Konkretisierung des SIPs und Grundsätze zur SIP-Bildung

Die Digitalisierung in Wirtschaft und Verwaltung hat einen grundsätzlichen Wandel bei der täglichen Arbeit in allen Lebensbereichen eingeleitet. Das für die jeweilige Tätigkeit notwendige Wissen wird nicht mehr über analoge Träger wie z.B. Papier fixiert, aufbewahrt und weitergegeben, sondern man nutzt nun digitale Medien und Speichermöglichkeiten dazu. Dasselbe gilt für Bücher und Zeitschriften sowie für diverse Arbeitsabläufe, die inzwischen hauptsächlich auf digitale Fachverfahren basieren und sich nicht mehr auf Laufzettel oder Karteikarten stützen.

Die dabei entstandenen Daten müssen zu großen Teilen – wie ihre analogen „Vorfahren" auch – aufgehoben werden, denn sie sind wichtig für die Datenproduzenten, für Behörden und ggf. auch für die Nachwelt. Um diese Informationen auch über Jahrhunderte hinweg lesbar und damit nutzbar zu halten, müssen sie fachgerecht in digitalen Magazinen von Archiven, Bibliotheken und anderen Institutionen aufbewahrt werden.

Mangels konkreter Vorgaben gibt es derzeit eine bunte Vielfalt von Datenpaketen, die diese archivierende Institutionen vor große Probleme bei der Datenübergabe und ihrer Erhaltung stellen. Für sie, genauso wie für Daten-Produzenten, IT-Dienstleiter und viele weitere Beteiligte wäre es praktisch und gleichzeitig wirtschaftlich, wenn die Datenpaketen nach allgemein gültigen Maßgaben erstellt würden, die deren Verarbeitung zur Routine machen könnten.

Ziel dieses Papiers ist es, Datenproduzenten, IT-Dienstleistern und archivierenden Institutionen eine Handreichung zu bieten, um über Vereinheitlichung der Datenpakete eine Vereinfachung der digitalen Archivierung zu erreichen.

Konkretisierung des SIP

Kommentar [G1]: Ergänzen:
- Informationstypen integrieren
- Einzelne Pakete noch definieren (Transferpaket, SIP etc.)

Die oben genannten Datenpakete der Abgeber werden entsprechen dem OAIS-Referenzmodell[1] Submission Information Package (SIP) genannt und wie folgt beschrieben:

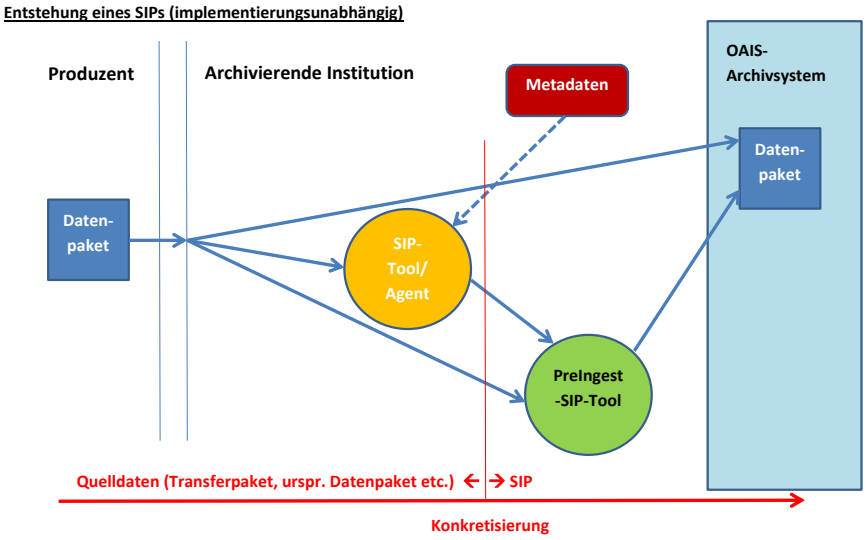
Das Übergabeinformationspaket (SIP)ist das Paket, das von einem Produzenten an das OAIS geschickt wird. Seine Form und sein genauer Inhalt werden typischerweise zwischen dem Produzenten und dem Archiv ausgehandelt (siehe die entsprechenden Standards in 1.5). Die meisten SIPs werden einige Inhaltsinformationen und einige Erhaltungsmetadatenenthalten.[2]

Allerdings ist die Realität deutlich vielschichtiger. Wie schon oben erwähnt, werden Datenpakete in verschiedenen Formen erzeugt und müssen in der Regel nach der Übergabe durch den Produzenten bei der archivierenden Institution noch weiterverarbeitet werden. So ist es oft notwendig, Metadaten zu ergänzen oder das abgegebene Datenpaket so umzuarbeiten, dass es anschließend verlustfrei und integer in das OAIS-konforme digitale Archiv übertragen und lesbar gehalten werden kann. Letztlich

[1] Open Archival Information System (ISO 14721): Siehe nestor-Materialen 16 unter https://www.langzeitarchivierung.de/Webs/nestor/DE/Publikationen/nestor_Materialien/nestor_materialien_node.html;jsessionid=D815D8283717133B03EAAEC6390E69A1.internet552 (zuletzt aufgerufen am 16.10.2019)

[2] Vgl. nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (Hrg.): Referenzmodell für ein Offenes Archiv-Informations-System -Deutsche Übersetzung2.0-, S. 23, online verfügbar unter https://d-nb.info/104761314X/34 (zuletzt aufgerufen am 22.10.2019).

muss man zumindest derzeit zwischen dem ursprünglichen Datenpaket und dem eigentlich SIP im Übergabeworkflow unterscheiden. Im folgenden Schaubild ist dies dargestellt:



Je größer die Annäherung zwischen dem ursprünglichen Datenpaket und dem SIP ist, desto einfacher wäre der Übergabe- und Archivierungsprozess. Um dies zu erreichen, können sich Datenproduzenten und archivierende Institutionen an folgenden Grundsätzen orientieren.

Grundsätze zur SIP-Bildung

Bei den Grundsätze zur SIP-Bildung handelt es sich um die für deutsche archivierende Institutionen angepasste und erweiterte Übersetzung der General Principles der beim EU-Projekt E-ARK entwickelten Common Specification for Information Packages (CSIP).[3]

Die Grundsätze gliedern sich in fünf Bereiche:

1. Allgemeine Grundsätzen
2. Grundsätze zur Identifikation eines Informationspakets
3. Struktur eines Informationspaktes
4. Metadaten eines Informationspakets
5. Authentizität und Integrität eines Informationspakets

Die in Großbuchstaben geschriebenen Modalverben drücken aus, welche Bedeutung die einzelnen Grundsätze bei der SIP-Erstellung besitzen:

- MUSS = Verpflichtend

[3] Siehe https://earkcsip.dilcis.eu/pdf/eark-csip.pdf (zuletzt aufgerufen am 16.10.2019)

- SOLL = Empfehlenswert (Abweichung ist zu begründen)
- KANN = Erlaubt (?)

Begriffe:
- Digitales Objekt = siehe Nestor-Handbuch, Punkt 7.2.
- Identifikator =
- Informationspaket = Transferpaket etc. und SIP (siehe oben)
- Informationstyp =
- Ingest = Übernahme in das OAIS-Archivsystem

1. Allgemeine Grundsätze

Grundsatz 1.1

Es MUSS möglich sein, beliebige digitale Objekte~~Daten~~ und~~oder~~ Metadaten ~~unabhängig von Typ oder Format~~ in ein Informationspaket aufzunehmen.

Dies ist eines der wichtigsten Prinzipien für das Informationspaket. Um wirklich „allgemein“ zu sein, DÜRFEN technische Implementierungen des Informationspakets KEINE Einschränkungen oder Restriktionen einführen, die nur für bestimmte digitale Objekte und~~Daten oder~~ Metadaten~~typen~~ gelten. Wenn die Implementierung eines Informationspakets dieses Prinzip nicht erfüllt, kann es nicht branchen- und werkzeugübergreifend eingesetzt werden, wodurch die praktische Interoperabilität eingeschränkt wird. Der archivierenden Institution steht es dabei frei, Vorgaben (z.B., bei Formaten oder Metadaten) zu machen. Ein Informationspaket SOLL auf frei zugängliche Standards oder Quasi-Standards basieren.

Grundsatz 1.2

Das Informationspaket DARF NICHT die Mittel, Methoden oder Werkzeuge für den Ingest~~Austausch~~ einschränken.

Tools und Methoden zum Übertragen von Informationspaketen zwischen Standorten werden ständig weiterentwickelt. Es ist auch möglich, dass für Pa~~e~~kete unterschiedlicher Größe unterschiedliche Methoden bevorzugt werden. Um sicherzustellen, dass ein Informationspaket wirklich plattformübergreifend funktioniert, DÜRFEN KEINE Einschränkungen oder Restriktionen eingeführt werden, die von bestimmten Tools oder Kanälen für den Informationsaustausch nicht erfüllt werden können.

~~Deswegen definiert die Spezifikation für das Informationspaket auch nicht das Prinzip, ein bestimmtes Transferpaket oder einen bestimmten Container zu verwenden. Die Vorgaben für das Informationspaket beschränken sich auf die Struktur und die Anforderungen für Daten und Metadaten innerhalb des Pakets. Jeder kann seine eigenen Methoden zusätzlich zu den Vorgaben des Informationspakets anwenden.~~

Grundsatz 1.3

Das Paketformat DARF NICHT den logischen inhaltlichen Umfang~~Umfang~~ der digitalen Objekte~~Daten~~ und Metadaten definieren, die ein Informationspaket bilden.

Kommentar [2]: Noch zu diskutieren

Formatiert: Schriftart:

Formatiert: Schriftart:

Formatiert: Aufgezählt + Ebene: 1 + Ausgerichtet an: 1,27 cm + Einzug bei: 1,9 cm

Formatiert: Schriftart:

Formatiert: Aufgezählt + Ebene: 1 + Ausgerichtet an: 1,27 cm + Einzug bei: 1,9 cm

Formatiert: Schriftart:

Kommentar [G3]:

Formatiert: Abstand Nach: 10 Pt.

Formatiert: Schriftart: 12 Pt.

Eines der Grundprinzipien ist, dass jedes einzelne digitale Archiv den ~~(intellek-tuellen~~logischen inhaltlichen Umfang~~)~~ ~~Umfang~~ eines Informationspakets und dessen Beziehungen zu anderen Objekten ~~zum dazugehörigen analogen Objekt zu den jeweiligen geltenden Rahmenbedingungen~~ definieren können MUSS. Daher MUSS es bei der Implementierung von Informationspaketen möglich sein, beispielsweise gleichermaßen entweder den gesamten Inhalt einer intellektuellen Einheit~~s ERMS~~ zu einem Informationspaket zusammenzuführen oder Teile davon~~jeden Datensatz und seine Metadaten einzeln aus dem ERMS~~ zu extrahieren und als jeweils ein Informationspaket zu verpacken.

~~Aus dem Vorhergehenden kann auch abgeleitet werden, dass eine Spezifikation für Informationspakete NICHT definieren DARF, dass beispielsweise ein SIP genau einem AIP entsprechen soll. Stattdessen MUSS das die Spezifikation die Aufnahme von "allem, was der Umsetzer/Programmierer als SIP, AIP oder DIP definieren möchte" und "beliebige Beziehungen (1-1; 1-n; n-1; nm) zwischen SIPs, AIPs" zulassen und DIPs ".~~

Grundsatz 1.4

Das Informationspaket ~~SOLLTE~~ MUSS skalierbar sein.

Eines der praktischen Anliegen für Informationspakete ist ihre Größe. Viele digitale Archive haben Probleme mit Datenobjekten und Metadaten ab einer gewissen Größe, was es besonders schwierig macht, Aufgaben im Zusammenhang mit der Daten- oder Metadatenvalidierung sowie der Identifizierung und Änderung auszuführen. Zum Beispiel können Informationspakete mit relationalen Datenbanken oder genuin digitalen 3D-Filmen leicht TB-Größen erreichen.

Infolgedessen muss jede aktuelle oder zukünftige Implementierung von Informationspaketen geeignete Mechanismen zur Skalierung vorsehen (z. B. über Aufteilung großer Datenmengen oder Metadaten).

Prinzip 1.5

Das Informationspaket MUSS maschinenlesbar und automatisierbar zu verarbeiten sein.

Um das Ziel der Automatisierung von Ingest-, Aufbewahrungs- und Zugriffsabläufen zu unterstützen, muss jede Implementierung von Informationspaketen maschinell umsetzbar sein. Dies bedeutet, dass Entscheidungen über die Verwendung von Metadatensyntax und -semantik sowie die physische Struktur explizit und klar ausgedrückt werden müssen. Dies ermöglicht wiederum, dass die Spezifikation in verschiedenen Tools und Umgebungen identisch implementiert wird.

Grundsatz 1.6

Das Informationspaket MUSS~~SOLLTE~~ technisch interpretierbar~~für Menschen lesbar~~ sein, um eine auch für den Menschen inhaltliche Deutung zu ermöglich.

Bei der Langzeitaufbewahrung müssen wir auch berücksichtigen, dass möglicherweise "vergessene" Informationspakete gefunden werden, zu denen es keine Informationen über

Kommentar [GS4]: Vgl. Protokoll der Sitzung vom 06.02.19: Die Frage, ob die Lesbarkeit mit menschlichem Auge überhaupt möglich ist (eigentlich braucht man dazu immer technische Hilfsmittel) und damit dieser Grundsatz sinnvoll ist, muss noch weiter diskutiert werden.

deren Implementierung oder zu den verwendeten Zugriffstools mehr gibt. Für diese Szenarien ist es entscheidend sicherzustellen, dass die Struktur und die Metadaten des Informationspakets mit minimalem Aufwand verständlich sind, indem einfache Tools wie Texteditoren und Datei-Viewer verwendet werden.
In der Praxis bedeutet dies, dass bei jeder Implementierung von Informationspaketen sichergestellt werden muss, dass die Namenskonventionen für Ordner und Dateien die Identifizierung der einzelnen Teile des Informationspakets durch den Benutzer ermöglichen und der Aufbau des Pakets eindeutig ist.

~~**Grundsatz 1.7**~~

~~*Das Informationspaket DARF NICHT eine bestimmte Bestandserhaltungsmethode vorschreiben*~~

~~Unterschiedliche archivierende Institutionen und unterschiedliche Daten verlangen unterschiedliche Methoden für die Langzeitaufbewahrung. Migration und Emulation sind dafür typische Beispiele. Die Implementierung von Informationspaketen DARF NICHT die Verwendung einer bestimmten Aufbewahrungsmethode vorschreiben. Stattdessen sollten Benutzer die Möglichkeit haben, beliebige Daten oder Metadaten zu dokumentieren und/oder hinzuzufügen, falls dies zu Aufbewahrungszwecken erforderlich ist.~~

Kommentar [GS5]: Vgl. Protokoll der Sitzung vom 06.02.19: Inwieweit spätere Möglichkeiten zur digitalen Bestandserhaltung bei einem SIP berücksichtigt werden können, muss noch genauer untersucht werden.

Grundsatz 1.7[4]

Die Spezifikation des Informationspakets MUSS offen und frei sein.

Spezifikationen (inkl. der dazugehörigen Informationstypen) zur Beschreibung von Informationspaketen müssen vollumfänglich frei zugänglich sein und die Verwendung unterliegt keinerlei Beschränkungen.

~~[Verwendungsfreiheit]~~

Grundsatz 1.8

Die Komplexität der Spezifikation eines Informationspakets SOLL angemessen sein.

Eine angemessene Komplexität begünstigt grundsätzlich Weiterentwicklungen und Verbreitung. Dies ist für jede Form von Institutionen in der Regel besser handhabbar.

~~[offene Formate]~~

~~Grundsatz 1.9~~

Kommentar [6]: Entwurf von Herrn Steidl folgt; ggf. über 1.2 abgedeckt?

~~[Robustheit]~~

2. Grundsätze zur Identifikation eines Informationspakets

[4] Neu, entspricht nicht Principle 1.7 im CSIP.

~~**Grundsatz 2.1**~~

~~*Der OAIS-Typ des Informationspakets (SIP, AIP oder DIP) MUSS eindeutig angegeben werden.*~~

~~Eine der ersten Aufgaben bei der Analyse eines Informationspakets besteht darin, seinen aktuellen Status im gesamten Archivierungsprozess zu ermitteln. Daher muss sich jedes Informationspaket ausdrücklich und einheitlich als SIP, AIP oder DIP ausweisen.~~

~~**Grundsatz 2.1[5]2**~~

~~*Jedes Informationspaket SOLLTE seinen MUSS den oder die Inhaltsinformationstypen mitteilen/en seiner Daten und Metadaten eindeutig identifizieren.*~~

~~Wie in Grundsatz 1.1 angegeben, MUSS jedes Informationspaket jede Art von Daten und Metadaten enthalten können. Gleichzeitig haben wir in früheren Abschnitten das Konzept der Inhaltsinformationstypen eingeführt, mit denen Benutzer eine detailliertere Kontrollsteuerung und eine fein abgestimmte Interoperabilität erreichen können. Daher MUSS jedes Informationspaket eine Aussage darüber enthalten, welche Inhaltsinformations-typspezifikation/en im Informationspaket befolgt wurden, oder im Gegenteil, es muss eindeutig angegeben werden, dass keine spezifische Inhaltsinformationstypspezifikation vorliegt.~~

~~—~~

~~Der Sinn der Prinzipien 1.1, 2.1 und 2.2 besteht darin, dass wir, sobald diese bei Implementierungen angewandt werden, entsprechende modulare Identifikations- und Validierungswerkzeuge und -workflows entwickeln können. Während allgemeine Komponenten unabhängig vom Content Information Type übergeordnete Aufgaben ausführen können, ist es hier möglich, automatisch zu erkennen, welche zusätzlichen inhaltsspezifischen Module ausgeführt werden müssen.~~

Grundsatz 2.2~~3~~[6]

Jedes Informationspaket MUSS einen in der archivierenden Institution eindeutigen und dauerhaften Identifikator ~~im digitalen Archiv~~ haben oder erhalten.

Um ein digitales Archiv zu verwalten und die Bereitstellung sicherzustellen, MUSS jedes ~~im digitalen Archiv~~ gespeicherte Informationspaket mindestens innerhalb des digitalen Archivs~~sen~~ eindeutig identifiziert werden. Gleichzeitig darf eine entsprechende Implementierung die Auswahl des Identifikatorensystems NICHT einschränken, solange dieses im gesamten digitalen Archiv konsistent implementiert ist.

~~**Grundsatz 2.3[7]b**~~

~~[5] Im CSIP Principle 2.2. Principle 2.1 ist weggefallen.~~

[6] Im CSIP Principle 2.3.

~~[7] Fehlt im CSIP.~~

Kommentar [GS7]: Vgl. Protokoll der Sitzung vom 06.02.29: Der Grundsatz kann sich bei der AG-Arbeit nur auf SIP und AIP (als Prozessende der SIP-Erstellung) beziehen.

Kommentar [ST8]: Kann gestrichen werden, da die Grundsätze nur für den Zustand „SIP" stehen und nicht für den Workflow von Quelldaten zu AIP/DIP

Kommentar [SG9]: Jetzt 3.5

Kommentar [GS10]: Vgl. Protokoll der Sitzung vom 06.02.29: Inwieweit spezielle Datenprofile zu berücksichtigen sind, muss noch eingehend diskutiert werden.

Kommentar [GS11]: Z.B. ERMS, Geodaten, SIARD; siehe https://earkcsip.dilcis.eu/pdf/eark-csip.pdf S. 12X

Es muss zwischen Informationstyp/Profil und Content unterschieden werden.

Kommentar [ST12]: Muss genauer nach Informationstypen ausgearbeitet werden.

~~*Jedes Informationspaket KANN einen beim Produzenten eindeutigen und dauerhaften Identifikator haben.*~~

~~*[...]*~~

Kommentar [SG13]: Jetzt Punkt 4.4

Grundsatz 2.4

Jedes Informationspaket SOLLTE einen Identifikator besitzen, der global eindeutig und dauerhaft ist.

Zusätzlich zum vorherigen Grundsatz wird empfohlen, dass das im digitalen Archiv verwendete Identifikatorensystem gewährleistet, dass die Identifikatoren weltweit eindeutig und persistent sind (z.B. UUID, URN, DOI). Solche Identifikatoren erlauben es digitalen Archiven, einfacher institutionenübergreifend Daten auszutauschen und wiederzuverwenden (z. B. bei nationalen oder internationalen Portalen oder archivübergreifender AIP-Duplizierung). Die Spezifikation der Informationspakete DARF jedoch NICHT die Auswahl des genauen Identifikatorensystems einschränken.

Grundsatz 2.5

Alle Teile eines Informationspakets ~~MÜSSEN~~ SOLLEN einen eindeutigen und dauerhaften Identifikator ~~im digitalen Archiv~~ haben.

Wie oben erwähnt, MUSS ein Informationspaket flexibel genug sein, um Daten oder Metadaten in Abhängigkeit von den Anforderungen des digitalen Archivs und seiner Nutzer aufnehmen zu können. Ein Informationspaket kann auch zusätzliche Unterstützungs-dokumente wie Metadatenschemata, Benutzerrichtlinien, kontextbezogene Dokumentation usw. enthalten. Unabhängig davon, welche und wie viele Teile ein vollständiges Informationspaket bilden, ~~MÜSSEN~~ SOLLEN alle Teile einen eindeutigen und persistenten Identifkator haben, der die sachgerechte Verknüpfung von Daten, Metadaten und allen anderen Teile gewährleistet. Dies ist wiederum einer der wichtigsten Aspekte, um Interoperabilität und die Integrität der Informationspakete zu ~~verbinden~~gewährleisten.

~~Erwähnenswert ist auch, dass bei jeder Implementierung die Eindeutigkeit und Persistenz des Identifikators nur im jeweiligen Informationspaket erreicht werden muss. Bei jeder Implementierung muss Eindeutigkeit und Persistenz des Identifikators nur im jeweiligen Informationspaket erreicht werden (nicht zwingend zwischen verschiedenen Informationspaketen).~~ Die internen IDs MÜSSEN nur innerhalb des SIPs eindeutig sein, da~~Wenn dies der Fall ist, kann~~ durch die Kombination des Paket-Identifikators (eindeutig gemäß Grundsatz 2.3) und des Identifikators des einzelnen Informationspaketsbestandteils auf einfache Weise eine archivinterne Eindeutigkeit erreicht werden. ~~Der Identifikator darf eine nachträgliche Erweiterung um Nutzdaten nicht behindern.~~

Kommentar [SG14]: Ggf. in einem Glossar zu „Identifikator“ erläutern

~~Die unterschiedlichen Teile eines Informationspakets werden im folgenden Abschnitt näher erläutert.~~

3. Struktur eines Informationspakets

Grundsatz 3.1

Das Informationspaket MUSS sicherstellen, dass Daten und Metadaten logisch voneinander getrennt sind.

Auf der höchsten Ebene kann jedes Informationspaket logisch in Daten und Metadaten unterteilt werden. Diese logische Trennung minimiert den Aufwand für die Identifizierung oder Validierung von Inhalten/Metadaten und vereinfacht die langfristige Bestandserhaltung. ~~Beispielsweise können Ingestanwendungen entweder separate, effiziente Skripte zur Identifizierung und Validierung von Metadaten implementieren oder zur Identifizierung und Anpassung von Inhaltsformaten.~~ Beispielsweise können Ingestanwendungen ~~Skripte~~Methoden implementieren, die Metadaten identifizieren und validieren oder Inhaltsformate identifizieren und anpassen. ~~Während der gesamten Lebensdauer eines Informationspakets können durch diese Trennung auch Bestandserhaltungsaufgaben vereinfacht werden, z. B. teilweise Metadaten- oder Datenaktualisierungen, die die Integrität des Pakets nicht gefährden. Unabhängig von der physischen Struktur eines Pakets MUSS das Informationspaket eine logische Trennung von Daten und Metadaten im Paketinhaltsverzeichnis (Paketmanifest) ermöglichen. Dies kann sich auch in einer Trennung im Dateisystem äußern (z.B. als einzelne Computerdateien oder klar getrennte Bitströme formatiert)~~

~~**Grundsatz 3.2**~~

~~*Das Informationspaket SOLLTE sicherstellen, dass Daten und Metadaten physisch voneinander getrennt sind.*~~

~~Zusätzlich zur logischen Trennung von Daten und Metadaten ist es vorteilhaft, diese physisch getrennt zu halten (d.h. als einzelne Computerdateien oder klar getrennte Bitströme formatiert). Auf diese Weise können Tools und Systeme zur digitalen Bestandserhaltung die jeweiligen Daten oder Metadaten eines Informationspakets aktualisieren, ohne die Integrität des gesamten Informationspakets zu gefährden.~~

Grundsatz 3.2[8]~~3~~

Die Struktur des Informationspakets SOLLTE die Trennung verschiedener Arten von Metadaten ermöglichen

~~Zusätzlich zum vorherigen Grundsatz wird empfohlen,~~ Metadaten SOLLTEN ~~klar~~ gegliedert werden können~~in noch spezifischere Untergruppen zu gliedern~~. Obwohl die Definitionen von Metadatentypen zwischen den Implementierungen sehr unterschiedlich sind, ~~empfehlen wir,~~ist es hilfreich, Metadaten ~~logisch und physisch~~ zumindest in Metadaten zur Beschreibung und Bestandserhaltung ~~beschreibende und konservierende Metadaten~~ zu unterteilen.

Grundsatz 3.3[9]~~4~~

Die Struktur des Informationspakets ~~MUSS~~ SOLLTE die Erstellung von Daten und Metadaten in mehreren Repräsentationen ermöglichen.

[8] Im CSIP Principle 3.3; Principle 3.2 ist weggefallen.

[9] Im CSIP Principle 3.4.

Kommentar [ST15]: Gestrichen, da durch Grundsatz 3.1 abgedeckt.

Kommentar [ST16]: „SOLLTE“, um klarzustellen, dass man bei der Ausarbeitung einer Spezifikation diesen Grundsatz nicht anwenden muss, wenn der Anwendungsfall dies nicht erfordert.

Das Konzept der Repräsentation ist einer der Grundbausteine der digitalen Bestandserhaltung. Da sich die Technologien weiterentwickeln und veralten, werden Daten und Metadaten ständig aktualisiert, um eine langfristige Bereitstellung zu gewährleisten. Dadurch werden neue Versionen oder Repräsentationen der Daten und Metadaten erstellt.

~~Repräsentationen innerhalb der logischen und physischen Struktur eines Informationspakets darzustellen, hilft den archivierenden Institutionen, die verschiedenen Ausprägungen der Information während ihres gesamten Lebenszyklus eindeutig zu verstehen, wodurch auch die langfristige Verwaltung und Wiederverwendung der Informationen verbessert wird.~~Die Repräsentationen SOLLEN~~sollen~~ in der ~~logischen und physischen~~ Struktur eines Informationspakets darstellbar sein~~gestellt werden~~. Dies hilft den archivierenden Institutionen, die verschiedenen Ausprägungen der Information während ihres gesamten Lebenszyklus‘ eindeutig zu verstehen, und verbessert die langfristige Verwaltung und Wiederverwendung der Information.

~~Die Metadaten sind über die verschiedenen Ebenen hinweg konsistent zu halten.~~

Grundsatz 3.4[10] ~~5~~

Die Struktur des Informationspakets ~~MUSS~~ SOLLTE die Möglichkeiten zum Hinzufügen zusätzlicher Daten zum Informationspaket explizit definieren.

Für archivierende Einrichtungen kann es notwendig sein, zu den bereits mit vorhandenen digitalen Objekte und Metadaten geformten Informationspaket zusätzliche digitale Objekte bzw. Metadaten einzufügen. So kann ein XML-Schema zur Validierung der Metadatenstruktur oder auch Dokumentationen zur Beschreibung der ursprünglichen, technischen Umgebung zum Informationspaket hinzugefügt werden.~~Zusätzlich zu bestehenden Daten und Metadaten müssen Institute möglicherweise zusätzliche Daten in ein Informationspaket aufnehmen. Beispielsweise könnte beschlossen werden, dass XML-Schemata zu Metadatenstrukturen und zusätzliche Binärdokumentation zur ursprünglichen IT-Umgebung zum Paket hinzugefügt werden müssen.~~

In diesem Anwendungsfall SOLLTE das Informationspaket nicht einschränken, welche zusätzlichen digitalen Objekte bzw. Metadaten in ein Informationspaket eingefügt werden. Im Anwendungsfall MUSS es klar definierte Erweiterungsmöglichkeiten für die Aufnahme der neu hinzukommenden digitalen Objekte bzw. Metadaten geben. Diese Erweiterung MUSS gleichzeitig so definiert sein, dass bereits vorhandene digitale Objekte bzw. Metadaten des Informationspakets hiervon nicht beeinträchtigt werden.~~In diesem Fall darf das Informationspaket NICHT einschränken, welche zusätzlichen Daten in ein Informationspaket aufgenommen werden können, und es MUSS klar definierte Erweiterungspunkte für die Übernahme dieser zusätzlichen Daten in das Informationspaket bieten. Gleichzeitig MÜSSEN diese Erweiterungsbereiche so definiert werden, dass andere Teile des Informationspakets nicht beeinträchtigt werden (d.h. die Erweiterungsbereiche MÜSSEN klar von anderen Teilen eines Informationspakets getrennt sein).~~

Grundsatz 3.5[11] ~~6~~

[10] Im CSIP Principle 3.5.

[11] Im CSIP Principle 3.6.

Kommentar [ST17]: „SOLLTE“, um klarzustellen, dass man bei der Ausarbeitung einer Spezifikation diesen Grundsatz nicht anwenden muss, wenn der Anwendungsfall dies nicht erfordert.

Kommentar [SG18]: Ursprünglich 2.1

Formatiert: Deutsch (Deutschland)

Jedes Informationspaket SOLLTE seinen Informationstypen mitteilen.

Ein definierter Informationstyp stellt ein Profil dar, dem das Informationspaket mit den digitalen Objekten und Metadaten folgt. Informationstypen ermöglichen Interoperabilität für gleiche Anwendungskontexte. Implementierungen können je nach Informationstyp auf gemeinsame Module zurückgreifen. ~~*Die Struktur des Informationspakets SOLLTE unabhängig von seiner technischen Umsetzung einer gemeinsamen konzeptionellen Struktur folgen in seiner konzeptionellen Struktur dokumentiert sein.*~~

~~Basierend auf den GrundsätzenPrinzipien 3.1 – 3.45 präsentieren wir eine gemeinsame Struktur für jedes Informationspaket (Abbildung 7).~~

~~Entsprechend GrundsatzPrinzip 3.1 MUSS das Paket eine übergeordnete Strukturkomponente für Metadaten enthalten, die mindestens relevante Metadaten für das gesamte Paket enthält. Darüber hinaus MÜSSEN die Repräsentationen intern zwischen Daten und Metadaten trennen (beim E-ARK-CSIP ist jedoch zu beachten, dass dort nicht vorgeschrieben ist, sowohl Daten als auch Metadaten in allen Repräsentationen verfügbar zu halten).~~

~~Gemäß Grundsatz 3.2 wird dringend empfohlen, diese logische Struktur als physische Ordnerstruktur auszuprägen.~~

~~Gemäß Grundsatz 3.3 wird dringend empfohlen, alle Paketmetadaten in separate Metadaten-untergruppen zu gliedern.~~

~~Gemäß Grundsatz 3.4 trennt die Struktur des Informationspakets die Repräsentationen von Daten und Metadaten klar in einen separaten Strukturteil.~~

~~Gemäß Grundsatz 3.5 können digitale Archive und ihre Nutzer zusätzliche Daten (z.B. Schemata und Dokumentation zur binären Unterstützung) entweder als Erweiterungen des gesamten Informationspakets oder in einer bestimmten Repräsentation hinzufügen.~~

~~Diese gemeinsame Struktur SOLLTE bei allen Implementierungen des CSIP befolgt werden.~~

~~Das oben dargestellte Strukturkonzept kann auf verschiedene Arten implementiert werden – zum Beispiel können die Einzelteile durch begleitende Metadaten oder durch eine eindeutige physische Struktur definiert werden. Es ist jedoch nicht sinnvoll, mehrere Implementierungen gleichzeitig verfügbar zu haben, da dies zu unnötiger Komplexität bei der Entwicklung interoperabler Tools zum Erstellen, Verarbeiten und Verwalten von Informationspaketen führen würde. Derzeit wird für CSIP dringend empfohlen, eine feste physische Ordnerstruktur in Kombination mit einem Inhaltsverzeichnis in Form eines METS-Dokuments (siehe Abschnitt 4 und Abschnitt 5) bei der Implementierung dieser Spezifikation zu verwenden.~~

~~Gleichzeitig ist klar, dass jede technische Implementierung mit der Zeit überholt sein wird, zum Beispiel wenn neue Übertragungsmethoden und Speicherlösungen Anwendung finden. Insofern verbietet diese Spezifikation nicht die Einführung neuer logischer oder physischer technischer Lösungen.~~

4. Metadaten eines Informationspakets

Kommentar [GS19]: Z.B. ERMS, Geodaten, SIARD; siehe https://earkcsip.dilcis.eu/pdf/eark-csip.pdf S. 12X

Es muss zwischen Informationstyp/Profil und Content unterschieden werden.

Kommentar [SG20]: Ggf. in einem Glossar zu erläutern

Formatiert: Schriftart: (Standard) Times New Roman, 12 Pt.

Kommentar [ST21]: Muss entsprechend angepasst werden.

Kommentar [SG22]: Ggf. in Einleitung: Die konkrete Auswahl und Bedeutung einzelner Metadaten in der digitalen Archivierung entwickelt sich während des Betriebs. Deshalb sind Standards und Informationsttypen stets so bald wie möglich diesen Entwicklungen anzupassen.

Grundsatz 4.1

Metadaten im Informationspaket SOLLEN~~MÜSSEN~~ einem etablierten Standard entsprechen.

Um Informationspakete interoperabel und automatisiert auszutauschen, zu validieren, zu verarbeiten und wiederzuverwenden, müssen die wichtigen Metadaten im Paket standardisiert vorliegen. Als in diesem Sinne “Wichtige Metadaten" werden hier sämtliche Kerninformationen über die Erstellung und Verwaltung des Paketinhalts (administrative- und Bestandserhaltungsmetadaten), zur eindeutigen Darstellung der Paketstruktur (strukturelle Metadaten) und technische Details der digitalen Objekte selbst (technische Metadaten) verstanden.~~Um Informationspakete interoperabel und automatisiert auszutauschen, zu validieren, zu verarbeiten und wiederzuverwenden, müssen wir standardisieren, wie wichtige Metadaten im Paket dargestellt werden. "Wichtige Metadaten" werden in dieser Spezifikation als Kerninformationen über die Erstellung und Verwaltung des Paketinhalts (Verwaltungs- und Bestandserhaltungsmetadaten), eindeutige Darstellung zur Paketstruktur (strukturelle Metadaten) und technische Details der Daten selbst (technische Metadaten).~~

Das Verwenden etablierter und zweckdienlicher Metadatenstandards wird zur einheitlichen und interoperablen Interpretation und Implementierung dringend empfohlen. ~~Um sicherzustellen, dass diese Metadaten in jedem Informationspaket auf eine einheitliche und interoperable Weise verstanden und implementiert werden, wird die Verwendung etablierter und weit verbreiteter Metadatenstandards dringend empfohlen. In der aktuellen Implementierung wird ein großer Teil dieser Metadaten von den weit verbreiteten METS- und PREMIS-Standards abgedeckt (siehe Abschnitt 5).~~

Grundsatz 4.2

Die exakte Verwendung der Metadaten SOLLTE in den Informationstypen erarbeitet werden.

Viele Metadatenstandards unterstützen i.d.R. mehrere Beschreibungsmöglichkeiten bestimmter Details eines Informationspakets. Solche Interpretationsmöglichkeiten können jedoch auch zu unterschiedlichen Implementierungen und letztendlich zum Verlust der Interoperabilität führen. Deswegen SOLLTEN Informationstypen (siehe Grundsatz 3.5) die exakte Verwendung der Metadaten festlegen.

~~*Metadaten im Informationspaket MÜSSEN SOLLTEN eine eindeutige unzweideutige Verwendung ermöglichen.*~~

~~Viele Metadatenstandards unterstützen i.d.R. mehrere Beschreibungsmöglichkeiten bestimmter Details eines Informationspakets. Solche Interpretationsmöglichkeiten können jedoch auch zu unterschiedlichen Implementierungen und letztendlich zum Verlust der Interoperabilität führen.~~

~~Um dieses Risiko zu überwinden, MUSS nach dem CSIP bei der Entwicklung einer spezifischen Implementierung der ausgewählte Metadatenstandard im Hinblick auf mögliche Mehrdeutigkeiten überprüft werden. Bei Bedarf MUSS der ausgewählte Metadatenstandard weiter verfeinert werden, um die Anforderungen an Interoperabilität und Automatisierung zu erfüllen.~~

Kommentar [SG23]: Neuer Grundsatz!

Formatiert: Schriftart: Times New Roman, 12 Pt., Kursiv

Formatiert: Schriftart: Times New Roman, 12 Pt., Kursiv

Formatiert: Schriftart: Kursiv

Kommentar [ST24]: „SOLLTE“, um klarzustellen, dass man bei der Ausarbeitung einer Spezifikation diesen Grundsatz nicht anwenden muss, wenn der Anwendungsfall dies nicht erfordert.

~~**Grundsatz 4.3**~~

~~*Das Informationspaket DARF SOLLTE NICHT das Hinzufügen zusätzlicher Metadaten einschränken.*~~

~~Die vorangegangenen Grundsätze legen die Bedeutung festgelegter Metadaten für Interoperabilitätszwecke fest. Gleichzeitig gilt das Gegenteil für andere Metadatenarten, insbesondere für die Bestandssuche (auch als beschreibende Metadaten bezeichnet) oder für Content Information Type spezifische technische und strukturelle Metadaten. Um die breite Akzeptanz des Informationspakets nicht einzuschränken, muss es jedem Programmierer möglich sein, neben den obligatorischen Metadaten, die für die Automatisierung und Interoperabilität auf Paketebene erforderlich sind, Metadaten hinzuzufügen.~~

~~Für den Fall, dass archivierende Institutionen weitere Details zu beschreibenden oder Content Information Type-spezifischen Metadaten als verpflichtend ansehen, um ein tieferes Maß an Interoperabilität zu erreichen, kann die oben beschriebene Methode der Content Information Type Spezifikationen verwendet werden.~~

~~Um die oben genannten Anforderungen aus einer mehr technischen Sicht zusammenzufassen, sieht das CSIP einen modularen Ansatz für die Metadaten bei Informationspaketen vor:~~
~~• Alle Informationspakete teilen einen gemeinsamen Kern von Metadaten, der die gemeinsame Entwicklung von fortgeschrittenen Tools zur Erstellung, Validierung, Identifizierung und Wiederverwendung von Informationspaketen ermöglicht.~~
~~• Die restlichen Metadaten im Informationspaket können zusätzlichen Vereinbarungen folgen, die getroffen wurden, um bestimmte Tools zu entwickeln, z. B. Tools zum Verwalten von Erschließungsdaten in EAD oder für bestimmte Content Information Types wie relationale Datenbanken in SIARD2 Format.~~

Grundsatz 4.3[12]

Jedes Informationspaket KANN beschreibende Metadaten (u.a. zusätzliche Identifkatoren des Produzenten) enthalten.

Es steht jedem Datenproduzenten frei, zusätzliche Metadaten in das SIP zu integrieren. Z.B. müssen digitale Archive oft die Herkunft eines Objekts aus der abgebenden Stelle zurückverfolgen können, um Authentizität und Integrität nachzuweisen (vgl. dazu auch Punkt 5).
Das Vorhandensein eines Identifikators ist Grundvoraussetzung für diese Aufgabe.
Der Identifikator, sofern vorhanden, wird in die Metadaten übernommen.

5. Authentizität und Integrität eines Informationspakets[13]

Grundsatz 5.1

Im Informationspaket SOLLEN Möglichkeiten enthalten sein, die Authentizität ~~und Integrität~~

[12] Fehlt im CSIP.

[13] Fehlt im CSIP.

Kommentar [SG25]: Wird über 4.1 und 4.2 abgedeckt.

Formatiert: Deutsch (Deutschland)

sicherzustellen.

Zu jeder Zeit SOLL~~MUSS~~ nachweisbar sein, wann und vom wem~~oder~~ das Datenpaket ursprünglich erstellt wurde~~stammt~~. Dies KANN ~~Wenn dies beim Übertrag nicht über Metadaten sicher dargestellt werden kann („unbroken chain of custody“),~~ über entsprechende Metadaten~~sollen~~ wie beispielsweise qualifizierten elektronischen Signaturen dargestellt werden~~verwendet werden~~.

~~[...]~~

Grundsatz 5.2

Im~~Ein~~ Informationspaket SOLLEN Möglichkeiten enthalten sein, die Integrität sicherzustellen~~I robust sein~~.

Die Integrität eines Informationspakets ~~während des Übertragungsprozesses oder während einer (längerfristigen) Speicherung~~ SOLL ~~soll~~ durch Nutzung geeigneter Verfahren wie z.B. Prüfsummen sichergestellt werden (~~weitere Verfahren wären Message Authentification Codes oder qualifizierte elektronische Signaturen~~). ~~Die Prüfsummen MÜSSEN einem geeigneten Sicherheitsgrad entsprechen.~~
Fehler müssen mindestens auffallen und sind im besten Fall korrigierbar.

Formatiert: Schriftart: Nicht Kursiv

Formatiert: Schriftart: Kursiv

Formatiert: Schriftart: Kursiv

Formatiert: Nur Text