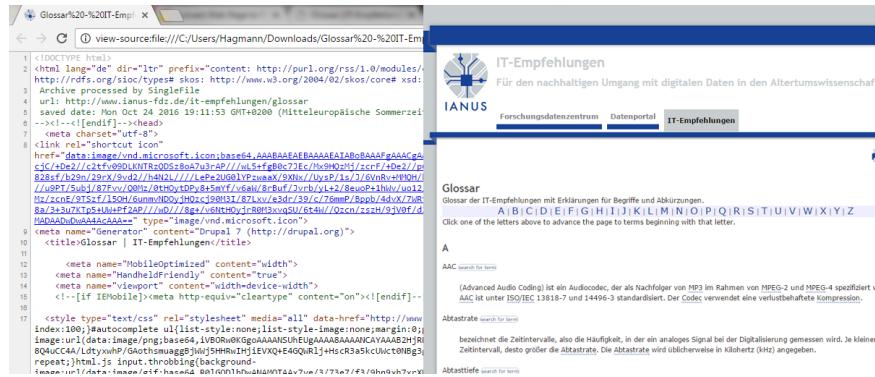


Abb. 3.43: Links der Quellcode einer HTML-Datei mit Data-URIs. Rechts die Darstellung im Browser. Eindeutig erkennbar ist die vollständige Übernahme des Inhalts samt der Hyperlinks sowie die vollständige Übernahme des Designs. Der blau unterstrichene Codeblock ist eine Data-URI eines Bildes.

archivierten Seite enthält, gespeichert werden. Da Websitearchivierungsdienste wie Internet Archive das WWW auch selbstständig durchsuchen und Websites archivieren, kann die zu archivierende Webseite bereits auf dieser Plattform gesichert worden sein. Dies wird durch die Archivierungsdienste jedoch gesondert ausgewiesen und hindert nicht daran, die Webseite zusätzlich ein weiteres Mal zu archivieren. Plug-ins wie „Archiveror“ für Google Chrome oder Mozilla Firefox ermöglichen die Archivierung einer Webseite direkt aus dem Browser heraus auf Internet Archive.

Dezidierte Softwarelösungen wie Wget oder Heritrix ermöglichen den automatisierten Abruf aller zu einer Website gehörenden Komponente. Sie wurden primär für Linux entwickelt, können aber auch auf anderen Betriebssystemen verwendet werden. Beide sind frei verfügbar und können die gefundenen Ressourcen als WARC-Datei speichern.

Google Chrome „Archiveror“:

<https://chrome.google.com/webstore/detail/archiveror/cpjdnckhgjdecpmjglkceghhijadpb>

Mozilla Firefox „Archiveror“:

<https://addons.mozilla.org/en-US/firefox/addon/archiveror>

Dienst der Bayrischen Landesbibliothek:

https://www.babs-muenchen.de/index.html?c=workflows_web

Internet Archive mit Wayback Machine:

<https://archive.org/web>

Wget:

<https://www.gnu.org/software/wget>

Heritrix:

<https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>

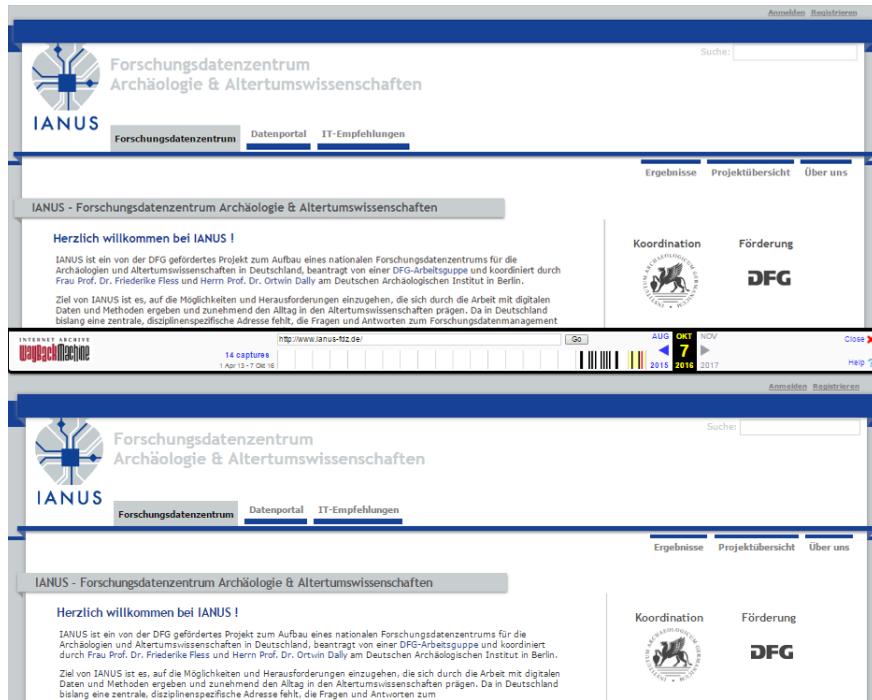


Abb. 3.44: Screenshot der Webseite von IANUS in Google Chrome im Vollbild (oben) und ein Abbild der Webseite auf Internet Archive. Eindeutig erkennbar ist die vollständige Übernahme des Inhalts samt der Hyperlinks sowie die vollständige Übernahme des Designs.

Screenshots von Webseiten Die Speicherung von Webseiten in der Form von Screenshots ist nicht für die Archivierung geeignet, ist aber ein gutes Hilfsmittel, um das ursprüngliche Aussehen der Webseite zu dokumentieren. Screenshots können mit Hilfe der Screenshot-Funktion des Computers, spezieller Screenshot Software (z.B. Microsofts Snipping Tool) oder durch eigene, auf die Verarbeitung von Webseiten spezialisierte Plug-ins für Webbrower erzeugt werden. Während für gewöhnlich mit der Screenshot Funktion des Computers der gesamte Bildschirm bzw. aktive Fenster und mit Screenshot Softwarelösungen zusätzlich einzelne Bildausschnitte als Grafiken gespeichert werden können, fertigen erwähnte Plug-ins einen Screenshot der gesamten Webseite oder eines ausgewählten Teiles davon an. Screenshots werden durch die erwähnten Funktionen und Programme in der Regel im PNG- oder JPEG-Format gespeichert.

Quellen

U. Ackermann – C. Berner – N. Elbert – J. Kett – K. K. Koçer – N. von der Hude – M. Wiegand, Policy für die Vergabe von URNs im Namensraum urn:nbn:de (2012)

<http://d-nb.info/1029114455/34>

N. Brügger, Archiving Websites. General Considerations and Strategies (Aarhus 2005)

http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf

Archaeology Data Service — Digital Antiquity, Documents and Digital Texts:
A Guide to Good Practice. Section 2. Creating Texts and Documents
http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_2

Archaeology Data Service — Digital Antiquity, Documents and Digital Texts:
A Guide to Good Practice. Section 3. Archiving Texts and Documents
http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_3

DOI (Hrsg.), DOI Handbook
<https://www.doi.org/hb.html>

A. Rauber – H. Liegmann, Webarchivierung zur Langzeiterhaltung von Internet-Dokumenten, in: H. Neuroth – A. Oßwald – R. Scheffel – S. Strathmann – K. Huth (Hrsg.) nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchiverung. Version 2.3 (2010) Kap. 17.9
<http://www.nestor.sub.uni-goettingen.de/handbuch>

S. M. Schafer, HTML, XHTML, and CSS Bible ⁴(Indianapolis 2008)

M. Schäfer, Einführung in JavaScript – Deutschsprachige Dokumentation der Programmiersprache JavaScript
<http://molily.de/js/>

SELFHTML-Wiki, Glossar
<https://wiki.selfhtml.org/wiki/Kategorie:Glossar>

M. Trognitz, Abschlussbericht Testbed „Persistent Identifiers“ (2013)
<http://www.ianus-fdz.de/attachments/download/560/Testbed-Persistent%20Identifiers.pdf>

W3C (Hrsg.), CSS
<https://www.w3.org/Style/CSS>

W3C (Hrsg.), DOM
<https://www.w3.org/DOM>

W3C (Hrsg.), HTML
<https://www.w3.org/html>

W3C
<https://www.w3c.org>

W3schools
<http://www.w3schools.com/default.asp>

Formatspezifikationen

HTML:

<https://www.w3.org/TR/html5>

XHTML:

<https://www.w3.org/TR/html5>

CSS:

<https://www.w3.org/TR/2011/REC-CSS2-20110607>

MHTML: RFC 2557

<https://tools.ietf.org/html/rfc2557>

WARC 0.9:
http://archive-access.sourceforge.net/warc/warc_file_format-0.9.html

WARC: ISO 28500
http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

WARC: Library of Congress
http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

MAFF:
http://maf.mozilla.org/maff-specification.html

Data-URI: RFC 2397
https://tools.ietf.org/html/rfc2397

WWW: RFC 7540
https://tools.ietf.org/html/rfc7540

FTP: RFC 959
https://tools.ietf.org/html/rfc959

HTTP: RFC 7540
https://tools.ietf.org/html/rfc7540

HTTP:
https://html.spec.whatwg.org/multipage

URI, URL : RFC 3986
https://tools.ietf.org/html/rfc3986

URN: RFC 2141
https://tools.ietf.org/html/rfc2141

Java:
https://docs.oracle.com/javase/specs

Java:
http://docs.oracle.com/javase/specs/jls/se8/html/index.html

JavaScript:
http://www.ecma-international.org/ecma-262/5.1

JavaScript:
http://ecma-international.org/ecma-402/1.0/index.html

Tools und Programme

Adobe Muse CC:
https://www.adobe.com/at/products/muse.html

Adobe Dreamweaver CC:
https://www.adobe.com/at/products/dreamweaver.html

Sublime Text:
https://www.sublimetext.com

Aptana Studio:
http://www.aptana.com

Seamonkey:
http://www.seamonkey-project.org

BlueGriffon:

<http://www.bluegriffon.org>

Mozilla Firefox „Firefox Web2PDF Converter“:

<https://addons.mozilla.org/en-US/firefox/addon/web2pdf-converter>

Mozilla Firefox „Save as PDF“:

<https://addons.mozilla.org/de/firefox/addon/save-as-pdf>

Mozilla Firefox „FireShot“:

<https://addons.mozilla.org/de/firefox/addon/fireshot>

Google Chrome „Save as PDF“:

<https://chrome.google.com/webstore/detail/save-as-pdf/kpdjmbiefanbdgnkceikhllpmjnllbbc>

Google Chrome „FireShot“:

<https://chrome.google.com/webstore/detail/capture-webpage-screensho/mcbpblocmgfnppndjkmgjaogfceg>

Web2PDF:

<http://www.web2pdfconvert.com>

Swiss Academic Software Citavi:

<https://www.citavi.com/de/index.html>

Adobe Acrobat Pro DC:

<https://acrobat.adobe.com/at/de/acrobat.html>

Google Chrome „Archiveror“:

<https://chrome.google.com/webstore/detail/archiveror/cpjdnckhgjdecpmjglkcegchhiijadpb>

Mozilla Firefox „Archiveror“:

<https://addons.mozilla.org/en-US/firefox/addon/archiveror>

Google Chrome „Save As MHTML“:

<https://chrome.google.com/webstore/detail/save-as-mhtml/eomfifclcdpkaghkehajpolkdnkmegfa>

Google Chrome „In Google Drive speichern“:

<https://chrome.google.com/webstore/detail/save-to-google-drive/gmbmikajggmnabiglmosfipeabaddhgne>

Mozilla Firefox „Mozilla Archive Format, with MHT and Faithful Save“:

<https://addons.mozilla.org/de/firefox/addon/mozilla-archive-format>

Dienst der Bayrischen Landesbibliothek:

https://www.babs-muenchen.de/index.html?c=workflows_web

Internet Archive mit Wayback Machine:

<https://archive.org/web>

Wget:

<https://www.gnu.org/software/wget>

Heritrix:

<https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>

4 | Forschungsmethoden

F. Schäfer, M. Trognitz

Die Ergebnisse aus verschiedenen in den Altertumswissenschaften angewandten Methoden setzen sich oft aus mehreren verschiedenen Dateien unterschiedlicher Formate zusammen. Dadurch gehen die Ansprüche an die Dokumentation und die erforderlichen Metadaten zu den einzelnen Dateiformaten über die des Kapitels zu Dateiformaten hinaus und unterscheiden sich von Methode zu Methode.

Sofern spezifische Dateiformate eine besondere Relevanz besitzen, sind diese entweder hier beschrieben oder es wird auf die entsprechenden Abschnitte im Kapitel Dateiformate auf Seite 52 verwiesen.

Da die Kapitel der einzelnen Forschungsmethoden von verschiedenen Spezialisten verfasst wurden, können sie deshalb und in Abhängigkeit von der jeweils beschriebenen Methode in Inhalt, Länge und Gliederung sehr unterschiedlich ausfallen.

Die vorhandenen Inhalte decken nur einen kleinen Teil der angewandten Forschungsmethoden in den Archäologien und Altertumswissenschaften ab. Weitere Inhalte sind geplant, wobei dafür auch Autoren gesucht werden. Die folgende Auflistung soll einen Eindruck der möglichen Inhalte vermitteln, wobei weitere Vorschläge und Ergänzungen willkommen sind.

3D-Scanning · Anthropologie · Archäobotanik · Archäometrie · Archäozoologie · Ausgrabung · Bauforschung · Datierungsmethoden · Geoarchäologie · Numismatik · Oberflächenbegehung (Survey) · RTI · Textanalyse

Bis wir Ihnen ausführlichere Hinweise zur Verfügung stellen können, können Sie sich auf den Seiten des Archaeology Data Service informieren, wo insbesondere folgende Inhalte zu Forschungsmethoden zur Verfügung stehen.

Dendrochronologie · Fernerkundung · GIS · Geophysik · Laserscanning · Photogrammetrie · UAV Survey · Unterwassersurvey

4.1 Geodäsie

K. Heine, U. Kapp

Alle Protokolle von Messungsrohdaten, mit Ergebnissen der Netzmessungen, der Ausgleichung und Festpunktkoordinaten inkl. Einmessungsskizzen müssen im ASCII-Format (TXT, CSV, DAT) gespeichert werden.

Vermessungswesen

Es ist ja eine Binsenweisheit, dass archäologische und bauforscherische Tätigkeiten nicht ohne fundierte vermessungstechnische Basis- und begleitende Arbeiten auskommen. Die Zeiten, in denen man im 5x5 Meter Planquadrat seine Erkenntnisse sammelte, gehören zwar nicht der Vergangenheit an, aber die Bedingung ist, dieses Planquadrat in einen globalen Zusammenhang zu bringen. Die Vermessung bekommt in diesem Fall eine andere Bedeutung, die von der bloßen Herstellung von rechtwinkligen Quadranten abweicht. Andere Unterlagen wie Satelliten Images und Luftbilder sowie genaueres Kartenmaterial oder auch Unterlagen von z. B. geomagnetischen Prospektionen geben der Vermessung eine neue Wertigkeit.

Aber: der Einsatz elektronischer und satellitengestützter Geräte führt leider wie auch in der digitalen Fotografie zu einer fast unkontrollierbaren Anhäufung von Daten. Diese Ergebnisse müssen kurz- und langfristig verwaltbar sein. Aber dafür gibt es keine Ideallösung, weil die Entwicklungen im Gerätbau und den damit zusammenhängenden Programmen sich überschlagen, und was heute noch als Standard gilt, ist morgen schon überholt.

Meine Empfehlung: die Ergebnisse möglichst schnell auswerten, solange sie noch lesbar sind, denn nach wie vor ist eine Langzeitarchivierung von digitalen Daten nicht gesichert.

Bezugssystem für Messdaten

Nach wie vor ist für viele archäologische Aktivitäten die Vermessungsgrundlage ein lokales Vermessungsnetz, das nicht in überörtliche oder gar in globale Systeme eingebunden ist. Für kleinere Messgebiete ist das legitim, solange sie nicht in einen Zusammenhang gebracht werden müssen.

Da das DAI aber in der Zwischenzeit viele Projekte mit großer räumlicher Ausdehnung bearbeitet, ist es sinnvoll, Messdaten in globalen Systemen zu erfassen, was durch die mittlerweile zur Verfügung stehenden Methoden und Geräte der Satelliten gestützten Vermessung deutlich vereinfacht wird. Durch die Einbeziehung von GPS-Messungen (Genauigkeit 5-10 m) und Differential GPS-Messungen (Zentimetergenauigkeit) kann der räumliche Zusammenhang von verschiedenen Arbeitsplätzen hergestellt werden.

Als Voraussetzungen für globale Positionierungsbestimmungen mit cm-Genauigkeit muss entweder ein Punkt im globalen Koordinatensystem (im Umkreis von 10 km) bekannt sein oder die Möglichkeit der Einbindung eines Satellitenpositionierungsservices (z. B. EGNOS) bestehen. Eine hinreichende Genauigkeit kann auch erzielt werden, wenn auf einem Punkt im Messgebiet entsprechend lange Beobachtungszeiten registriert werden. Grundsätzlich sollten für die Anlage von Vermessungsnetzen im globalen Koordinatensystem mittels GPS Zweifrequenzempfänger benutzt werden. Alternativ können Einfrequenzempfänger mit langer Beobachtungszeit (mindestens 20 min) genutzt werden. Eine Netzauswertung sollte im Post Processing erfolgen. Für die Ermittlung von Höhen sind ein Referenzpunkt sowie ein Geoid-Modell notwendig.

Geografisches System

Grundsätzlich sollte als geografisches Bezugssystem das System WGS 84 (World Geodetic System 84) gewählt werden. Dieses ist auch Bezugssystem für das Satellitenavigationssystem GPS und zukünftig auch für das europäische System Galileo. WGS 84 ist ein räumliches, globales, geozentrisches System mit den rechtwinkligen Koordinaten x,y,z oder wahlweise auch geografischen Koordinaten geografische Länge (longitude) L und geografische Breite (latitude) B ($^{\circ}$, $'$, $''$). Die Koordinaten lassen sich nach Bedarf in andere Koordinatensysteme transformieren.

Geodätisches System

Für praktische Zwecke der Aufnahme mit der Totalstation und Darstellung von Objekten in Zeichnungen und Plänen, insbesondere für großmaßstäbige Darstellungen (großmaßstäbig bedeutet hier: großer Abbildungsmaßstab, demzufolge

i.d.R. kleineres Gebiet) ist dieses räumliche System in ein Abbildungssystem (geodätisches Koordinatensystem, projected coordinate system) zu transformieren. Hierfür eignen sich insbesondere transversale Zylinderabbildungen wie die UTM-Abbildung (Universal Transversal Mercator projection). Mit der UTM-Abbildung können auch lokale Systeme realisiert werden, indem als Mittelmeridian (central meridian) ein durch das Aufnahmegebiet verlaufender Meridian (Längengrad) und als Maßstab der Wert 1,0000 festgelegt wird. Die Angabe der UTM-Koordinaten erfolgt als East (E)- und North (N) - Wert.

Nationale Systeme

In vielen Ländern existieren ein oder mehrere amtliche Systeme, denen unterschiedliche Bezugsellipsoide (z.B. Hayford, Bessel, Krassowski o.a.) mit unterschiedlichem Datum (Lagerung) und verschiedene Abbildungsmethoden (stereografisch, Lambert, Gauß-Krüger, UTM o.a.) zugrundeliegen. Sofern Ellipsoidparameter, Lagerung und Abbildungsverfahren bekannt sind, kann auch in diesen Systemen gearbeitet werden, da diese dann jederzeit in andere Systeme transformierbar sind. Sofern die Parameter nicht bekannt sind, ist zu empfehlen, ein eigenes System unter Verwendung des WGS 84 und der UTM-Abbildung (s.o.) zu wählen und bestehende Karten im unbekannten System über Passpunkte in das neue System zu integrieren.

Momentan verfügen die meisten europäischen Staaten und auch die meisten Bundesländer noch über eigene amtliche Systeme. Für die nächsten Jahre ist aber der Übergang zum einheitlichen System ETRS89 (European Terrestrial Reference System 89) geplant. ETRS89 basiert auf dem Ellipsoid GRS80 (für praktische Zwecke mit WGS 84 identisch) sowie der europäischen Lagerung ETRF (European Terrestrial Reference Frame) von 1989. Als Abbildungsverfahren wird die UTM-Abbildung mit 6° Streifenbreite und dem Maßstabsfaktor 0.9996 gewählt. Messungen in Deutschland sowie in den betroffenen europäischen Staaten sollten dann grundsätzlich ebenfalls in diesem System erfolgen, da die amtlichen Koordinaten für Festpunkte sowie Geobasisdaten dann ebenfalls in diesem System vorliegen.

EPSG

International werden Bezugssysteme nach dem System der EPSG (European Petroleum Survey Group) angegeben. Die EPSG stellt die Parameter für die meisten Bezugssysteme bereit. Diese sind als Download auf den Webseiten der EPSG (www.epsg.org) zu beziehen.

Örtliche Systeme für die Bauforschung

In der Bauforschung ist auch ein objektbezogenes Koordinatensystem, welches beispielsweise parallel zur längsten Bauwerksachse ausgerichtet ist und einen beliebigen Koordinatenursprung besitzt, anwendbar. Dies ist für große und komplexe Bauwerke unter Umständen praktikabler als die Verwendung eines globalen Systems. Um die Messungen bzw. Pläne später in solch ein übergeordnetes System einhängen zu können, sollten mindestens drei Passpunkte bestimmt werden, deren Koordinaten sowohl im bauwerksbezogenen, als auch im globalen System (bestimmbar beispielsweise mittels GPS) bekannt sind.

Einsatz von Hard- und Software in der Praxis

Da es eigentlich nur noch zwei wesentliche Hersteller von Vermessungsgeräten gibt - Leica und Trimble, die alle anderen Hersteller integriert haben - liegt der Entschluss, welches Gerät man wählt, eigentlich nur in der Beziehung, die zum Verkäufer besteht, denn preislich findet man die Unterschiede nicht mehr so groß. Auch die Handhabung ist nicht so unterschiedlich, da sie einem Schema folgt: Gerät aufbauen, Orientieren, Messen, Auslesen. Bei Neuanschaffungen sollte darauf geachtet werden, dass eine Systemreihe fortgeführt wird, wenn nicht zwingende Gründe dagegen sprechen, denn mit jedem Gerätesystemwechsel geht auch ein Softwarewechsel einher.

Eine Beschreibung der Handhabung der Geräte geht meistens aus der Gebräuchsanleitung hervor und würde den Rahmen dieses Leitfadens sprengen.

Messungsdaten

Ich möchte hier nicht das Handaufmaß einbeziehen, sondern nur die elektronischen Messweisen mit Totalstation und GPS. Laserscan-Daten fallen beim DAI noch nicht an, da es noch kein eigenes Gerät gibt. Zum Handaufmaß gehören auch Daten, die mit Laser-Distomat gemessen, also nicht registriert, gespeichert werden.

Grundsätzlich muss man die eigentliche Vermessung von der Weiterverarbeitung im Computer trennen, da es für viele Vermessungsgeräte nach wie vor keine direkte Schnittstelle zum Computer gibt, sieht man von einigen Programmen ab, wie z.B. TachyCad als Aufsatz für AutoCad. Doch damit tauchen dann auch andere Probleme auf, die hardwaremäßig bedingt sind.

Sämtliche Vermessungsdaten der gängigen Totalstationen lassen sich mit den vom Hersteller des Geräts beigefügten Programmen in ASCII, TXT- oder DXF-Dateien in den Computer übertragen. Für die Weiterverarbeitung in CAD-Programmen bedarf es nun eines Transferprogramms, damit – besonders bei AutoCad – diese Daten erkannt und als Objekte behandelt werden können.

Es gibt einige Profi-Programme. Ich benutze seit Jahren BB-VermCad und bin, auch weil ich es gut kenne, damit zufrieden. Nur hat es, wie jedes Aufsatzprogramm den Nachteil, dass vorhandene Versionen nie kompatibel mit neueren Versionen von AutoCad sind, man also theoretisch mit jeder neueren Ausgabe von AutoCad auch eine neue Version von BB kaufen müsste; oder man behält die laufende Version - auch von AutoCad - zur Bearbeitung der Vermessungsdaten auf dem Rechner und lädt dann die erhaltenen Ergebnisse in die neue Version. Die Frage ist, ob man diese neuen Versionen braucht? Meine Antwort: leider ja, da viele Kollegen sich diesem Trend nicht ausschließen, und dann plötzlich Ergebnisse im Austausch nicht mehr kompatibel sind, weil sie mit verschiedenen Versionen abgespeichert wurden (down-kompatibel ist fast immer möglich, up-kompatibel fast nie!).

AutoCad, MicroStation, ArchiCad sind nur drei Programme, die mir im Laufe meines Arbeitslebens als Vermesser begegnet sind. Von diesen ist AutoCad am verbreitetsten, und zusätzliche Programme (BB, TachyCad) werden auch nur für ACAD gemacht. Auch „handgeschriebene“ Programme beziehen sich fast nur auf ACAD, da AutoCad eine eigene Programmiersprache¹, die relativ leicht zu erlernen sein soll, beinhaltet. Es ist nicht nur eine Sache des guten Ge-

¹ AutoCad lisp (*.lsp)

schmacks sondern auch eine Antwort auf die Frage: welches Programm benutzt mein Kollege, und der Kompromiss ist AutoCad.

Das einzige Austauschformat zwischen unterschiedlichen Programmen - *.dxf - ist hier mit Vorsicht zu genießen, da es auch versionsabhängig ist. Häufig wird die gesamte *.dxf-Datei nicht von einem anderen Programm vollständig gelesen, besonders bei 3D-Dateien kann es hier zu Irritationen führen.

Es gilt also, einen Standard aufzubauen, der sowohl Geräte als auch Computerprogramme einschließt.

Netzmessungen

Eine dauerhafte Sicherung der Messungsrohdaten scheint nur für die Vermessung der Grundlagennetze sinnvoll. Dies gilt sowohl für terrestrische als auch für GPS-Messungen. Für jedes Projekt sollten die Festpunktkoordinaten und die Festpunktbeschreibungen inkl. der Einmessungsskizzen gespeichert werden.

Objektpunkte

Ob es sinnvoll ist, die Koordinaten aller aufgemessenen Punkte dauerhaft zu sichern ist fraglich, denn entweder werden bei der Aufmessung alle relevanten Informationen zu einem Punkt mit eingegeben (was sehr aufwendig ist) und abgespeichert, oder es muss, um die Messung nachvollziehen zu können eine eindeutige Skizze (Feldriss) vorhanden sein, die dann ebenfalls dauerhaft (digitalisiert) gespeichert werden müsste.

Wenn eine dauerhafte Sicherung erfolgen soll, sollte dies im ASCII-Datenformat (*.txt, *.dat, *.csv o.ä.) erfolgen, da die meisten CAD-Programme diese in irgendeiner Form erkennen. Häufig müssen die Daten mit Zusatzprogrammen (s.o.) CAD-gerecht aufbereitet werden. Es gibt in diesem Fall auch keine Empfehlung, da die Programme nur bedingt kompatibel sind. Wie schon erwähnt, ist auch das *.dxf-Format nur mit Vorsicht zu benutzen, EXCEL-Dateien müssen auch wieder in *.txt-Dateien gespeichert werden, usw.

Man erreicht einen Standard nur dann, wenn alle Beteiligten die gleichen Programme der gleichen Version benutzen.

Um Zeichendateien, bei AutoCad *.dwg auszutauschen, empfiehle ich die Dateien in einer Version (2000, 2002) abzuspeichern, die ohne Schwierigkeiten und Verluste auch mit einer neueren Version zu öffnen ist. Die meisten 3D-Daten lassen sich so verlustfrei austauschen.

Quellen

- Bauer, M.: Vermessung und Ortung mit Satelliten. Wichmann Verlag
Heidelberg
- Schimke, A.: Kartenprojektionen und Koordinatentransformationen
- Zeiske, K.: Vermessen leicht gemacht. Leica Publikation
- Herrmann K.: Bautechnische Vermessung. Dümmers Verlag Bonn
- Seckel/Hell/Schnuchel: Vermessungskunde und Bauaufnahme für Architekten.
Herbert Wichmann Verlag Karlsruhe
- GPS Basics, Einführung in die GPS Vermessung, Leica Publikation
- Gebrauchsanweisungen für Geräte beinhalten häufig gute Ratschläge
- <http://www.olanis.de/knowhow/mappj/index.shtml>
- www.epsg.org

4.2 Georeferenzierung

K. Heine, U. Kapp

GeoReferenzierung, d.h. Zuordnung von Koordinaten zu Bildpunkten auf Fotos und Plänen (Karten), die in Digitaler Form vorliegen. Mit der GeoReferenzierung erreicht man gleichzeitig eine Entzerrung dieser Grundlagen. Voraussetzung ist ein Programm, mit dem man diese Arbeit durchführen kann. Jedes GIS (Geoinformationssystem) hat notwendigerweise ein Modul für GeoReferenzierung.

Einsatz von Hard- und Software in der Praxis

Die GeoReferenzierung von Topographischen Karten ist in der Regel einfach, da Koordinaten normalerweise in Form eines Rasters auf den Karten vorhanden sind. Es wird komplizierter, wenn man mit diesen Koordinaten nicht arbeiten möchte. Dann wird es erforderlich, entweder vor Ort Vergleichspunkte (Referenzpunkte) zu messen oder mit anderen Programmen die vorhandenen Koordinaten in das gewünschte System zu transformieren. Viele GIS-Programme erlauben auch das. Ich habe Erfahrungen mit dem einfachen FUGAWI, das für die Bearbeitung der GPS-Daten der Handgeräte vorhanden ist.

Eine andere Möglichkeit, Referenzpunkte zu bekommen, bildet im Moment GOOGLE EARTH. Besonders in den Bereichen mit hoher Auflösung bekommt man eine ausreichend genaue Informationen. Zu berücksichtigen ist natürlich der Maßstab, in dem man später arbeiten möchte.

Die Anzahl der Referenzpunkte richtet sich nach dem Inhalt des Images. Liegt eine flächenfüllende Karte vor, reichen mindestens 4 Punkte in den Ecken der Karte. Zu empfehlen sind Kontrollpunkte; bei einer Karte würde ich die doppelte Anzahl an Punkten wählen, um auch bei der Entzerrung sicher zu sein.

Bei Fotos, die kein Raster vorgegeben haben, muss man daher mehr Punkte wählen, die das Foto bis an die Ränder der Darstellung bearbeiten können oder beschneidet das Image auf die Größe, die benötigt wird. Die Punkte bezieht man aus Messungen vor Ort oder aus einer Karte oder anderen Medien (GOOGLE EARTH). Es müssen auch Punkte innerhalb des Images bestimmt werden, um die Kontrolle über das Bild zu erhalten. 4 Punkte sind da nicht ausreichend, ich empfehle die doppelte Anzahl. Das richtet sich nach der Geometrie des Fotos.

Wenn verfügbar, sollten digitale Orthophotos bzw. Luftbildkarten an Stelle von Luftbildern verwendet werden, da diese bereits auf ein digitales Geländemodell entzerrt sind. Bei ebenem Gelände kann eine Entzerrung gemäß Tabelle 1 vorgenommen werden, ohne dass dabei größere Lagegenauigkeiten hervorgerufen werden. Durch eine Entzerrung von Luftbildern von unebenem Gelände mit polynomialem Ansätzen höherer Ordnung können keine Unstetigkeiten im Gelände wie beispielsweise steile Felshänge o.ä. berücksichtigt werden. Hier ist eine Entzerrung auf ein digitales Geländemodell vorzunehmen (s. Tab. 1).

Natürlich mangelt es auch hier nicht an Programmen, die diese Arbeiten durchführen können. Jedes GIS Programm hat ein Modul für die Georeferenzierung von Rasterdaten. Die Ergebnisse sind untereinander austauschbar, da das gleiche Dateiformat benutzt werden kann. Man kann also ein „world-tiff“ oder „jpg“, in ein anderes GIS Programm exportieren.

AutoCad Map oder der AutoCad Aufsatz PhotoToPlan bieten auch die Möglichkeit der Affinen Transformation oder sogar der Polynominalen Entzerrung. Leider sind die neu entstandenen Rasterdaten nur in diesen Programmen verwendbar. Beim Export dieser Rasterdaten in GIS Programme müssen sie wieder neu referenziert werden.

Auch bei Anwendung der Digitalen Bildentzerrung bei der Bauaufnahme oder bei Grabungsflächen oder Grabungsprofilen kann eine Georeferenzierung durchgeführt werden. Die Referenzpunkte müssen eingemessen werden, wobei darauf zu achten ist, dass die Digitale Bildentzerrung nur 2-D einsetzbar ist. Um Strukturen mit unterschiedlicher Höhe mit diesem Verfahren zu bearbeiten, ist es notwendig, die Ebenen separat zu definieren und zu entzerren. PhotoToPlan erlaubt die Montage eines Bildverbands, solange genügend verbindende Passpunkte vorhanden sind.

Die Tabelle mit der Übersicht der unterschiedlichen Entzerrungs- und Referenzierungsmethoden entnehmen Sie bitte Seite 28 des IT-Leitfadens des DAIs (http://www.dainst.org/sites/default/files/medien/de/IT-Leitfaden_Teil2_Praxisratgeber_v1_0_0-ENTWURF.pdf?ft=all)

Quellen

Alle Handbücher der GIS Programme, AutoCad Map, PhotoToPlan etc.

<http://www.giub.uni-bonn.de/gistutor/>

4.3 Oberflächen- (DOM) und Geländemodellierung (DGM)

K. Lambers, M. Reindel

Bei Geländemodellen ist zwischen primären und sekundären Geländemodellen zu unterscheiden:

- a) primäre Geländemodelle

Aus Messung entstandene Geländemodelle sind über die Vermessungsdaten zu speichern.

- b) sekundäre Geländemodelle

Aus primären Geländemodellen abgeleitete Geländemodelle (z.B. erworbene Geländemodelle) sind als Rasterdaten im TIFF-Format abzuspeichern.

Eine kurze Einführung

Um Oberflächen maßstabsgerecht darstellen zu können, müssen sie vorher durch geeignete Vermessungen aufgenommen werden. Dies kann durch Handvermessung geschehen oder aber durch den Einsatz moderner Messinstrumente. Je nach Art und Größe des Untersuchungsobjektes kommen unterschiedliche Methoden und Techniken zur Anwendung. Die Dokumentation der Oberflächen aller dieser Objekte erfolgt prinzipiell nach den gleichen Methoden. Die folgenden Ausführungen konzentrieren sich auf die Dokumentation und Darstellung von Geländeoberflächen.

Als Messinstrumente kommen für mittlere und große Objekte Theodolithe bzw. Tachymeter (mit elektronischer Distanzmessung) zum Einsatz. Moderne Messverfahren bedienen sich auch der Photogrammetrie (Vermessung in stereoskopisch aufgenommenen Bildern) und neuerdings des Laserscannings oder

LIDAR (engl. light detection and ranging, s. u.). Allen Aufnahmeverfahren ist gemeinsam, dass sie Lagekoordinaten der vermessenen Objekte in einem zuvor definierten Vermessungssystem liefern. Für eine planimetrische Darstellung (ohne Höhenangaben) reichen Messwerte in X- und Y-Richtung aus. Sie ermöglichen zweidimensionale Darstellungen (2D). Will man jedoch die Oberfläche eines Objektes oder von Gelände räumlich darstellen, benötigt man auch die Höheninformation, das heißt einen Z-Wert. Die Grundlage einer räumlichen Darstellung einer Geländeoberfläche ist somit eine Sammlung von Punkten mit X-, Y- und Höheninformation in einem Koordinatensystem.

Im Gegensatz zu einer echten dreidimensionalen Darstellung (3D), in der X-, Y- und Höhenkoordinaten zu einem Volumenmodell verarbeitet werden, werden in sogenannten 2,5D-Darstellungen 2D-Daten mit einer zusätzlichen Höheninformationen versehen, die nicht als Koordinate, sondern als Attribut gespeichert wird. Dies bedeutet, dass jedem XY-Punkt nur ein Höhenwert Z zugeordnet werden kann, was die Modellierung von Senkrechten oder Überhängen ausschließt. Geoinformationssysteme, die ursprünglich für großflächige geographische Anwendungen entwickelt wurden, verarbeiten bis heute in der Regel nur 2,5D-Daten, während in CAD-Anwendungen, die für die Objektmodellierung entwickelt wurden, 3D-Informationen verwendet werden. Die Entwicklung geht jedoch in Richtung einer Konvergenz beider Systeme, so dass in Zukunft voll 3D-fähige GIS zum Standard werden dürften.

Vermessung eines Grabungsgeländes Die Aufnahme von Messdaten mit dem Ziel, Gelände zu modellieren, spielt in der archäologischen Praxis zum einen bei der Dokumentation von Grabungsorten vor dem Beginn oder während der Ausgrabungen und zum anderen bei der Durchführung von Regionalstudien mit Hilfe von Geoinformationssystemen eine Rolle. Grabungsgelände werden in der Regel durch tachymetrische Vermessung aufgenommen, während die Topographie größerer Gebiete durch erworbene digitalen Kartendaten, Fernerkundungsdaten oder von Luftbildern erarbeitet wird.

Der erster Schritt bei der Vermessung eines Grabungsgeländes ist die Anlage eines Messnetzes. Dieses wird zumeist genordet, oft aber auch an den zu erwartenden Grabungsbefunden ausgerichtet. In der Regel wird zunächst ein Ausgangspunkt und eine Grundlinie angelegt und von dort das Messnetz aufgebaut. In anderen Fällen wird von dem Ausgangspunkt ein Polygonzug angelegt, von dem aus weitere Geländepunkte eingemessen werden.

Das Messnetz sollte an ein bestehendes übergeordnetes Messnetz angebunden werden. Bezugspunkte können vermarkte Messpunkte eines Landesvermessungsnetzes oder in abgelegeneren Regionen Triangulationspunkte von Kartierungsprojekten oder ähnliches sein. Die Koordinaten der Hauptmesspunkte können auch durch präzise GPS-Messungen bestimmt werden. Häufig werden bei der Anlage von Grabungen willkürliche Messpunkte festgelegt, die durch spätere Messungen an die übergeordneten Messnetze angebunden werden (siehe hierzu auch Kapitel 2).

Ausgehend von den Basispunkten des Messnetzes werden mit dem Tachymeter Geländepunkte aufgemessen. Bei der Auswahl der Punkte ist es wichtig sich zu vergegenwärtigen, welche Art und welcher Maßstab der Darstellung des Geländes angestrebt wird. Grundsätzlich sollten diejenigen Punkte aufgenommen werden, an denen sich die Geländeneigung oder der Verlauf einer gedachten

Höhenlinie (Linie, die sich aus Punkten gleicher Höhe zusammensetzt) ändert. Das heißt, je unregelmäßiger die Geländeoberfläche ist und je detaillierter die Geländeoberfläche wiedergegeben werden soll, umso mehr Punkte pro gegebener Fläche sind aufzunehmen.

Die Punkte sind so dicht zu setzen, dass die relevanten Geländeformen in der graphischen Darstellung im angestrebten Maßstab wiedergegeben werden. Es ist einleuchtend, dass ein etwa 30 cm breiter Mauerzug, der sich im Gelände durch eine leichte Erhebung abzeichnet, nicht durch Punkte abgebildet werden kann, die im Abstand von 1 m gemessen wurden. Vielmehr muss zumindest der Beginn der Geländewölbung, der Scheitelpunkt über der Mauer und der Endpunkt der Geländewölbung vermessen werden. Bruchkanten müssen gesondert vermessen werden. Sie können bei der Auswertung der Messdaten mit Computerprogrammen speziell behandelt werden.

Regionale Geländemodelle Ist das Untersuchungsgebiet größer als ein Grabungsgelände, werden dafür benötigte Geländemodelle üblicherweise nicht selbst erzeugt, sondern aus anderen Quellen erworben. Bezugsquellen für digitale Geländemodelle sind je nach Land staatliche Vermessungsämter, Universitäts- oder Forschungsinstitute für Geographie, Geodäsie, Kartographie oder Geophysik, Luft- und Raumfahrtagenturen oder das Militär.

Grundlage für die Erzeugung von Geländemodellen sind Fernerkundungsdaten, die entweder mit passiven Sensoren (Luft- und Satellitenbilder) oder mit aktiven Sensoren (Radar- oder Lidardaten) aufgenommen werden (siehe hierzu die Beiträge von Sabine Reinhold). Bilddaten werden mit photogrammetrischen Methoden ausgewertet. In überlappenden Bereichen benachbarter Bilder werden stereoskopisch homologe Punkte an der Erdoberfläche gemessen (engl. matching). Entweder werden ähnlich wie bei der terrestrischen Geländebevermessung gezielt Einzelpunkte und ggf. Bruchkanten gemessen, oder es werden bei automatisierten Verfahren Punkte in gleichmässigen Abständen gemessen. Photogrammetrische Auswertungen von Luftbildern bilden bis heute in vielen Ländern die Grundlage für die Produktion und Nachführung von topographischen Karten und für die Generierung von Geländemodellen.

Radar- und Lidardaten liefern demgegenüber direkte Informationen über die Distanz der gemessenen Punkte zum Sensor. Ein auf Radar (engl. radio detection and ranging) beruhendes Geländemodell der Shuttle Radar Topography Mission (SRTM) ist praktisch weltweit frei verfügbar und in vielen Gegenden die einfachste Möglichkeit, digitale Geländemodelle zu beziehen. Die geringe Auflösung von 90 m sowie Lücken und große Höhenfehler in gebirgigem Gelände sind jedoch limitierende Faktoren für die Verwendbarkeit. Vermessungen mit flugzeuggestütztem LIDAR (engl. light detection and ranging), bei dem die Erdoberfläche in hoher Dichte durch einen Laserstrahl gescannt wird, werden in Europa immer mehr zu einem Standardverfahren der Landesvermessung, sind jedoch in anderen Ländern oft noch nicht verfügbar.

Zur korrekten Georeferenzierung der Fernerkundungsdaten dienen GPS-Messungen, oft in einer Kombination von Positionsmeßungen an Bord der Sensorplattform (z.B. GPS an Bord eines Flugzeuges gekoppelt mit INS, engl. inertial navigation system) mit Einmessungen von Kontrollpunkten am Boden, die in den Daten erkennbar sein müssen. Die GPS-Messungen dienen der Überführung der gemessenen Punkte von einem lokalen in ein globales Koordinatensystem (Ge-

oreferenzierung).

Erworbenen Geländemodelle sind rechnerische Ableitungen von tatsächlich erhobenen Messdaten. Wie schon bei anderen Messdaten, sollten nach Möglichkeit auch die Rohdaten gespeichert werden. Beim Kauf von Standard-Produkten ist dies üblicherweise nicht möglich, wohl aber bei der Beauftragung von speziellen Geländemodellen, z.B. Lidar-Überfliegungen.

Modellierung Zur Modellierung einer durchgehenden Oberfläche auf der Grundlage der vermessenen Punkte müssen die zwischen den Messpunkten liegenden Geländepunkte durch rechnerische Verfahren verknüpft werden. Dazu werden zunächst die Einzelpunkte zu einer aus Dreiecken bestehenden Oberfläche vermascht. Diese liegt im Vektorformat vor (engl. TIN: triangulated irregular network). Durch Interpolation werden die Vektordaten ggfs. in ein Rasterformat (z.B. Grid) mit einer festen Maschenweite überführt. Hierzu werden verschiedene mathematische Verfahren angewandt. Die Einzelheiten dieser Verfahren werden in den relevanten Computerprogrammen und Lehrbüchern erläutert, müssen jedoch vom Archäologen nicht im Detail nachvollzogen werden können. Wichtig ist jedoch, die Modellierungen mit den verschiedenen Verfahren zu testen, um entscheiden zu können, welche Methode die tatsächliche Geländeoberfläche am realistischsten darstellt. Die Maschenweite oder Zellengröße ist ein übliches Maß für die Auflösung eines Geländemodells. Während einige GIS-Programme (z.B. ArcGIS) auch TINs verarbeiten können, können viele andere nur mit Raster-Geländemodellen umgehen.

Je nach Platzierung der zu Grunde liegenden Messpunkte bilden Geländemodelle verschiedene Dinge ab. Als Oberbegriff dient die Bezeichnung Digitales Höhenmodell / DHM (engl. digital elevation model / DEM). Man unterscheidet zwei Typen von DHMs: 1. das Digitale Oberflächenmodell / DOM (engl. digital surface model / DSM), das die Erdoberfläche inklusive aller darauf befindlichen Objekte wie Häuser, Vegetation, etc. abbildet, und 2. das Digitale Geländemodell / DGM (engl. digital terrain model / DTM), das nur die Erdoberfläche abbildet. Bei der eigenen Vermessung des Grabungsgeländes wird man üblicherweise ein DGM erzeugen. Auch bei erworbenen Geländemodellen sind DGMs den DOMs vorzuziehen, da in archäologischen Untersuchungen üblicherweise die Geländegestalt ohne moderne Bebauung und Vegetation interessiert. Bei Geländemodellen mit geringer Auflösung wie dem SRTM-DHM sind die Unterschiede jedoch zu vernachlässigen.

Einsatz von Hard- und Software in der Praxis

Für die Verarbeitung und die Interpolation der Messpunkte stehen unterschiedliche Computer-Programme zur Verfügung, darunter auch speziell auf die Messgeräte angepasste Software der Hersteller von Vermessungsgeräten (Zeiss, Leica, Wild, Sokkia, Trimble, etc.). In der archäologischen Praxis wird zur Verarbeitung und Umzeichnung von echten 3D-Messdaten zumeist das Programm AutoCAD verwendet (<http://www.autodesk.de>). Zur Generierung von Höhenlinien und 3D-Modellen können Zusatzmodule oder erweiterte Versionen erworben werden. Zur Modellierung von 2,5D-Geländeoberflächen kann auch jedes GIS-Programm verwendet werden. Im Rahmen unserer Arbeiten wurden die Höhenlinien und Geländemodelle in der Regel mit dem Programm Surfer (<http://www.ssg-surfer.com/>) generiert. Surfer ist einfach zu handhaben und

bietet eine Vielzahl von Möglichkeiten der Verarbeitung und Analyse archäologischer Daten. Unter anderem können auch Daten geophysikalischen Prospektions einfach eingelesen und über die Geländemessdaten projiziert werden. Surfer ermöglicht auch die schnelle und einfache Überprüfung der Messergebnisse nach der Übertragung der Daten auf den Computer, wenn der Tachymeter nicht ohnehin mit einem Bildschirm ausgestattet ist. Die Koordinatenlisten lassen sich nachbearbeiten, Bruchkanten einfügen und nicht benötigte Messpunkte ausblenden. Die Höhenlinienpläne können als DXF-Dateien in AutoCAD exportiert werden. Schattierte oder farblich gestaltete Geländemodelle lassen sich in verschiedenen Graphikformaten (tiff, jpeg, bmp, etc.) exportieren und als solche ebenfalls in AutoCAD einbinden.

Erworbenen Geländemodelle liegen üblicherweise in einem Rasterformat vor, in dem einer durch eine XY-Eckkoordinate definierten Zelle mit einer festgelegten Seitenlänge ein Höhenwert zugeordnet wird. Damit gleichen Raster-Geländemodelle im Prinzip Bilddaten, nur dass statt eines Farbwertes ein Höhenwert gespeichert wird. Deshalb können Geländemodelle z.B. auch als TIFF vorliegen. Ein anderes übliches Format ist ASCII; daneben gibt es proprietäre Formate wie z.B. ESRI Grid. Die Transformation von Raster-Geländemodellen von einem Format in ein anderes in einem GIS-Programm stellt üblicherweise kein unüberwindbares Problem dar, so dass hier kein Format empfohlen wird, zumal bei Standard-Produkten das Format oft bereits festgelegt ist.

Je nach gewähltem Ausgabemedium werden verschiedene Verfahren der Darstellung der Geländeoberflächen gewählt. Soll das Ergebnis auf Papier, also in 2D auf einem Plan als Arbeitsdokument oder für die Publikation ausgegeben werden, wird die Darstellung in Form von Höhenlinien, mit Schattierungen oder unterschiedlicher Farbgebung gewählt. AutoCAD-Dateien (*.dwg) lassen sich häufig nur mit Schwierigkeiten in andere vektorbasierte Programme (z.B. CorelDraw oder Adobe Illustrator) zur weiteren Bearbeitung exportieren. Für die Einbindung in Publikationen können AutoCAD-Pläne und -Zeichnungen jedoch als PDF-Dateien ausgegeben, in JPEG-Dateien umgewandelt und mit Bildbearbeitungsprogrammen ergänzt oder verändert werden. Am Computer können die räumlichen Modelle dargestellt, bewegt und von unterschiedlichen Standpunkten aus betrachtet werden.

GIS-Programme bieten die Möglichkeit, die in Geländemodellen enthaltene Höheninformation in verschiedener Form in 2D darzustellen, z.B. als Höhenlinien, Grauwerte oder Farbabstufungen. Das Layout von Karten ist in den verschiedenen Programmen unterschiedlich einfach zu handhaben, jedoch sind die grundlegenden Funktionen wie die automatische Platzierung eines Koordinatenrahmens, eines Maßstabes etc. überall vorhanden, so dass Karten im GIS-Programm erzeugt und entweder direkt ausgedruckt oder in ein Standard-Graphikformat (Vektor oder Raster) exportiert werden können.

Quellen

Archaeology Data Service 2007: Guides to Good Practice.

<http://ads.ahds.ac.uk/project/goodguides/g2gp.html>

Bill, Ralf 1999: Grundlagen der Geo-Informationssysteme, Band 2: Analysen, Anwendungen und neue Entwicklungen. 2. Aufl. Heidelberg: Wichmann.

Kapitel 2.5: Digitales Geländemodell (DGM).

Conolly, James & Mark Lake 2006: Geographical Information Systems in

- Archaeology (Cambridge Manuals in Archaeology). Cambridge: Cambridge University Press. Kapitel 6: Building Surface Models.
- Howard, Phil 2007: Archaeological Surveying and Mapping: Recording and Depicting the Landscape. London: Routledge.
- Wheatley, David & Mark Gillings 2002: Spatial Technology and Archaeology: The Archaeological Applications of GIS. London: Taylor & Francis. Kapitel 5: Digital Elevation Models.

4.4 Rekonstruktion

A. Müller, U. Wulf-Rheidt

3D-Modelle

Einfache 3D-Modelle sind relativ schnell angelegt und können zur visuellen Kommunikation einen erheblichen Beitrag leisten. Nach Auswertung erster Forschungs- oder Grabungsergebnisse entstehen oft erste, vereinfachte Massenmodelle, die z. B. der Erläuterung von Rekonstruktionsideen dienen können. Es ist dann kein Problem, die erstellten Volumina in georeferenzierte Systeme, wie beispielsweise Google Earth zu integrieren oder in präzisere Geländemodelle einzubauen. Im Verlaufe des Projektes wird das 3D-Modell in der Regel immer differenzierter ausgearbeitet und es werden mehr Datensätze und Informationen angehängt, in dem einzelne Gebäudeteile und Räume differenziert dargestellt, Bauteile ausgearbeitet, u.U. Bilder, Geodaten, Befund-Daten und Literaturhinweise angehängt und schließlich die zur Visualisierung notwendigen Texturen und Lichtquellen sowie dynamische und statische Schnitte generiert werden. Derart komplexe Aufgaben erfordern eine gut durchdachte Datenverwaltung, da ihr Verwaltungs- und Pflegeaufwand exponentiell zur Datenmenge anwächst. Deshalb ist es wichtig, sich vor Anlegen eines 3D-Modells über die Ziele, die verwirklicht werden sollen, klar zu werden. Wenn z.B. ein Modell nur benötigt wird, um einige Perspektiven für eine Publikation zu visualisieren, genügt es, die Kubatur des Gebäudes zu erstellen, sie zu texturieren, Lichtquellen zu setzen und Ansichtsdefinitionen festzulegen. Dementsprechend gering bleibt der Sicherungs- und Dokumentationsaufwand. Je genauer der Auftraggeber eines 3D-Modells die Ziele, zur Verwendung des Modells definiert, desto geringer kann der Aufwand der Datenerstellung und Datenhaltung gehalten werden.

Deshalb sind Mindestanforderungen bei der Auftragsvergabe nur auf sehr niedrigem Niveau möglich. Zurzeit beschränken sich diese Forderungen in der Regel auf die Sicherung in einem gut dokumentierten Format wie z.B. dxf oder Wavefront OBJ-Dateien, mit allen Nachteilen, die sich dabei aus einer eventuellen Datenkonvertierung ergeben.

Praxis

Es ist derzeit verbreitet, die Wahl der Software dem Zufall zu überlassen: Softwarekenntnisse des Bearbeiters, das Vorhandensein von Software oder die Frage, welche Software preiswert zu erwerben ist, bestimmen meist die Wahl.

Das bedeutet zwar nicht, dass eine spätere Entscheidung für eine andere Software zu totalem Datenverlust führen muss, denn grundsätzlich ist die Kon-

vertierung beinahe jedes Zeichenformates in ein anderes möglich. Dies ist jedoch häufig mit erheblichen Informationsverlusten oder/und gewaltigem Arbeits- und Kostenaufwand verbunden.

Wurde ein Modell beispielsweise in einer Visualisierungssoftware wie Rhinoceros, Maya oder 3ds Max erstellt, so haben die Objekte hauptsächlich Metainformationen hinsichtlich Geometrie und Textur. Mit dieser Software sind sehr hochwertige Bilder und Animationen generierbar, die Verwaltung hinsichtlich bautechnisch sinnvoller Einheiten wie Wände und Fenster ist dagegen nicht vorgesehen und ist deshalb schwierig, zeitaufwändig und umständlich.

Architektursoftware bietet erhebliche Vorteile hinsichtlich der Datenverwaltung, da Objekte von vornherein in üblichen Klassifizierungen, wie Wänden, Decken, Stützen, Böden und Öffnungen usw. verarbeitet werden können. Die Eingabe ist aufwändig, da sie parametrisch erfolgt. Sie bietet aber den Vorteil der Nachvollziehbarkeit. Außerdem besteht die Möglichkeit, zwischen verschiedenen Maßstäben hin- und herzuschalten und standardmäßig 2D-Schnitte und -Grundrisse darzustellen bzw. zu exportieren. Nachteile liegen vor allem im Preis der Anschaffung und der Notwendigkeit regelmäßig upzudaten, verlustfreier Datentransfer ist - wenn überhaupt - meist nur innerhalb der Produktgruppe des selben Software-Entwicklers möglich. Standardisierte Objekte wie Wände sind starr definiert und können nicht frei modelliert werden, eine verformungsgerechte Darstellung ist damit nicht möglich. Allerdings ist in den meisten Fällen eine verformungsgerechte Nachbildung des Baukörpers im Modell nicht nötig und verkompliziert unnötig die Geometrie und vergrößert die Datenbanken.

Offene Systeme wie AutoCAD, VectorWorks und MicroStation bieten viele Vorteile hinsichtlich der Konvertierbarkeit und haben Schnittstellen zu beinahe jeder Art von Software. Sie ermöglichen jedoch meist keine normierten Klassifizierungen (Wände, Türen etc), es sei denn, es wird eine spezifische Applikation (z.B. für Architektur oder GIS) verwendet. Diese Vorgabe muss daher vor Anlegen des 3D-Modells festgelegt werden. Diese Systeme haben neben den genannten Vorteilen vor allem den Vorzug, weit verbreitet, gut dokumentiert und als Schulungsversionen relativ kostengünstig zu sein. Nachteile liegen in der Datenverwaltung und dem Bearbeitungskomfort.

Ein wichtiges Kriterium stellt ein gut dokumentiertes und leicht lesbares Format dar, das sämtliche Erfordernisse des Projekts mit einem Datenformat erfüllt. Ein gut dokumentiertes Format ist dabei unbedingt einem proprietären Binärformat, welches außer von einer kleinen Community nicht genutzt wird, vorzuziehen.

Wegen ihrer Verbreitung und den ubiquitär verfügbaren Importern sollte als erstes die Verwendung der ASCII-Version von DXF oder eines reinen Wavefront OBJ-Files in Betracht gezogen werden.

Zudem sollte das Datenformat rekursives Laden von Daten ermöglichen, also die Zerlegung eines Modells in mehrere Dateien, die dann wiederum an ein datenbankunabhängiges Bauwerksinformationssystem angehängt werden können.

Auswahl geeigneter Software Bei den 3D-Architekturmodellen im archäologischen Kontext ist es sinnvoll, drei Kategorien zu unterscheiden:

- 1. Arbeitsmodelle
- 2. Visualisierungsmodelle

- 3. Interaktive Modelle

Arbeitsmodelle müssen maßstäblich und ggfls. auch georeferenziert sein. Dabei können sehr unterschiedliche Fragestellungen im Vordergrund stehen, z. B. stadträumliche oder topografische Beziehungen, Massenermittlungen, bei Einzelbefundrekonstruktionen aufgrund eines Versturzes z.B auch Lagebestimmungen, statische und konstruktive Bedingungen, bis hin zu einfachen Licht- und Schatten- Untersuchungen. Aus solchen Modellen sollten maßstäbliche und vermaßbare technische Pläne generierbar sein, die der Kommunikation exakt nachvollziehbarer Zusammenhänge wie Höhen, Längen, Breiten und Massen dienen. Die Geometrie kann hierbei stark vereinfacht sein, grundsätzlich gilt dabei, dass Genauigkeit und Nachvollziehbarkeit vor Schönheit stehen sollte. Je nach gewünschter (mess- und abfragbarer) Informationsdichte des Modells sollte sorgfältig überlegt werden, welche Software bei welchem Arbeitsaufwand die gewünschten Standards bietet. Je nach Komplexität der angestrebten Informationsdichte besonders bei Rekonstruktionsmodellen ist zusätzlich zu beachten, dass auch der Zeichner zumindest architektonische Vorkenntnisse haben sollte.

Geeignete Software sind vor allem Architekturprogramme, wie z.B. Archicad, AutocadArchitectural, Nemetschek, Spirit usw. aber auch offene Systeme wie Microstation, Vectorworks und AutoCad. Architekturprogramme bringen logische Klassifizierungen, wie Wände, Stützen, Unterzüge, Treppen, Fenster, Dächer etc. mit sich, die Eingabe erfolgt dabei parametrisch und alle Parameter sind in Projekt-Datenbanken erfasst. Ob solche Datenbanken nur programmintern abgefragt oder wie sie in geeignete übergreifende Systeme exportiert werden können, muss sorgfältig geklärt werden.

Mit allen diesen Programmen lassen sich auch mehr oder weniger realistische Bilder generieren, häufig jedoch nur von mäßiger Qualität. Die Modellierung hochkomplexer geometrischer Körper, wie z. B. eines korinthischen Kapitells, ist jedoch kaum möglich.

Bei Visualisierungsmodellen ist vor allem Art und Qualität der angestrebten Visualisierung zunächst zu klären. Dabei gilt „je höher die Qualität, desto höher der Arbeitsaufwand und die aufzuwendenden Kosten. Von einfachen Modellen ohne Textur bis hin zu fotorealistischen Bildern ist heute alles möglich. Für fotorealistische Bilder, wenn diese nicht als animierte Szenen gewünscht werden, kann auch überlegt werden, ob eine Kombination von Einzelbildsynthese und Bildmontage in einem Bildbearbeitungsprogramm (z.B. Photoshop) einfacher und schneller zum Ergebnis führt.

Zur Anwendung können sehr einfache, z.T. auch Freewareprogramme kommen, wenn das Interesse in erster Linie der Geometrie gilt und diese eine überschaubare Polygon - Anzahl umfasst. Sollen anspruchsvolle, realistische oder gar fotorealistische Szenen erstellt werden, ist die Wahl einer professionellen Visualisierungs- Software ratsam. Da Visualisierungsprogramme in erster Linie der Erstellung maßstabloser, möglichst ansprechender Bilder dienen, ist präzises, maßstäbliches Zeichnen nur bedingt möglich. Die Definition der Körper erfolgt bei den einfacheren Programmen über ihre absolute Geometrie, bei aufwändigen Programmen wie Rhino, Cinema4d, MAYA oder 3dsMAX über Parameter für die Ausgangsgeometrie und ggfls. verschiedene Modifikatoren für Geometrie und Animation.

In Visualisierungsprogrammen ist - mit entsprechendem Aufwand - die Herstellung jeder denkbaren Qualität von Bildern, sowohl hinsichtlich der Auflösung

wie der Detailqualität und des Realismus, möglich. Außerdem können Filmszenen generiert werden, die z.B. für die Visualisierung der Funktionsweise von Maschinen, für dynamische Schnitte oder die Darstellung von Bauabläufen sehr sinnvoll sein können.

Da das Erstellen fotorealistischer Bilder mit 3d-Software hochkomplex ist, ist eine Verknüpfung mit bau-, geo- oder grabungsspezifischen Datenbanken nicht sinnvoll. Es gilt ganz allgemein der Satz: Schönheit und/oder visuelle Aussage vor Genauigkeit. Solche Modelle dienen allein der visuellen Kommunikation über Standbilder und Filme.

Interaktive Modelle können mit Datenbanken verknüpft werden. Diese sind ohne Programmierkenntnisse bislang kaum sinnvoll erstellbar. Es stellt sich dabei - abgesehen von der Qualität der verknüpften Datenbanken und ihren Fra gestellungen - vor allem das Problem der Ladezeiten. Deshalb sind besondere Render- Engines unverzichtbar, die Geometrie und Texturen mit intelligenten Algorithmen reduzieren, je nachdem, ob sich diese im Vorder-, Mittel- oder Hintergrund befinden.

In der Regel werden solche Modelle von einem Team aus Zeichnern und Programmierern erstellt. Wegen der sehr spezifischen Anforderungen an die Benutzeroberflächen und den Workflow ist auch an Personen mit Kenntnissen in der Erstellung von 3d-Spielen und/oder Webdesigner zu denken.

Allgemein gilt die Prämisse: Workflow, einfache Steuerung, Übersichtlichkeit und intuitive Verständlichkeit der Oberfläche. Komplexe Inhalte können über Links abgefragt werden und sollten nicht auf der Oberfläche platziert sein.

Auswahl der Hardware Bei der Erstellung von 3d- Modellen sind naturgemäß sehr hohe Anforderungen an die grafische Rechenleistung zu stellen. Aufwändige Visualisierungsarbeiten erfordern eine entsprechend gute Ausstattung mit Hard- und Software, die teuer ist. Beim Echtzeit- Rendering im virtuellen 3d- Raum müssen riesige Datenmengen in sehr kurzer Zeit bearbeitet werden. Dabei sind Einzelplatz- Rechner meist auf die Render- Parameter, wie sie für eine gute Performance in 3d- Spielen erforderlich sind, ausgelegt. Sie sind für die Produktion hochwertig texturierter Einzelbilder oder Filme nicht geeignet. Man spricht heutzutage von Render- Farmen und Netzwerk- Rendering. Das bedeutet, das Rechenleistungen auf viele Prozessoren in einem Netzwerk verteilt werden können. Dabei stellt sich die Frage, ob solche Konstellationen vom Administrator zu gewährleisten sind. Außerdem muss die Software Netzwerkrendering unterstützen. Multiple Grafikprozessoren, Multikern- Hauptprozessoren und adäquat ausgelegte Arbeitsspeicher sind sinnvoll (z.B. 128 GB). Dabei ist darauf zu achten, dass die Hardware auch vom Betriebssystem unterstützt wird (speziell beim Arbeitsspeicher: Windows hat hier bislang die größten Nachteile) und von der eingesetzten 3d- Software gleichermaßen sinnvoll verwaltet werden kann. Riesige Festplatten sind nur bei der Erstellung von Filmen notwendig. Hier empfiehlt sich die Auslagerung auf externe Medien. Wegen der umfangreichen Paletten ist ein großer Monitor mit zusätzlichem Kontrollmonitor unverzichtbar.

Quellen

Da Software- und Hardwaremärkte sehr dynamisch sind, ist es empfehlenswert, sich in aktuellen, Zeitschriften zu informieren, z.B.

<http://utils.exeter.ac.uk/german/media/archited.html> und Messen, wie die Bautech zu besuchen.

Außerdem kann man sich auf Wikipedia.de eine gute Orientierung über die diversen Probleme bei der Erstellung von 3d- Modellen verschaffen. Dort finden sich auch Links zu weiterführender Literatur die i.d. Regel recht aktuell sind: <http://de.wikipedia.org/wiki/Kategorie:Bildsynthese>, sehr empfehlenswert ist auch der Artikel:

<http://de.wikipedia.org/wiki/Grafik-Engine>

Viele Software- Hersteller unterhalten eigene Foren oder verweisen zumindest auf ihren Seiten auf User- Foren, z.B.: <http://www.autodesk.de/>.

Gerade im Bereich der 3d-Modellierung gibt es viele einschlägige Foren, z.B.: 3D-Ring:

<http://www.3d-ring.de/forum.php>

3Dlinks:

<http://www.3dlinks.com/>

Grafiker:

<http://www.grafiker.de/forum/forumdisplay.php?f=39>

3Dcenter:

<http://www.forum-3dcenter.org/vbulletin/>

3Dcafe:

<http://www.3dcafe.com/>

Zur spezifischen Problemstellung interaktiver oder Web- basierter Modelle:

ki-smile:

<http://www.ki-smile.de/kismile/view145,11,1424.html>

garybeene:

<http://www.garybeene.com/3d/3d-pure.htm>

Quality management systems - Requirements(Iso- Norm) Edition: 4 ; Stage: 60.60 ; TC 176/SC 2 ICS: 03.120.10 Beuth- Verlag,Berlin

nestor edition 5: Langzeiterhaltung von 3D-Röntgen-Computertomographien in der Archäologischen Denkmalpflege. Sayer-Degen, Leila (2012)
<http://files.dnb.de/nestor/edition/05-sayer-degen.pdf>

4.5 Reflectance Transformation Imaging (RTI)

M. Trognitz

Reflectance Transformation Imaging (RTI) ist eine computergestützte Fotografiermethode, mit der von einem Objekt mehrere Bilder mit fixierter Kameraposition und variablen Beleuchtungspositionen gemacht werden. Anschließend werden die Aufnahmen vom Computer zu einem Polynomial Texture Map (PTM) zusammengerechnet. In der resultierenden Datei kann die Position der Lichtquelle verändert werden, um beispielsweise die Oberfläche des aufgenommenen Objekts im Schräglicht untersuchen zu können.

RTI wurde erst 2001 entwickelt und hat inzwischen eine weite Verbreitung für die Dokumentation von Objekten mit flachen Oberflächen, wie etwa Felsbilder, Inschriften oder Münzen, gefunden, da hochauflösende Ergebnisse mit vergleichsweise günstiger Ausrüstung erzielt werden können.

Langzeitformate Die Aufnahmen für RTI erfolgen mit einem Fotoapparat, weshalb man es bei der Archivierung hauptsächlich mit Rastergrafiken zu tun hat. Ausführliche Hinweise zu geeigneten Langzeitformaten sind in dem Kapitel Rastergrafiken ab Seite 78 zu finden.

Zusätzlich müssen die resultierenden softwareabhängigen Dateien für die Archivierung berücksichtigt werden. Die Formate für RTI wurden in den Hewlett-Packard-Laboratories entwickelt und von Cultural Heritage Imaging (CHI) erweitert und ergänzt.

Ausgangsformat	Begründung
✓ DNG	Originale Aufnahmen in RAW, sollten als DNG gespeichert werden.
TIFF	Originale Aufnahmen, die nicht in RAW aufgenommen wurden, sollten als TIFF gespeichert werden. Vor der Prozessierung erheblich bearbeitete Aufnahmen (z.B. Freistellung des Objektes), sollten in einem gesonderten Ordner ebenfalls als TIFF gespeichert werden.

Prozessierungsformat	Begründung
✓ JPG	Programme zur Erzeugung von RTI-Dateien benötigen meist JPG-Dateien. Auch die von den Programmen erzeugten Zwischenbilder werden als JPG gespeichert. Für diese Dateien ist die Archivierung im JPG-Format zulässig.
XML	Alle Einstellungen werden in einer automatisch erzeugten XML-Datei gespeichert.
LP	In der textbasierten LP-Datei (light position file) werden die Positionen der Lichtquelle für jedes einzelne Bild gespeichert. Sie wird automatisch erzeugt, kann aber auch manuell erstellt werden.

Ergebnisformat	Begründung
PTM ✓	Das Format Polynomial Texture Map wurde in den Hewlett-Packard-Laboratories entwickelt und speichert für jedes Pixel eine Funktion, mit der die Farbwerte des Pixels in Abhängigkeit der Lichtquelle errechnet werden können.
RTI	Das RTI-Format wurde von CHI entwickelt und bietet im Vergleich zu PTM erweiterte Funktionalitäten im <i>RTIViewer</i> von CHI.
XMP	Der <i>RTIViewer</i> von CHI erlaubt das Annotieren der Dateien, um beispielsweise besondere Bildbereiche hervorzuheben. Diese Informationen werden in einer textbasierten XMP-Datei gespeichert.

Ordnerstruktur Zwar kann eine PTM- oder RTI-Datei ohne zusätzliche Dateien verwendet werden, jedoch sollte für die Nachnutzbarkeit und Nachvollziehbarkeit das gesamte Datenpaket archiviert werden. Dabei sollten alle Daten nicht alle in einem einzigen Ordner, sondern in einer sinnvollen Ordnerstruktur abgelegt werden. In der folgenden Tabelle wird eine solche Struktur vorgeschlagen und auch die entsprechenden Dateitypen darin aufgelistet.

Die Benennung der Ordner ist hier als Vorschlag zu betrachten. Dabei muss berücksichtigt werden, dass einige Ordner und Dateien von der Software erzeugt werden und diese auch so belassen werden sollten, um wieder mit dem Programm geöffnet werden zu können. Die Struktur orientiert sich an den Vorgaben von Cultural Heritage Imaging (CHI) und der von dem Programm *RTIBuilder* erzeugten Dateistruktur.

Die gesamten Aufnahmen sollten in einem übergeordneten Verzeichnis mit dem Projektnamen gespeichert werden. Für jedes aufgenommene Objekt muss ein eigener Unterordner angelegt werden, der idealerweise eine eindeutige Bezeichnung des Objektes, wie etwa die Inventarnummer, in der Benennung enthält. In dem übergeordneten Verzeichnis können Dokumentationsdateien abgelegt werden.

Ordnerstruktur und enthaltene Dateiformate

■ Projekt RTI-Aufnahmen
□ dokumentation
■ Inventarnummer 001
■ Inventarnummer 002
□ .xml
■ original-captures
□ .dng / .tiff
■ edited-captures
□ .tiff
■ jpeg-exports
□ .jpg
■ cropped-files
□ .jpg
■ assembly-files
□ .jpg
□ .lp
■ finished-files
□ .ptm
□ .rti
□ .xmp
■ Inventarnummer 003

In dem übergeordneten Verzeichnis sind alle Aufnahmen enthalten. Außerdem können hier auch Dokumentationsdateien abgelegt werden. Pro Aufnahme entsteht ein Ordner. Neben den verschiedenen Unterordnern enthält dieser auch eine Projektdatei im XML-Format.

'original-captures' enthält die unveränderten Originalaufnahmen im DNG- oder TIFF-Format.

'edited-captures' (optional) enthält überarbeitete Aufnahmen im TIFF-Format.

'jpeg-exports' enthält die Bilder, aus denen das Ergebnis erzeugt wird. Die JPG-Bilder werden entweder aus den Dateien in 'original-captures' oder in 'edited-captures' erzeugt.

'cropped-files' wird automatisch erzeugt. Er enthält JPG-Dateien, die bei der Verwendung der Zuschnittsfunktion als Zwischenergebnis gespeichert werden.

'assembly-files' wird automatisch erzeugt. Er enthält JPG-Dateien, welche die Ausleuchtung des Objektes repräsentieren und eine LP-Datei, in der die Position der Lichtquellen gespeichert werden.

'finished-files' wird automatisch erzeugt und enthält die resultierenden RTIs und PTMs. Werden die Dateien im RTIViewer annotiert, entstehen zusätzlich XMP-Dateien.

Dokumentation Neben den allgemeinen minimalen Angaben zu Einzeldateien, wie sie in dem Abschnitt Metadaten in der Anwendung ab Seite 36 beschrieben sind, erfordern RTI-Daten Angaben, die Aufschluss über die Aufnahmemethode, deren Umstände und über die Prozessierung der Daten geben. Zusätzlich können weitere Angaben zu dem Projekt als ganzes notwendig sein, um etwa genauere Angaben zu den aufgenommenen Objekten zu machen. Hinweise hierfür sind ebenfalls in dem Abschnitt Metadaten in der Anwendung ab Seite 36 beschrieben.

Speziell für die freihändige RTI-Aufnahme mittels Highlight-RTI gibt es von Cultural Heritage Imaging einen *Shooting-Log*, der in angepasster Form auch für andere Aufnahmemethoden verwendet werden kann. Eine Vorlage hierfür wird online in den IT-Empfehlungen zur Verfügung gestellt. Es empfiehlt sich diese Übersicht schon während der Aufnahme der einzelnen Objekte auszufüllen.

Metadatum	Beschreibung
Kamera und Objektiv	Modell und Name der verwendeten Kamera sowie Details zum verwendeten Objektiv.
Kamerasfilter	Auflistung aller für die Aufnahme verwendeten Filter, wie etwa Graufilter oder Infrarotfilter.
Aufnahmemethode	Angabe der Methode, die für die RTI-Aufnahme verwendet wurde. Beispielsweise feste Kuppel oder Highlight-RTI (H-RTI).
Radius oder Schnurlänge	Radius der verwendeten Kuppel oder des Armes. Wenn H-RTI angewandt wurde, muss die Länge der Schnur angegeben werden. Die Angabe erfolgt in cm.
Größe der Sphäre	Durchmesser in cm oder mm der verwendeten Sphäre, wenn H-RTI angewandt wurde.
Lichtquelle	Anzahl und Typ der Lichtquelle, sowie deren Einstellungen. Beispielsweise Canon Speedlite 580 EX II, 1/4 Leistung.
Projektordner	Name des Verzeichnisses in dem alle Aufnahmeordner gespeichert werden.
Verzeichnisname	Der Name des Verzeichnisses in dem die originalen Aufnahmen, prozessierten Bilder und Ergebnisdateien liegen.
Objektbezeichnung	Eindeutige Bezeichnung des Objektes, wie beispielsweise die Inventar- oder Fundnummer.
Beschreibung	Kurze Beschreibung des Objektes.
Farbkarte	Wenn eine Farbkarte verwendet wurde, soll angegeben werden, ob sie in jedem Bild vorhanden ist oder nur in einem separaten Bild. Wenn ein Farbprofil verwendet wurde, kann dieses hier angegeben werden.
Testbilder	Angabe der Dateinamen aller nicht für die Erstellung der RTI-Datei verwendeten Bilder, die jedoch weiterhin beibehalten werden, wie beispielsweise das Bild mit der Farbkarte.
Aufnahmeteam	Vor- und Nachnamen der Personen, die an der Aufnahme beteiligt waren.
Rechte	Lizenz der Aufnahme, z.B. CC-BY 3.0
Notizen	Weitere Angaben zum Aufnahmeaufbau und besonderen Vorkommnissen.
Software	Name und Version der zur Erzeugung der RTI-Daten verwendeten Software, beispielsweise <i>RTIBuilder</i> Version 2.0.2.
Anzahl der Bilder	Zahl der für das finale RTI verwendeten Bilder.
Weitere Dateien	Liste weiterer Dateien, wie beispielsweise des Shooting-Logs oder Dokumentation der verwendeten Software.

Vertiefung

Reflection Transformation Imaging (RTI) erzeugt Polynomial Texture Maps (PTMs), in denen für jedes Pixel Informationen über die Reflexionseigenschaften der Oberfläche gespeichert werden. Die erzeugten interaktiven Dateien werden aus einer Vielzahl von Bildern errechnet, wobei in jedem Bild die Kameraposition unverändert ist, während die Richtung, aus der das Licht kommt, jeweils wechselt.

Mit speziellen Viewer-Programmen können die Daten angesehen werden und die aufgenommenen Objekte interaktiv aus jeder Richtung beleuchtet werden. Da es sich aber nicht um eine 3D-Datei handelt, können die Objekte nicht bewegt werden. Jedoch bieten die Programme zusätzlich unterschiedliche Algorithmen, um bestimmte Oberflächeneigenschaften, wie Kanten oder Risse, optisch hervorzuheben, die unter natürlichen Bedingungen nicht oder nur schwer erkennbar sind.

Für die Erstellung von RTI-Daten können unterschiedliche Aufnahmemethoden verwendet werden, die unterschiedliche Ausrüstung voraussetzen, wobei das Grundprinzip dahinter jeweils immer das gleiche ist.



Abb. 4.1: Ausschnitt der Rückseite der Grabstele TA 1029 mit reichsaramäischer Inschrift des Verstorbenen (Tayma, Saudi-Arabien). Links ein Foto. In der Mitte der gleiche Ausschnitt als RTI mit 'Specular Enhancement', rechts mit 'Static Multi Light'. (Foto: Mirco Cusin; RTI: Martina Trognitz, Max Haib; DAI-Orient-Abteilung)

Polynomial Texture Map Ein Polynomial Texture Map (PTM) ist eine Repräsentationsform von Bildern mit Hilfe von Funktionen, statt einzelner Farbwerte. Im Gegensatz zu Rastergrafiken werden für die einzelnen Pixel von PTMs nicht nur feste Farbwerte gespeichert, sondern zusätzlich eine Funktion, die mit Hilfe der Parameter l_u und l_v die Leuchtdichte der Oberfläche berechnet. Die Leuchtdichte bestimmt wie das menschliche Auge eine Oberfläche wahrnimmt, also ob sie besonders hell, dunkel, spiegelnd oder matt erscheint.

Die Parameter l_u und l_v spezifizieren, wo sich eine punktförmige Lichtquelle befindet. Wird die Lichtquelle bewegt, ändern sich die Parameter und somit auch der errechnete Farbwert für den jeweiligen Pixel. Somit kann ein PTM simulieren, wie ein Objekt bei wechselnder Beleuchtungsrichtung aussieht.

Ganz vereinfacht lässt sich das Prinzip hinter RTI in der untenstehenden Abbildung visualisieren. Für jeden Punkt einer dreidimensionalen Oberfläche kann eine Normale bestimmt werden. An dieser Normale wird einfallendes Licht im gleichen Winkel reflektiert (Reflexionsgesetz). Da die Kamera sich in einer fest

ten Position befindet, die Richtung aus der das Licht kommt ebenfalls bekannt ist und es einen Bilderstapel mit unterschiedlichen Lichtpositionen gibt, kann für jeden Bildpunkt eine Oberflächennormale berechnet werden. Wird die fertige RTI-Datei betrachtet und die Beleuchtung geändert, können die angezeigten Farbwerte der einzelnen Punkte anhand des gespeicherten Farbwertes, der errechneten Normalen und dem Einfallswinkel berechnet und ausgegeben werden.

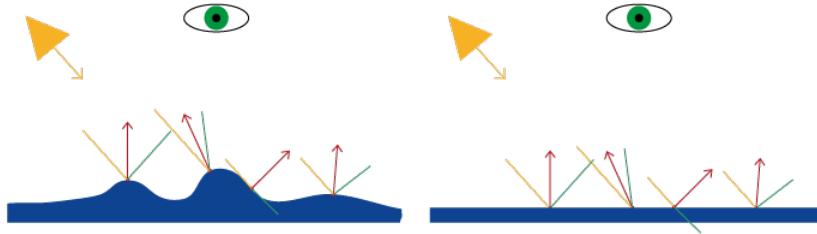


Abb. 4.2: Rechts sind auf einer dreidimensionalen Oberfläche für vier Punkte die Normalen (rot) eingezeichnet, an denen einfallendes Licht (gelb) reflektiert wird (grün). Links ist diese Oberfläche als PTM dargestellt, in dem für die Punkte jeweils die Oberflächennormalen gespeichert sind. So entsteht ein optischer 3D-Effekt, der an sich aber nicht in der Datei hinterlegt ist.

Aufnahmemethoden Das Grundprinzip für RTI-Aufnahmen besteht in der immer gleich bleibenden Position der Kamera in Bezug auf das Objekt und der wechselnden Position der Lichtquelle. Bei der Lichtquelle muss der Abstand zu dem Objekt immer der gleiche sein, um eine konstante Lichtintensität zu gewährleisten, und zusammen betrachtet sollten die jeweiligen Positionen so gewählt werden, dass sie gleichmäßig um das Objekt herum verteilt sind. Für die Positionierung der Lichtquellen wurden unterschiedliche Methoden und Ausrüstungsteile entwickelt, die in der nebenstehenden Abbildung zusammengefasst dargestellt sind.

Bei der Aufnahme mit einer Kuppel (engl. *dome*) wird eine Kuppel so über das Objekt platziert, dass dieses in der Mitte liegt. Die Kamera wird ebenfalls mittig der Kuppel direkt über dem Objekt platziert. An der Kuppel können mehrere Lichtquellen, wie etwa LEDs befestigt sein, wobei bei jedem Bild jeweils nur eine eingeschaltet wird. Es gibt auch Kuppeln, die statt fest eingebauter Lichtquellen nur Markie-



Abb. 4.3: Oben links eine Kuppel für RTI-Aufnahmen. Rechts ein gebogener Arm, der um das Objekt rotiert. Unten ist in den Aufbau von CHI eine imaginäre Kuppel projiziert. (Hewlett-Packard-Laboratories und CHI)

rungen für eine manuelle Platzierung der Lichtquelle haben. Die Position der Lichtquellen ergeben sich aus den Abmessungen der Kuppel und müssen in dem Programm eingetragen werden.

Eine vereinfachte Form der Kuppel sind gebogene Arme, mit daran befestigten Lichtquellen, wie etwa externe Blitzgeräte. Dieser Arm kann um das Objekt herum rotiert werden, so dass am Ende der Aufnahme eine kuppelförmige Verteilung der Lichter erzielt wird. Auch hier kann die Position der Lichtquellen aus den Abmessungen des Armes und dem Rotationswinkel berechnet werden.

Eine sehr flexible Methode, Highlight-RTI, wurde von Cultural Heritage Imaging entwickelt. Hierfür wird nur ein Blitzgerät (oder eine andere Lichtquelle) benötigt, an dem eine Schnur befestigt wird, mit deren Hilfe man das Gerät so ausrichten kann, dass der Abstand immer gleich ist und das Licht auch direkt auf das Objekt gerichtet ist. Auf diese Weise kann man sich an einer imaginären Kuppel entlang orientieren. Um die Position der Lichtquelle errechnen zu können, wird neben das Objekt eine schwarz glänzende Kugel platziert, die auch auf jedem Bild zu sehen sein muss. Aus den Reflexionspunkten des Lichtes an der Kugel kann das Programm dann die Richtung, aus der das Licht kommt, berechnen.

Praxis

In diesem Abschnitt werden Programme vorgestellt, mit denen RTI-Daten erstellt und betrachtet werden können.

Erstellung von RTI-Daten Für die Aufnahme und Erstellung von RTI-Daten gibt es zwei frei verfügbare Programme. Das eine ist das *PTM Fitter Programm*, welches an den Hewlett-Packard-Laboratories entwickelt wurde und zugleich auch das erste Programm ist, dass diese Art von Daten erstellen konnte. Es kann auf allen gängigen Betriebssystemen verwendet werden. Das zweite Programm, *RTIBuilder*, stammt von CHI und ist für Windows und Mac verfügbar. CHI stellt weitere Materialien zur Erstellung von RTI-Daten zur Verfügung und deren Webangebot dient als zentrale Anlaufstelle für alles rund um RTI. Sie bieten ebenfalls ein Forum für aktive Anwender an.

RTIBuilder:
http://culturalheritageimaging.org/What_We_Offer/Downloads/Process/index.html

Weiteres Material von CHI:
http://culturalheritageimaging.org/What_We_Offer/Downloads/

PTM Fitter Program:
<http://www.hpl.hp.com/research/ptm/downloads/download.html>

Ansicht von RTI-Daten RTI-Daten können nur mit speziellen Viewer-Programmen betrachtet werden. Von den Hewlett-Packard-Laboratories gibt es die frei verfügbare *PTM Viewer Application*, die auf Windows, Mac und Linux verwendet werden kann. Das Team von CHI bietet den *RTIViewer* an, der ebenfalls frei verfügbar ist und auf Windows und Mac läuft. Er bietet erweiterte Funktionalitäten, wie beispielsweise die Annotation von Dateien.

Für eine Präsentation der Daten auf Webseiten kann der *WebRTIViewer* empfohlen werden, der an dem Visual Computing Laboratory von CNR-ISTI in Pisa entwickelt wurde.

PTM Viewer Application:

<http://www.hpl.hp.com/research/ptm/downloads/download.html>

RTIViewer:

http://culturalheritageimaging.org/What_We_Offer/Downloads/View/index.html

WebRTIViewer:

<http://vcg.isti.cnr.it/rti/webviewer.php>

Quellen

Cultural Heritage Imaging:

<http://culturalheritageimaging.org/>

Hewlett-Packard-Laboratories, PTM:

<http://www.hpl.hp.com/research/ptm/>

S. Duffy – P. Bryan – G. Earl – G. Beale – H. Pagi – E. Kotouala, Multi-light Imaging for Heritage Applications (2013)

<https://www.historicengland.org.uk/images-books/publications/multi-light-imaging-heritage-applications/>

T. Malzbender – D. Gelb – H. Wolters, Polynomial Texture Maps, in:

Proceedings of ACM Siggraph 2001 (2001)

<http://www.hpl.hp.com/research/ptm/papers/ptm.pdf>

Formatspezifikationen

PTM:

<http://www.hpl.hp.com/research/ptm/downloads/PtmFormat12.pdf>

RTI:

http://forums.culturalheritageimaging.org/index.php?app=core&module=attach§ion=attach&attach_id=81

Tools und Programme

RTIBuilder:

http://culturalheritageimaging.org/What_We_Offer/Downloads/Process/index.html

Weiteres Material von CHI:

http://culturalheritageimaging.org/What_We_Offer/Downloads/

PTM Fitter Program:

<http://www.hpl.hp.com/research/ptm/downloads/download.html>

PTM Viewer Application:

<http://www.hpl.hp.com/research/ptm/downloads/download.html>

RTIViewer:

http://culturalheritageimaging.org/What_We_Offer/Downloads/View/index.html

WebRTIViewer:

<http://vcg.isti.cnr.it/rti/webviewer.php>

4.6 Satellitenmessungen

S. Reinhold

Einsatz von Hard- und Software in der Praxis

Auf allgemeiner Ebene sind momentan die unter GoogleEarth einsehbaren Satellitenbilder eine gute Alternative zu kommerziellen Satellitenbildern. Die bei GoogleEarth verwandten Bilder stammen aus verschiedenen Quellen und haben eine unterschiedliche Auflösung. Diese schwankt von 30-15m pro Pixel im Fall der auf Landsat-Aufnahmen basierenden Grundbilder bis zu einer Auflösung von etwa 1m pro Pixel bei Detailausschnitten. Diese Bilder stammen vom Satelliten Quickbird und werden laufend in GoogleEarth integriert. Es lohnt sich daher immer wieder die interessierende Region anzuschauen.

Eine Übersicht über die Quickbird-Aufnahmen sowie den Zeitpunkt der Aufnahme lässt sich unter Layers/DigitalGlobe Coverage im GoogleEarth Menü abrufen.

GoogleEarth hat den Nachteil, dass die Bilder nur in einer verhältnismäßig schlechten Auflösung abgespeichert werden können und sie nicht georeferenziert sind. Eine Lösung ist ein Upgrade auf GoogleEarthPlus. Dieses kostet 20\$ pro Jahr und erlaubt es, eine bessere Auflösung abzuspeichern und GPS-Daten sowie geographische Positionen können aus .CSV-Dateien importiert werden.

Für die Nutzung von georeferenzierten Satellitenbildern besteht die Möglichkeit, Bilder der Landsat Satelliten 4,5 und 7 im Internet frei zu laden. Die Daten des Satelliten Landsat ETM+ stammen aus dem Jahr 2000. Diejenigen von Landsat TM aus den 1990'er Jahren und die Bilder von Landsat MMS aus dem Zeitraum von 1972-1978.

BILD Aus: Eurimage Products and Services Guide

Um Bilder zu laden steuert man auf der Homepage landsat.org unter dem link 'Search for imagery' die freien LandsatOrtho Daten an. Mit dem dort vorgegebenen Path-Row Finder kann man die entsprechenden Angaben zur interessierenden Region über ein interaktives Kartenmodul ermitteln. Wichtig ist es dazu den Popblocker (Internetexplorer: Extra/Popupblocker; Firefox: Extras-Einstellungen-Inhalt/Pop-up-Fenster blockieren) zu deaktivieren. Wichtig ist auch im Suchmodus die Kästchen WRS I und WRS II anzukreuzen und refresh map zu aktivieren. Nur so lassen sich dann mit dem Informations-Tool die entsprechenden Flugbahndaten ermitteln. Eine alternative Suchmöglichkeit bietet sich unter: http://landsat.usgs.gov/tools_wrs-2_shapefile.php. Dort sind ESRI-shp files herunter zu laden, die ebenfalls Path/Row Informationen enthalten.

Hat man die Angaben lassen sich bei <http://www.landsat.org/ortho/> die entsprechenden Daten ansteuern und kopieren. Die gezippten Daten müssen dann entzipped werden und man erhält verschiedene Datensätze im Format Geotiff. Diese sind georeferenziert im System WGS 1984.

Die Auflösung der Landsat Bilder ist unterschiedlich und liegt bei 30m bzw. 15m per Pixel.

BILD Aus: Eurimage Products and Services Guide

Für detailliertere Satellitenbilder mit hoher Auflösung stehen momentan nur kommerzielle Anbieter zur Verfügung. Dort können georeferenzierte Bilder mit

einer Auflösung bis unter 1m pro Pixel erworben werden. Die folgende Tabelle gibt einen Überblick über die vorhandenen Systeme und die Quellen, über die Bildausschnitte und Preisinformationen zu erhalten sind.

Fotografische Sensoren

Satelliten	Beschreibung	Internet-Quelle
CORONA	Satellitenaufnahmen aus den 1960er und frühen 1970er Jahren der amerikanischen Spionagesatelliten CORONA, LANYARD und ARGON. CORONA Bilder der Missionen KH4 und 6 haben eine Auflösung von 2-3m. Es empfiehlt sich die Angaben in den Missionsdokumentationen zu lesen. Und Bilder in 7 Micron scannen zu lassen. Nicht georeferenziert!	Allgemeine Angaben finden sich unter: http://edc.usgs.gov/products/satellite/declass1.html#description Bildauswahl unter: http://edcsns17.cr.usgs.gov/ EarthExplorer/ dort finden sich die CORONA Bilder unter der Angabe „declassified images“. Preis (2007): 30\$ pro Filmstreifen (die DVD ist jedoch recht teuer) http://edcsns17.cr.usgs.gov/helpdocs/prices.html#CORONA
KFA-1000	Der russische Spionagesatellit KFA-1000 hat eine Auflösung von 1.5-3m. Die Bilder stammen aus den 1990er Jahren. Nicht georeferenziert!	Übersichtskarten und Bildauswahl sind erhältlich über http://www.sovinformsputnik.com/

Digitale Sensoren

Satelliten	Beschreibung	Internet-Quelle
LANDSAT	Amerikanische Multispektral-Satelliten, die seit 1972 aktiv sind. Die LANDSATS 1-3 hatten einen Multispektral Sensor (MSS) mit einer Auflösung von 80m. Die LANDSATS 4 and 5 haben MSS und Thematic Mapper (TM) Sensoren mit einer Auflösung von 30 oder 15m.	Übersicht bei http://landsat.org Kommerzielle Anbieter sind u.a. http://www.eurimage.com/index.html oder http://www.digitalglobe.com/ Preise (2007) ab 250 EURO http://www.eurimage.com/products/docs/eurimage_price_list.pdf
SPOT	SPOT. Französische Satelliten mit Multispektral (XS) Sensoren und einer Auflösung von 20m bis 10m.	Bilder und Preisangaben bei http://www.spotimage.fr/html/___.php
IKONOS 2	Satellit mit hochauflösenden Kameras, die Graustufen- und Farbbilder aufnehmen. Jedes Bild bildet eine Fläche von mindestens 11 x 11 km mit einer Auflösung von bis zu 0,82m ab. Es können auch Flächen von 60 x 60 km in einem Überflug aufgenommen werden können.	Bilder und Preisangaben bei http://www.euspaceimaging.com/ Preis ca. 30\$ pro qkm
QuickBird	QuickBird 2 ist ein kommerzieller Satellit zur Erdbeobachtung. Er wird von DigitalGlobe betrieben. Er hat eine Auflösung von 0,6m. Die Detailbilder in GoogleEarth sind Quickbirdaufnahmen seit 2001.	Bilder und Preisangaben bei http://www.digitalglobe.com/ oder http://www.eurimage.com/index.html Preis 16-41 \$ pro qkm (mindestens 272 qkm)

Quellen

http://www.dainst.org/medien/de/Nasca-BMBF_Programm.pdf
 David L. Kennedy, Declassified satellite photographs and archaeology in the Middle East: case studies from Turkey. *Antiquity* 72, Nr. 277 1998, 553-561.
 Graham Philip/Daniel Donoghue/Anthony Beck/N. Galiatsatos, CORONA

- satellite photography: an archaeological application from the Middle East. *Antiquity* 76, Nr. 291 2002, 109-118.
- Jason Ur, CORONA satellite photography and ancient road networks: A Northern Mesopotamia case study. *Antiquity* 77, Nr. 295 2003, 102-105.
- Wouter Gheyle/Raf Trommelmans/Jean Bourgeois/Rudi Goosens/Ignace Bourgeois/Alain De Wulf/Tom Wilkens, Evaluation CORONA: A case study in the Altai Republic (South Siberia). *Antiquity* 78, Nr. 300 2004, 391-403.
- M. Altaweel, The use of ASTER satellite imagery in archaeological contexts. *Archaeological Prospection* 12, 2005, 151-166.
- K.N. Wilkinson/Anthony Beck/Graham Philip, Satellite imagery as a resource in the prospection for archaeological sites in central Syria. *Geoarchaeology* 21, 2003, 735-750.
- Anthony Beck/Graham Philip/Maamun Abdulkarim/Daniel Donoghue, Evaluation of CORONA and IKONOS high resolution satellite imagery for archaeological prospection in western Syria. *Antiquity* 81, Nr. 311 2007, 176-190.

5 | Archivierung bei IANUS

M. Trognytz

In diesem Abschnitt sind Informationen zu finden, die sich speziell auf die langfristige Archivierung und Bereitstellung von altertumswissenschaftlichen Forschungsdaten in IANUS beziehen. Da die entsprechenden technischen Systeme zur Zeit noch im Aufbau begriffen sind, professionelle Workflows definiert und umgesetzt werden müssen und notwendige Vorgaben zusammen mit Hilfestellungen noch zu präzisieren sind, können hier aktuell nur allgemeine Hinweise gegeben werden. Diese werden in den nächsten Wochen und Monaten kontinuierlich weiter ausformuliert, so dass wir Sie bei Interesse bitten, sich regelmäßig an dieser Stelle zu informieren.

Sofern Sie als Datenproduzent Forschungsdaten besitzen, die Sie bereits jetzt gerne IANUS zur Kuratierung und Veröffentlichung anvertrauen möchten, können Sie dies gerne tun, indem Sie sich entweder per mail an ianus-fdz@dainst.de wenden oder die beiden Projektkoordinatoren direkt telefonisch kontaktieren.

Dateiformate Nicht jedes Dateiformat ist für die Langzeitarchivierung geeignet, weshalb dies ein zentrales Thema der IT-Empfehlungen ist. In dem Kapitel „Dateiformate“ ab Seite 52 werden entsprechende Hinweise für jeden Dateityp gegeben, wobei zu beachten ist, dass das Kapitel noch nicht abgeschlossen ist. Außerdem können in Abhängigkeit der Entstehung der Datei Abweichungen auftreten, was in dem Kapitel „Forschungsmethoden“ ab Seite 194 thematisiert werden soll.

Für die bereits überarbeiteten und abgeschlossenen Abschnitte sind die präferierten und akzeptierten Dateiformate in der folgenden Tabelle dargestellt, die separat auch als Download zur Verfügung steht. Präferierte Dateiformate sind Formate, die bereits für die Langzeitarchivierung geeignet sind, während akzeptierte Formate noch einer Konvertierung bedürfen. Die Tabelle steht auch als Download zur Verfügung.

Präferierte und akzeptierte Dateiformate

PDF-Dokumente

PDF/A-1	pdf	präferiert
PDF/A-2	pdf	präferiert
PDF/A-3	pdf	akzeptiert
andere PDF-Varianten	pdf	akzeptiert

Texte/Dokumente

Portable Document Format (PDF/A)	pdf	präferiert
andere PDF-Varianten	pdf	akzeptiert
OpenDocument Format	odt	präferiert
Microsoft Office XML	docx	präferiert

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Präferierte und akzeptierte Dateiformate

Microsoft Word	doc	akzeptiert
Rich Text Format	rtf	akzeptiert
Open Office XML	sxw	akzeptiert
Reiner Text, plain text	txt	präferiert
Strukturierter Text, Markup	xml, sgml, html etc. + dtd, xsd etc.	präferiert

Rastergrafiken

Baseline TIFF v. 6, unkomprimiert	tiff, tif	präferiert
Adobe Digital Negative	dng	präferiert
Portable Network Graphics	png	akzeptiert
Joint Photographic Expert Group	jpeg, jpg	akzeptiert
Graphics Interchange Format	gif	akzeptiert
Bit-Mapped Graphics Format (Microsoft)	bmp	akzeptiert
Photoshop (Adobe)	psd	akzeptiert
CorelPaint	cpt	akzeptiert
JPEG2000	jp2, jpx	akzeptiert
RAW image format	nef, crw etc.	akzeptiert

Vektorgrafiken und CAD-Daten

Scalable Vector Graphics	svg	präferiert
WebCGM	cgm	akzeptiert
Drawing Interchange Format, 2010 (AC1024)	dxf	akzeptiert
Drawing, 2010 (AC1024)	dwg	akzeptiert
PDF/A-1, PDF/A-2	pdf	akzeptiert
CGM	cgm	akzeptiert
Postscript, Encapsulated Postscript	ps, eps	akzeptiert
Illustrator, InDesign	ai, indd	akzeptiert
DWF	dwf.	akzeptiert

Tabellen

Comma Separated Values	csv	präferiert
Tab Separated Values	tsv	präferiert
OpenDocument Format	ods	präferiert

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Präferierte und akzeptierte Dateiformate

Microsoft Office XML	xlsx	präferiert
Strukturierter Text, Markup	xml, html etc. + dtd, xsd etc.	präferiert
Portable Document Format (PDF/A)	pdf	akzeptiert
Open Office XML	sxc	akzeptiert
Microsoft Excel	xls	akzeptiert
<hr/>		
Datenbanken		
SIARD, SIARD2	SIARD2	präferiert
SQL	sql	präferiert
XML	xml	präferiert
CSV	csv	akzeptiert
JSON	json	akzeptiert
Microsoft Access	mdb, accdb	akzeptiert
FileMaker	fp5, fp7, fmp12	akzeptiert
ODB	odb	akzeptiert
DBF	dbf	akzeptiert
BAK, DB, DMP	bak, db, dmp	akzeptiert
<hr/>		
Video		
Matroska	mkv	präferiert
Motion JPEG 2000	mj2	akzeptiert
MPEG-4	mp4	akzeptiert
Material eXchange Format	mxaf	akzeptiert
MPEG-2	mpeg, mpg	akzeptiert
andere MPEG-Varianten	mpeg, mpg	akzeptiert
Audio Video Interleave	avi	akzeptiert
MOV	mov	akzeptiert
ASF/WMF	asf, wmv	akzeptiert
Ogg	ogg, ogv, ogx, ogm, spx	akzeptiert
Flash	flv, f4v	akzeptiert
<hr/>		
Audio		
Free Lossless Audio Codec	flac	präferiert
Waveform Audio File Format	wav	präferiert
Broadcast Wave Format	bwf	präferiert
Matroska	mkv	akzeptiert

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Präferierte und akzeptierte Dateiformate

RF64/MBWF	wav	akzeptiert
Advanced Audio Coding/MP4	aac, mp4	akzeptiert
MP3	mp3	akzeptiert
Audio Interchange File Format	aiff, aif	akzeptiert
Windows Media Audio	wma	akzeptiert
Ogg	ogg, ogg, opus	akzeptiert
3D-Daten/Virtual Reality		
X3D	x3d + avi, mpg, jpeg	präferiert
COLLADA	dae + avi, mpg, jpeg	präferiert
OBJ	obj + jpeg	präferiert
Polygon File Format (PLY)	ply + jpeg	präferiert
Virtual Reality Modeling Language (VRML)	vrml + avi, mpg, jpeg	akzeptiert
Universal 3D Format	u3d + avi, mpg jpeg	akzeptiert
STL	stl + jpeg	akzeptiert
DXF	dxf + jpeg	akzeptiert
Webseiten		
PDF/A-1, PDF/A-2	pdf	präferiert
HTML, XHTML	html, htm, xhtml, xhtm, css etc.	präferiert
MHTML	mhtml	präferiert
Web Archive	warc	präferiert
MAFF	maff	akzeptiert
HTML mit Data URIs	html, htm	akzeptiert
PDF Varianten	stl + jpeg	akzeptiert
Screenshots	png, jpeg, tiff	akzeptiert
Webarchive Safari	webarchive	akzeptiert

Sollten Sie Dateien in anderen Formaten vorliegen haben, kontaktieren Sie uns. Für weitere Dateitypen finden Sie außerdem auf den Seiten des ADS Anhaltspunkte:

<http://archaeologydataservice.ac.uk/advice/FilelevelMetadata#section-FilelevelMetadata-FileLevelMetadataRequirements>

Die Informationen der Seiten des ADS sind in einer erweiterten Tabelle aggregiert, die ebenfalls als Download zur Verfügung steht.

Metadaten Neben dem geeigneten Format sind zusätzliche Informationen zur Datei erforderlich, die sogenannten Metadaten. Ausführliches dazu ist in dem Abschnitt „Dokumentation“ ab Seite 28 zu finden. Speziell die tabellarischen Aufstellungen im Unterabschnitt „Metadaten in der Anwendung“ ab Seite 36 sind dabei zu berücksichtigen.

Dabei wird zwischen projektbezogenen, dateibezogenen und methodenbezogenen Metadaten unterschieden, für die Sie im folgenden die tabellarischen Übersichten finden, die auch gesondert als Download zur Verfügung stehen. Zu beachten ist dabei, dass in Abhängigkeit des Dateityps und der verwendeten Methode weitere spezifische Metadaten erforderlich sein können, die in den jeweiligen Kapiteln angegeben sind.

Projektbezogene Metadaten	
Bezeichnung	Kurzdefinition
Identifizierung – Projekttitle	Verbindliche Kurzbezeichnung des Projektes.
Identifizierung – Alternativtitel	Ggf. alternative Titel für ein Projekt.
Identifizierung – Projektnummer(n)	Nummern oder Kennungen, die z.B. innerhalb der durchführenden Organisation oder von Mittelgebern verwendet wird, um das Projekt eindeutig identifizieren zu können.
Kurzbeschreibung	Knappe Angaben zur Fragestellung, zum Verlauf und Ergebnis des Projektes sowie Skizzierung der Datensammlung (insgesamt ca. 100-300 Worte)
Schlagworte – Fachdisziplinen	Stichworte, die die beteiligten Disziplinen und Fächer benennen. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.
Schlagworte – Inhalt	Stichworte, die den Inhalt der Datensammlung benennen., z. B. zu Materialgruppen, Fundstellen-Klassifizierung, Quellenarten, Kulturgruppen etc. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.
Schlagworte – Methoden	Stichworte, die die eingesetzten Forschungsmethoden beschreiben. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Bezeichnung	Kurzdefinition
Ausdehnung – Geografisch-1	Detailierte Angaben zur räumlichen Ausdehnung oder zum Fundort des untersuchten Gegenstandes mittels geografischer Koordinaten. Die maximale Ausdehnung kann als Bounding Box angegeben werden.
Ausdehnung – Geografisch-2	Sprachliche Beschreibung des untersuchten Gegenstandes mittels Ortsangaben mit Land, Stadt, Kreis, Straße, Gemarkung etc. Sofern Namen sich im Lauf der Zeit geändert haben, dies gesondert vermerken. Sofern eine Referenz zu einer Geo-Ressource oder einem Gazetteer existiert, sollte diese ebenfalls angegeben werden.
Ausdehnung – zeitlich	Chronologische Angaben zum untersuchten Gegenstand, entweder als Periodenbezeichnung und/oder mit groben/genauen Jahresangaben. Sofern die Stichworte auf publizierten Standards oder internen Thesauri beruhen, müssen diese mitangegeben werden.
Primärforscher – Person	Personen, die entweder für das Projekt als Ganzes, für das Datenmanagement oder für die Erzeugung bestimmter Datenarten zentral bzw. verantwortlich sind. Hier ist eine Kontaktadressen erforderlich und die aktuelle/letzte institutionelle Zugehörigkeit, damit die Personen bei Rückfragen erreicht werden kann.
Eigentümer – Organisation	Organisation, der die unter „Primärforscher“ genannten Personen angehören, oder die nach Ausscheiden derselben für die Daten verantwortlich ist, im weitesten Sinne also Eigentümer der Daten ist. Hier ist eine Kontaktadresse erforderlich, damit die Organisation bei Rückfragen erreicht werden kann.
Finanzierung	Nennung der Organisation(en) / (Dritt-)Mittelgeber, durch die das Projekt finanziert wurde. Es sollte jeweils der Zeitraum der Finanzierung angegeben werden.
Veröffentlichung – Projektdaten	Wenn die hier beschriebene Datensammlung des Projektes bereits an anderer Stelle veröffentlicht / online gestellt wurde, bitte entsprechende Angaben machen, z. B. durch Nennung der Organisationen, Datenarchive, Online-Ressourcen etc.
Veröffentlichung – Ergebnisse	Analoge oder digitale Publikationen zu Ergebnissen des Projektes oder zur Datensammlung des Projektes, ausführliche bibliographische Angaben (ohne fachspezifische Abkürzungen) unter Nennung des Verlages erforderlich.
Dauer – Projekt	Anfangs- und Enddatum des Projektes.

Fortsetzung auf der nächsten Seite

Fortsetzung der vorhergehenden Seite

Bezeichnung	Kurzdefinition
Dauer – Datenbestand	Anfangs- und Enddatum der Erzeugung oder Verarbeitung digitaler Daten im Rahmen des Projektes.
Rechtliches – Urheberrechte	Name des Inhabers der Urheber-, Nutzungs- und Verwertungsrechte; i. d. R. die Organisation, an der der Primärforscher beschäftigt war.
Rechtliches – Lizenzgeber	Angabe der Person, die i. d. R. als Vertretung für eine Organisation für die Lizenzierung von Daten zur Nachnutzung verantwortlich und berechtigt ist, einen Datenübergabevertrag abzuschließen.
Rechtliches – Datenschutz	Angaben, ob in der Datensammlung datenschutzrelevante Informationen enthalten sind. Wenn ja, in welchem Umfang.
Quellen – Ältere	Ältere Quellen oder existierende Ressourcen, auf denen die Daten aufbauen.
Quellen – Zugehörige	Sofern während des Projektes Informationen, Datensammlungen, (un-)publizierte Dokumente, Online-Ressourcen etc. verwendet oder erzeugt wurden, die nicht Teil der hier beschriebenen Datensammlung sind, aber für deren Verständnis wichtig sind, bitte entsprechende Angaben zu Art und Umfang dieser Quellen machen.
Sprache	Die in den Dokumenten und Dateien verwendete(n) Sprache(n). Sprachkennungen nach ISO 639 angeben.
Art der Daten	Kurzcharakterisierung der Daten, z. B. ob es sich um Rohdaten, verarbeitete Daten, Interpretationen, Ergebnisse, Abschlussberichte etc. handelt.
Vollständigkeit	Aussagen zur Vollständigkeit der Projektdaten, z. B. ob bestimmte Datenarten noch fehlen und warum.
Dateiformate	Auflistung der Dateiformate, die in der Datensammlung vorkommen, ggf. unter Nennung der verwendeten Programme und Zeichenkodierungen.
Zugriffsrechte	Festlegung der gewünschten Zugriffsrechte für die Daten, sofern diese für den gesamten Projekt-Datenbestand gelten sollen; differenzierte Regelungen müssen auf Dateiebene vorgenommen werden.
Signatur Metadaten	Angabe darüber, wer die o. g. Metadaten wann ausgefüllt hat.

Dateibezogene Metadaten	
Bezeichnung	Kurzdefinition
Dateiname	Eindeutiger Name der Datei.
Dateiformat	Format, in der die Datei abgespeichert ist.
Software – Dateierstellung	Software, mit der die Datei erstellt wurde.
Hardware – Dateierstellung	Hardware, mit der die Datei erstellt wurde, v. a. bei technischen Geräten wie Kameras, GPS-Geräten, Vermessungsinstrumenten, Laserscanner etc.
Betriebssystem – Dateierstellung	Betriebssystem, das verwendet wurde als die Datei erstellt wurde.
Erstellungsdatum	Datum, an dem die Datei erstellt wurde. Datum und Zeit in UTC nach ISO 8601.
Letzte Aktualisierung	Datum, an dem die Datei zuletzt bearbeitet wurde. Datum und Zeit in UTC nach ISO 8601.
Dateiversion	Angabe der Versionsnummer der Datei.
Referenzdateien	Referenzen auf Dateien, die für das Verständnis einer anderen Datei zentral sind, v. a. bei zusammenhängenden, komplexen Dateien oder wenn auf eine Ursprungsdatei verwiesen werden soll.

Methodenbezogene Metadaten	
Bezeichnung	Kurzdefinition
Prozessnummer	Eindeutige Nummer eines Prozesses oder einer Methode.
Ausgangsformat(e)	Format der Dateien, die am Anfang eines gesamten Prozesses stehen und den Ausgangspunkt bilden.
Zwischenformat(e)	Format der Dateien, die im Verlauf eines Prozesses erzeugt werden und den Ausgangspunkt für weitere Prozesse bilden.
Zielformat(e)	Format der Dateien, die am Ende eines gesamten Prozesses erzeugt werden und den Zielpunkt definieren.
Durchführende(r)	Person(en), die den Prozess durchgeführt hat (haben).
Prozessbeschreibung	Beschreibung des Prozesses/der Methode, vor allem der Ausgangssituation und der Zielvorstellung.
Prozessbeginn	Datum, an dem der Prozess begonnen wurde. Datum und Zeit in UTC nach ISO 8601.
Prozessende	Datum, an dem der Prozess beendet wurde. Datum und Zeit in UTC nach ISO 8601.
Software	Software, mit der der Prozess durchgeführt wurde.
Hardware	Hardware, auf der der Prozess durchgeführt wurde.

Datenmanagement Wenn Sie noch am Beginn Ihres Projektes stehen, haben Sie bereits jetzt die Möglichkeit die spätere Archivierung der Daten ein-

zuplanen. Um dies zu erleichtern, ist die Konzeption und Verwendung eines Datenmanagementplanes hilfreich. In dem Kapitel „Datenmanagement“ ab Seite 16 finden Sie weitere Angaben dazu.

A | Autorenverzeichnis

- Reinhard Förtsch, DAI Berlin
- Philipp Gerth, DAI Berlin
- Peter Grunwald, ehemals DAI Berlin
- Katja Heine, BTU Cottbus
- Dominik Hagmann, Universität Wien
- Maurice Heinrich, DAI Berlin - IANUS
- Kerstin Hofmann, Topoi Berlin
- Sabine Jahn, DAI Berlin - IANUS
- Ulrich Kapp, ehemals DAI Berlin
- Rainer Komp, DAI Berlin
- Susanne Kurz, HKI Köln
- Karsten Lambers, Universität Bamberg
- Armin Müller, DAI Berlin
- Knut Rassmann, RGK Frankfurt
- Markus Reindel, KAAK Bonn
- Sabine Reinhold, DAI Berlin
- Felix Schäfer, DAI Berlin - IANUS
- Anne Sieverling, DAI Berlin - IANUS
- Manfred Thaller, HKI Köln
- Martina Trognitz, DAI Berlin - IANUS
- Ulrike Wulf-Rheidt, DAI Berlin

B | Glossar

AAC (Advanced Audio Coding) ist ein Audiocodec, der als Nachfolger von MP3 im Rahmen von MPEG-2 und MPEG-4 spezifiziert wurde. AAC ist unter ISO/IEC 13818-7 und 14496-3 standardisiert. Der Codec verwendet eine verlustbehaftete Kompression.

Abtastrate bezeichnet die Zeitintervalle, also die Häufigkeit, in der ein analoges Signal bei der Digitalisierung gemessen wird. Je kleiner das Zeitintervall, desto größer die Abtastrate. Die Abtastrate wird üblicherweise in Kilonertz (kHz) angegeben.

Abtasttiefe bezeichnet die Messgenauigkeit, mit der ein analoges Signal bei der Digitalisierung gemessen wird, also wie genau der gemessene Wert dem ursprünglichen Signal entspricht. Die Abtasttiefe wird mittels der Anzahl an Bits bestimmt. Je höher die Bitzahl, desto größer ist auch die Abtasttiefe. Die Abtasttiefe ist vom Konzept her vergleichbar mit der Farbtiefe.

Adobe RGB ist ein häufig verwendeter RGB-Farbraum, der 1998 von Adobe veröffentlicht wurde.

AIFF (Audio Interchange File Format) ist ein proprietäres Containerformat zum Speichern von Audiodaten. Es wurde von Apple entwickelt und hat vor allem auf Apple-Systemen eine weite Verbreitung gefunden.

Alphakanal wird bei Rastergrafiken verwendet, um Transparenz zu speichern.

AMP Apache-Server, MySQL oder PostgreSQL Datenbank, PHP, Perl oder Python-Skriptsprache ist ein Akronym für den kombinierten Einsatz von Programmen, um dynamische Webseiten zur Verfügung zu stellen. Die einzelnen Komponenten des AMP-Systems können durch ähnliche Komponenten ersetzt werden.

ASCII (American Standard Code for Information Interchange) ist eine Zeichenkodierung, deren Zeichensatz aus insgesamt 128 Zeichen besteht und mit jeweils einem Byte gespeichert werden. ASCII enthält keine diakritischen Zeichen oder gar andere Schriften, weshalb verschiedene Erweiterungen der ASCII-Kodierung entwickelt wurden, um insgesamt 256 verschiedene Zeichen zu kodieren.

ASF/WMV ist ein proprietäres Containerformat für Videos, das von Microsoft für das Streaming von Multimediadaten entwickelt wurde.

Auflösung gibt die Punktdichte für ein physikalisches Wiedergabemedium an. Sie wird üblicherweise in Punkten (dots per inch; dpi), Pixeln (pixel per inch; ppi) oder Linien (lines per inch; lpi) pro Zoll (inch) angegeben.

Auszeichnungselement (Tag) wird in Auszeichnungssprachen verwendet, um bestimmte Bereiche zu etikettieren. Beispielsweise kann durch die Verwendung eines Tags die Zeichenkette "24-28" als eine Größenangabe gekennzeichnet werden: <groesse>24-28</groesse>

Auszeichnungssprache (Markup Language) beschreibt mit Hilfe von Auszeichnungselementen den Inhalt von textbasierten Formaten. Dadurch kann das Aussehen eines Textdokumentes von dessen Struktur und Inhalt getrennt werden. Außerdem können implizite Informationen, die bisher nur für den menschlichen Leser verständlich waren, explizit gekennzeichnet werden und somit auch maschinell verarbeitbar gemacht werden. Beispiele sind: SGML, XML, HTML und TEI.

AutoCAD ist Teil der CAD-Produktpalette von Autodesk zum Erstellen von 2D- und 3D-Zeichnungen mit einem PC.

AVI (Audio Video Interleave) ist ein einfaches und robustes, jedoch proprietäres Containerformat für Videos. Es wurde von Microsoft entwickelt.

B-Rep (Boundary Representation), Begrenzungsflächenmodell, beschreibt ein 3D-Modell durch dessen begrenzende Oberflächen.

Baseline TIFF ist der kleinste gemeinsame Nenner aller TIFF-Formate, das von jedem TIFF-fähigen Programm verarbeitet werden kann und daher für die Langzeitarchivierung geeignet ist. Es kann beispielsweise keine Ebenen speichern und erlaubt auch keine Kompression wie LZW. Ursprünglich war das Format als TIFF 6.0, Teil 1: Baseline TIFF bekannt.

Batchverarbeitung siehe Stapelverarbeitung.

Bildformat wird für Videos durch die Anzahl der Pixel und dem Seitenverhältnis angegeben, um die Darstellungsgröße zu ermitteln.

Bildfrequenz gibt an wie viele Bilder pro Sekunde in einem Video abgespielt werden.

Bildpunkt siehe Pixel.

Bit ist die kleinste digitale Informationseinheit und kann entweder den Wert 0 oder 1 haben. Der Name ist aus *binary digit* zusammengesetzt.

Bitrate siehe Datenrate

Bitstream ist eine Sequenz von Bits, die einen Informationsfluss repräsentiert und eine unbestimmte Länge hat.

BLOB (Binary Large Object) ist ein großes binäres und damit für die Datenbank nicht weiter strukturiertes Objekt, beziehungsweise Felddaten (z.B. Bilddateien).

BOM (Byte Order Mark, Bytereihenfolge-Markierung) ist das Unicode-Zeichen U+FEFF am Anfang eines in Unicode kodierten Dokumentes, das angibt in welcher Reihenfolge die Bytes angeordnet sind. BOM ist bei

der Verwendung von UTF-16 und UTF-32 zwingend in der Datei erforderlich. Zusätzlich kann das BOM ein Hinweis auf die Verwendung von UTF-Kodierungen sein, jedoch wird von dessen Verwendung außer für UTF-16 und UTF-32 abgeraten.

Bounding Box, auch Minimum Bounding Box, (engl. minimal umgebende Box) ist der kleinste mögliche Quader, der eine bestimmte Menge an Punkten oder Objekten umschließt. Eine Bounding Box kann sowohl zweidimensional als auch dreidimensional sein.

BWF (Broadcast Wave Format) ist ein von der EBU entwickeltes Containerformat für Audiodaten, das auf WAVE aufbaut. Es unterstützt die Einbettung von weiteren Metadaten und kann maximal zwei Tonkanäle speichern.

Byte ist eine Folge von 8 Bits.

CAD (Computer Aided Design) Rechnerunterstützte Konstruktion.

CBR (engl. *constant bitrate*, konstante Bitrate) überträgt pro Sekunde immer die gleiche Datenmenge. Siehe auch Datenrate.

CHI (Cultural Heritage Imaging) ist eine gemeinnützige Firma, die visuelle Dokumentationsmethoden für Kulturgut entwickeln und einsetzen. CHI hat die Methode des Highlight-RTI entwickelt und stellt kostenlose Programme für die Erstellung von RTIs zur Verfügung.

Chroma steht für den Farbanteil eines Bildpunktes, bezogen auf Farbmodelle für selbstleuchtende Geräte.

Chrominanz siehe Chroma.

CIDOC (Comité International pour la Documentation) ist das internationale Komitee für Dokumentation des internationalen Museumsverbandes (ICOM) und setzt Standards für die Dokumentation in Museen. Es hat außerdem eine beratende Funktion. Das Komitee ist verantwortlich für CIDOC CRM.

CIDOC CRM (CIDOC Conceptual Reference Model) ist eine erweiterbare Struktur, um Begriffe und Informationen im Bereich des Kulturerbes in kontrollierter Form austauschen zu können. Neben der physischen Beschreibung von Objekten können auch zeitliche und örtliche Informationen beschrieben werden. Es ist die Norm ISO 21127:2006. Seit 2014 ist ebenfalls eine neue Version normiert: ISO 21127:2014.

CMYK ist ein subtraktives Farbmodell. Es bildet die technische Grundlage des Vierfarbendrucks. Die Abkürzung steht für die Farben Cyan, Magenta, Gelb (Yellow) und den Schwarzanteil für die Farbtiefe (Key). CMYK-Farträume enthalten weniger Farben, als RGB-Farträume.

CNR-ISTI (Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione) ist ein in Pisa angesiedeltes informationstechnisches Institut, das dem Nationalen Forschungsrat angehört.

Codec Der Begriff Codec setzt sich aus den englischen Wörtern *coder* (Codierer) und *decoder* (Decodierer) zusammen. Codecs definieren wie Datenströme gespeichert und gelesen werden, können also als Format verstanden werden. Codecs bieten immer eine Kompression der Daten, die entweder verlustbehaftet oder verlustfrei erfolgen kann. Im Bereich Video und Audio werden die visuellen und auditiven Inhalte jeweils in einem eigenen Codec gespeichert und anschließend in einem Containerformat zusammengeführt.

Container Eine Containerdatei kann unterschiedliche Dateien und Dateitypen enthalten. Beispielsweise werden in einem ZIP-Container mehrere Dateien in einer Archivdatei zusammengefasst. Als Containerformat werden Containerdateien bezeichnet, die eine bestimmte Strukturierung der beinhalteten Dateien voraussetzen. Beispielsweise werden in dem Containerformat Matroska visuelle und auditive Inhalte zusammengeführt, um ein Video zu speichern. Es gibt eine große Zahl an Containerformaten, die für unterschiedliche Anforderungen entwickelt wurden und daher auch jeweils einen unterschiedlichen Funktionsumfang aufweisen.

CSG (Constructive Solid Geometry) , Konstruktive Festkörpergeometrie, beschreibt 3D-Modelle anstatt durch Eckpunkte, Flächen oder Kurven, mit Hilfe von geschlossenen geometrischen Körpern. Dabei kann die Vereinigung, die Differenz oder die Schnittmenge der Körper gebildet werden.

CSV (Comma Separated Values) ist ein textbasiertes Dateiformat für tabellarische Daten. Hierbei werden einzelne Zellen durch das ein bestimmtes Trennzeichen, z.B. durch ein Komma, voneinander getrennt. Jede Zeile einer CSV-Datei entspricht der Zeile einer Tabelle. Da das Trennzeichen auch als Inhalt der Zellen erlaubt ist, werden in diesem Fall zusätzlich Textbegrenzungszeichen (z.B. Anführungszeichen) vor und nach den Zellen eingefügt, falls das Trennzeichen auch als Wert innerhalb der Zelle verwendet wird. Ein De-facto-Standard für CSV-Dateien ist in RFC 4180 beschrieben.

DAT (Data File) ein Dateiformat mit sehr unterschiedlicher Verwendung, meist als Container für binäre Daten.

Datenrate gibt an wie viele Daten in einer bestimmten Zeitspanne übertragen oder gelesen werden. Sie wird meist in Bits pro Sekunde gemessen, weshalb auch von Bitrate gesprochen wird.

dBASE-Format dBASE war das erste weit-genutzte Datenbankverwaltungssystem für Mikrocomputer. Die Basisidee des dBASE-Systems ist, die Tabellen einer Datenbank in speziell strukturierten Files zu halten.

DBF (Data Base File) Datenbank im dBASE-Format.

Demultiplexing bezeichnet das Aufsplitten der Inhalte eines Containers für Videos in die einzelnen Bestandteile.

Demuxing siehe Demultiplexing.

Detaillierungsgrad gibt an wie detailliert ein zweidimensionales Bild oder ein dreidimensionales Objekt dargestellt ist. In 3D-Szenen können unterschiedliche Objekte unterschiedliche Detailstufen aufweisen.

Dirac/Schroedinger ist ein vom BBC entwickelter Codec für Videos, der eine verlustfreie Komprimierung bietet, jedoch nicht sehr performant ist.

DNG (Digital Negative) ein offenes Format, das von Adobe speziell für die Speicherung und Langzeitarchivierung von Kamerarohdaten (RAW) entwickelt wurde. Es werden die Metadaten-Standards Exif, IPTC-NAA und XMP unterstützt.

DOCX ist ein offenes XML-basiertes Format, das von Microsoft zur Speicherung von Textdokumenten entwickelt wurde.

Dokumenttypdefinition siehe DTD.

Dokumenttypdeklaration steht zu Beginn eines SGML-, XML- oder HTML-Dokuments. Es enthält den Verweis auf die verwendete Dokumenttypdefinition oder auf das verwendete XML Schema.

dpi (dots per inch) gibt an wie viele Bildpunkte pro Zoll eine Rastergrafik enthält. Je größer die Punktdichte, desto größer auch die Auflösung des Bildes.

DTD (Dokumenttypdefinition) beinhaltet für SGML-Dateien die Regeln für die zu verwendende Auszeichnungselemente und deren Kombinationsmöglichkeiten, beschreibt also deren Struktur. Es handelt sich um eine separate Datei, die in der SGML-Datei in der Dokumenttypdeklaration angegeben wird.

Dublin Core ist ein Metadatenschema, das 15 Kernfelder beinhaltet. Es wurde entwickelt, um möglichst viele verschiedene Objekttypen beschreiben zu können. Es ist standardisiert als ISO 15836-2003, ANSI/NISO Z39.85-2007 und IETF RFC 5013.

DV (Digital Video) ist ein Codec für digitale Videos auf Videokassetten, der mit der Ablösung von Kassetten durch andere Speichermedien obsolet wird.

DWG (Drawing File) ein AutoCAD Dateiformat für Zeichnungen, das kontinuierlich an die Anforderungen der jeweiligen AutoCAD-Versionen angepasst und erweitert wird. Die eingeschränkte Kompatibilität des DWG-Dateiformates zu älteren AutoCAD-Versionen kann durch die Verwendung des DXF-Dateiformates und den Einsatz von externen Konverterprogrammen teilweise umgangen werden.

DXF (Drawing Exchange File) Die DXF-Schnittstelle ist eine proprietäre Schnittstelle des Herstellers Autodesk und unterliegt keinem neutralen Normungsausschuss, die Dokumentation für DXF ist aber frei verfügbar. Sie ist ein in ASCII-Zeichen lesbares Abbild der binär abgespeicherten DWG.

EBML (Extensible Binary Meta Language) ist das binäre Äquivalent zu XML.

EBU (European Broadcasting Union) ist die Europäische Rundfunkunion, die einen Zusammenschluss von über 70 Rundfunkanstalten bildet. Die Organisation beteiligt sich an der Forschung und Entwicklung von neuen Standards, wie etwa EBUCore.

EBUCore ist ein von der EBU entwickeltes Metadatenschema für Film und Ton.

Ecma (eigentlich Ecma International, früher ECMA) ist eine private Normungsorganisation mit Sitz in Genf, die sich insbesondere auf Informations- und Kommunikationssysteme konzentriert. Sie wurde 1961 gegründet.

Editor ist ein Programm für die Erstellung und Bearbeitung von bestimmten Dateiformaten. Beispielsweise bieten Texteditoren spezielle Funktionen für das Arbeiten mit Textdateien.

Einbetten bezeichnet in der IT das integrieren von externen Ressourcen in eine Datei. Beispielsweise kann eine Schriftart in ein PDF-Dokument eingebettet werden, um das Erscheinungsbild des Dokumentes auch auf anderen Systemen zu erhalten, wo die Schriftart nicht installiert ist.

encoding siehe Zeichenkodierung.

ERD (Entity Relationship Diagram) die grafische Darstellung von Entitätsarten und Beziehungstypen wird Entity-Relationship-Diagram (ERD) oder ER-Modell genannt. Das ER-Modell kann bei der Erstellung von Datenbanken genutzt werden. Hierbei wird mit Hilfe von ER-Modell zunächst die Konzeption der Datenbank vorgenommen, auf deren Grundlage dann die Implementierung der Datenbank erfolgt.

ESRI-Shape Das Dateiformat Shapefile (oft Shapedaten oder Shapes genannt) ist ein von ESRI ursprünglich für ArcView entwickeltes Format für Geodaten. Ein Shapefile ist keine einzelne Datei, es besteht aus mindestens drei Dateien:

- .shp dient zur Speicherung der Geometriedaten
- .shx dient als Index der Geometrie zur Verknüpfung der Sachdaten (auch Attributdaten genannt)
- .dbf Sachdaten im dBASE-Format

Exif (Exchangeable Image File Format) ist ein Standard der Japan Electronic and Information Technology Industries Association (JEITA) für die Speicherung von Metadaten in digitalen Bildern. Exif-Daten können in TIFF-, DNG- und JPEG-Formaten gespeichert werden.

Farbmodell ist ein mathematisches Modell, das üblicherweise mit Hilfe von Zahlentupeln (Zahlenpaare) beschreibt wie Farben dargestellt werden können. Alle Farben eines Farbmodells können dreidimensional als Farbraum dargestellt werden.

Farbraum ist die dreidimensionale Abbildung eines Farbmodells.

Farbtiefe gibt die Anzahl der Bits an, die den Farbwert eines Pixels speichern. Ein Bit kann dabei 2^1 , also zwei Farbwerte (z.B. Schwarz und Weiß) speichern. Die Anzahl der darstellbaren Farbwerte steigt mit der Anzahl der Bits exponentiell. So können mit 8 Bits bereits 2^8 also 256 Farbwerte und mit 24 Bits schon $2^{24} = 16.777.216$ Farbwerte dargestellt werden.

Farbunterabtastung (engl. *chroma subsampling*) macht sich zunutze, dass das menschliche Auge Farben im Vergleich zu Helligkeiten schlechter auflösen kann. Die Farbunterabtastung bedeutet, dass von dem Kanal für Chrominanz und den zwei Kanälen für Luminanz nur die Chrominanz vollständig gespeichert wird, während die Chrominanz horizontal und vertikal geringer aufgelöst wird. Wenn die drei Kanäle zusammengerechnet werden, weist das Gesamtbild optisch kaum sichtbare Unterschiede auf. Die Farbunterabtastung wird üblicherweise mittels drei Ziffern angegeben, wie etwa 4:2:0.

FFmpeg ist ein Projekt, das aus einer Reihe von frei verfügbaren Programm-bibliotheken besteht, um digitale Videos und Audioaufnahmen aufzunehmen, zu konvertieren und streamen. FFmpeg enthält eine umfangreiche Sammlung von Audio- und Videocodecs, wie beispielsweise FFV1 für Videos.

FFV1 (FFmpeg Video Codec 1) ist ein Codec für Videos, der eine verlustfreie und schnelle Intraframe-Kompression verwendet. Er wurde in dem Projekt FFmpeg entwickelt und hat eine weite Verbreitung gefunden.

FLAC (Free Lossless Audio Codec) ist ein offen dokumentierter und frei verfügbarer Audiocodec, der verlustfrei komprimiert.

Flash ist ein Containerformat für das Streaming von Multimediadaten. Es wurde von Macromedia entwickelt und dann von Adobe übernommen.

Font steht für Schriftart.

FourCC (Four Character Code) ist ein vier Byte langer Bezeichner aus ASCII-Zeichen, der in dem Header einer Videodatei angibt, welcher Codec verwendet wird. Beispielsweise wird *FFV1* verwendet um den Codec FFV1 zu identifizieren.

Generationsverlust bezeichnet die Abnahme der Qualität von einer Kopie zur nächsten. Gerade bei verlustbehafteten Formaten, wie beispielsweise JPEG entstehen beim Speichern Fehler, die sich mit jeder weiteren Kopie der Kopie fortpflanzen und durch neue ergänzt werden. Vorbeugen kann man diesem Phänomen, indem man möglichst verlustfreie Formate verwendet.

GeoTIFF ist ein georeferenziertes TIFF. Die Besonderheit gegenüber dem normalen TIFF-Format liegt darin, dass spezielle Daten über die Georeferenz zusätzlich zu den sichtbaren Rasterdaten in die Bilddatei eingebettet werden. Dazu zählen Koordinaten zur Georeferenzierung des Bildausschnitts sowie zur verwendeten Kartenprojektion: Die Datei enthält spezifische Angaben über das Koordinatenreferenzsystem.

GIF (Graphics Interchange Format) ist ein proprietäres Dateiformat zur Speicherung von Bilddaten, das von CompuServe entwickelt wurde. Dieses Format ist im Internet besonders verbreitet und bietet eine verlustfreie Komprimierung und nur eine Farbtiefe von 8 Bit, also 256 Farben, an. Metadaten können nur eingeschränkt in diesem Dateiformat gespeichert werden. Es wird empfohlen stattdessen PNG oder TIFF zu verwenden.

GIS (Geoinformationssystem) ein rechnergestütztes Informationssystem, mit dem raumbezogene Daten digital erfasst und redigiert, gespeichert und reorganisiert, modelliert und analysiert sowie alphanumerisch und grafisch präsentiert werden können.

H-RTI siehe Highlight-RTI.

H.264 (H.264/MPEG-4 AVC) ist ein Videocodec, mit dem eine verlustfreie Kompression möglich ist. Der Codec bietet eine hohe Bildqualität, ist Teil des MPEG-4-Standards und wird in dem Containerformat MP4 verwendet.

H.265 (H.265/MPEG-H (High Efficiency Video Coding (HEVC)) ist ein Videodecoder der seit 2013 existiert und als Nachfolger von H.264 gedacht ist. Er bietet eine verlustfreie Kompression, ist jedoch noch nicht vollständig spezifiziert und weist patentrechtliche Einschränkungen auf. Daher ist dieser Codec *nicht* für die Langzeitarchivierung geeignet.

Hardware beschreibt in der IT die physische Ausrüstung, also Geräte und Geräteteile, wie beispielsweise Festplatten, Grafikkarten oder Computer. Der Begriff kann auf alle weiteren Geräte erweitert werden, die digitale Daten produzieren, wie etwa Digitalkameras oder Tachymeter.

Header bezeichnet in der IT den Dateikopf, also die Daten am Anfang eines Datenblocks, in dem üblicherweise Metadaten gespeichert werden. Beispielsweise werden Metadaten für Bilder mit Hilfe des Exif-Formats in den Header der Bilddatei geschrieben. Für XML- oder HTML-Dateien gibt es eigene Auszeichnungselemente für Informationen im Header, der am Anfang der jeweiligen Datei zu finden ist.

hexadezimal (Hexadezimalsystem) beschreibt Zahlen auf Basis von 16. Im Unterschied zu dem üblichen Dezimalsystem gibt es für die Zahlen 0 bis 15 eigene Werte: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F.

Highlight-RTI ist eine von CHI entwickelte Aufnahmemethode für die Erstellung von RTI-Dateien. Hierfür wird nur ein Blitzgerät benötigt, das für jede Aufnahme an einer imaginären Kuppel entlang bewegt wird. Um die Position der Lichtquelle errechnen zu können, wird neben dem Objekt eine schwarz glänzende Kugel platziert. Aus den Reflexionspunkten des Lichtes an der Kugel kann das Programm dann die Richtung, aus der das Licht kommt, berechnen.

HTML (Hypertext Markup Language) ist eine Auszeichnungssprache, die einen Anwendungsfall von SGML darstellt. HTML hat als Grundlage von Webseiten eine große Verbreitung gefunden und wird vom W3C und der Web Hypertext Application Technology Working Group (WHATWG) gepflegt und entwickelt. Die aktuellste Version ist HTML5.

HuffYUV ist ein verlustfreier Codec für Videodaten, der eigens für Windows-Systeme entwickelt wurde. HuffYUV ist mit Lagarith verwandt.

IANA (Internet Assigned Numbers Authority) ist für die Zuordnung von Namen und Nummern im Internet zuständig. Darunter fallen beispielsweise IP-Adressen oder Medientypen.

IASA (International Association of Sound and Audiovisual Archives) ist eine 1969 gebildete Vereinigung, um als Mittler für internationale Kooperationen zwischen Archiven für Ton- und Videoaufnahmen zu fungieren.

ICC (International Color Consortium) wurde 1993 von acht Industrieunternehmen gegründet um eine Vereinheitlichung der Farbmanagementsysteme für alle Betriebssysteme und Softwarepakete zu erreichen. Weithin bekanntes Ergebnis der Bemühungen des ICC ist ein Standard zur Beschreibung von Farbprofilen.

IEC (International Electrotechnical Commission) ist eine 1906 gegründete Organisation für die Erstellung und Veröffentlichung von internationalen Standards im Bereich von Elektrotechnik und Elektronik. Einige Normen entwickelt sie auch gemeinsam mit der ISO.

IETF (Internet Engineering Task Force) ist eine Organisation, die in mehreren Arbeitsgruppen Standards für das Internet erstellt.

IIM (Information Interchange Model) Ein Standard zur Speicherung von Metadaten in Bilddateien, der 1990 vom International Press Telecommunications Council und der Newspaper Association of America (NAA) entwickelt wurde und ein Vorläufer des IPTC-NAA Standards ist. Die letzte Version, 4.1, stammt von 1997.

Interframe-Kompression bezieht sich bei Videos auf die Komprimierung über die ganze Sequenz hinweg, wobei im Gegensatz zur Intraframe-Kompression mehrere Bilder zusammengefasst werden können. Dies hat den Nachteil, dass eventuell auftretende Fehler nicht nur ein Einzelbild, sondern eine ganze Sequenz an Bildern korrumpern.

Interpolation bezeichnet in der Mathematik Methoden, um für gegebene Werte eine stetige Funktion zu finden. Anwendung findet dies beispielsweise bei der Vergrößerung von Rastergrafiken. Dabei werden die fehlenden Bildpunkte mittels Interpolation berechnet, wobei dies immer nur eine Annäherung sein kann.

Intraframe-Kompression beschränkt sich bei Videos auf die Komprimierung der Daten jedes Einzelbildes für sich, im Gegensatz zur Interframe-Kompression.

IPTC-NAA (kurz auch nur IPTC) ist ein Standard zur Speicherung von Metadaten in Bilddateien. Er wurde vom International Press Telecommunications Council (IPTC) zusammen mit der Newspaper Association of America (NAA) entwickelt. Vorläufer dieses Standards ist der IIM-Standard.

ISBN (Internationale Standardbuchnummer) ist eine Nummer, die verwendet wird, um Bücher eindeutig zu kennzeichnen. Eine ISBN ist ein persistenter Identifikator.

ISO (International Organization for Standardization) eine internationale Vereinigung von Normungsorganisationen. Sie erarbeitet internationale Normen in allen Bereichen mit Ausnahme der Elektrik und der Elektronik, für die die Internationale elektrotechnische Kommission (IEC) zuständig ist, und mit Ausnahme der Telekommunikation, für die die Internationale Fernmeldeunion (ITU) zuständig ist.

ISO 639 XXX

ISO 8601 ist eine internationale Norm, die Empfehlungen für numerische Datumsformate und Zeitangaben enthält. Der Titel der Norm ist *Data elements and interchange formats – Information interchange – Representation of dates and times* (dt. Datenelemente und Austauschformate – Informationsaustausch – Darstellung von Datum und Uhrzeit). Eine Datums- und Zeitangabe kann beispielsweise folgendermaßen aussehen: 2015-06-03T12:45:30+02:00. Diese Zeichenfolge steht für den 03. Juni 2015 um 14:45 (und 30 Sekunden) in Berlin. Die Zeichenfolge *+02:00* gibt dabei die Differenz zur UTC an. Auch Zeitspannen können mit der Norm dargestellt werden, wobei beispielsweise *P3Y6M4DT12H30M5S* für 3 Jahre, 6 Monate, 4 Tage, 12 Stunden, 30 Minuten und 5 Sekunden steht, oder die Zeichenfolge *T2H2M* für 2 Stunden und 2 Minuten steht.

IT (Informationstechnologie) ein Oberbegriff für die Informations- und Datenverarbeitung sowie für die dafür benötigte Hard- und Software. Häufig wird auch die englisch ausgesprochene Abkürzung IT verwendet.

JPEG (Joint Photographic Experts Group) ist ein Dateiformat zur Speicherung von Bilddaten, das verlustbehaftete Komprimierung verwendet und eine Farbtiefe von 32 Bit anbietet. Es wird in der Norm ISO/IEC 10918-1 bzw. CCITT Recommendation T.81 beschrieben und eignet sich *nicht* für die Langzeitarchivierung. Metadaten können im Exif- oder IPTC-NAA-Format gespeichert werden.

JPEG 2000 ist ein Dateiformat zur Speicherung von Bilddaten. Es ist als ISO/IEC 15444 zertifiziert. Im Gegensatz zu JPEG ermöglicht die verwendete diskrete Wavelet-Transformation auch eine verlustfreie Kompression, weshalb dieses Format für die Langzeitarchivierung verwendet werden kann.

JPG siehe JPEG.

Khronos Group wurde 2000 von als Konsortium von Industrievertretern gegründet, um sich für die Erstellung und Verwaltung von offenen Standards im Multimediacbereich einzusetzen.

Kompression siehe Komprimierung.

Komprimierung (auch Kompression) ist die Reduzierung der Dateigröße mit Hilfe verschiedener Algorithmen. Es gibt verlustfreie und verlustbehaftete Komprimierungsmethoden.

Konvertierung ist die Überführung eines Dateiformates in anderes Format. Konvertierung kann verlustbehaftet sein.

Lagarith ist ein verlustfreier Codec für Videodaten, der eigens für Windows-Systeme entwickelt wurde. Lagarith ist mit HuffYUV verwandt.

LOD (Level of Detail) siehe Detaillierungsgrad.

LP (light position file) ist eine textbasierten Datei für die Erstellung von RTIs. In der Datei werden die Positionen der Lichtquelle für jedes einzelne Bild gespeichert.

LPCM (Lineare Puls-Code-Modulation) verwendet für das Abtasten eines Signals einen gleich groß bleibenden Wertebereich bei der Quantisierung (also immer die gleiche Messskala).

lpi (lines per inch) gibt an wie viele Zeilen mit Bildpunkten pro Zoll eine Rastergrafik enthält. Je größer die Punktdichte, desto größer auch die Auflösung des Bildes.

Luma steht für die Helligkeit eines Bildpunktes, bezogen auf Farbmodelle für selbstleuchtende Geräte.

Luminanz siehe Luma.

LZW (Lempel-Ziv-Welch-Algorithmus) ist ein Algorithms zur verlustfreien Komprimierung von Bilddateien. Bis 2004 war diese Methode nur eingeschränkt verwendbar, da Patente darauf bestanden. Diese Patente sind inzwischen ausgelaufen, weshalb LZW-Komprimierung frei verwendet werden kann. Allerdings wurde die Verwendung in der Langzeitarchivierung noch nicht ausreichend erprobt.

M-JPEG (Motion JPEG) ist ein Videocodec, bei dem jedes Einzelbild separat als JPEG-Bild komprimiert wird. Er eignet sich *nicht* zur Langzeitarchivierung, da die Kompression verlustbehaftet ist.

Matroska (MKV für Video oder MVA für Audio) ist ein offenes Containerformat, das eine große Bandbreite von Codecs und ergänzenden Inhalten unterstützt. Das Format basiert auf einer binären Variante von XML, nämlich EBML, was eine zukünftige flexible Erweiterung erlaubt, jedoch auch sicher sicher stellt, dass ältere Programme weiterhin damit umgehen können. Das Format ist fehlertolerant und kann bis zu einem gewissen Grad auch beschädigte Dateien wiedergeben. Für die Langzeitarchivierung wird die Verwendung von den Codecs FFV1 für Video und FLAC für Audio empfohlen.

Metadaten sind Informationen über eine Datei oder ein Dokument. Sie umfassen sowohl technische (Dateiname, Dateigröße etc.) als auch inhaltliche Angaben, wie z.B. Schlagworte. Bei einem Foto beinhalten Metadaten beispielsweise Informationen über den Fotografen, die verwendete Ausrüstung, den Aufnahmeort etc.

MIME-Type (Internet Media Type) wird im Internet verwendet, um in einer Datei zu spezifizieren um welches Format es sich handelt. MIME-Types werden von der IANA verwaltet.

Motion JPEG 2000 (MJ2) ist ein speziell für die Archivierung entwickeltes Containerformat für Videos. MJ2 ist unter ISO/IEC 15444-3:2007 zertifiziert und als MIME-Type video/mj2 registriert. Das Format verwendet einen eigenen Codec, der jedes Einzelbild separat und verlustfrei als Bild in JPEG 2000 speichert.

MOV ist ein proprietäres Containerformat für Videos, das von Apple entwickelt wurde und beispielsweise in QuickTime verwendet wird.

MP3 (MPEG-1 Audio Layer III) ist ein Audiocodec, der in MPEG-1 spezifiziert wurde und seit 1991 unter ISO/IEC 11172 zertifiziert ist. Der Codec hat eine weite Verbreitung gefunden, verwendet jedoch verlustbehaftete Kompression.

MPEG (Moving Picture Experts Group) (engl. „Expertengruppe für bewegte Bilder“) ist eine Gruppe von Experten, die verschiedene Standards für Video und Audio erstellen. Es gibt mehrere Generationen von Standards, darunter MPEG-1, MPEG-2, MPEG-4 oder MPEG-7. Die Standards beschreiben nicht nur Containerformate, sondern auch Codecs für Video und Audio oder Vorgaben für die Dokumentation von Multimedia-Dateien (MPEG-7). Viele MPEG-Standards sind auch ISO/IEC zertifiziert.

Multiplexing bezeichnet das Zusammenführen der einzelnen Komponenten einer Videodatei in ein Containerformat.

Muxing siehe Multiplexing.

MXF (Material eXchange Format) ist ein offenes Containerformat für Videos, das von der SMPTE standardisiert wird. Es ist als MIME-Type application/mxf registriert und kann, zusammen mit der verlustfreien Variante des Codecs JPEG 2000, für die Langzeitarchivierung verwendet werden. MXF ist eine Teilmenge des Advanced Authoring Format (AAF) und wird vor allem im Bereich Kino und Fernsehen in spezialisierten Varianten verwendet.

NISO (National Information Standards Organization) ist eine Standardisierungsorganisation aus den USA. Sie wurde 1939 gegründet und entwickelt technische Standards für den bibliografischen und bibliothekarischen Bereich. Die Namen der Standards beginnen jeweils mit "ANSI/NISO Z39.". Die NISO ist auch im entsprechenden Komitee der ISO vertreten.

Nyquist-Frequenz ist nach Harry Nyquist benannt und besagt, dass das originale (analoge) Signal exakt reproduziert werden kann, wenn die Abtastrate doppelt so hoch wie die höchste Signalfrequenz ist.

OAIS (Open Archival Information System) Offenes Archiv-Informations-System (ISO-Standard 14721:2003). Der Grund für die Entwicklung dieses Modells bestand in der Einsicht, dass elektronisch archivierte Dokumente nach längerer Zeit aus vielfältigen Gründen nicht mehr lesbar sein könnten. Das Referenzmodell beschreibt ein Archiv als Organisation, in dem

Menschen und Systeme zusammenwirken, um einer definierten Nutzer-
schaft Archivgut verfügbar zu machen. Die Implementierung eines OAIS-
konformen Archivs ist dabei jedoch nicht festgelegt. „Offen“ bedeutet, dass
die Entwicklung des OAIS sich in offenen Foren abspielt. Dies bezieht sich
nicht auf den uneingeschränkten Zugang zu einem Archiv.

OASIS (Organization for the Advancement of Structured Information Standards)
ist eine Organisation, die sich mit der Weiterentwicklung von Standards
im Bereich IT-Sicherheit, Internet der Dinge und weiteren beschäftigt.

OCR (Optical Character Recognition, optische Zeichenerkennung) ist
die Bezeichnung für den automatisierten Prozess, bei dem Text in digitalen
Bildern (z.B. Scans von Dokumentseiten) automatisch in einzelne Buch-
stabenzeichen umgewandelt wird.

ODF (Open Document Format) ist ein offenes XML-basiertes Dateiformat
für Textdokumente (ODT), Tabellenkalkulationen (ODS), Präsentationen
(ODP), Datenbanken (ODB), Grafiken (ODG) und mathematische For-
meln (ODF). Es wurde von einem technischen Komitee unter der Lei-
tung der Organization for the Advancement of Structured Information
Standards (OASIS) entwickelt und ist seit 2006 als internationale Norm
ISO/IEC 26300 verfügbar.

ODS ist ein Teil von ODF zur Speicherung von Tabellenkalkulationen. Es ist
offen verfügbar und basiert auf XML.

ODT ist ein Teil von ODF zur Speicherung von Textdokumenten. Es ist offen
verfügbar und basiert auf XML.

Ogg ist ein offenes Containerformat für Audio und Video , das von der Xiph.Org
Foundation entwickelt wurde. Es verwendet Codecs mit verlustbehafteter
Kompression und ist vor allem für das Streaming von Multimediadaten
gedacht.

Opus ist ein Audiocodec, der in dem Format Ogg verwendet wird. Er kompri-
miert verlustbehaftet.

PBCore ist ein in den USA von der Public Broadcasting entwickeltes Metada-
tenschema für Film und Ton.

PCM (Puls-Code-Modulation) bezeichnet das Abtasten eines Signals in ei-
ner vorgegebenen Abtastrate und die anschließende Quantisierung des ge-
messenen Wertes. Es handelt sich um ein Verfahren, um beispielsweise ein
analoges Audiosignal in ein digitales Signal umzuwandeln.

PDF (Portable Document Format) wurde 1993 von Adobe Systems entwi-
ckelt, um den Datenaustausch zu erleichtern. Es ist ein plattformunabhän-
giges, offenes Dateiformat, das 2008 mit der Version 1.7 als ISO-Standard
zertifiziert wurde und seitdem von der ISO weiter gepflegt wird.

PDF/A ist gezielt als stabiles, offenes und standardisiertes Format für die
Langzeitarchivierung unterschiedlicher Ausgangsdateien entwickelt wor-
den. Insgesamt sind aktuell drei, aufeinander aufbauende Versionen von

PDF/A-Formaten zu unterscheiden, von denen PDF/A-1 (seit 2005, ISO 19005-1) und PDF/A-2 (seit 2011) echte Archivformate sind, da sie das Objekt in einem stabilen Zustand konservieren. Bei PDF/A-3 (seit 2012) handelt es sich hingegen um einen Container, in den beliebige Dateiformate eingebettet werden können. Letzteres ist nur dann für die Langzeitarchivierung geeignet, wenn alle eingebetteten Dateien in einem anerkannten Archivformat vorliegen.

persistenter Identifikator (PID) ist eine einzigartige Zeichenkette, die einem Etwas dauerhaft zugeordnet wird, um dieses Etwas eindeutig zu identifizieren oder zu referenzieren. Ein PID wird exakt einmal vergeben. Beispielsweise ist die ISBN-Nummer ein persistenter Identifikator für Bücher.

Pillarboxing bezeichnet bei Videodateien das Einfügen eines Bildes in ein Breitbildformat, wobei schwarze Ränder in Kauf genommen werden. Ein typisches Beispiel ist die Wiedergabe von älteren Videos im Format 4:3 auf neueren Wiedergabemedien mit dem Format 16:9.

Pixel ist ein Bildpunkt einer Rastergrafik und das kleinste Element darin, das kontrolliert werden kann. Auch digitale Videos bestehen in ihren kleinsten Elementen aus Pixeln, wobei diese im Gegensatz zu denen bei Rastergrafiken, nicht immer quadratisch sind, sondern aufgrund des analogen Ursprungs von einigen Videoformaten auch rechteckig sein können.

plain text (reiner Text) beschreibt Textdokumente, die nur Text (und Zahlen) enthalten und meist als TXT-Datei gespeichert werden. Formatierungsangaben, Tabellen, Bilder etc. sind nicht in plain-text-Dateien enthalten.

PNG (Portable Network Graphics) ist ein Dateiformat zur Speicherung von Bilddaten. Es bietet eine verlustfreie Komprimierung, hat eine Farbtiefe von 32 Bit und ermöglicht die Speicherung von Transparenzen. Das Format ist primär für die Verwendung im Internet gedacht, weshalb auch nur das RGB-Farbmodell unterstützt wird. Exif-Daten können nicht gespeichert werden. Das Format eignet sich nicht für digitale Fotos.

ppi (pixel per inch) gibt an wie viele Bildpunkte, Pixel, pro Zoll eine Rastergrafik enthält. Je größer die Punktdichte, desto größer auch die Auflösung des Bildes.

Profil fasst Konfigurationsmerkmale für bestimmte Anwendungsbereiche zusammen. Siehe z.B. Videoprofil.

PTM (Polynomial Texture Map) ist eine Repräsentationsform von Bildern mit Hilfe von Funktionen, statt einzelner Farbwerte. Im Gegensatz zu Rastergrafiken werden für die einzelnen Pixel von PTMs nicht nur feste Farbwerte gespeichert, sondern zusätzlich eine Funktion, die mit Hilfe der Parameter l_u und l_v die Leuchtdichte der Oberfläche berechnet.

Quantisierung bezeichnet bei der Digitalisierung eines analogen Signals das Runden des gemessenen Wertes auf den nächstliegenden Wert der verwendeten Messskala. Da die verwendete Messskala nur bestimmte Werte abdeckt, kann die digitale Repräsentation des Signals vom Original abweichen.

Rastergrafik (auch Pixelgrafik) ist ein digitales Bild, das mittels rasterförmig angeordneter Pixel beschrieben wird. Jedem Pixel ist dabei ein Farbwert zugeordnet. Rastergrafiken haben eine fixe Größe und Auflösung und sind daher nicht skalierbar. Rastergrafiken können sein: Digitale Fotografien jeder Art, Satellitenbilder, digitalisierte Bilder (Scans), Screenshots sowie digitale Originalbilder und -grafiken.

Rastern bezeichnet den Vorgang der Umwandlung einer Vektorgrafik in eine Rastergrafik.

RAW ist ein jeweils kamera-, modell- und herstellerspezifisches Dateiformat von Digitalkameras, bei dem die Kamera die Daten nach der Digitalisierung weitgehend ohne Bearbeitung auf das Speichermedium schreibt. Die Schreibweise RAW hat sich analog zur Schreibweise Raw eingebürgert.

RF64/MBWF ist eine Erweiterung des BWF-Formats, das von der EBU entwickelt wurde, um Dateien zu speichern, die größer als 4GB sind oder mehr als zwei Tonkanäle beinhalten.

RFC (Request for Comments) sind Dokumente in denen Methoden, Ein geschäften, Forschungen oder Innovationen im Zusammenhang mit dem Internet und der damit verbundenen Systeme veröffentlicht werden. Sie können als Richtlinien dienen oder werden sogar von der *Internet Engineering Task Force* zu Standards erhoben.

RGB ist ein additives Farbmodell, in dem die Farben durch mischen der drei Grundfarben Rot, Grün und Blau dargestellt werden. Übliche RGB-Farträume sind sRGB und Adobe RGB.

RIFF (Resource Interchange File Format) ist ein von Microsoft und IBM entwickeltes Containerformat für Multimediadateien. Je nachdem ob es sich um Audio- oder Videoinhalte handelt, kann es sich bei dem Container um eine WAVE- oder AVI-Datei handeln.

RTI (Reflectance Transformation Imaging) ist eine computergestützte Fotografiermethode, mit der von einem Objekt mehrere Bilder mit fixierter Kameraposition und variablen Beleuchtungspositionen gemacht werden. Die Bilder werden zu einem PTM zusammengerechnet. In der resultierenden Datei kann die Position der Lichtquelle verändert werden, um beispielsweise die Oberfläche des aufgenommenen Objekts im Schräglicht untersuchen zu können.

Screenreader ist ein Programm, das Sehbehinderten und Blinden Informationen vom Bildschirm auf anderem Wege, meistens akustisch, vermittelt.

Screenshot, auch Bildschirmfoto, bezeichnet die Abbildung des Inhalts eines Computerbildschirms. Ein Screenshot ist immer eine Rastergrafik.

SD PAL beschreibt ein bestimmtes Seitenverhältnis und eine bestimmte Bildgröße von Videodateien.

Seitenverhältnis bezieht sich auf die Darstellung eines Videobildes und wird als Verhältnis von Breite zu Höhe angegeben, wie beispielsweise 4:3.

SGML (Standard Generalized Markup Language) ist eine Auszeichnungssprache, die seit 1986 als ISO-Standard ISO 8879 besteht.

SHP eine zur Speicherung der Geometriedaten. Es ist Bestandteil des Dateiformats ESRI-Shape.

SHX dient als Index der Geometrie zur Verknüpfung der Sachdaten (auch Attributdaten genannt) eines ESRI-Shapefiles.

Skalieren bedeutet z.B. ein digitales Bild oder 3D-Objekt zu vergrößern oder zu verkleinern.

SMPTE (Society of Motion Picture and Television Engineers) ist eine internationale Organisation im Bereich der professionellen Film- und Videotechnik, die 1916 gegründet wurde. Die Organisation kümmert sich um Standards, Praxisempfehlungen und Richtlinien im Film-, Video- und Audiobereich.

Software meint Programme, die auf einer Hardware laufen, beziehungsweise deren Funktionalität erst herstellen. Dies kann beispielsweise das Betriebssystem auf einem Computer oder einer Totalstation sein. Es kann auch ganz speziell ein bestimmtes Programm gemeint sein, wie etwa das Programm zur Verarbeitung von Messdaten.

SQL (Structured Query Language) eine Datenbanksprache zur Definition, Abfrage und Manipulation von Daten in relationalen Datenbanken. SQL ist von ISO standardisiert und wird von fast allen gängigen Datenbanksystemen unterstützt.

SQL-DDL (SQL-Data Definition Language) eine Computersprache, die verwendet wird, um Datenstrukturen und verwandte Elemente zu beschreiben, zu ändern oder zu entfernen. Als Datenbanksprache ist DDL die Datenbeschreibungssprache einer Datenbank. In SQL liegt sie in Form englischer Befehlsklauseln vor (z. B. CREATE TABLE PROJECT (PROJNO DECIMAL(6,0) NOT NULL PRIMARY KEY, ...)).

sRGB (Standard-RGB) ist ein häufig verwendeteter RGB-Farbraum.

Stapelverarbeitung bezeichnet die automatisierte Verarbeitung von mehreren Dateien mittels eines Programms. Beispielsweise können Bilddateien im Stapel gedreht und skaliert werden.

Streaming bezeichnet in der elektronischen Datenverarbeitung die kontinuierliche Übertragung eines Datenstroms. Beispielsweise können mittels Audiostreaming Audiodaten vergleichbar zu einer Radiofunkübertragung gesendet und empfangen werden.

SVG (Scalable Vector Graphics) ist ein offenes XML-basiertes Format zur Speicherung von Vektorgrafiken, das vom W3C empfohlen wird.

Tabellenkalkulation kann sowohl das Programm (z.B. OpenOffice Calc), als auch die damit erstellte Datei mit einer oder mehreren Tabellen meinen.

Tag siehe Auszeichnungselement.

TEI (Text Encoding Initiative) ist eine Initiative, die speziell für Geistes-, Sozial- und Sprachwissenschaften ein auf XML basierendes Dokumentenformat entwickelt, das den Austausch von maschinenlesbaren Texten unterstützen und standardisieren soll. Die aktuelle Version ist P5.

Textbasiertes Format ist ein Format, das seinen Inhalte in Form von Darstellbaren Zeichen beschreiben. Textbasierte Formate sind auf eine bestimmte Weise strukturiert oder verwenden Auszeichnungssprachen. Im Gegensatz zu Binärdateien, können textbasierte Formate beispielsweise mit einem Texteditor geöffnet und auch von Menschen gelesen und verstanden werden. Beispiele für textbasierte Formate sind: XML, CSV, PLY.

Textbegrenzungszeichen wird in textbasierten Tabellenformaten vor und nach den Zellen eingefügt, falls das Trennzeichen auch als Wert innerhalb der Zelle verwendet wird.

Textdatei ist eine Datei, die nur darstellbare Zeichen enthält. Sie enthält keine Formatierungssangaben oder externe Medien wie Bilder oder Tabellen.

Textdokument ist eine Datei, die neben darstellbaren Zeichen auch Formatierungssangaben und externe Medien enthält. Für Textdokumente gibt es spezielle Speicherformate wie DOCX oder ODT.

Texteditor ist ein Editor, der speziell für die Bearbeitung von Textdateien entwickelt wurde.

Textverarbeitungsprogramm ist ein Programm, um Textdokumente zu erstellen und zu bearbeiten. Es bietet Funktionalitäten, um Formatierungssangaben zu machen und Medien wie Bilder oder Tabellen einzubetten. Frei verfügbare Beispiele wären OpenOffice Writer und LibreOffice Writer.

Theora ist der für Ogg-Dateien entwickelte Videocodec, der allerdings nur eine verlustbehaftete Kompression bietet und daher *nicht* für die Langzeitarchivierung zu empfehlen ist.

TIFF (Tagged Image File Format) ist ein Dateiformat zur Speicherung von Rastergrafiken. TIFF bietet verschiedene Speicheroptionen an, wie die Verwendung von LZW-Komprimierung, mehrere Seiten, Ebenen und eingebettete Metadaten. Der kleinste gemeinsame Nenner aller Speicheroptionen bildet das unkomprimierte Baseline TIFF.

Tonkanal entspricht in einer Audiodatei einer Tonspur. Eine Audiodatei kann eine, zwei oder mehr Tonspuren enthalten und diese können bei der Wiedergabe, je nach verwendetem Tonsystem, an unterschiedliche Ausgabegeräte geleitet werden.

Tonsystem gibt an, wie die Anordnung der Ausgabegeräte für die Tonkanäle einer Audiodatei intendiert ist.

Transcodierung bezeichnet für Audio- und Videodateien die Konvertierung von einem Codec in einen anderen.

Trennzeichen ist in textbasierten Tabellenformaten (wie z.B. CSV) das Zeichen, das verwendet wird, um die einzelnen Zellen voneinander zu trennen.

True Color ist eine Bezeichnung für Bilder, die einen natürlichen Eindruck beim Betrachter erwecken. Man spricht auch von Echtfarben. Dabei hat jedes Pixel 256 mögliche Werte (also 8 Bit) für jede der Farbkomponenten. True Color RGB-Bilder weisen daher eine Farbtiefe von 24 (3×8) und CMYK-Bilder von 32 (4×8) Bit auf.

TrueType ist ein Standard zur Darstellung von Schriftarten. Die einzelnen Zeichen werden dabei als Vektorgrafiken abgelegt, was eine Skalierbarkeit ermöglicht.

TSV (Tab-Separated Values) ist ein textbasiertes Dateiformat für tabellarische Daten. Hierbei werden einzelne Zellen durch das Tabulator-Zeichen (U+0009) voneinander getrennt. Jede Zeile einer TSV-Datei entspricht der Zeile einer Tabelle. Das Tabulator-Zeichen ist nicht als Inhalt der Zellen erlaubt. TSV ist ein Standard der IANA und ist als MIME-Type *text/tab-separated-values* registriert.

TXT ist ein einfaches reines Textformat (*plain text*).

U3D (Universal 3D Format) ist ein 2005 von der ECMA standardisiertes 3D-Format, das vom 3D Industry Forum mit Mitgliedern wie Intel und Adobe Systems entwickelt wurde. Dieses Format ist nur für die Integration von 3D-Modellen in ein 3D-PDF relevant.

Unicode ist ein Zeichensatz, in dem aktuell für 113.021 Zeichen aus 123 Schriftsystemen eindeutige Codepunkte (*code points*) zugewiesen werden. Die Codepunkte werden mittels einer hexadezimalen Zahl und einem vorangestellten *U+* dargestellt, wie beispielsweise *U+00C4* für ä. Zugleich stellt dieser Zeichensatz die Umsetzung von dem in ISO 10646 beschriebenen universellen Zeichensatz *Universal Character Set* dar. Um den Unicode-Zeichensatz in einem System anwenden zu können, wurden Zeichenkodierungen definiert, die unter dem Namen *Unicode Transformation Format* (UTF) subsumiert werden.

UTC (koordinierte Weltzeit) wurde 1972 eingeführt und ist die gültige Weltzeit. Die mitteleuropäische Zeit ergibt sich aus der Weltzeit durch die Addition von einer (Winterzeit) oder zwei (Sommerzeit) Stunden. Mit Hilfe der UTC können nach ISO 8601 Datum und Zeit angegeben werden, wie beispielsweise 2015-06-03T12:45:30+02:00 für den 03. Juni 2015 um 14:45 (und 30 Sekunden) in Berlin. Die Zeichenfolge " '+02:00'" gibt dabei die Differenz zur UTC an, die an dem Tag wegen der geltenden Sommerzeit bei zwei Stunden lag.

UTF (Unicode Transformation Format) ist eine Menge von Zeichenkodierungen für den Unicode-Zeichensatz. Zu den häufigsten gehören dabei UTF-8 und UTF-16, die im Web und in verschiedenen Betriebssystemen eine große Verbreitung gefunden haben. Der Unterschied besteht dabei in der Zahl der pro Zeichen verwendeten Bytes. Eine Besonderheit von UTF-8 besteht darin, dass die Bytedarstellungen der ersten 128 Zeichen denen der 128 Zeichen des ASCII-Zeichensatzes entspricht. Weitere Varianten von UTF sind UTF-1, UTF-7 und UTF-32.

Valide (Validität) bezieht sich im Zusammenhang von Auszeichnungssprachen auf ein Dokument, das sich an die angegebene Dokumenttypdefinition, XSD oder andere angegebene Grammatik hält. Beispielsweise ist ein XML-Dokument valide wenn die in der XSD definierte Struktur eingehalten wird. Wird keine Dokumentgrammatik für das Dokument spezifiziert, dann das Dokument auch nicht validiert werden.

VBR (engl. *variable bitrate, variable Bitrate*) überträgt, je nach Komplexität des Inhaltes, pro Sekunde variierende Datenmengen. Da ruhige und einfache Übergänge und Inhalte mit einer niedrigen Datenrate gespeichert werden können, kann der Speicherplatzverbrauch verringert werden. Siehe auch Datenrate.

Vektorgrafik ist ein digitales Bild, das mittels Parametern wie Linien, Kreisen, Polygonen oder Kurven beschrieben wird. Vektorgrafiken sind stufenlos skalierbar.

Verwaistes Werk (engl. *orphan work*) ist ein Werk an das keine Metadaten mehr geknüpft sind. Beispielsweise lässt ein Bild ohne eingebettete, bzw. verknüpfte Metadaten keinerlei Rückschlüsse auf dessen Urheber und Inhalt zu.

Videoprofil, auch einfach nur Profil. Da Videodateien auf den unterschiedlichsten Medien wiedergegeben werden können, die hardwareseitig stark differierende Leistung bieten, gibt es für einige Containerformate Profile, die Konfigurationsmerkmale für bestimmte Anwendungsbereiche zusammenfassen, wie beispielsweise die Profile *Baseline*, *Extended* oder *Main* für MPEG-4.

Ein Profil stellt eine Zusammenfassung bestimmter Merkmale dar und ist jeweils auf verschiedene Anwendungen zugeschnitten, wie beispielsweise mobile Applikationen, professionelle Videobearbeitung oder Videostream per Internet. Dazu definiert ein Profil einen Satz an unterstützten Eigenschaften, wie etwa verlustfreie Kompression oder eine bestimmte Farbunterabtastung.

Zusätzlich kann ein Profil durch Level weiter definiert werden. Die verschiedenen Level definieren maximale Werte für etwa Bitrate, Bildfrequenz oder Bildformat, um beispielsweise die Wiedergabe auf Systemen mit schwacher Leistung zu gewährleisten.

Vorbis ist ein Audiocodec, der in dem Format Ogg verwendet wird. Er komprimiert verlustbehaftet.

VRML (**Virtual Reality Modelling Language**) ist das Vorgängerformat von X3D für 3D-Daten. Die jüngste Version wurde 1997 unter dem Namen VRML97 veröffentlicht.

W3C (**World Wide Web Consortium**) ist das wichtigste Gremium zur Standardisierung des World Wide Web, das 1994 gegründet wurde. Das W3C entwickelt technische Spezifikationen und Richtlinien und stellt diese Vorgänge auch transparent auf deren Webseiten dar. <http://www.w3.org/>

WAVE (Waveform Audio File Format) ist ein Containerformat zur Speicherung von Audiodaten, das von Microsoft und IBM als Teil von RIFF entwickelt wurde. Das Format ist proprietär, aber offen dokumentiert und kann verschiedene Audio codecs speichern. Von der IASA wird WAVE mit linearem PCM als Archivformat empfohlen und es hat sich als Standard de facto etabliert.

Web3D Consortium ist eine 1997 gegründete Organisation, um die Entwicklung des X3D-Formates zu fördern.

WMA (Windows Media Audio) ist ein proprietäres Containerformat für Audiodaten, das von Microsoft entwickelt wurde und verlustbehaftete Kompression verwendet.

Wohlgeformt bezieht sich im Kontext von Auszeichnungssprachen auf Dokumente, welche die Regeln der jeweils verwendeten Auszeichnungssprache einhalten.

X3D (extensible 3D Graphics) wurde vom Web3D Consortium als Format zur Speicherung von 3D-Daten entwickelt und ist seit 2006 ISO-zertifiziert. Es eignet sich sowohl zur Speicherung von einzelnen 3D-Modellen, als auch komplexer 3D-Inhalte, wie etwa Virtual Reality. Es darf nicht mit dem proprietären Format 3DXML verwechselt werden.

XLSX ist ein offenes XML-basiertes Format, das von Microsoft zur Speicherung von Tabellenkalkulationen entwickelt wurde.

XML (Extensible Markup Language) ist eine Auszeichnungssprache, die eine Teilmenge von SGML bildet und die Definition von eigenen Auszeichnungselementen erlaubt, um beliebige Strukturen annotieren zu können. De facto wurde SGML von der einfacher anwendbaren XML verdrängt. XML wird vom W3C gepflegt und entwickelt. Es bildet die Grundlage von vielen weiteren Dateiformaten wie ODT, DOCX, SVG etc. Für XML-Dateien gibt es als Alternative zu einer Dokumenttypdefinition (DTD) die Möglichkeit der Verwendung eines XML Schemas.

XML Schema siehe XSD.

XMP (Extensible Metadata Platform) ist ein ISO-Standard (seit 2012, 16684-1), um Metadaten in digitale Medien einzubetten oder sie dateibegleitend zu speichern.

XSD (XML Schema Definition) dient zur Beschreibung der Struktur und der verwendeten Elemente einer XML-Datei. XSD ist eine Empfehlung des W3C und ist selbst ein XML-Dokument.

YCbCr ist ein Farbmodell für selbstleuchtende Geräte. Es hat drei Kanäle, von der erste (Y) die Helligkeit (Luma) der Bildpunkte beschreibt und für die Übertragung von Schwarz-Weiß-Bildern verwendet wird. Der Farbannteil (Chroma) wird mit den Kanälen Cb und Cr repräsentiert.

YPbPr ist ein Farbmodell für selbstleuchtende Geräte. Es hat drei Kanäle, wovon der erste (Y) die Helligkeit (Luma) der Bildpunkte beschreibt und für die Übertragung von Schwarz-Weiß-Bildern verwendet wird. Der Farbannteil (Chroma) wird mit den Kanälen Pb und Pr repräsentiert.

YUV ist ein Farbmodell für selbstleuchtende Geräte. Es hat drei Kanäle, wovon der erste (Y) die Helligkeit (Luma) der Bildpunkte beschreibt und für die Übertragung von Schwarz-Weiß-Bildern verwendet wird. Der Farbannteil (Chroma) wird mit den Kanälen U und V repräsentiert.

Zeichenkodierung kann als abstrakte Tabelle verstanden werden, in der einer bestimmten Zeichensetzung, dem Zeichensatz, Zahlenwerte zugeordnet werden. Beispielsweise hat der Buchstabe *A* in dem *American Standard Code for Information Interchange* (ASCII) den dezimalen Zahlenwert von 65. In ISO 8859-1 stellt der Wert 228 das Zeichen ä dar, während der gleiche Wert in ISO 8859-7 als δ interpretiert wird. Die Angabe der verwendeten Zeichenkodierung ist entscheidend dafür, ob auf dem Bildschirm öå÷ic oder τεχνη dargestellt wird.

Zeichensatz definiert eine Menge von Zeichen, wie beispielsweise das deutsche Alphabet. Die Darstellung eines Zeichensatzes im Computer wird mit einer Zeichenkodierung realisiert. Beispielsweise kann der Zeichensatz Unicode mit der Zeichenkodierung UTF-8 dargestellt werden.