nestor

Grundsätze zur SIP-Bildung

Handreichung der nestor Arbeitsgruppe SIP-Konkretisierung

nestor-materialien 23

Grundsätze zur SIP-Bildung

Handreichung der
nestor-Arbeitsgruppe
SIP-Konkretisierung

nestor-materialien 23

nestor - Kompetenznetzwerk Langzeitarchivierung
und Langzeitverfügbarkeit Digitaler Ressourcen
für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de

nestor - Partner:

- Bayerische Staatsbibliothek
- Bibliotheksservice-Zentrum Baden-Württemberg
- Bundesarchiv
- Deutsche Kinemathek – Museum für Film und Fernsehen
- Deutsche Nationalbibliothek
- FernUniversität Hagen
- Georg-August-Universität Göttingen / Niedersächsische Staats- und Universitätsbibliothek Göttingen
- GESIS Leibniz Institut für Sozialwissenschaften
- HeBIS - Hessisches BibliotheksinformationsSystem
- Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen
- Humboldt-Universität zu Berlin
- Institut für Deutsche Sprache
- Institut für Museumsforschung (Stiftung Preußischer Kulturbesitz)
- Landesarchiv Baden-Württemberg
- Landesarchiv Nordrhein-Westfalen
- Leibniz-Institut für Deutsche Sprache
- Münchner Stadtbibliothek
- Rechenzentrum der Universität Freiburg
- Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden
- TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek
- ZB MED - Informationszentrum Lebenswissenschaften
- ZBW – Leibniz-Informationszentrum Wirtschaft (ZBW)
- Zuse-Institut Berlin / Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)

© 2020 nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

Der Inhalt dieser Veröffentlichung darf vervielfältigt und verbreitet werden, sofern der Name des Rechteinhabers "nestor - Kompetenznetzwerk Langzeitarchivierung" genannt wird.
Eine kommerzielle Nutzung ist nur mit Zustimmung des Rechteinhabers zulässig.

URN: urn:nbn:de:0008-2020071606

http://nbn-resolving.de/urn:nbn:de:0008-2020071606

Grundsätze zur SIP-Bildung

Handreichung der nestor Arbeitsgruppe SIP-Konkretisierung

Einleitung

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

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

Derzeit gibt es mangels konkreter Vorgaben eine bunte Vielfalt von Informationspaketen, die zur Übernahme angeboten werden. Die archivierenden Institutionen sehen sich dabei großen Problemen bei der Datenübergabe und ihrer Erhaltung gegenüber, denn für jede Datenvariante müssen spezielle Lösungen gefunden werden. Für sie, genauso wie für Datenproduzenten, IT-Dienstleiter und viele weitere Beteiligte wäre es nur praktisch und gleichzeitig wirtschaftlich, wenn die Informationspakete nach allgemein gültigen Maßgaben erstellt würden, die deren Verarbeitung zur Routine machen könnten.

Ziel dieser Handreichung ist es deswegen, Datenproduzenten, IT-Dienstleistern und archivierenden Institutionen eine Richtschnur zu bieten, um über Vereinheitlichung der Informationspakete eine Vereinfachung der digitalen Archivierung zu erreichen.

Konkretisierung eines SIP

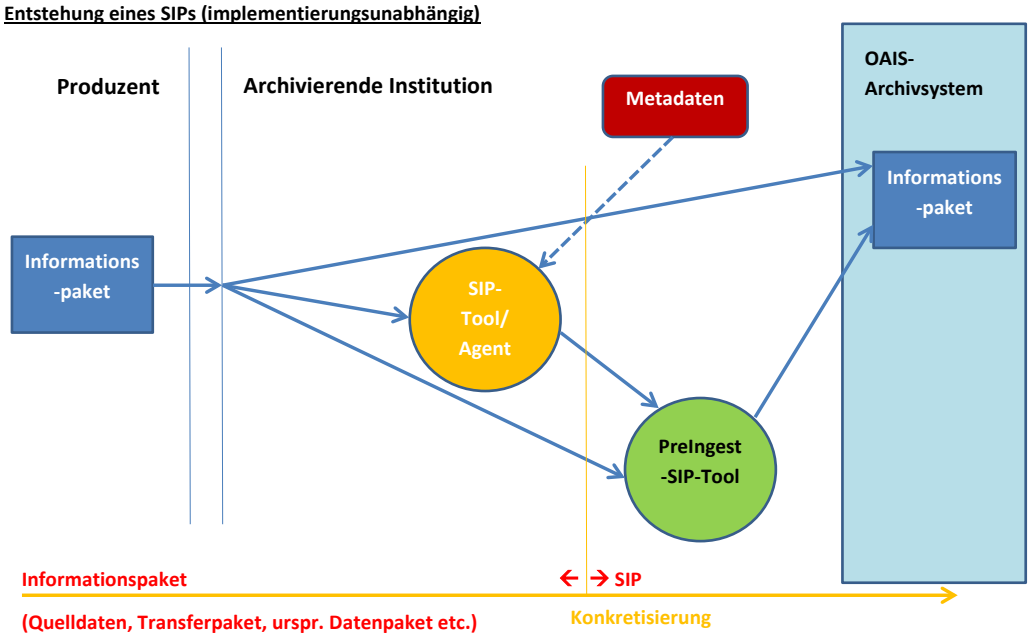
Die Informationspakete der Abgebenden werden entsprechend dem OAIS-Referenzmodell[1] Submission Information Package (SIP) genannt und wie folgt beschrieben:

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

[1] Open Archival Information System (ISO 14721): Siehe nestor-Materialen 16 unter http://nbn-resolving.de/urn:nbn:de:0008-2013082706

Produzenten und dem Archiv ausgehandelt (siehe die entsprechenden Standards in 1.5). Die meisten SIPs werden einige Inhaltsinformationen und einige Erhaltungsmetadaten enthalten.[2]

Allerdings ist die Realität deutlich vielschichtiger. Informationspakete (Quelldaten, Transferpakete, ursprüngliche Datenpakete etc.) werden 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 Informationspaket so umzuarbeiten, dass es anschließend verlustfrei und integer in das OAIS-konforme digitale Archiv übertragen und lesbar gehalten werden kann. Letztlich muss man zwischen dem ursprünglichen Informationspaket und dem eigentlichen SIP im Übergabeworkflow unterscheiden. Im folgenden Schaubild ist dies dargestellt:



Je größer die Annäherung zwischen dem ursprünglichen Informationspaket und dem SIP ist, desto einfacher wäre der Übergabe- und Archivierungsprozess. Standards sind für die Paketkonzeption immer zu bevorzugen. Für bestimmte Anwendungsbereiche wie z. B. Datenbank-Daten kann diese noch zusätzlich über Profile für Informationstypen ausdifferenziert werden.[3]

Um dies zu erreichen, können sich Datenproduzenten und archivierende Institutionen an folgenden Grundsätzen orientieren.

[2] Vgl. nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (Hrg.): Referenzmodell für ein Offenes Archiv-Informations-System - Deutsche Übersetzung 2.0 -, S. 23.

[3] Die konkrete Auswahl und Bedeutung einzelner Metadaten in der digitalen Archivierung entwickelt sich dabei in der Regel während des Betriebs. Deshalb sind Standards und Profile für Informationstypen stets so bald wie möglich diesen Entwicklungen anzupassen.

Grundsätze zur SIP-Bildung

Die Grundsätze[4] gliedern sich in fünf Bereiche:

1. Allgemeine Grundsätze
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/DARF NICHT = Verpflichtend
- SOLL = Empfehlenswert (Abweichung ist zu begründen)
- KANN = Erlaubt

Begriffe:
- Digitales Objekt = Ein aus einer Reihe von Bit-Sequenzen zusammengesetztes Objekt[5]
- Informationspaket = Transferpaket etc. und SIP (siehe oben)
- Informationstyp = Gruppe gleichartiger digitaler Objekte mit weitgehend gleichen Eigenschaften (z. B. Text, Bild etc.), wobei sich die konkreten Werte unterscheiden[6]
- Ingest = Übernahme in das OAIS-Archivsystem

1. Allgemeine Grundsätze

Grundsatz 1.1

Es MUSS möglich sein, beliebige digitale Objekte und Metadaten in ein Informationspaket aufzunehmen.

Um wirklich „allgemein" zu sein, dürfen technische Implementierungen des Informationspakets keine Einschränkungen oder Restriktionen einführen, die nur für bestimmte digitale Objekte und Metadaten gelten. Wenn die Implementierung eines Informationspakets dieses Prinzip nicht erfüllt, kann es nicht branchen- und werkzeugübergreifend eingesetzt werden, wodurch die Interoperabilität eingeschränkt wird. Die archivierende Institution kann dabei Vorgaben (z. B. bei Formaten oder Metadaten) machen. Ein Informationspaket soll auf frei zugänglichen Standards oder Quasi-Standards basieren.

[4] Die Grundsätze lehnen sich an die General Principles der beim EU-Projekt E-ARK entwickelten Common Specification for Information Packages (CSIP) an. Siehe https://earkcsip.dilcis.eu/pdf/eark-csip.pdf
[5] Vgl. Neuroth, Heike u. a. (Hrg.): nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3, Boizenburg o. J., S. 7.3-7.6.
[6] Vgl. nestor-Arbeitsgruppe Digitale Bestandserhaltung (Hrg.): Nestor – Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung. Version 2.0 (= nestor-materialien 15), Frankfurt am Main 2012, S. 9.

Grundsatz 1.2

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

Tools und Methoden zum Übertragen von Informationspaketen zwischen Standorten werden ständig weiterentwickelt. Es ist auch möglich, dass für Pakete 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.

Grundsatz 1.3

Das Paketformat DARF NICHT den logisch-inhaltlichen Umfang der digitalen Objekte und Metadaten definieren, die ein Informationspaket bilden.

Jedes einzelne digitale Archiv muss den logisch-inhaltlichen Umfang eines Informationspakets und dessen Beziehungen zu anderen Objekten definieren können. Daher muss es beispielsweise bei der Implementierung von Informationspaketen möglich sein, gleichermaßen entweder den gesamten Inhalt einer intellektuellen Einheit zu einem Informationspaket zusammenzuführen oder Teile davon zu extrahieren und als jeweils ein Informationspaket zu verpacken.

Grundsatz 1.4

Das Informationspaket MUSS skalierbar sein.

Viele digitale Archive haben Probleme mit Datenobjekten und Metadaten ab einer gewissen Größe der Informationspakete, wodurch Aufgaben im Zusammenhang mit der Daten- oder Metadatenvalidierung sowie der Identifizierung und Änderung nur schwer auszuführen sind. 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 interpretierbar sein, um eine auch für den Menschen inhaltliche Deutung zu ermöglichen.

Bei der Langzeitaufbewahrung muss auch berücksichtigt werden, dass möglicherweise "vergessene" Informationspakete gefunden werden, zu denen es keine Informationen über 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

Die Spezifikation des Informationspakets MUSS offen und frei sein.

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

Grundsatz 1.8

Die Komplexität der Spezifikation eines Informationspakets SOLL angemessen sein.

Eine angemessene Komplexität begünstigt grundsätzlich Weiterentwicklungen und Verbreitung und ist in der Regel besser handhabbar.

2. Grundsätze zur Identifikation eines Informationspakets

Grundsatz 2.1

Jedes Informationspaket MUSS einen im archivierenden Archiv eindeutigen und dauerhaften Identifikator haben oder erhalten.

Um ein digitales Archiv zu verwalten und die Bereitstellung sicherzustellen, muss jedes gespeicherte Informationspaket mindestens innerhalb des digitalen Archivs eindeutig identifiziert werden. Gleichzeitig darf eine entsprechende Implementierung die Auswahl des Identifikatorensystems nicht einschränken.

Grundsatz 2.2

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

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

Grundsatz 2.3

Alle Teile eines Informationspakets SOLLEN einen eindeutigen und dauerhaften Identifikator haben.

Wie in Grundsatz 1.1 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. Unabhängig davon, welche und wie viele Teile ein vollständiges Informationspaket bilden, sollen alle Teile einen eindeutigen und dauerhaften Identifikator 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 gewährleisten.

Die internen Identifikatoren müssen nur innerhalb des SIPs eindeutig sein, da durch die Kombination des Paketidentifikators (eindeutig gemäß Grundsatz 2.1) und des Identifikators des einzelnen Informationspaketsbestandteils auf einfache Weise eine archivinterne Eindeutigkeit erreicht werden.

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 Methoden implementieren, die Metadaten identifizieren und validieren oder Inhaltsformate identifizieren und anpassen.

Grundsatz 3.2

Die Struktur des Informationspakets SOLL die Trennung verschiedener Arten von Metadaten ermöglichen.

Metadaten sollen gegliedert werden können. Obwohl die Definitionen von Metadatentypen zwischen den Implementierungen sehr unterschiedlich sind, ist es hilfreich, Metadaten zumindest in Metadaten zur Beschreibung und Bestandserhaltung zu unterteilen.

Grundsatz 3.3

Die Struktur des Informationspakets SOLL die Erstellung von Daten und Metadaten in mehreren Repräsentationen ermöglichen.

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.

Die Repräsentationen sollen in der Struktur eines Informationspakets darstellbar sein. 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.

Grundsatz 3.4

Die Struktur des Informationspakets SOLL die Möglichkeiten zum Hinzufügen zusätzlicher Daten zum Informationspaket explizit definieren.

Für archivierende Einrichtungen kann es notwendig sein, zu dem bereits mit vorhandenen digitalen Objekten und Metadaten geformten Informationspaket zusätzliche digitale Objekte bzw. Metadaten einzufügen (siehe Grundsatz 1.1). So kann beispielsweise ein XML-Schema zur Validierung der Metadatenstruktur oder auch Dokumentationen zur Beschreibung der ursprünglichen technischen Umgebung zum Informationspaket ergänzt werden.

In diesem Anwendungsfall soll das Informationspaket nicht einschränken, welche zusätzlichen digitalen Objekte bzw. Metadaten in ein Informationspaket eingefügt werden. Es muss 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.

Grundsatz 3.5

Jedes Informationspaket SOLL seinen Informationstypen mitteilen.

Für Informationstypen können Profile definiert werden, denen die Informationspakete mit den digitalen Objekten und Metadaten folgen. Solche Profile ermöglichen Interoperabilität für gleiche Anwendungskontexte. Zum Beispiel kann es ein definiertes METS-Profile für Informationspakete in Webarchiven geben. Implementierungen können je nach Informationstyp auf gemeinsame Module zurückgreifen.

4. Metadaten eines Informationspakets

Grundsatz 4.1

Metadaten im Informationspaket SOLLEN 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.

Das Verwenden etablierter und zweckdienlicher Metadatenstandards wird zur einheitlichen und interoperablen Interpretation und Implementierung dringend empfohlen.

Grundsatz 4.2

Die exakte Verwendung der Metadaten SOLLTE in Profilen für 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 Profile für Informationstypen (siehe Grundsatz 3.5) die exakte Verwendung der Metadaten festlegen.

Grundsatz 4.3

Jedes Informationspaket KANN beschreibende Metadaten enthalten.

Es steht jedem Datenproduzenten frei, zusätzliche Metadaten in das SIP zu integrieren. Z. B. müssen digitale Archive oft die Herkunft eines Objekts zurückverfolgen können, um Authentizität und Integrität nachzuweisen (vgl. dazu auch Bereich 5).

5. Authentizität und Integrität eines Informationspakets

Grundsatz 5.1

Im Informationspaket SOLLEN Möglichkeiten enthalten sein, die Authentizität sicherzustellen.

Zu jeder Zeit soll nachweisbar sein, wann und vom wem das Datenpaket ursprünglich erstellt wurde. Dies kann über entsprechende Metadaten wie beispielsweise qualifizierte elektronische Signaturen dargestellt werden.

Grundsatz 5.2

Im Informationspaket SOLLEN Möglichkeiten enthalten sein, die Integrität sicherzustellen.

Die Integrität eines Informationspakets soll durch Nutzung geeigneter Verfahren wie z. B. Prüfsummen sichergestellt werden.

Fehler müssen mindestens auffallen und sind im besten Fall korrigierbar.