MICHAEL REICHE[1], DIANA TILLMANN[1], ALEXANDER GROSSMANN[1], DAVID BÖHM[1]

Zur Situation des strukturierten Austauschs von Metainformationen über die Herstellungs- und Distributionsprozesse von Open-Access-Publikationen

[1] HTWK Leipzig - Leipzig University of Applied Sciences, 04251 Leipzig, Germany

Diese Arbeit ist im Rahmen des Forschungsprojektes OA-STRUKTKOMM (Open-Access-Strukturierte-Kommunikation) entstanden und vom Bundesministerium für Bildung und Forschung (BMBF) gefördert.

KURZFASSUNG

Open Access Publikationen sollen von allen Menschen, unabhängig vom Ort und unabhängig von finanziellen, gesetzlichen oder technischen Restriktionen, problemlos rezipiert werden können. Um dies sicherzustellen, müssen neben Sprachbarrieren auch die Metadatenbarrieren überwunden werden. Metadaten werden in artifiziellen Sprachen verfasst, die wie die natürlichen Sprachen einer ständigen Entwicklung unterworfen sind. Kommen bei natürlichen Sprachen Dolmetscher zum Einsatz, um zwischen den Sprachen zu vermitteln, sollen Metadaten mittels Standards harmonisiert und ausgetauscht werden. Da die Bemühungen um Standardisierung gleichzeitig in verschiedenen Domains stattfinden, führt dies notwendigerweise zu einer Heterogenität der Metadatenstandards. Die Open-Access-Publikationsworkflows sind in diese heterogene Metadatenlandschaft eingebettet. Das heißt, abhängig vom Veröffentlichungsweg müssen verschiedene Metadatensätze erstellt werden. Dies führt zu einem Transformationsaufwand bei der Metadatenerstellung. Der vorliegende Artikel untersucht die Schnittmengen und die Transformationsaufwände in einer Stichprobe aktiv genutzter Metadatenstandards.

Schlagworte: Open Access, Open Science, Publikationsworkflow, Metadaten, Standards, Interoperabilität

ABSTRACT

Open Access publications should be easily accessible to everyone, regardless of location and regardless of financial, legal or technical restrictions. To ensure this, not only language barriers but also metadata barriers must be overcome. Metadata are written in artificial languages which, like natural languages, are subject to constant development. While natural languages use interpreters to mediate between languages, metadata should be harmonised and exchanged by means of standards. Since standardisation efforts take place simultaneously in different domains, this necessarily leads to heterogeneity in metadata standards. Open Access publishing workflows are embedded in this heterogeneous metadata landscape. This means that different metadata sets have to be created depending on the publication route.

This leads to a transformational effort in metadata creation. This article examines the intersections and transformation efforts in a sample of actively used metadata standards.

Keywords: Open Access, Open Science, publication workflow, metadata, standards, interoperability

Hintergrund

Kommunikation ist eine der wichtigsten Triebkräfte der menschlichen Innovationskraft. Nach der Entwicklung der Sprachfähigkeit, die uns Menschen sogar physiologisch formte, trat mit der partiellen Sesshaftwerdung und der damit verbundenen Organisation des Menschen in größeren Gruppen die Notwendigkeit der Verschriftlichung von Sprache auf. Schrift und Schreiben wurden wohl mehrfach und in verschiedenen Weltgegenden erfunden. Seitdem wird nach Indizien gesucht, ob es eine frühere Ursprache gegeben haben könnte, die durch verschiedene Ursachen, seien es eine angenommene babylonische Sprachverwirrung oder doch eher unterschiedliche Entwicklungsbedingungen, in verschiedene Sprachen und Schriftsysteme unterschiedlichster Komplexität und Leistungsfähigkeit aufgespalten wurde. Wie man die Entwicklung der Sprachen auch betrachten möchte, Fakt ist, dass aktuell etwa 7000 verschiedene Sprachen identifiziert werden können[1]. Seit historisch kurzer Zeit kommen zu den vielen natürlichen Sprachen, die auf der Welt gesprochen werden, die artifiziellen hinzu. Dazu gehören vor allem Programmiersprachen, Metamarkupsprachen bzw. die mit letzteren konstruierten konkreten Markupsprachen. Artifizielle Sprachen erweitern damit die Artikulationsmöglichkeiten des Menschen in den von ihm kontrollierten technikverbundenen Bereichen. Eine der zu leistenden Aufgaben ist dabei die Formalisierung der dialogorientiert angelegten menschlichen Kommunikation mit vielen inhärenten Freiheitsgraden auf eine möglichst strukturierte technische Kommunikation mit eindeutigen Formulierungen, welche für die artifiziellen Sprachen notwendigerweise zu eingeschränkten Freiheitsgraden führen muss.

Mit der menschlichen Artikulationsfähigkeit erschien auch der Drang zur Beschreibung der Welt, vorrangig als Versuch, deren Komplexität zu begegnen. Lange war man der Meinung, dass sich die Welt als Ganzes beschreiben ließe, wende man nur das richtige Ordnungsprinzip an. Dem Ziel, dieses zu finden, widmeten sich Generationen von Philosophen. Letztlich setzte sich die Erkenntnis durch, dass das immer progressiver wachsende Menschheitswissen keinen Zeitpunkt kennt, an dem es als vollständig erkannt (bzw. beschrieben) definiert werden kann. Dieser recht pessimistischen Erkenntnis setzen die Wissenschaften neue Methoden, wie Data Mining und KI entgegen. Gleichsam synonym zur Entwicklung der natürlichen Sprachen, von einfachsten Artikulationen zu komplexesten Grammatiken, vollzieht sich jene der artifiziellen parallel zur Leistungsfähigkeit der Rechentechnik. Lange waren Speicherplatz rar, die Variablen der Programmiersprachen in ihrer Länge beschnitten und die Möglichkeiten von Markupsprachen beschränkt. SGML[2] wurde als mächtige Metamarkupsprache entwickelt, war aber auf der zeitgenössischen Technik nicht flächendeckend anwendbar. Erst das verschlankte XML konnte etwas später auf leistungsfähigeren Computern seinen

[1] vgl. Eberhard et al. 2023
[2] vgl. Library of Congress 2023

DOI: 10.33968/9783966270526-00

Durchbruch feiern. Gleiches gilt für die Entwicklung der konkreten Markupsprachen zur Beschreibung von Metadaten. Wurde 1994 im amerikanischen Dublin noch nach fünfzehn Termen gesucht[3], um möglichst jede Ressource in einem elektronischen Metadatensatz beschreiben zu können, sehen wir uns heute, bei nahezu unbeschränkten Kapazitäten zur Datenspeicherung und -übertragung, einer unüberschaubaren Vielzahl von Metadatenstandards gegenüber. Hier geschah wohl in historisch kurzer Zeit, wofür die natürlichen Sprachen Jahrtausende benötigten: mit wachsenden Ressourcen und in unterschiedlichen Domains bildeten sich lokale Sprachen heraus, welche in dem, was sie beschreiben, einen hohen semantischen Überdeckungsgrad, aber auch technische Unterschiede aufweisen.
Wie bei verschiedenen natürlichen Sprachen den Kommunikationsproblemen mit einem Dolmetscher abgeholfen werden kann, wird zwischen den artifiziellen Sprachen mittels Transformationsschnittstellen vermittelt. Aber, sowohl die Menschen als auch die Maschinen (bzw. die Ingenieure, welche diese Maschinen entwickeln), haben tief in ihrem Inneren die Sehnsucht nach dieser einen Sprache, welche die Kommunikation dramatisch vereinfachen, Missverständnisse vermeiden und den Aufwand reduzieren würde. Diese Sehnsucht kann ein Leitmotiv für neue Forschungsansätze sein, nicht nur für Geschichtswissenschaftler und Linguisten, sondern auch für Ingenieure.

In der wissenschaftlichen Publikationslandschaft und in den unterschiedlichsten Bereichen der gesamten Wertschöpfungskette von wissenschaftlichen Werken wurde unter den oben beschriebenen Prämissen bereits eine Vielzahl existierender Standards, Normen und Spezifikationen entwickelt und in die Praxis überführt. Mit der fortschreitenden Digitalisierung und Automatisierung von Publikationsprozessen ist die Verwendung von Standards unumgänglich geworden, um Prozesse ökonomisch und effizient zu gestalten. Auch der Austausch und die Weiterverarbeitung von Informationen über Systemgrenzen[4] hinweg macht die Verwendung von Standards notwendig. Diese Notwendigkeit hat zu den verschiedensten Standardisierungskampagnen geführt. So, wie sich die natürlichen Sprachen territorial entwickelt haben, geschah und geschieht dies auch für bestimmte abgegrenzte Bereiche, Domains, die sich hier nicht örtlich, aber thematisch abgrenzen. Und wie die natürlichen Sprachen die Kommunikation nur dahingehend regeln, was die regionalen Bedingungen verlangen, werden die artifiziellen Sprachen nur jene Elemente enthalten, die dem zu vereinheitlichenden Zweck dienen. Dies führt zu verschiedenen Arten von Standards: technischen Standards, bibliographischen Standards, (verlags-)herstellerspezifischen Standards, um nur beim Herstellungs- und Publikationsworkflow von wissenschaftlichen Monografien zu bleiben. Und wie bei den natürlichen Sprachen lässt es sich nicht vermeiden, dass die entstandenen Standards idealerweise aufeinander aufbauen bzw. sich ergänzen, oft aber einander durchdringen und nur in bestimmten Schnittmengen übereinstimmen. Zudem gilt für artifizielle wie für natürliche Sprachen, dass sie sich entwickeln und ihren „Duden“ regelmäßig überarbeiten. Notwendigerweise führt dies, wiederum wie bei den Sprachen und Alphabeten der Welt, zu einer heterogenen Landschaft an Standards. Besonders im wissenschaftlichen Publizieren verstärkt sich die Diversität an Standards, da die zu publizierenden Werke in den verschiedensten Wissenschaftsdisziplinen entstehen und sich daher inhaltlich und

[3] vgl. Dublin Core™ Metadata Initiative n. d.
[4] Die Grenzen bestehen hier zwischen den Bereichen der Stakeholder im Publikationsprozess. Dies sind die Kontributoren der Werke, die Hochschulbibliotheken und –verlage, die Dienstleister und Distributoren sowie andere Multiplikatoren.

DOI: 10.33968/9783966270526-00

strukturell stark voneinander unterscheiden. Dabei verfügt jede Wissenschaftsdisziplin über spezifische Besonderheiten und Anforderungen, sodass sich oftmals ein eigener Standard im jeweiligen Bereich entwickelt und etabliert hat. Für geisteswissenschaftliche Werke gilt beispielsweise TEI als Quasi-Standard. In den Naturwissenschaften, bei denen eine Aufbereitung und Veröffentlichung neuester Forschungsergebnisse in Form von Artikeln üblich ist, wird überwiegend JATS als Standard für die Inhaltserfassung verwendet. Die Komplexität wird zusätzlich gesteigert, da zwischen den verschiedenen Standards Schnittmengen bestehen, die beispielsweise dadurch entstehen, wenn sich mehrere Standards aus einem gemeinsamen "Mutter"-Standard herausgebildet haben und sich auf diesen beziehen. Um diese Komplexität in ihrer Dimension überhaupt quantifizieren zu können, ist es notwendig, einen Überblick über die angewandten Standards im wissenschaftlichen Publizieren[5] zu schaffen.

Analyse

Allein für den Bereich der Metadatenkommunikation wurden dabei 27 Standards identifiziert[6]. Die Zahl der im Feld aktiv angewandten Standards dürfte noch höher sein, da die Untersuchung einerseits keinen Anspruch auf Vollständigkeit erhebt und andererseits Metadatensammlungen, die noch keinen Standardisierungsprozess durchlaufen haben, nicht erfasst wurden. Zusätzlich wurden in die weitere Untersuchung auch häufig eingesetzte Struktur- und Dokumentenformate[7] in die Analyse einbezogen, da diese in ihren Strukturdefinitionen ebenfalls bereits Metadatenauszüge enthalten, vornehmlich von Metadaten, welche im eigentlichen Produktionsprozess entstehen und verändert werden. Die Untersuchung ergibt bereits einen ersten Blick auf die Komplexität des Metadatenuniversums für die Kommunikation von Open-Access-Publikationen. Es kann eingeschätzt werden, dass die Datengrundlage für die Betrachtung aller am Produktions- und Distributionsprozess beteiligten Metadaten gegen eine (vorläufige) Vollständigkeit tendiert[8].

Die in der Untersuchung identifizierten Metadatenstandards bilden die Grundlage für die Beschreibung der Interoperabilität der am OA-Publikationsprozess beteiligten Stakeholder. Diese soll beschrieben werden als die Fähigkeit, von am wissenschaftlichen Publikationsprozess beteiligten Menschen, Institutionen und Systemen, unabhängig voneinander zu interagieren und unter Verwendung von offenen Standards so miteinander zu kommunizieren, dass Daten und Informationen effizient, fehler- und verlustfrei ausgetauscht werden können[9]. Es wird von den Annahmen ausgegangen, dass die betrachteten Standards in ihrer Gesamtheit einerseits nahezu alle Metadaten in allen denkbaren Workflows beinhalten und andererseits diese Metadaten stochastisch über alle diese Standards verteilt sind.

[5] Böhm et al. 2021

[6] Zur Vervollständigung der Information: von den identifizierten Standards wurden 21 als Struktur- und Dokumentenformate klassifiziert, drei dem Bereich Validierung, zwei dem Bereich Formatierung und ebenfalls zwei dem Bereich Archivierung zugeordnet. Vier Standards aus dem Bereich Barrierefreiheit finden in OA-Publikationsworkflows Anwendung, sechs Standards wurden dem Bereich XML-Technologien zugeordnet und weitere zwölf sind Standards für die Identifikation von Dokumenten.

[7] Wie beispielsweise BITS, JATS, DocBook und DITA.

[8] Die aktuell diskutierten Möglichkeiten des Einsatzes von KI-Werkzeugen wie ChatGPT werden notwendigerweise neue Metadaten erzeugen, die künftig in den strukturierten Kommunikationen beachtet werden müssen. Auch darum ist es notwendig, bei den Entwicklungen von Kommunikationsstandards auf Offenheit und Erweiterbarkeit zu achten.

[9] Definition für die Verwendung innerhalb des Forschungsprojektes OA-STRUKTKOMM.

DOI: 10.33968/9783966270526-00

Weiterhin soll angenommen werden, dass sich die Standards mit ihren Metadatenbeziehungen in einer Matrix veranschaulichen lassen. Um diese Annahmen zu untersuchen, sollen zuerst aus den o.a. voneinander unabhängigen Standards in einem ersten Schritt alle Einzel-Metadatenelemente selektiert und hinsichtlich ihrer semantischen Übereinstimmung in einer offenen Liste sortiert werden. Dabei sollten folgende Bedingungen gelten:

$M = M = \sum_{x=0}^{n} S_x$	Matrix aller betrachteten Standards
$S = (S_{x0}, S_{X1}, \ldots, S_{xn})$	alle Datenelemente eines Standards
i	Index der Elemente eines Standards
j	Index des jeweiligen Standards im Vergleich (j <= m)
m	Anzahl der betrachteten Standards
$S_{jj} = \{0,1\}$	0 = keine Übereinstimmung; 1 = Übereinstimmung

0:1-Beziehungen werden nicht betrachtet[10]

Zuerst musste ein Standard ausgewählt werden, der eine Metadatenmenge anbietet, welche als Bezugs- bzw. Vergleichsmenge herangezogen werden kann. Dazu eignet sich grundsätzlich jeder Standard mit einem ausreichenden Umfang. Da mit dem medienneutralen sowie kosten- und personaleffizienten Publikationsworkflow für OA-Monografien[11] ein Workflowmodell vorliegt, welches den Anspruch erhebt, alle signifikanten Prozesse für die Herstellung und Distribution von OA-Monografien und damit notwendigerweise die damit verbundenen Daten zu beschreiben, ist es naheliegend, diesen Datensatz als Vergleichsbasis heranzuziehen. Dabei ist zu beachten, dass dieser gegenüber den meisten anderen als nicht standardisiert zu betrachten ist. Dieser Vergleichsdatensatz soll S_0 genannt werden.

S_0 Metadatensatz des OA-HVerlag-Workflowmodells

Der Datensatz des OA-HVerlag-Workflowmodells wird in sechs Datenkategorien unterteilt, welche jeweils eine verschiedene Menge an Datenelementen enthalten:

1. Daten zu Werk & Lizenzen (33 Elemente)
2. Daten zu Mitwirkenden (25 Elemente)
3. Daten zu Produktion (23 Elemente)
4. Daten zu Klassifikation (9 Elemente)
5. Daten zu Peer Review (8 Elemente)
6. Daten zu Barrierefreiheit (6 Elemente)

Aus der Menge der in der o.a. Untersuchung identifizierten Standards wurden sieben der oben erwähnten Kommunikationsstandards und fünf Standards, die buchartige Dokumente strukturieren, ausgewählt. Die Auswahl wurde durch die Vertrautheit der Analysten mit den gewählten Standards vorgenommen, um eine möglichst genaue Zuordnung der zu vergleichenden Metadatenelemente zu erreichen.

m = 12 (5 Standards, die den Struktur- und Dokumentenformaten zuzurechnen sind und 7 Standards, die zu den Kommunikationsstandards gehören)

Für diese wurden alle enthaltenen Metadatenelemente mit den Elementen im Referenzstandard

[10] Jeder Standard, welcher mit dem Metadatensatz des OA-HVerlag-Workflowmodells verglichen wird, kann mehr oder weniger Metadatenelemente enthalten. Enthält dieser mehr Metadatenelemente, werden diejenigen, welche kein Pendant in einem der Metadatenelemente des Metadatensatzes des OA-HVerlag-Workflowmodells haben, nicht betrachtet. Insofern ist die entstehende Matrix nur für die Betrachtung der Kommunikationsbeziehungen bei Verwendung des o.a. Workflowmodells als vollständig anzusehen. Für die Wahl eines anderen Standards für S0 würden abweichende Ergebnisse zu erwarten sein.

[11] https://oa-hochschulverlag.htwk-leipzig.de/forschungsprojekt

DOI: 10.33968/9783966270526-00

S_0 verglichen. Wurde im ersten Durchgang eine semantische Übereinstimmung festgestellt, wurde die entsprechende Stelle in der Matrix mit einer „1“ belegt.

$S_{ij} = 1$ wenn $S_{ij} \in S_0 \cap \in S_j$; sonst 0

Die Ergebnismatrix weist folgende, in Tab. 1 dargestellte, Struktur auf (Ausschnitt).

Tab. 1 Analyse der Überscheidungen der Metadatenelemente auf semantischer Ebene

	Struktur- und Dokumentenformate					Komm. -Standards	
	PS1	**PS2**	**PS3**	**PS4**	**PS5**	**MS1**	**...**
Daten Kategorie 1							
1.01	1	1	1	1	1	1	...
1.02	1	1	1	1	1	1	...
1.03	1	1	1	1	1	1	...
1.04	1	1	1	1	1	1	...
1.05	0	1	1	1	0	1	...
1.06	1	1	1	1	1	1	...
1.07	1	1	1	1	1	1	...
1.08	1	1	1	1	1	1	...
1.09	1	1	0	1	1	0	...
1.10	0	1	0	1	1	0	...
1.11	0	1	0	1	1	0	...
...	...	...	...	...	...	...	...
Daten Kategorie 2							
2.01	1	1	1	1	1	1	...
2.02	1	1	1	1	1	1	...
2.03	1	1	1	1	1	1	...
2.04	0	1	1	1	1	0	...
...	...	...	...	...	...	...	...

Damit wurde im ersten Analysedurchgang eine Matrix erstellt, welche die semantischen Überschneidungen zwischen den betrachteten Standards darstellt. Diese Überschneidungen waren erwartbar und sind ursächlich aus der Historie der weitestgehend unabhängigen Entwicklungen der betrachteten Standards abzuleiten. Verschiedene Interessengruppen haben mit unterschiedlichen Foki die in ihrer jeweiligen betrachteten Domäne notwendigen Metadatensätze aufgestellt und standardisiert.

Darüber hinaus ist zudem einerseits von einer evolutionären Entwicklung auszugehen, d.h., die Metadatensätze wurden und werden in aufeinanderfolgenden Versionen angereichert[12] und andererseits können bereits entstandene Metadatensätze in andere Strukturen inkludiert oder von diesen adressiert werden. Damit musste notwendigerweise eine ebenso heterogene Landschaft entstehen, wie sie schon für die Publikationslandschaft festgestellt werden konnte.

Auswertung

Bei einer Auswertung der Matrix können folgende Feststellungen getroffen werden: 7 der betrachteten Elemente aus S_0 finden in allen 12 betrachteten Standards eine semantische Entsprechung. Dies entspricht 7,07%. Beispielhaft soll

[12] sowie ggf. ab einem gewissen Komplexitätsgrad bereinigt und verschlankt, in seltenen Fällen auch in eine andere technische Struktur überführt

DOI: 10.33968/9783966270526-00

hier das semantische Element „Titel“ betrachtet werden, das erwartungsgemäß in allen Standards in seiner semantischen Bedeutung als „kennzeichnender Name eines Buches, einer Schrift, eines Kunstwerks o. Ä.“[13] enthalten ist. Die 12 Metadatenelemente, für welche die Bedeutung des Metadatums 1.01[14] „Titel“ im Datensatz S_0 semantisch zugeordnet werden können, sind Tabelle 2 zu entnehmen.

Tab. 2 Metadatenelemente zum Element „Titel“

S_0	1.01 Werk-Titel
PS_1	<titel>
PS_2	<book-title>
PS_3	<title>
PS_4	<booktitle>
PS_5	<title>
MS_1	<b203>
MS_2	<datafield tag=“245“>
MS_3	<dc: title>
MS_4	<title>
MS_5	<title>
MS_6	<title>
MS_7	<dc: title>

Auf der anderen Seite der Skala finden sich 8 Elemente in S_0, für die in keinem der untersuchten Standards ein adäquates Element gefunden werden konnte. Dazu gehören zum Beispiel Angaben zur Ausstattung einer gedruckten Monografie, wie Farbigkeit, Bindungsart und Veredelungsoptionen.

Tab. 3 Semantische Zuordnung der Datenelemente aus S_0 zu den betrachteten Standards in Zahlen

Anzahl Auftreten	Elemente	Prozent
12	7	7,07%
11	10	10,10%
10	4	4,04%
9	6	6,06%
8	4	4,04%
7	10	10,10%
6	5	5,05%
5	7	7,07%
4	7	7,07%
3	14	14,14%
2	11	11,11%
1	6	6,06%
0	8	8,08%

Interessant ist in diesem Zusammenhang die Betrachtung der Überschneidung der Metada
Interessant ist in diesem Zusammenhang die
Betrachtung der Überschneidung der Metadatenensätze der einzelnen betrachteten Standards mit den jeweiligen Datengruppen. Diese sind in der nachfolgenden Tabelle dargestellt:

Tab. 4 Verteilung der Überschneidungen der einzelnen betrachteten Standards mit der Liste, aufgeschlüsselt nach Datengruppen

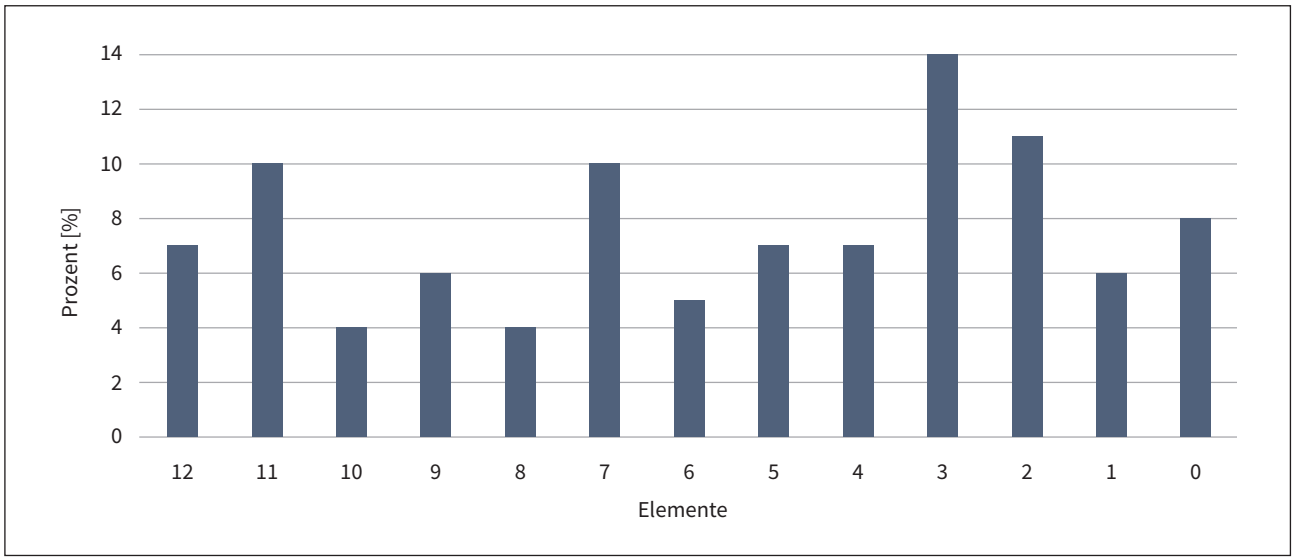
	PS_1	PS_2	PS_3	PS_4	PS_5	MS_1	MS_3
1.	48%	73%	55%	64%	58%	79%	18%
2.	36%	84%	60%	72%	88%	40%	20%
3.	15%	25%	10%	5%	20%	65%	25%
4.	0%	33%	22%	22%	22%	100%	22%
5.	0%	0%	0%	63%	75%	0%	0%
6.	100%	0%	0%	0%	0%	67%	0%

[13] lt. Duden 2023
[14] Daten Kategorie 1 (Daten zu Werk & Lizenzen), Element 01 (Titel)

DOI: 10.33968/9783966270526-00

Die Aufstellung bestätigt die oben angeführte Vermutung, dass die einzelnen Standards historisch mit speziellen Foki entwickelt wurden. Es ergeben sich zumeist sehr hohe oder sehr geringe prozentuale Übereinstimmung. D.h., lagen die Foki der verglichenen Metadatensätze auf gleichen Fragestellungen, ergibt sich eine hohe Übereinstimmung. Sichtbar wird dies zum Beispiel am Fokus auf Barrierefreiheit, der eine vollständige semantische Übereinstimmung der diesbezüglichen Metadaten in S_0 und dem Metadatensatz PS1 ergab. Mit anderen Metadatensätzen, die aus verschiedensten Gründen Barrierefreiheit (noch) nicht adressieren, finden sich keine korrespondierenden Metadaten.

Abb. 1 Semantische Zuordnung der Datenelemente aus S_0 zu den betrachteten Standards



Die semantische Übereinstimmung zwischen den Datentermen der verglichenen Standards geben allerdings nur einen vorläufigen Hinweis auf die Komplexität des Datenaustausch zwischen diesen. Zwischen Datenstandards, und dies betrifft auch die betrachteten Metadatenstandards, treten gemeinhin Konflikte aufgrund technischer, semantischer, struktureller, syntaktischer Ungleichheiten sowie Variationen in den Datenmodellen auf[15], welche bei der Analyse der Metadatensätze auch festgestellt wurden (ebd.). Die Datenmodellheterogenität beschreibt den Fall, wenn Abweichungen zwischen den Schemata oder Sprachen festgestellt werden, in denen die Daten hinterlegt sind. Diese liegt in den verglichenen Standards vor. Als syntaktische Heterogenität wird die unterschiedliche Darstellungsweise von Informationen bezeichnet. Bei der semantischen Heterogenität werden Datenmodelle auf verschiedene Art gedeutet und verwendet (ebd., S. 73). Die strukturelle Heterogenität[16] letztendlich bezieht sich auf Unterschiede in der Struktur, in der die Daten erfasst werden[17]

[15] vgl. Leser, Naumann 2007

[16] Welche die schematische Heterogenität einbeziehen soll, welche auftritt, wenn das gleiche Datenmodell verwendet wird, die Daten allerdings unterschiedlich angelegt werden (beispielsweise einmal als Attribut, einmal als Element innerhalb von BITS).

[17] Unterschiede in der Granularität von Metadatenstandards existieren beispielsweise bei der Auszeichnung von Namen, die in einem Standard innerhalb eines Elementes erfasst werden (z. B. in Dublin Core mit <dc:creator>), in einem anderen jedoch mit mehreren Elementen unterteilt werden (z. B. in BITS mit <surname> und <given-names> innerhalb von <name>). Durch diese Granularität wird auch die Hierarchie beeinflusst: So befinden sich Namenselemente bei BITS auf der sechsten Ebene (book/book-meta/contrib-group/name/*), bei Dublin Core auf der zweiten (metadata/dc:creator).

DOI: 10.33968/9783966270526-00

und kann nochmals untergliedert werden in Konflikte der Granularität, der Hierarchie, dem erlaubten minimalen/maximalen Vorkommen (Occurence) und der (nicht) verpflichtenden Angabe (Obligation), den Eigenschaften (Properties) sowie der verwendeten Schemata. Aus dieser, in Kombination mit der semantischen Heterogenität, leitet sich allgemein der Transferaufwand bei der Übertragung zwischen verschiedenen Datenstrukturen ab[18].

Für die oben beschriebene Matrix wurden mit m=7 Matrix[19] erneut alle Zellen S_{ij}=1 auf die beschriebenen Heterogenitäten untersucht. Dabei wurden die Heterogenitäten zwischen den jeweiligen Metadaten-Elementen folgendermaßen bewertet: Den Wert „3“ erhielten alle Zellen, für welche keine semantische und technische Heterogenität festgestellt werden konnte (1:1-Übertragung möglich), den Wert „2“ alle Zellen, welche semantisch übereinstimmen, aber einen Strukturunterschied aufweisen (Übertragung mit kleinem Aufwand möglich), den Wert „1“ alle Zellen, welche eine semantische Übereinstimmung aufweisen, aber strukturell und datentechnisch differieren[20]. Zellen mit dem Wert „0“ (keine Metadatenbeziehung) wurden nicht betrachtet. Die Ergebnisse werden in der folgenden Matrix dargestellt (Ausschnitt):

Tab. 5 Analyse der Komplexität der Transferleistung zwischen verschiedenen Metadatenstandards

	Struktur- und Dokumentenformate					Komm. -Standards	
	PS_1	PS_2	PS_3	PS_4	PS_5	MS_1	...
Daten Kategorie 1							
1.01	3	3	3	2	3	2	...
1.02	3	3	2	2	3	2	...
1.03	3	3	2	2	3	2	...
1.04	3	3	2	2	3	2	...
1.05	0	3	3	1	0	2	...
1.06	3	3	3	3	3	2	...
1.07	1	1	1	1	1	1	...
1.08	3	3	3	3	3	2	...
1.09	3	3	0	3	2	0	...
1.10	0	3	0	3	2	0	...
1.11	0	2	0	3	2	0	...
...	...	...	...	...	...	...	...
Daten Kategorie 2							
2.01	3	3	3	3	3	3	...
2.02	3	3	3	3	3	3	...
2.03	3	3	3	3	3	3	...
2.04	0	3	3	3	3	0	...
...	...	...	...	...	...	...	...

[18] Die Angaben zur Occurence und Obligation variieren in den Standards stark – BITS gibt hier nur wenige Einschränkungen vor, Crossref hingegen fordert genau 1–10 Publikationsdaten. Diese Unterschiede im Verpflichtungsgrad sowie dem Vorkommen können dazu führen, dass teilweise keine Daten vorliegen, diese in anderen Standards jedoch benötigt werden. Entscheidend für die Kompatibilität von Standards sind zudem die Properties. Es kann vorkommen, dass bestimmte Metadaten in einem Standard hinterlegt sind, es allerdings kein Äquivalent in dem anderen Standard gibt. Dieser Fall wird auch zur strukturellen Heterogenität gezählt.

[19] Auf Grund des Aufwandes wurde nur ein Ausschnitt der zwölf oben analysierten Standards hinsichtlich der Heterogenität ausgewertet.

[20] Von technischen Heterogenitäten ist die Rede, wenn sich die Zugriffsschnittstellen unterscheiden, beispielsweise XPath oder MySQL für die Verschlüsselung verwendet werden.

DOI: 10.33968/9783966270526-00

Betrachtet man die Aufwände bei der Transferleistung bei der Konvertierung der einzelnen betrachteten Standards gegenüber der Liste S_0, lassen sich zwei Fälle unterscheiden. Für den Fall 1 werden auch alle möglichen Transferaufwände zwischen S_0 und den betrachteten Standards bewertet, also auch die Fälle, für die keine semantische Übereinstimmung gefunden werden konnte ($S_{ij} = 0$). Bei der Konvertierung dieser Metadatenterme müssen Regeln angewendet werden, die es ermöglichen, entweder einen Datenterm der sendenden Datenstruktur zu behandeln, für den kein Pendant in der empfangenden Struktur vorhanden ist (Datum kann nicht übertragen werden) bzw. eine der nicht betrachteten 0:1-Beziehungen. Fall 2 betrachtet nur Fälle $S_0 > 0$, d.h. der Metadatentransfer kann erfolgen.

Für Fall 1 könnten 39,04% mit geringem Aufwand, 2,83% mit mittleren und 1,13 % der Daten von S_0 mit hohem Aufwand in die anderen ausgewählten Standards transferiert werden. Für 57,00% der Metadatenterme wäre dies nicht möglich. Für F2 betragen die Transferaufwände für 3: 90,79%, für 2: 6,58% und für 1: 2,63%. D.h., die Elemente, für die eine semantische Beziehung zu S_0 hergestellt werden kann, können zu 90% mit keinem bzw. wenigem Aufwand in den jeweiligen Zielstandard übertragen werden. Für ca. 10% muss von erhöhtem Konvertierungsaufwand ausgegangen werden. Diese Angaben müssen als Tendenz aufgefasst werden, da die Bewertung der Transferaufwände zwischen S_0 und den einzelnen Elementen der Metadatenstandards hochgradig von den individuellen Kenntnissen der Bewertenden abhängig ist und durch verschiedene Fachexperten auch anders eingeschätzt werden können.

Zusammenfassung

Ziel der Untersuchung war es, ausgehend von den Annahmen, dass einerseits in den im Feld vorhandenen Standards im Umfeld der Open-Access-Publikation, speziell von Monografien, einerseits eine weitestgehende Vollständigkeit der notwendigen Metadaten vorliegt und dass für die angenommene heterogene Verteilung der Metadaten in den einzelnen Standards historische bzw. domainspezifische Gründe vorliegen.

Dazu wurden zwölf Metadatensätze in Bezug zum Metadatensatz des OA-HVerlag-Workflowmodells gesetzt und ausgewertet. Dabei wurde festgestellt, dass bis auf acht, meist ausstattungsbezogene Metadaten, für alle weiteren in wenigstens einem der untersuchten Standards eine semantische Entsprechung gefunden werden kann. 7 Elemente (7,07%) der Elemente aus S_0 kommen in allen untersuchten Standards vor. Schlüsselt man die semantischen Überschneidungen nach Datenkategorien auf, stellt man einerseits fest, dass in den Datenkategorien Werks- und Autorendaten eine durchschnittlich große Schnittmenge zwischen den Standards besteht. Dies war so zu erwarten. Zu spezielleren Datengruppen, wie Peer-Review-Stammdaten und Barrierefreiheits-Stammdaten sind entweder sehr kleine oder sehr hohe ten. Dies ist als Indiz für die Domainabhängigkeit der verglichenen Standards zu werten.

Weiterhin wurde für m=7 der Transformationsaufwand der Elemente bei der Übertragung aus S_0 in die anderen Standards ermittelt. Dabei wurde festgestellt, dass 43,00% der Metadatenterme in S_0 in einen der betrachteten Metadatensätze übertragen werden können und für 90,79% dieser übertragbaren Elemente der Transferaufwand als gering eingeschätzt werden kann, d.h., die Informationen können direkt oder mit nur geringem Aufwand

DOI: 10.33968/9783966270526-00

transformiert werden. Obwohl diese Aussage für den Transfer von S_0 in alle untersuchten Standards getroffen wird, kann der Aufwand in seiner Dimension auch für die Übertragung der semantisch übereinstimmenden Elemente zwischen allen untersuchten Standards angenommen werden. D.h., wenn es übertragbare Elemente gibt, können diese zum überwiegenden Teil problemlos übertragen werden.

Fazit

Die Untersuchung zeigt, dass im Moment der allumfassende Metadatenstandard nicht existiert. Die angenommene Heterogenität der betrachteten Metadatenstandards kann vor allem durch die Angaben in Tabelle 3 bestätigt werden. Damit ist der aktuelle Zustand der Kommunikation über Metadaten in den OA-Publikationsprozessen im Feld hinreichend beschrieben. Wie mit dieser Situation in Zeiten zunehmender, auch datentechnischer, Verschlankung der Publikationsprozesse umzugehen ist, soll an anderer Stelle untersucht werden.

Literatur

Böhm et al. (2021): Überblick über offene Standards im wissenschaftlichen Publizieren/ Overview of open standards in scientific publishing, Science Open, DOI 10.14293/S2199-1006.1.SOR-.PPNKUIH.v1

Dublin Core™ Metadata Initiative (o. J.): DCMI History. In: Dublin Core, https://www.dublincore.org/about/history/, Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0), Letzter Zugriff: 05.06.2023

Duden (2023): Titel, In: Duden. https:/www.duden.de/rechtschreibung/Titel, Letzter Zugriff: 27.03.2023

Eberhard/Simons/Fennig, C. D. (Hrsg.) (2023): How many languages are there in the world? In: Ethnologue: Languages of the World, Twenty-sixth edition, Dallas, Texas: SIL International, https://www.ethnologue.com/insights/how-many-languages/, Letzter Zugriff: 05.06.2023

Leser/Naumann (2007): Informationsintegration. Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen, Heidelberg: dpunkt.verlag GmbH, ISBN 3-8986-4400-6, S. 60ff.

Library of Congress (26.04.2023): Standard Generalized Markup Language (SGML). ISO 8879:1986, In: Sustainability of Digital Formats: Planning for Library of Congress Collections, https://status.iso.org/incidents/dldzn680scx2, Letzter Zugriff: 05.06.2023

DOI: 10.33968/9783966270526-00

OA HVerlag
Open Access Hochschulverlag