Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Einschlägige Normen für Collections. Metadaten und Annotationen

Deliverable C3.2

Das vorliegende Dokument wurde im Rahmen des Konsortiums Text+ im Kontext der Arbeit des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. verfasst. NFDI wird von der Bundesrepublik Deutschland und den 16 Bundesländern finanziert, und das Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 460033370. Die Autor:innen bedanken sich für die Förderung sowie Unterstützung. Ein Dank geht außerdem an alle Einrichtungen und Akteur:innen, die sich für den Verein und dessen Ziele engagieren.

This document was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Version	1.0
Redaktion	30.09.2023
Redaktionsteam	Florian Barth, Marcel Fladrich, Philippe Genêt, Susanne Haaf, Timm Lehmberg, Felix Rau, Thorsten Trippel, Andreas Witt
Projekt	Text+ - Sprach- und textbasierte Forschungsdateninfrastruktur
Bezeichnung	C3.2 Publication on relevant standards for Collections (metadata and annotations)
Förderung	DFG Förderkennzeichen 460033370
Projektlaufzeit	01.10.2021 bis 30.09.2026

Inhalt

1. Einführung und Motivation

Als Konsortium der bundesweiten Initiative zum Aufbau der nationalen Forschungsdateninfrastruktur (NFDI) hat der Verbund Text+ zum Ziel, text- und sprachbasierte Forschungsdaten langfristig zu erhalten und ihre breite Nutzung in der Wissenschaft zu ermöglichen. Dabei konzentriert sich die Text+ Infrastruktur zunächst auf drei Datendomänen: Korpora und digitale Textsammlungen (Collections), lexikalische Ressourcen und Editionen.

Zur Datendomäne Collections tragen 19 Institutionen bei, die umfangreiche Sammlungen sprach- und textbasierter Daten in Text+ einbringen. Es handelt sich dabei um Sammlungen geschriebener, gesprochener oder gebärdeter Sprache und Texte sowie sprach- und textbezogene Experimental- oder Messdaten, die auf Grundlage wissenschaftlicher Kriterien gesammelt wurden. Dazu gehören: Textsammlungen (z. B. von literarischen Texten, Sachtexten, Zeitungs- und Zeitschriftentexten, Interviews, Inschriften, Handschriften, Drucken), mono- und multimodale Aufnahmen z. B. von spontaner und formaler Sprache (z. B. von Reden, Dialogen, Nachrichten, Interviews, Interaktion im Alltag), Sensordaten (z. B. EEG, Eyetracking, Artikulographie), Befragungen, Reaktionszeiten etc.

Die Heterogenität dieser Sammlungen spiegelt sich in den Daten wider: Die Texte, Ton-, Bild- oder Bewegtbildaufnahmen sowie zahlreiche andere Daten liegen samt ihrer jeweiligen Metadaten in verschiedensten Formaten vor. Der Grad der Erschließung und Annotation variiert von Institution zu Institution und von Sammlung zu Sammlung. Angestrebt wird zudem die Integration und Nutzbarmachung weiterer Daten aus Quellen, die bisher noch nicht Teil von Text+ sind.

Die Text+ Infrastruktur möchte möglichst umfangreiche Bestände an Forschungsdaten in dezentralen Repositoren versammeln und entlang der FAIR-Kriterien zentral über ein Webportal verfügbar machen. Insbesondere zur Gewährleistung von Auffindbarkeit, Interoperabilität und Nachnutzbarkeit in einer ortsverteilten Infrastruktur sind einheitliche Formate und Standards über die Repositorien hinweg unerlässlich.

Das vorliegende Dokument ist eine Momentaufnahme der aktuell verwendeten Metadaten- und Objektdatenformate in der Datendomäne Collections sowie der dort angewandten Standards. Es richtet sich zum einen an Forschende und Institutionen, die einschlägige Daten in die Infrastruktur von Text+ integrieren möchten[1]. Ihnen soll das Dokument Orientierung geben, indem es die bevorzugten und akzeptierten Formate und Standards beschreibt, mit denen derzeit in den Datenzentren von Collections gearbeitet wird. Die Wahl dieser Formate und Standards vereinfacht die Datenintegration erheblich. Wenn jedoch neu zu integrierende Ressourcen es erfordern, sind die Zentren offen, neue Ansätze und Formate in ihre Arbeitsweise zu integrieren und das Portfolio an Formaten und Standards zu erweitern.

Zum anderen dient diese Zusammenstellung auch den Datenzentren der Datendomäne Collections selbst als Richtschnur, in welchen Formaten und nach welchen Standards sie ihre Daten üblicherweise ausliefern, und wofür Ingestprozesse vorzuhalten sind. So wird in den folgenden Kapiteln in Bezug auf Metadaten, Objektdaten und Containerdaten jeweils beschrieben, mit welchen Formaten und Standards die Datenzentren in der Datendomäne Collections in Text+ bevorzugt arbeiten.

[1] Die Vorgehensweise zur Integration von Sammlungen in die Infrastruktur von Text+ wird in der Publikation *Leitlinie für das Integrieren von Daten in Text+/NFDI* beschrieben, siehe https://zenodo.org/doi/10.5281/zenodo.12744055

2. Daten

2.1 Metadaten

Metadaten im Sinne dieses Abschnitts sind Repräsentationen von zugrundeliegenden Daten in anderer Form, die bestimmten Zwecken dienen. Dazu gehören Daten, die verwendet werden, um die Daten zu katalogisieren und in Nachweissysteme aufzunehmen.

Metadaten können in verschiedenen Strukturen erstellt werden, darunter einfache Attribut-Wert-Strukturen (AV-Strukturen), Tripel und Baumstrukturen. Die AV-Strukturen, die von den meisten Metadatenstandards wie Dublin Core und OLAC verwendet werden, folgen einem einfachen Attribut-Wert-Schema. Diese flache Struktur ermöglicht eine einfache Speicherung in relationalen Datenbanken, wobei die Beziehung zwischen Attribut und Wert als ISA-Beziehung dargestellt wird. Tripel, eine Weiterentwicklung von AV-Strukturen, bieten eine Lösung für das vordefinierte Beziehungsproblem, indem sie eine explizite Angabe der Relation zwischen Attribut, Wert und Prädikat ermöglichen. Das Resource Description Framework (RDF) des World Wide Web Consortium (W3C) verwendet Tripel als zugrundeliegendes Modell. Eine weitere Strukturform sind Baumstrukturen, die hierarchische Beziehungen zwischen Metadatenkategorien beschreiben. Diese Hierarchien können taxonomischer Natur sein, wobei Informationen von übergeordneten Kategorien auf untergeordnete übertragen werden. Der CMDI-Standard (ISO 24622-1 und 24622-2) und der TEI-Metadaten-Header nutzen Baumstrukturen für die Klassifizierung von Kategorien in Sprachressourcen. Über eine formale Definition der Semantik von Datenkategorien und die Einführung von Inhaltsmodellen könnten Anforderungen an Mehrsprachigkeit, verschiedene Schriftsysteme, multiple Annotationsebenen und unterschiedliche Modalitäten adressiert werden.

Zur reichen Beschreibung von Forschungsdaten im Sinne der FAIR-Prinzipien sind aussagekräftige Beschreibungskategorien notwendig, die vom Ressourcentyp abhängen, auch wenn für Archivierungszwecke allgemeine bibliographische Kategorien für alle Typen Anwendung finden können. Eine herangehensweise ist daher, basierend auf einem Klassifikationssystem für Forschungsdaten und möglichen Prototypen für diese Beschreibungsmuster zu definieren, die in Abhängigkeit vom Typus der Forschungsdaten auf die jeweiligen Forschungsdatensätze angewendet werden können. Diese Beschreibungsmuster bilden Profile oder Schemata für Metadaten. Da einige Beschreibungskategorien für unterschiedliche Ressourcentypen verwendet werden können, müssen die Metadatenprofile oder -schemata nicht überschneidungsfrei sein. Beispielsweise können bibliographische Informationen häufig von verschiedenen Ressourcentypen verwendet werden.

Übliche Standards für Metadatenrepräsentationen sind Dublin Core oder MARC 21, DataCite, JSON-LD oder CMDI. Metadaten in einem oder mehreren dieser Formate werden häufig über Standardschnittstellen von datenhaltenden Institutionen bereitgestellt, so dass sie von anderen Institutionen nachgenutzt oder von Suchmaschinen indiziert werden können. Daneben werden die Metadatenschemata auch für unterschiedliche Forschungsdaten verwendet.

2.1.1 Dublin Core

Der am häufigsten verwendete Metadatenstandard, der auch als Referenz verwendet wird, wurde von der Dublin Core Metadata Initiative entwickelt und definiert verschiedene Kernkategorien (siehe Dublin Core, 2003, Coyle und Baker, 2008)).[2]

Der Standard der unqualifizierten Dublin Core Metadaten besteht aus 15 Kernelementen, die optional

[2] Die Absätze zu Dublin Core, OLAC und TEI sind angelehnt an Kapitel 5 von Trippel, Thorsten. *The Lexicon Graph Model: A generic Model for multimodal lexicon development*. Saarbrücken: AQ-Verlag, 2006.

sind und beliebig oft wiederholt werden können. Diese Kategorien sind (Untergliederung nach Hillmann, 2000):

1. Content:
    a. resource title: title
    b. resource description in prose: description
    c. related resources: relation
    d. coverage of the resources: coverage
    e. resource source: source
    f. type of the resource, to be selected from a controlled vocabulary: type
    g. resource subject: subject

2. Intellectual property rights and editorial information:
    a. contributors to the resource: contributor
    b. resource creator: creator
    c. publisher: publisher
    d. copyright: rights

3. Instance:
    a. resource language: language
    b. resource identifier (e.g. a URI): identifier
    c. resource format (e.g. Mime Type): format
    d. resource date: date

Weitere Elemente und Verfeinerungen sind spezifiziert worden (siehe Dublin Core, 2003) im Bereich

- der anderen Elemente, die Folgendes umfassen: *audience, alternative, tableOfContents, abstract, created, valid, available, issued, modified, extent, medium,*
- des Kodierungsschemas, das sich auf die Kodierung von Datenkategorien wie Datum, Sprache und Ländercodes bezieht, sowie Verweise auf bestehende Ontologien, darunter ISO639-2, URI, ISO3166,
- der Dublin Core Metadata Initiative (DCMI) Typenvokabulare wie Collection, Dataset, Image, Software, Sound, Text.

Für Forschungsdaten ist Dublin Core aber nur sehr begrenzt einsetzbar. So können mehrsprachige Dokumente mit den Metadaten nicht ausreichend beschrieben werden. Obwohl es möglich ist, auf mehr als eine Sprache zu verweisen, ist nicht definiert, wie Ressourcen mit einer Mischung von Sprachen zu beschreiben sind. Für traditionelles Material in Bibliotheken mag dies nicht der Normalfall sein, aber für linguistische Studien und Ressourcen, z.B. in einem vergleichenden Umfeld, trifft dies nicht zu. Zudem sind viele Datenkategorien unzureichend spezifiziert. So definiert etwa die Kategorie Datum nicht, welches Datum einer Ressource gemeint sein könnte – wie das Datum der Veröffentlichung, das der Erstellung oder das der Aufnahme.

Dennoch herrscht weitgehend Einigkeit darüber, dass die Dublin Core-Elemente den Kern der Metadaten bilden. Daher werden die Elemente auch in anderen Metadatenschemata auch für Forschungsdaten verwendet.

2.1.2 DataCite

Das DataCite Metadatenschema ist ein generisches Metadatenformat für Forschungsdaten, das seit 2010 vom DataCite Konsortium entwickelt und gepflegt wird. Version 1.0 des Schemas wurde 2011 veröffentlicht. Aktuell liegt Version 4.5 vor (DataCite Metadata Working Group, 2024). Der *DataCite – International Data Citation Initiative e.V.* ist ein gemeinnütziges internationales Konsorti-um, das

sich zum Ziel gesetzt hat, einfachen Zugang zu wissenschaftlichen Forschungsdaten zu ermöglichen. DataCite ist darüber hinaus eine besonders in Europa sehr oft genutzte Registrierungsagentur für Digital Object Identifier (DOI). Da bei der Registrierung eines DOI die Metadaten in Form einer DataCite-XML-Datei übergeben werden müssen, ist dieses Metadatenformat sehr verbreitet.

Als ein spezifisch für wissenschaftliche Forschungsdaten entwickeltes Format ist das DataCite Metadatenschema hervorragend für eine generelle Beschreibung von Forschungsdaten geeignet. Allerdings ist sein genereller Charakter auch ein Nachteil, da keine tiefere fachliche Beschreibung mit dem DataCite Metadatenschema möglich ist. Dafür sind die zwanzig DataCite Metadata Properties so gewählt, dass sie auf fast jedes Forschungsdatenset sinnvoll anzuwenden sind. Auch die semantische Definition der Kategorien ist hochwertig und klar, so dass eine einheitliche Interpretation und Nutzung des Schemas gesichert ist.

1. Identifier
2. Creator
3. Title
4. Publisher
5. PublicationYear
6. Subject
7. Contributor
8. Date
9. Language
10. ResourceType
11. AlternateIdentifier
12. RelatedIdentifier
13. Size
14. Format
15. Version
16. Rights
17. Description
18. GeoLocation
19. FundingReference
20. RelatedItem

Einige dieser Kategorien sind in sich komplexe Objekte, die nicht nur assoziierte Attribute haben, sondern untergeordnete Properties. So ist die Property *2. Creator* wie folgt strukturiert:

2. Creator
    - 2.1 creatorName
        - 2.1.a nameType
    - 2.2 givenName
    - 2.3 familyName
    - 2.4 nameIdentifier
        - 2.4.a nameIdentifierScheme
        - 2.4.b schemeURI
    - 2.5 affiliation
        - 2.5.a affiliationIdentifier
        - 2.5.b affiliationIdentifierScheme
        - 2.5.c schemeURI

Die DataCite Schemadokumentation bietet ein Mapping von ausgewählten DataCite Properties auf Dublin Core Kategorien (siehe Anhang A). Die Schemadefinition liefert auch sechs Listen mit kontrollierten Listenwerten für Attribute:

- contributorType
- dateType
- resourceTypeGeneral
- relatedIdentifierType
- relationType
- descriptionType

DataCite ist ein im akademischen Bereich weit verbreitetes und akzeptiertes Metadatenformat, das sich zur generellen Beschreibung von Forschungsdaten deutlich besser eignet als das weiter verbreitete Dublin Core.

2.1.3 MARC 21

Das Metadatenformat MARC (MAchine-Readable Cataloging) ist vor allem im bibliothekarischen Umfeld gebräuchlich. Entwickelt und laufend aktualisiert von der Library of Congress, dient es zur Erstellung maschinenlesbarer bibliografischer Datensätze. Ein MARC-Datensatz enthält die Beschreibung des Mediums, Haupt- und zusätzliche Einträge sowie Schlagwörter und Klassifikations- oder Signaturnummern. Die Beschreibung folgt im deutschsprachigen Raum den Erschließungsregeln aus RDA DACH. Die Schlagwörter stammen üblicherweise aus standardisierten Listen wie der Gemeinsamen Normdatei.

MARC-Datensätze stellen strukturierte Datenfelder für jede bibliografische Information bereit. Diese Felder ermöglichen Flexibilität bei der Katalogisierung verschiedener Medien und erlauben eine effiziente Datenabfrage und -anzeige.

Das Datenmodell von MARC besteht aus

- einem *Leader* aus 24 Bytes mit festgelegter Bedeutung je Position,
- einer Liste von *Control fields* mit Feldnummer und Feldinhalt und
- einer Liste von *Data fields* mit Feldnummer, Indikatoren und einer Liste von Unterfeldern als Feldinhalt, wobei jedes Unterfeld aus Unterfeldcode und -inhalt besteht.

Der Dateninhalt ist in variable Felder organisiert, die mit dreistelligen Kennungen (sogenannten Tags) kategorisiert werden. Für verschiedene Datentypen (z.B. bibliografische Daten, Normdaten oder Bestandsdaten) sind jeweils individuelle Tag-Blöcke vorgesehen. Das folgende Beispiel ist der Tag-Block für bibliografische Daten:

0XX = Kontrollfelder, Codes
1XX = Haupteintrag
2XX = Titel, Ausgabe, Verlag
3XX = Physische Beschreibung etc.
4XX = Reihenangaben
5XX = Notizen
6XX = Schlagworte, Genreangaben etc.
7XX = Namen etc., zusätzliche (Reihen-)Angaben, Links
8XX = Weitere Angaben zu Reihen, Bestand und Ort
9XX = Für individuelle Nutzung vorbehalten

Weitere Informationen zum MARC 21-Format sind auf der Website der Deutschen Nationalbibliothek zu finden, ebenso die vollständige deutschsprachige Dokumentation zu MARC 21 für bibliografische Daten, für Normdaten und für Bestandsdaten.

Seinem Ursprung entsprechend eignet sich MARC 21 besonders für die Erfassung von Katalogsmetadaten. Wenn Datensätze in Katalogen von Bibliotheken, Bibliotheksverbünden oder FID-Portalen erfasst werden sollen, ist dieses Format die beste Wahl.

2.1.4 OLAC

Die Open Language Archive Community (OLAC) (Simons und Bird, 2002) versucht, die Unzulänglichkeiten der Dublin Core-Metadatenkategorien für sprachliche Ressourcen zu überwinden, um die Bildung eines weltweit zugänglichen Datenrepositoriums für sprachliche Daten zu ermöglichen. Um einen konstanten Informationsbestand innerhalb des Katalogs zu erhalten, wird der Inhalt der Datenelemente weitgehend durch ein kontrolliertes Vokabular festgelegt. Ein Mapping der OLAC- und Dublin Core-Elemente findet sich in Anhang B. Dieses Mapping verdeutlicht, dass OLAC eine echte Erweiterung von Dublin Core ist.

Die Erweiterungen umfassen zwei für Daten im sprachlichen Umfeld entscheidende Merkmale:

1. Format:
   Das Datenformat hängt sowohl von technischen Faktoren als auch von der linguistischen Theorie ab. Da elektronisch verfügbare Daten- und Sprachressourcen für eine effiziente Verarbeitung elektronisch verfügbar sein müssen (vgl. Gibbon et al., 1997c), sind Informationen über die verwendete Plattform (insb. CPU, Betriebssystem und verwendete Software) erforderlich. Die technische Ausgestaltung impliziert den Hinweis auf technische Grenzen und Möglichkeiten. Markup besteht in diesem Bereich aus einer technischen Beschreibung eines Datenformats und muss ebenfalls aufgezeichnet werden, um die Struktur der Daten zu definieren.

2. linguistische Kategorien:
   Angaben zur Funktionalität und zur Sprache, die Gegenstand der Ressource ist, verweisen auf sprachlich relevante Informationen. Unterschiedliche Textgattungen, aber auch unterschiedliche Sprachen, die Gegenstand der Ressource sind, sind vor allem in Materialien von Bedeutung, die diese Unterschiede behandeln.

Der OLAC-Metadatensatz stellt eine Verbesserung für linguistische Daten dar, weist aber nach wie vor Unzulänglichkeiten auf:

- Datenkategorien bleiben weiterhin unterspezifiziert
- Die Beziehung zwischen verschiedenen Sprachen in mehrsprachigen Ressourcen ist nicht leicht zu beschreiben
- Für Ressourcen, die auf Signaldaten basieren, gibt es keine Möglichkeit die technischen Eigenschaften der Signalaufnahme zu beschreiben.

2.1.5 TEI

Die strukturelle Auszeichnung von verschiedenen Textsorten mit gemeinsamen Strukturen war das Ziel der Text Encoding Initiative (TEI). Das TEI-Consortium stellt daher ein Framework für die Definition von Annotationsschemata in XML für konkrete Textsorten und Dokument-Arten zur Verfügung. Zur Unterscheidung von Textsorten und zur Katalogisierung wurde ein allgemeiner Metadaten-Header entwickelt, der unter Umständen für einzelne Dokumententypen eingeschränkt wird. Der TEI-Metadaten-Header enthält folgende Informationen:

- Dateibeschreibung, die eine bibliographische Beschreibung einer Ressourcendatei enthält, wie z.B. Autorschaft, Eigentumsrechte, Quelle und Größe
- Kodierungsbeschreibung, einschließlich Zeichensätze, Format der Maßeinheiten usw.
- Inhaltsbeschreibung, z.B. Schlüsselwörter, Erstellungsdatum, Sprache
- Revisionsinformationen, z.B. Änderungen an der Ressource mit Datum und Umfang.

Ein mögliches Mapping von Teilen des TEI-Headers auf Dublin Core findet sich in Anhang C.

Der TEI-Metadatensatz ist vor allem für Textdaten in der TEI-Kodierung relevant, so dass keine Informationen über Datenformate erforderlich sind, da diese bereits Teil der Dokumentgrammatik sind. Allerdings gelten für TEI-Metadaten ähnliche Unzulänglichkeiten wie für Dublin Core und OLAC:

- das Problem der Unterspezifizierung von Metadatenkategorien
- der Umgang mit mehrsprachigen Ressourcen
- die Einbeziehung von signalbasierten Ressourcen.

2.1.6 ISO 24622-1/2 (Component MetaData Infrastructure)

Die Component MetaData Infrastructure (CMDI, siehe Broeder et al., 2010, siehe auch Barkey, et al., 2011) stellt ein Framework zur Verfügung, mit dessen Hilfe gezielt Forschungsdaten gemäß ihrem Typus beschrieben werden können, wobei gemeinsame Beschreibungskategorien verwendet werden. In diesem Framework werden zusammengehörige Datenkategorien und -strukturen zu *Komponenten* zusammengefasst. Komponenten sind dabei zunächst Mengen von beschreibenden Datenkategorien. Diese wiederum können selbst zu größeren Komponenten kombiniert werden, um schließlich für einen Ressourcentyp als ein Beschreibungsprofil Verwendung zu finden. Komponenten werden so Bausteine für Profile, wobei die gleichen Komponenten innerhalb verschiedener Profile enthalten sein können.

Das CMDI Framework wurde im Rahmen des EU-Projektes CLARIN zur systematischen Verwendung von Komponenten entwickelt. Diese Arbeiten sind in die Entwicklung von ISO 24622-1 und ISO 24622-2 eingeflossen, in denen CMDI als Standard beschrieben wird. Neben einer Beschreibungssprache für Profile und Komponenten enthält diese Infrastruktur dazu auch weitere Werkzeuge, sowohl Editoren als auch Analysewerkzeuge. Diese operieren unabhängig vom Ressourcentyp auf bestimmten Datenkategorien. Bestehende Metadatenstandards wie Dublin Core, OLAC oder der TEI-Header (TEI P5) können als Profile oder auch als Komponenten dargestellt werden, sodass ein Komponentenmodell mit Profilen als Obermenge bestehender Metadatenschemas angesehen werden kann. So werden die bibliographischen Informationen in den Metadaten einer Ressource für Archiv- und Bibliothekskataloge verwendbar. Andere Datenkategorien dagegen, wie z.B. die Angabe von Annotationstypen bei linguistischen Korpora, werden von allgemeinen Kataloganwendungen ignoriert, aber von spezialisierten Suchmaschinen oder Diensten verwendet.

Um auch institutionsübergreifend die Verwendung gleicher Komponenten und Profile zu ermöglichen, wurde im Rahmen von CMDI die Component Registry veröffentlicht. Dabei handelt es sich um ein Verzeichnis, das zentral Komponenten und Profile sowohl zur Weiterverwendung in Institutionen und Projekten als auch zur Validierung konkreter Instanzen zur Verfügung stellt. Die Komponenten erhalten dort einen persistenten Identifikator (Persistent Identifier oder PID, siehe ISO 24619), auf den sowohl von anderen Komponenten als auch Instanzen verwiesen werden kann und der über ein Handle-System zu einer URL aufgelöst wird.

Innerhalb der Komponenten werden die Datenkategorien mit einer Referenz auf bereits standardisierte oder im Standardisierungsprozess befindliche Datenkategorien verwendet, die in einem Verzeichnis definiert und nachhaltig dokumentiert werden.

In den Komponentendefinitionen von CMDI können zudem kontrollierte Vokabulare angegeben werden. Diese können ebenfalls dazu beitragen, das Problem des Tag Abuse zu minimieren, da Datenkategorien durch das kontrollierte Vokabular formal auf ihre Konsistenz geprüft werden können. Gleichzeitig gibt es auch Freitextfelder wie Zusammenfassungen und Beschreibungen, deren Inhalt nicht genauer reglementiert wird. Der Gebrauch von Datenmodellen ist nach Maßgabe der zugrundeliegenden Schemasprache möglich. Im Rahmen des CMDI-Datenmodells ist dies mit der Verwendung von XSchema sehr weitgehend umgesetzt worden, angefangen von Datumsformaten bis zu regulären Ausdrücken für Zeichenkettendefinitionen.

CMDI ist vor allem für Forschungsdaten verschiedener Ressourcentypen relevant, die trotz unterschiedlicher Charakteristiken möglichst ähnlich beschrieben werden sollen. Dabei wird eine möglichst hohe semantische Interoperabilität angestrebt. Als Framework ist CMDI dabei aber sehr variabel und kann von unterschiedlichen Institutionen auch verschieden eingesetzt werden.

2.1.7 Beispielimplementierung von Metadaten in einem lokalen Repositorium: TextGrid-Metadaten

Neben den bereits vorgestellten etablierten Standards existieren auch repositorienspezifische Metadatenschemata. TextGrid, eine virtuelle Forschungsumgebung für die Geisteswissenschaften, betrieben von der Niedersächsischen Staats- und Universitätsbibliothek Göttingen, verwendet eine solche individuelle Lösung. Diese soll im Folgenden als Beispiel erläutert werden.

TextGrid verwaltet eigens definierte Objekte und die Arten von Relationen zwischen ihnen (TextGrid-Konsortium, 2018). Das Metadatenschema von TextGrid orientiert sich dabei an den Functional Requirements for Bibliographic Records (FRBR) (IFLA Study Group, 2009, TextGrid-Metadaten, 2010). Alle TextGrid-Objekte müssen eindeutig identifizierbar sein, um Funktionalitäten wie zum Beispiel eine Suchfunktion zu ermöglichen, weshalb alle Objekte durch Metadaten beschrieben werden (TextGrid-Konsortium, 2018). Für folgende Objekte gibt es in TextGrid spezifische Metadaten-Dateien mit eigenen MIME-Types:

2.1.7.1 Aggregation: text/tg.aggregation+xml

Eine Aggregation ist ein TextGrid-Objekt, das aus einer sortierten Liste von Referenzen auf andere TextGrid-Objekte besteht, die wiederum selbst Aggregationen sein können (TextGrid-Konsortium, 2015a). Aggregationen eignen sich zur Organisation von TextGrid-Objekten. Sie werden ähnlich wie Dateiordner verwendet, sind aber flexibler – zum Beispiel kann ein TextGrid-Objekt von mehreren Aggregationen referenziert werden (TextGrid-Konsortium, 2015a). Editionen oder Kollektionen sind ebenfalls Aggregationen, die durch ein charakteristisches Set von Metadaten definiert sind (TextGrid-Konsortium, 2015a).

2.1.7.2 Edition (Form der Aggregation): text/tg.edition+tg.aggregation+xml

Eine Edition ist die Manifestation eines Werks und eine spezielle Form einer Aggregation (TextGrid-Konsortium, 2015b). Sie eignet sich zur Beschreibung der Originalquelle (beispielsweise eine Ausgabe eines Romans) und ihre Metadaten enthalten Felder, um u.a. beteiligte Personen und Organisationen zu erfassen (TextGrid-Konsortium, 2015b). Eine Edition ist immer eine bestimmte Ausgabe eines Werks und muss daher mit mindestens einem Werk assoziiert sein (TextGrid-Konsortium, 2015b, TextGrid-Konsortium, 2015d).

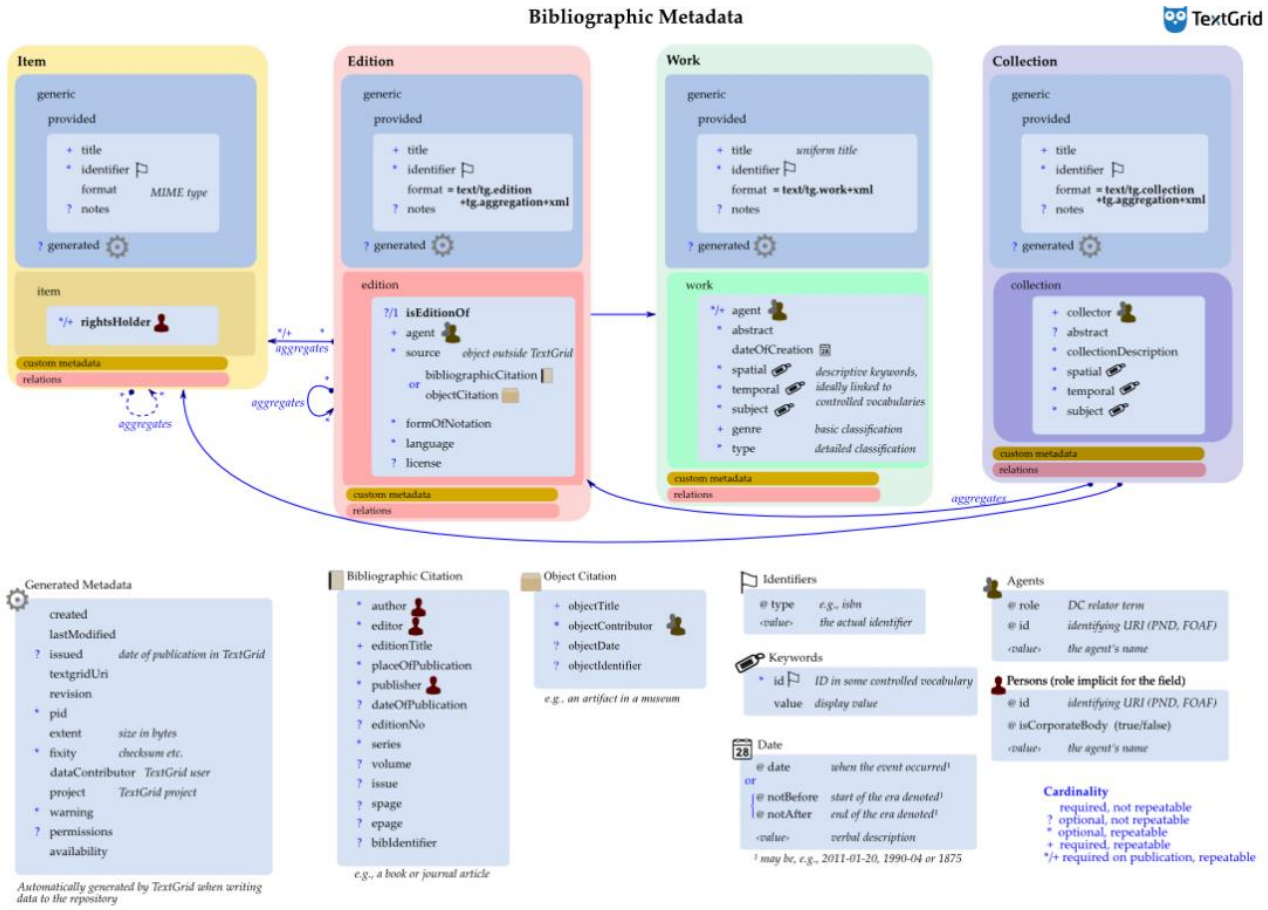
2.1.7.3 Kollektion (Form der Aggregation): text/tg.collection+tg.aggregation+xml

Eine Kollektion ist eine Anhäufung von TextGrid-Objekten, die aufgrund eines bestimmten Themas aggregiert werden sollen (TextGrid-Konsortium, 2015c). Beispielsweise können auf diese Weise mehrere Editionen mit digitalisierten Drucken des achtzehnten Jahrhunderts unter der Kennzeichnung *VD 18* kombiniert werden. (TextGrid-Konsortium, 2015c).

2.1.7.4 Werk: text/tg.work+xml

Ein Werk ist eine individuelle Schöpfung, zum Beispiel ein literarisches Werk, das in verschiedenen Editionen vorliegen kann, beispielsweise als Taschenbuchausgabe, als Teil einer Sammlung von Werken, als Theateraufführung oder als Hörbuch (TextGrid-Konsortium, 2015d). Werk-Objekte in TextGrid beinhalten Metadaten wie einen eindeutigen Titel, das Erstellungsdatum oder das Werk beschreibende Schlüsselwörter (TextGrid-Konsortium, 2015d). Zum Zeitpunkt der Veröffentlichung muss jedes Editions-Objekt mit einem Werk-Objekt verknüpft sein (TextGrid-Konsortium, 2015d).

2.1.7.5 Metadata Cheatsheet



2.2 Objektdaten

Unter Objektdaten werden die Daten verstanden, die primär im Forschungskontext von Forschenden ausgewertet und bearbeitet werden. In diesem Sinne sind etwa Transkriptionen einer Audio-Datei keine Metadaten, auch wenn sie streng genommen eine andersartige Repräsentation von Informationen in den Audio-Daten sind. Annotationen von Texten werden ebenfalls zu den Objektdaten gezählt, aber auch die Textdateien selbst. Im Folgenden werden einige Objektdaten-typen vorgestellt, die in Text+ verwendet werden.

2.2.1 Geschriebene Sprache

Für die standardkonforme Aufbereitung genuin geschriebener Sprache kann ein vierstufiger Aufbereitungsprozess skizziert werden: (1) die Akquise, oftmals Transkription und Bereinigung der Quellen, (2) die Annotation grammatischer Informationen, (3) die Annotation nach textstrukturellen, semantischen und inhaltlichen Gesichtspunkten sowie (4) die Erfassung von Metadaten. Punkt 4 war Gegenstand von Kap. 2.1; die Punkte 1 bis 3 werden im Folgenden näher erläutert.

2.2.1.1 Akquise der Textgrundlage

Der Schritt, wie Textdaten gewonnen werden, unterscheidet sich je nachdem, ob es sich um genuin analoge oder genuin digitale (born-digital) Daten handelt. Historische Texte, die primär analog entstanden sind (egal, welches Schreibmaterial zur betreffenden Zeit zur Verfügung stand) müssen zunächst in das digitale Format gebracht werden. Die Transkription kann manuell oder mithilfe

automatisierter Verfahren (Optical Character Recognition, OCR) erfolgen[3]. Der erfasste Text sollte im UTF-8-Format kodiert sein, damit auch Sonderzeichen im Zuge der Nachnutzung korrekt entschlüsselt und wiedergegeben werden können. Abweichungen der Transkription von der Vorlage sind zu dokumentieren, entweder via Annotation im Text oder, wenn es sich um systematische Abweichungen handelt, an geeigneter Stelle in den Metadaten. Der erfasste Text ohne weitere Annotationen sollte im TXT-Format vorliegen. Für weitere Annotationen bieten sich XML-Formate an.

2.2.1.2 Annotation

Für Annotationen gilt grundsätzlich, dass eine Dokumentation notwendig ist, welche die einzelnen Auszeichnungen erläutert. XML als Annotationsstandard bringt eine erhebliche Menge an Tools für die Weiterverarbeitung und Auswertung von Daten mit sich und hat sich daher als Standard für die Textaufbereitung etabliert. Da XML selbst jedoch nur wenige Regeln für die konkreten Annotationen enthält, müssen auch XML-basierte Formate dokumentiert werden. Es hat sich daher als sinnvoll erwiesen, dass bestehende Regelsysteme nachgenutzt werden, da damit 1. bereits Dokumentationen vorhanden sind und 2. Daten eher über Projekte hinaus interoperabel sind bzw. ihre Nachnutzung wesentlich erleichtert wird.

Für die Annotation von Layout- und logischen Strukturen haben sich die P5-Guidelines der Text Encoding Initiative (TEI) als de facto-Standard etabliert. Da die Variabilität innerhalb dieses Regelsystems allerdings weiterhin groß ist, wurden auf Grundlage der TEI-P5-Guidelines Formate gebildet, die weitere Einschränkungen vornehmen mit dem Ziel, Interoperabilität zwischen Dokumenten und Korpora zu gewährleisten und so den Aufwand für die Nachnutzung in neuen Kontexten zu minimieren. Für historische Daten des Deutschen (mit Anwendbarkeit für andere europäische Sprachen) wird das DTA-Basisformat (DTABf; Haaf, Geyken & Wiegand, 2014; Haaf & Thomas, 2016)[4] bereitgestellt und in unterschiedlichen Fachkontexten für die Nachnutzung vorgeschlagen[5]. Weitere verbreitete generische TEI-Formate sind etwa TEI Lite oder TEI SimplePrint[6]. Auch das Deutsche Referenzkorpus (DeReKo) verwendet mit I5 (siehe Lüngen & Sperberg-McQueen, 2012) ein in TEI P5 definiertes Format. Daneben existieren zahlreiche weitere TEI-Formate, welche z.B. gebunden sind an inhaltliche Gegebenheiten (vgl. Bański & Hedeland, 2022: 325).

OCR-Verfahren bieten als Output in der Regel ebenfalls XML-basierte Formate an (verbreitet sind ALTO-XML und hOCR). Diese enthalten u.a. basale Layout-Informationen (z. B. über Spaltensatz, Abbildungen u. dgl.) sowie Informationen über die Platzierung der erfassten Zeichen auf der jeweiligen Seite. Da diese Formate ebenfalls standardisiert sind, ist ihre Weiterverarbeitung und Transformation z.B. nach TEI-P5 grundsätzlich möglich.

Annotationen, die gegenüber Formaten wie DTABf und TEI Lite domänenspezifisch weiter in die Tiefe gehen, können innerhalb der TEI z.B. mit Feature Structures realisiert werden[7]. Auch hier ist es von großer Relevanz, dass verständliche Bezeichnungen gewählt und ausführliche Dokumentationen beigegeben werden, um die Nachvollziehbarkeit und Nachnutzung zu ermöglichen.

---

[3] Für Empfehlungen zur Wahl einer OCR-Software s. https://www.berd-nfdi.de/limesurvey/index.php/996387.

[4] Das DTABf ist dokumentiert unter: https://deutschestextarchiv.de/doku/basisformat/. Das Format wird entwickelt unter: https://github.com/deutschestextarchiv/dtabf. Die Entwicklung wird gegenwärtig durch eine Steuerungsgruppe betreut (https://deutschestextarchiv.de/doku/basisformat/steuerungsgruppe.html) und folgt festgelegten Leitlinien (https://deutschestextarchiv.de/doku/basisformat/leitlinien.html). Das Schema für gedruckte Texte ist unter http://www.deutschestextarchiv.de/basisformat.rng für die direkte Integration in Dokumente zugänglich.

[5] Vgl. etwa die Empfehlungen des DFG-Fachkollegiums 104 “Sprachwissenschaften" 2019 (https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf) sowie die DFG-Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft 2015 (https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf).

[6] Vgl. https://tei-c.org/guidelines/customization/.

[7] Vgl. https://tei-c.org/Vault/P5/4.6.0/doc/tei-p5-doc/en/html/FS.html#FSBI.

Annotationen grammatischer Informationen betreffen standardmäßig Informationen zu Lemma (Grundform) und Part-of-Speech (Wortart) auf Token-Ebene. Hinzu kommen können auf Token-Ebene Informationen zu orthographischen Unregelmäßigkeiten sowie morphosyntaktische Informationen (z.B. zur Flexion). Weitere Auszeichnungen betreffen in der Regel die Satzebene, und dabei sowohl die Auszeichnung nach Sätzen als auch nach syntaktischen Informationen, etwa Satzbestandteilen und Satzstrukturen als Ergebnissen eines Syntax-Parsing oder -Chunking. Auch darüber hinaus sind aus linguistischer Perspektive Annotationen denkbar, etwa zu Sprachhandlungen, Satztypen u.a. Während linguistische Informationen auf Token-Ebene noch im TEI-Text mit übermittelt werden können (Bański, Haaf & Mueller, 2018)[8], werden üblicherweise hierfür und insbesondere für die Auszeichnung von Informationen auf Satz- oder Textebene Stand-off-Formate genutzt. Als linguistisches Austauschformat wurde seitens CLARIN das XML-basierte TCF-Format bereitgestellt, welches inhaltlich mit den Formaten LAF und GrAF kompatibel ist und für die verschiedenen Annotationsebenen eigene Bereiche bereitstellt (Eckart, 2012)[9]. Verbreitet sind außerdem CoNLL-Formate, welche Annotationen in Tab-separierten Tabellen kodieren, wobei jede Zeile ein Wort und jede Spalte eine Annotation beinhaltet. Gängig ist hier z.B. das auf CoNLL-X basierende CoNLL-U-Format, welches u.a. Informationen zu Dependenzrelationen enthält[10]. Korpusannotations- und Analyseplattformen nutzen zum Teil spezifische XML-Schemata (WebAnno bzw. INCEpTION mit XMI[11]) oder spezifisches TEI als Exportformat (CATMA[12]).

2.2.2 Gesprochen-Sprachliche Daten

Transkriptionen gesprochener Sprache stellen im Bereich objektsprachlicher Daten in mehrerlei Hinsicht besondere Anforderungen an die zu definierenden Standards und Datenformate.

1. *Zeitalignierung*. Sammlungen gesprochensprachlicher Objektdaten bestehen aus Transkripten mit jeweils mindestens einer (oft mehreren) dazugehörigen Audio- und/oder Videoaufnahmen. Ohne eine feingranulare Referenzierung der transkribierten Einheiten zu den korrespondierenden Zeitpunkten sind die Sammlungen in der überwiegenden Mehrheit der Fälle zu wissenschaftlichen Zwecken nicht nutzbar.
2. *Überlappungen*. Sammlungen gesprochensprachlicher Objektdaten beinhalten in der Regel Aufnahmen und Transkripte, an denen mehrere Individuen teilweise synchrone Äußerungen tätigen. Diese Überlappungen sind hochrelevant für eine Vielzahl von Analysen und müssen in den gewählten Dateiformaten entsprechend modelliert werden.
3. *Nicht-standardisierte Einheiten und Segmentierungen*. Aufgrund der besonderen Beschaffenheit gesprochener Sprache weichen deren Transkriptionen in ihrer Form stark von geschriebenen standardorthographischen Texten ab. Etablierte Transkriptionssysteme verwenden zudem neben eigenen Konventionen zur Repräsentation von Phänomenen gesprochener Sprache auch individuelle Klassifikationen linguistische Einheiten. Beispielhaft zu nennen sind die Begriffe *Äußerung* (HIAT[13]) oder *Intonationsphrasen* (GAT[14]) gegenüber standardorthographischen Satzeinheiten.

Ziel der Standardisierungsbemühungen von Initiativen wie Text+, die auf nachhaltige und breite Nutzbarkeit der Forschungsdatendatensammlungen ausgerichtet sind, muss daher sein, den o.g.

[8] Vgl. https://tei-c.org/Vault/P5/4.6.0/doc/tei-p5-doc/en/html/ref-att.linguistic.html.

[9] Vgl. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

[10] Vgl. https://universaldependencies.org/format.html, https://universaldependencies.org/ext-format.html.

[11] Vgl. https://www.omg.org/spec/XMI

[12] Vgl. https://catma.de/documentation/access-your-project-data/tei-export-format/

[13] Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004) Handbuch für das computergestützte Transkribieren nach HIAT. In: Arbeiten zur Mehrsprachigkeit, Folge B 56, 1 ff.
https://www.exmaralda.org/files/azm_56.pdf

[14] Selting, Margret et al. (2009): "Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)". In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10 (2009), 353–402.
http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf

Spezifika so weit wie möglich zu entsprechen und gleichzeitig die Anbindung und Interoperabilität mit den parallel verwendeten Standards und Dateiformaten zu gewährleisten.

Insbesondere in Bezug auf die unter 1. und 2. genannten Merkmale bedienen sich alle in den vergangenen Jahren im Bereich der Transkription und Analyse gesprochener Sprache etablierten Werkzeuge graphenbasierter Repräsentationen im XML-Format.
In diesem Zusammenhang zu nennen sind das EXMARaLDA-System, sowie die Werkzeuge ELAN und Praat, die in unterschiedlichen Fachcommunities große Vorbereitung besitzen und für die bereits seit fast zwei Jahrzehnten Konvertierungsroutinen zwischen den verwendeten Dateiformaten existieren.

Einen entscheidenden auf dieser Entwicklung fußenden Fortschritt stellt die Etablierung des Standards *ISO 24624:2016 Language resource management — Transcription of spoken language* dar, der als erster ISO-Standard Regeln für die Darstellung von Transkriptionen von audio- und video-aufgezeichneten gesprochenen Interaktionen in XML-Dokumenten spezifiziert. Er ermöglicht unter anderem die Verwendung von auf unterschiedlichen Transkriptionssystemen basierenden Segmentierungen bei gleichzeitiger Kompatibilität zu den Richtlinien der TEI.
Zum Zeitpunkt der Entstehung dieses Dokumentes ist eine Anpassung verschiedener Softwareprogramme an diesen Standard zu beobachten[15].
Basierend auf den Entwicklungen der vergangenen Jahre und den Erfahrungen der beteiligten Datenzentren (UHH/HZSK, AdWHH und BAS) kann davon ausgegangen werden, dass für die Integration in die Infrastruktur von Text+ alle Formate zur Repräsentation (zeitalignierter) gesprochen-sprachlicher Daten problemlos genutzt werden, die in einem Format vorliegen, das nachweislich in ISO 24624 überführt werden kann.
Grundsätzlich können alle Datenformate, für die Konvertierungsroutinen in eines der oben genannten Formate und Werkzeuge bestehen, aufgenommen und zugänglich (d.h. online durchsuchbar und visuell zugänglich) gemacht werden.

3. Datenpaketformate

Datenpaketformate (data packaging formats) sind ein Bereich des Datenmanagements, der in den letzten Jahren zunehmend Aufmerksamkeit erhalten hat. Datenpakete sind eine Möglichkeit, Daten und Metadaten zusammenzufassen, um sie gemeinsam zu verwalten und zu archivieren. Eine in manchen Datenpaketformaten obligatorische Manifest-Datei liefert ein komplettes Inventar des Paketes und ermöglicht es, die Integrität der Datenpakete zu überprüfen. Datenpakete können auch als Container für die Übertragung von Daten zwischen Systemen verwendet werden. Typische Pakete sind bereits im OAIS-Modell (ISO 14721:2012) für die Archivierung beschrieben, dort noch abstrakt z.B. als Submission Information Package (SIP), Archive Information Package (AIP) oder Dissemination Information Package (DIP).

Der Impuls für die Entwicklung von Datenpaketformaten entstammt nicht nur der praktischen Arbeit an Datenrepositorien, sondern ist auch die direkte Konsequenz der Diskussion zu Digital Objects (Kahn & Wilensky, 2006; ursprünglich 1996) und in der Folge der FAIR Prinzipien (Wilkinson et al., 2016) zu FAIR Digital Objects (FDO; De Smedt, Koureas, und Wittenburg, 2020). Datenpakete sind eine Möglichkeit, FDOs zu realisieren. Datenpaketformate werden in der Praxis sowohl als Submission Information Package (SIP) und Exchange Formate aber als auch als Archival Information Package (AIP) verwendet. Sie sind in der Regel nicht auf bestimmte Datenarten beschränkt, sondern können für beliebige Daten verwendet werden. Es gibt jedoch auch Datenpaketformate, die speziell für bestimmte Datenarten entwickelt wurden, wie z.B. für Sprachdaten (Cross-Linguistic Data Formats).

[15] z. B. Release des EXMARaLDA-Systems vom 20. Juli 2023

Ein weitverbreitetes Datenpaketformat ist das BagIt File Packaging Format (Kunze et al., 2018). BagIt ist ein einfaches Format, das ursprünglich für die Übertragung von Daten zwischen Systemen entwickelt wurde. Es wird inzwischen auch häufig als Archivierungsformat eingesetzt. BagIt ist ein Datenpaketformat, das aus einer Sammlung von Dateien besteht, die in einem Verzeichnisbaum organisiert sind. Die Dateien werden mit einer Hash-Funktion in eine Manifest-Datei aufgenommen, die die Integrität des Datenpakets sicherstellt. BagIt ist ein simples Format, dass nur minimale Metadaten erfordert, keine Versionierung oder andere weiterreichenden Optionen spezifiziert, aber auch keine speziellen Anforderungen an die Struktur der Daten stellt.

Oxford Common File Layout (OCFL, Hankinson et al., 2019) ist ein komplexeres Format. Es basiert konzeptuell auf BagIt und bietet wie dieses eine Manifest-Datei mit Datei-Hashes. OCFL definiert darüber hinaus aber auch eine Struktur für die Daten, die eine Versionierung der Daten ermöglicht. OCFL ist ein Format, das explizit für die Archivierung von Daten in Repositorien entwickelt wurde. Es ist daher weniger für die Übertragung von Daten zwischen Systemen geeignet.

Bei FDOs ist die Kombination von Digital Objects mit Metadaten und Persistent Identifiern (PIDs) zentral. BagIt und OCFL definieren aber kein weitergehendes Metadatenformat. Für die Verwendung von BagIt und OCFL als FDOs müssen daher zusätzlich ein umfassenderes Metadatenformat und ein PID-System verwendet werden.

Andere Datenpaketformate, wie z.B. Research Object Crate (RO-Crate; Soiland-Reyes et al., 2022), definieren ein umfassenderes Metadatenformat als Teil des Datenpakets. Durch Einbeziehung des Metadatenformats in die Datenpaketdefinition sind die im Paket enthaltenen Metadaten und deren Serialisierung vorgegeben. Dadurch werden die Pakete maschinell leichter nutzbar und FAIRer. RO-Crate ist ein Datenpaketformat, das explizit für die Verwendung als FDO entwickelt wurde und enthält Metadaten in Form einer JSON-LD-Datei (Kellogg, Champin, Longley, 2019), die an Schema.org angelehnt ist. Die Spezifikation der RO-Crates schließt allerdings keine Manifest-Datei ein, so dass die Integrität der Datenpakete nicht automatisch sichergestellt ist.

RO-Crates können mit BagIt- oder OCFL-Objekten kombiniert werden, um die Integrität der Datenpakete sicherzustellen, gleichzeitig standardisierte Metadaten zu verwenden und so die Vorteile der verschiedenen Datenpaketformate zu verbinden.

Für Sprachressourcen gibt es darüber hinaus die Cross-Linguistic Data Formats (CLDF). CLDF (Forkel et al., 2018) ist ein Datenpaketformat, das speziell für Sprachdaten entwickelt wurde. Es basiert auf den *World Wide Web Consortium (W3C) recommendations Model for Tabular Data and Metadata on the Web* (Kellogg & Tennison, 2015) und dem *Metadata Vocabulary for Tabular Data* (Tennison & Kellogg, 2015). Es kann vor allem für tabellarische Daten verwendet werden und ist zum Beispiel mit BagIt kombinierbar.

Je nach Einsatzzweck und Anforderungen an die Datenpakete können unterschiedliche Datenpaketformate verwendet werden. Die Wahl des Datenpaketformats ist eine wichtige Entscheidung, die die Möglichkeiten der Datenverwaltung und -archivierung erheblich beeinflusst, nicht zuletzt aus dem Grund, dass Repositoriensysteme unterschiedliche Anforderungen für Ingestprozesse haben und nur mit bestimmten Datenpaketformaten umgehen können. Die Wahl des Datenpaketformats sollte daher sorgfältig abgewogen werden. In der Regel ist es aber immer sinnvoll ein Datenpaketformat zu verwenden, da es die Verwaltung der Daten erleichtert und die Integrität der Daten sicherstellt.

4. Formate für andere Dokumenttypen

Die bisherigen Ausführungen in diesem Papier widmeten sich verschiedenen Dokumenttypen auf dem Gebiet der geschriebenen und gesprochenen Sprache, das den größten Teil des Zuständigkeitsbereichs von Text+ ausmacht. Jenseits davon beschäftigt sich Text+ aber auch mit weiteren Dokumenttypen, darunter insbesondere Experimentaldaten (von Fragebogenstudien bis EEG-

Messungen) oder (große) Sprachmodelle. Das Feld ist – ja nach Forschungsfrage und -gegen-stand – sehr heterogen und in Teilen hochdynamisch. Daher ist es zum aktuellen Zeitpunkt für Text+ nicht möglich, allgemeinen Vorschläge für Standards abzugeben. Im Hinblick auf Metadaten können jedoch auch für diese Art von Daten die in Kapitel 1 vorgestellten Modelle verwendet werden.

5. Fazit

Zusammenfassend kann festgehalten werden, dass Text+ die Verwendung von Standardformaten grundsätzlich empfiehlt, wo immer sie allgemein etabliert sind und zum Anwendungskontext passen. Sei es bei den Objektdaten selbst, die direkter Gegenstand der Forschung sind, als auch bei den beschreibenden Metadaten, die zur Auffindbarkeit und Katalogisierung der Objektdaten beitragen, und schließlich auch den Containerdaten, die eine Zusammenfassung der Objekt- und der zugehörigen Metadaten ermöglichen.

Ist geplant, die Forschungsdaten in einem Datenzentrum von Text+ zu archivieren und so über die Infrastruktur von Text+ für andere auffind- und nachnutzbar zu machen, ist es von großem Vorteil, von vornherein möglichst die Standards und Formate zu verwenden, die vom Ziel-Datenzentrum bevorzugt werden. Orientierung und Rat bei der Suche nach dem geeignetsten Zentrum bietet die oben bereits erwähnte Publikation *Leitlinie für das Integrieren von Daten in Text+/NFDI* sowie der Text+ Helpdesk.

Sollten spezifische Anforderungen, die aus der Natur der Daten selbst oder der Forschungsfrage herrühren, Erweiterungen der angewendeten Standards notwendig machen, ist besonders wichtig, diese so standardkonform wie möglich vorzunehmen. Ebenso unabdingbar für die weitere Verwendung und Nachnutzung der Daten ist eine ausführliche Dokumentation solcher Erweiterungen.

Anhang:
Mappings zwischen verschiedenen Metadatenschemata und Dublin Core

A DataCite

DataCite-Property	Dublin Core Qualified
1. Identifier	dc.identifier
1.a identifierType	–
2. Creator	dc.creator
2.1 creatorName	dc.creator
2.1.a nameType	–
2.2 givenName	–
2.3 familyName	–
2.4 nameIdentifier	dc.creator.pid
2.4.a nameIdentifierScheme	–
2.4.b schemeURI	–
2.5 affiliation	dc.contributor
2.5.a affiliationIdentifier	dc.contributor.pid
2.5.b affiliationIdentifierScheme	–
2.5.c schemeURI	–
3. Title Mapped by 3.a titleType:	dc.title
• AlternativeTitle	dc.title.alternative
• Subtitle	dc.title 1
• TranslatedTitle	dc.title.alternative
• Other	dc.title.alternative
3.a titleType	–
4. Publisher	dc.publisher
4.a publisherIdentifier	dc.publisher.pid
4.b publisherIdentifierScheme	–
4.c schemeURI	–
5. PublicationYear	dc.date.issued
6. Subject	dc.subject
6.a subjectScheme	–
6.b schemeURI	–
6.c valueURI	dc.subject.pid
6.d classificationCode	dc.subject

7. Contributor	dc.contributor
7.a contributorType	–
7.1 contributorName	dc.contributor
7.1.a nameType	–
7.2 givenName	–
7.3 familyName	–
7.4 nameIdentifier	dc.contributor.pid
7.4.a nameIdentifierScheme	–
7.4.b schemeURI	–
7.5 affiliation	dc.contributor
7.5.a affiliationIdentifier	dc.contributor.pid
7.5.b affiliationIdentifierScheme	–
7.5.c schemeURI	–
8. Date Mapped by 8.a dateType:	dc.date
• Accepted	dc.date.accepted
• Available	dc.date.available
• Copyrighted	dc.date.copyrighted
• Collected	dc.date
• Created	dc.date.created
• Issued	dc.date.issued
• Submitted	dc.date.submitted
• Updated	dc.date.modified
• Valid	dc.date.valid
• Withdrawn	dc.date
• Other	dc.date
8.a dateType	–
8.b dateInformation	dc.description
9. Language	dc.language
10. ResourceType	dc.type
10.a resourceTypeGeneral	dc.type
11. AlternateIdentifier	dc.identifier
11.a alternateIdentifierType	–
12. RelatedIdentifier Mapped by 12.b relationType:	dc.relation
• IsReferencedBy	dc.relation.isReferencedBy
• References	dc.relation.references
• IsVersionOf	dc.relation.isVersionOf

• HasVersion	dc.relation.hasVersion
• IsVariantFormOf	dc.relation.isFormatOf
• IsPartOf	dc.relation.isPartOf
• HasPart	dc.relation.hasPart
• IsObsoletedBy	dc.relation.isReplacedBy
• Obsoletes	dc.relation.replaces
• IsDerivedFrom	dc.source or dc.relation.source
• *Other relationTypes*	dc.relation
12.a relatedIdentifierType	–
12.b relationType	–
12.c relatedMetadataScheme	–
12.d schemeURI	–
12.e schemeType	–
12.f resourceTypeGeneral	–
13. Size	dc.format.extent
14. Format	dc.format
15. Version	dc.title 2
16. Rights	dc.rights
16.a rightsURI	dc.rights.license
16.b rightsIdentifier	dc.rights
16.c rightsIdentifierScheme	–
16.d schemeURI	–
17. Description Mapped by 17.a descriptionType:	dc.description
• Abstract	dc.description.abstract
• Methods	dc.description
• SeriesInformation	dc.description
• TechnicalInfo	dc.description
• TableOfContents	dc.description.tableOfContents
• Other	dc.description
17.a descriptionType	–
18. GeoLocation	dc.coverage.spatial
18.1 geoLocationPoint	dc.coverage.spatial
18.1.1 pointLongitude	dc.coverage.spatial
18.1.2 pointLatitude	dc.coverage.spatial
18.2 geoLocationBox	dc.coverage.spatial
18.2.1 westBoundLongitude	dc.coverage.spatial

18.2.2 eastBoundLongitude	dc.coverage.spatial
18.2.3 southBoundLatitude	dc.coverage.spatial
18.2.4 northBoundLatitude	dc.coverage.spatial
18.3 geoLocationPlace	dc.coverage.spatial
18.4 geoLocationPolygon	dc.coverage.spatial
18.4.1 polygonPoint	dc.coverage.spatial
18.4.1.1 pointLongitude	dc.coverage.spatial
18.4.1.2 pointLatitude	dc.coverage.spatial
18.4.2 inPolygonPoint	dc.coverage.spatial
18.4.2.1 pointLongitude	dc.coverage.spatial
18.4.2.2 pointLatitude	dc.coverage.spatial
19. FundingReference	–
19.1 funderName	dc.contributor
19.2 funderIdentifier	dc.contributor.pid
19.2.a funderIdentifierType	–
19.2.b schemeURI	–
19.3 awardNumber	dc.relation
19.3.a awardURI	dc.relation.pid
19.4 awardTitle	dc.relation
20. RelatedItem Mapped by 20.b relationType as above for 12. RelatedIdentifier.	dc.relation 3
20.a relatedItemType	–
20.b relationType	–
20.1 relatedItemIdentifier	dc.relation
20.1.a relatedItemIdentifierType	–
20.2 creator	–
20.2.1 creatorName	–
20.3 title	–
20.3.a titleType	–
20.4 publicationYear	–
20.5 volume	–
20.6 issue	–
20.7 number	–
20.7.a numberType	–
20.8 firstPage	–
20.9 lastPage	–
20.10 publisher	–

20.11 edition	–
20.12 contributor	–
20.12.a contributorType	–
20.12.1 contributorName	–

B OLAC

OLAC-Elements	Dublin Core Elements
Contributor	Contributor
Coverage	Coverage
Creator	Creator
Date	Date
Description	Description
Format	Format
Format.cpu	–
Format.encoding	–
Format.markup	–
Format.os	–
Format.sourcecode	–
Identifier	Identifier
Language	Language
Publisher	Publisher
Relation	Relation
Rights	Rights
Source	Source
Subject	Subject
Subject.language	–
Title	Title
Type	Type
Type.functionality	–
Type.linguistic	–

C TEI-Header

TEI-Header							**Dublin Core**
Title Stmt	Edition Stmt	Extent	Publication Stmt	Series Stmt	Notes Stmt	Source Desc	Dublin Core
Author			Authority				Creator
Funder							Contributor
sponsor							
principal							
respStmt	resp tmt		respStmt	respStmt			Rights
			Publisher/ Distributor				Publisher
Title							Title
							Relation
					Notes		Description
	Edition	Extent					Coverage
						Bibl	Source
						listBibl	
						biblFull	
			Idno	idno			identifier
			Availability				–
							type
							subject
							date
							format
							language

Bibliographie

Bański, Piotr, Susanne Haaf, Martin Mueller: Lightweight Grammatical Annotation in the TEI: New Perspectives. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 7.–12. Mai 2018, Miyazaki (Jp), S. 1795–1802 http://www.lrec-conf.org/proceedings/lrec2018/pdf/422.pdf.

Bański, Piotr & Hanna Hedeland: Standards in CLARIN. In: Fišer, Darja & Andreas Witt (Eds.): CLARIN. The infrastructure for language resources. Berlin/Boston: de Gruyter, 2022. Pp. 307-339. (Digital Linguistics 1) DOI: https://doi.org/10.1515/9783110767377

Barkey, Reinhild, Erhard Hinrichs, Christina Hoppermann, Thorsten Trippel, Claus Zinn, J. Griesbaum, T. Mandl, und C. Womser-Hacker. „Komponenten-basierte Metadatenschemata und Facettenbasierte Suche: Ein flexibler und universeller Ansatz“. Gehalten auf dem 12. Internationales Symposium für Informationswissenschaft (ISI 2011), Hildesheim, 2011

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Thorsten Trippel, Twan Goosen, Victoria Arranz, Daan Broeder, u. a. „CMDI: a Component Metadata Infrastructure“. In Proceedings of the LREC 2012 Workshop Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources. Istanbul, Türkei, 2012.

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, und Andreas Witt. „A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI).“ Gehalten auf der Presented at Balisage: The Markup Conference 2011, Montréal, Canada, 2. August 2011.

Coyle, K.; Baker, Thomas (2009): Guidelines for Dublin Core Application Profiles. Dublin Core Metadata Initiative, 2009-05-18, http://dublincore.org/documents/2009/05/18/profile-guidelines/.

DataCite Metadata Working Group. (2024). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.5. DataCite e.V. https://doi.org/10.14454/g8e5-6293

De Smedt, Koenraad, Dimitris Koureas, and Peter Wittenburg. 2020. “FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units.” Publications 8 (2): 21. https://doi.org/10.3390/publications8020021.

Dublin Core (2003). DCMI metadata terms. http://dublincore.org/documents/2003/03/04/dcmi-terms/ , 2003

Eckart, Kerstin: Chapter 3. Resource annotations. In: CLARIN-D User Guide. Written by CLARIN-D AP 5. Version 1.0.1. https://www.clarin-d.net/en/help/user-handbook.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” Scientific Data 5 (October):180205. https://doi.org/10.1038/sdata.2018.205.

Hankinson, Andrew, Donald Brower, Neil Jefferies, Rosalyn Metz, Julian Morley, Simeon Warner, and Andrew Woods. 2019. “The Oxford Common File Layout: A Common Approach to Digital Preservation.” Publications 7 (2): 39. https://doi.org/10.3390/publications7020039.

Diane I. Hillmann. Using Dublin Core. http://dublincore.org/documents/2000/07/16/usageguide, July 2000. Working Draft

IFLA Study Group on the Functional Requirements for Bibliographic Records. 2009. “Functional Requirements for Bibliographic Records”. Accessed July 1, 2023. https://repository.ifla.org/bitstream/123456789/811/2/ifla-functional-requirements-for-bibliographic-records-frbr.pdf.

ISO 14721. „Space Data and Information Transfer Systems — Open Archival Information System (OAIS) — Reference Model.” September 2012

ISO 24619. „Language resource management - Persistent identification and sustainable access (PISA)“. International Standard. Geneva: International Organization for Standardization (ISO), Mai 2011.

ISO24622-1. „Language resource management -- Component Metadata Infrastructure (CMDI) -- Part 1: The Component Metadata Model“. Internationale Norm. Genf: International Organization for Standardization (ISO), 20. Januar 2015.

ISO 24622-2. „Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language“. Internationale Norm. Geneva: International Organization for Standardization (ISO), Juli 2019.

ISO 24624. Language resource management — Transcription of spoken language International Organization for Standardization (ISO), August 2016.

TextGrid-Konsortium. “Nutzerhandbuch 2.0”. Aggregationen. 2015a. Accessed July 1, 2023. https://doc.textgrid.de/Aggregationen_40220430Literatur:

Susanne Haaf, Alexander Geyken, Frank Wiegand (2014/15): *The DTA “Base Format”: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources.* In: jTEI 8 (Selected Papers from the 2013 TEI Conference). http://jtei.revues.org/1114.

Susanne Haaf, Christian Thomas (2016): *Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format.* jTEI 10 (Selected Papers from the 2015 TEI Conference). https://journals.openedition.org/jtei/1650.

Kahn, Robert, and Robert Wilensky. 2006. “A Framework for Distributed Digital Object Services.” *International Journal on Digital Libraries* 6 (2): 115–23. https://doi.org/10.1007/s00799-005-0128-x.

Gregg Kellogg, Pierre-Antoine Champin, Dave Longley. 2019. JSON-LD 1.1 – A JSON-based Serialization for Linked Data (W3C Working Draft). [Technical Report] W3C.

Kellogg, Gregg, and Jeni Tennison. 2015. “Model for Tabular Data and Metadata on the Web.” W3C Recommendation. W3C.

Kunze, John, Justin Littman, Elizabeth Madden, John Scancella, and Chris Adams. 2018. “The BagIt File Packaging Format (V1.0).” Internet Engineering Task Force. https://tools.ietf.org/html/rfc8493.

Harald Lüngen and C. M. Sperberg-McQueen (2012): *A TEI P5 Document Grammar for the IDS Text Model*. Journal of the Text Encoding Initiative, Issue 3.

Soiland-Reyes, Stian, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, et al. 2022. “Packaging Research Artefacts with RO-Crate.” Edited by Silvio Peroni. Data Science 5 (2): 97–138. https://doi.org/10.3233/DS-210053.

Gary Simons and Steven Bird. OLAC metadata. http://www.language-archives.org/OLAC/metadata.html, 05 2008.

TEI P5. (2021): P5: Guidelines for electronic text encoding and interchange (version 4.2.1. last updated on 1st march 2021, revision 654a5c551). Technical report, Text Encoding Initiative.

Tennison, Jeni, and Gregg Kellogg. 2015. “Metadata Vocabulary for Tabular Data.” W3C Recommendation. W3C.

TextGrid-Konsortium. “Nutzerhandbuch 2.0”. Editionen. 2015b. Accessed July 1, 2023. https://doc.textgrid.de/Editionen_40220434.

TextGrid-Konsortium. “Nutzerhandbuch 2.0”. TextGrid: Kollektionen. 2015c. Accessed July 1, 2023. https://doc.textgrid.de/Kollektionen_40220436.

TextGrid-Konsortium. “Nutzerhandbuch 2.0”. TextGrid-Objekte. 2018. Accessed July 1, 2023. https://doc.textgrid.de/TextGrid-Objekte.

TextGrid-Konsortium. “Nutzerhandbuch 2.0”. Werke. 2015d. Accessed July 1, 2023. https://doc.textgrid.de/Werke_40220432.

TextGrid-Konsortium. “TextGrid-Metadaten“. XML Schema. 2010. Accessed July 3, 2023. https://textgrid.info/namespaces/metadata/core/2010

TextGrid-Konsortium. “TextGrid Metadata Cheatsheet“. 2010. Accessed July 3, 2023. https://doc.textgrid.de/attachments/metadata/cheatsheet.pdf

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

“Model for Tabular Data and Metadata on the Web.” n.d. Accessed September 25, 2020. https://www.w3.org/TR/tabular-data-model/.