Bewertungskriterien für die Qualität von angelieferten Datensätzen in Mo|Re data

1. Hintergrund

1.1 Qualitätsmerkmale nach Wang & Strong (1996)

2. Bewertungskriterien für die Qualität von angelieferten Datensätzen in Mo|Re data

3. Quellen und weiterführende Literatur

1. **Hintergrund**

Der wohl größte Einflussfaktor auf den Wert und wissenschaftlichen Nutzen von Daten im Projektkontext und für die Nachnutzung ist die Datenqualität (vgl. Bertelmann et al., 2014). Naumann (2007) definiert Datenqualität als eine Menge von Qualitätsmerkmalen. Die Qualität von Daten, auch “Informationsqualität”, wird oft als die Eignung der Daten für die jeweilige datenverarbeitende Anwendung definiert. Daten von schlechter Qualität enthalten Datenfehler, Dubletten, fehlende Werte, falsche Formatierungen, Widersprüche, etc. (vgl. Naumann, 2007). Die Auswahl der relevanten Merkmale für die Datenqualität und die genaue Definition der Merkmale bleiben den Experten aus den unterschiedlichen Forschungsbereichen vorbehalten (vgl. Naumann, 2007). Bertelmann et al. (2014) stellen ausgewählte Faktoren vor, welche die Datenqualität beeinflussen:

1. **Objektivität:** sind die Daten genau, konsistent und verlässlich?
2. **Integrität:** wurden alle vorgenommenen Änderungen dokumentiert?
3. **Verständlichkeit:** ist transparent und nachvollziehbar wie die Daten entstanden sind?
4. **Formate:** Das Format beeinflusst wesentlich, wie mit Daten gearbeitet werden kann (Auswertung, Weiterverarbeitung). Handelt es sich bei einem verwendeten Format um einen Standard innerhalb der Community?
5. **Dokumentation:** sind ausreichend Kontextinformationen zum Forschungsprozess verfügbar? Ist die vorhandene oder geplante Dokumentation dazu geeignet, Datenerhebung und Analyse bzw. den gesamten Forschungsprozess transparent und nachvollziehbar zu machen.

1.1.Qualitätsmerkmale nach Wang & Strong (1996)

Datenqualität kann sich nicht nur auf den einzelnen Datensatz beziehen, sondern auch abstraktere Merkmale beinhalten, welche sich auf die ganze Datenmenge beziehen, z.B. die Verständlichkeit einer Datenmenge, deren Vollständigkeit oder auch die Reputation der Datenquelle. Naumann (2007) nennt als meist zitierte Aufstellung solcher Informationsqualitätsmerkmale die Aufstellung von Wang & Strong (1996). Sie befragten Datenkonsumenten in größeren Unternehmen und filterten aus 179 Merkmalen die in Tabelle 1 genannten 15 Qualitätsmerkmale heraus.

Tab.1: Qualitätsmerkmale nach Wang & Strong 1996 (entnommen aus Naumann, 2007, S.27)

Intrinsische Qualitität	Glaubhaftigkeit
	Genauigkeit
	Objektivität
	Reputation
Kontextuelle Datenqualität	Mehrwert
	Relevanz
	Zeitnähe
	Vollständigkeit
	Datenmenge
Repräsentationelle Datenqualität	Interpretierbarkeit
	Verständlichkeit
	Konsistenz der Darstellung
	Knappheit der Darstellung
Zugriffsqualität	Verfügbarkeit
	Zugriffsicherheit

Die Beurteilung der Datenqualität in Mo|Re data erfolgt in Orientierung an den Qualitätsmerkmalen nach Wang & Strong (1996).

2. Bewertungskriterien für die Qualität von angelieferten Datensätzen in Mo|Re data

Für die Bewertung der Qualität der angelieferten Daten wurde ein eigenes „Bewertungsschema“ entwickelt, welches unterschiedliche Ebenen von Qualitätskriterien berücksichtigt u.a. die von Wang & Strong (1996) dargestellten Kriterien, aber auch statistische Kennwerte (Verteilungen, deskriptive Kenngrößen, statistische Ausreißer-Identifikation, fehlende Werte und weitere Datenfehler auf der Schema- und Datenebene). Im Folgenden wird die Datenqualitätsprüfung im Rahmen des *Mo|Re data* Projektes beschrieben. Abbildung 2 verdeutlicht den Ablauf der Qualitätsprüfung.

1.) Der Datensatz wird vom Creator anonymisiert in die Datenbank hochgeladen

2.) Die Daten werden vom *Mo|Re data* Team und weiteren Experten auf ihre Qualität hinüberprüft. Eventuelle Überarbeitungsschritte zur Optimierung der Datenqualität sind notwendig, ggf. findet eine Rücksprachen mit dem Creator finden.

3.) Der Datensatz wird angenommen, bekommt eine DOI und wird in *Mo|Re data* abgelegt.

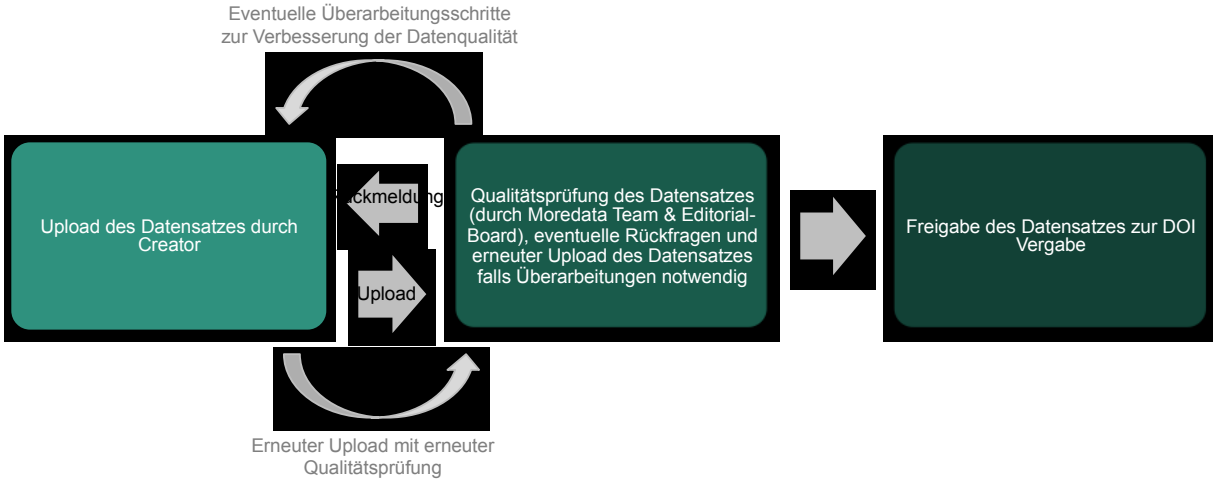

Abb.2: Ablauf der Qualitätsprüfung in Mo|Re data

In diesem Papier wird der Prozess der Qualitätsprüfung (Punkt 2) genauer dargestellt. Die Mo|Re data Qualitätsprüfung enthält 4 unterschiedliche Qualitätskategorien:

1. Durchführungsqualität
2. Datenqualität
3. Dokumentationsqualität
4. Expertengutachten

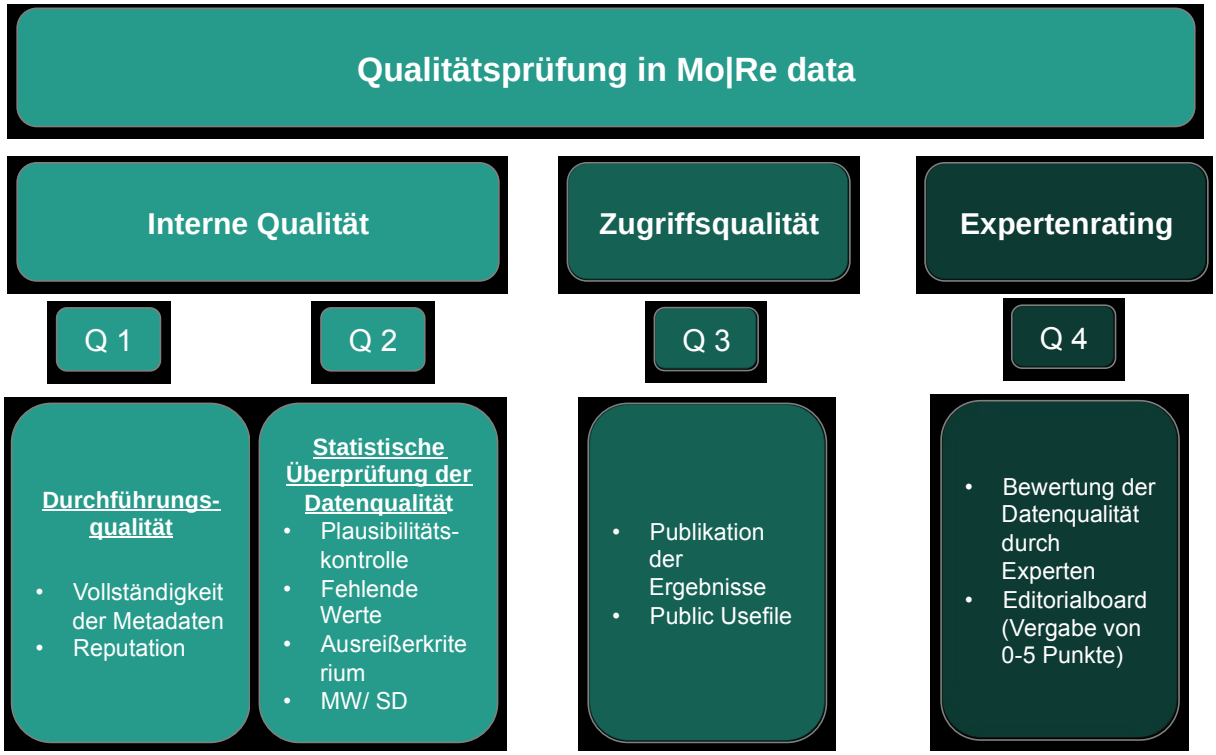

Abb. 3: Übersicht der Qualitätskriterien in Mo|Re data

Die Qualität der Daten an und für sich, im Folgenden auch als interne Qualität[1] bezeichnet, wird im großen Maße davon bestimmt, ob ein Datensatz fehlerfrei ist. Rahm & Do (2000) erstellten beispielsweise eine Klassifikation von Datenfehlern, in der sie unterscheiden, ob der Fehler auf Schemaebene oder auf Datenebene angesiedelt ist, und ob der Fehler bereits in einer einzigen Datensammlung besteht oder erst durch die Integration mehrerer Datensammlungen entsteht. Die Überprüfung der Datenqualität in Mo|Re data soll unter anderem dazu dienen diese Datenfehler zu erkennen und entsprechende Handlungen abzuleiten (Korrektur der Fehler oder Abweisung des Datensatzes).

Qualitätskriterium 1 und 2 werden der internen Qualität zugeordnet.

Die Durchführungsqualität wird aufgrund der vom Datenlieferant/ Creator angegebenen Metadaten überprüft. Diese müssen vollständig sein und möglichst detaillierte Informationen über die Studiendurchführung enthalten (z.B. durchführende Institution, Stichprobenbeschreibung, etc.). Bei fehlenden Informationen besteht die Möglichkeit mit dem Datenlieferant/ Creator Rücksprache zu halten, um die benötigten Informationen zu vervollständigen.

1 Die interne Qualität entspricht der intrinsischen Qualität bei Wang und Strong (1996)

Die Datenqualität an und für sich wird anhand statistischer Überprüfungen vorgenommen. Für Mo|Re data werden das Minimum, das Maximum, der Mittelwert, die Standardabweichung, Schiefe und Kurtosis, Ausreißer-Kriterien und Boxplot analysiert. Der Datensatz wird zunächst auf fehlende und unplausible Werte überprüft. Hierbei müssen fehlende Werte als Missings definiert sein und nicht als numerische Werte. Die der Ausreißer-Kontrolle wird anhand definierter Plausibilitätsgrenzen vorgenommen. Für jede motorische Testvariable bestehen aus wissenschaftlichen Studien abgeleitete Plausibilitätsgrenzen, welche nicht überschritten werden dürfen (sind Minimum und Maximum plausible Werte?). Bei normal verteilten Daten liegt das Ausreißer-Kriterium bei Werten, die größer (kleiner) als Mittelwert + (-) 3 Standardabweichungen sind, dies bedeutet, dass die Wahrscheinlichkeit, dass ein Wert so groß ist 0.3% beträgt. Es muss inhaltlich entschieden werden, ob es sich wirklich um einen „Ausreißer“ handelt, d.h. stammt der Wert aus einer anderen Population als die anderen Werte? Dabei wird zum Beispiel auch überprüft wie groß die Differenz-Intervalle zwischen den letzten drei Mittelwerten und den ersten drei Mittelwerten sind.
Die Qualitätskriterien 1-3 werden von Mitarbeitern des Mo|Re data Projektes manuell geprüft (Folgenantrag Verbesserung der Automatisierung der Datenprüfung). Die Prüfung der Datenqualität Q2 wird mit statistischen Auswertungstools und Programmen (z.B. SPSS) vorgenommen. Die folgende Übersichtstabelle (Tabelle 2) verdeutlicht die unterschiedlichen Qualitätskriterien und die Bewertungsskala.

Tab. 2: Qualitätsbewertungsschema Mo|Re data

Bewertungsraster der Qualität der angelieferten Datensätze in MoRe data

	Qualitätskategorie	Beurteilungskriterium	Antwort/ Kategorie	Punkte	max. Summenscores
Interne Qualität	Q1: Durchführungsqualität (Studienqualität)	Vollständigkeit der Metadaten	ja	1	
			nein	0	
		Reutation: Gehören die Daten zu einer Studie welche an einer wissenschaftliche Institution durchgeführt wurde?	ja	2	
			nein	0	
					3
	Q2: statistische Überprüfung der Datenqualität (Plausibilitätskontrolle für Rohdaten)	Minimum	keine Auffälligkeiten	1	Wenn nicht alle 9 Punkte erreicht werden wird Kontakt zum Datenerheber"Creator" hergestellt Wenn Fehlende Werte nicht als Missings definiert, muss dies noch vorgenommen werden
			Auffälligkeiten vorhanden	0	
		Maximum	keine Auffälligkeiten	1	
			Auffälligkeiten vorhanden	0	
		Mittelwert	keine Auffälligkeiten	1	
			Auffälligkeiten vorhanden	0	
		Standardabweichung	keine Auffälligkeiten	1	
			Auffälligkeiten vorhanden	0	
		Schiefe	keine Auffälligkeiten	1	
			Auffälligkeiten vorhanden	0	
		Kurtosis	keine Auffälligkeiten	1	
			Auffälligkeiten vorhanden	0	
		Ausreißerkriterium: % Werte > bzw. < 3 Stabw und Differenz-Intervall zwischen letzten drei und ersten drei MWs	keine Ausreißer	1	
			Ausreißer vorhanden	0	
		Boxplot	keine Auffälligkeiten	1	
			Auffälligkeiten vorhanden	0	
		fehlende Werte	keine fehlenden Werte	1	
			fehlende Werte	0	
					8 (8 von 8 Pflicht)
Zugriffsqualität	Q3: Dokumentationsqualität/ Zugriffsqualität	Publikation der Ergebnisse vorhanden	ja und die Beschreibung der eingesetzten Messinstrumente, Hilfsmittel und verwendeten Methoden, Teilnehmer-Anzahl, genaue und übersichtliche Darstellung der Zwischen- und Endergebniss	3	
			nein, dennoch genaue Dokumentation der Datenerfassung (Messinstrumente, verwendete Methoden) Homepage oder nach Rücksprache mit Datenerhebern/Creator	2	
			nein	0	
		"Datenpublikation (Public Usefile)"	ja	1	
			nein	0	
					4
Experten-einschätzung	Q4:Punkte Score des Editorial Boards (Experten-Score)	Beurteilung der Datenqualität durch Experten (hier ist das Ziel Experten von anderen Unis als Reviewer zu rekrutieren, ähnlich wie Review-Prozess bei Zeitschriften) * *im Aufbau	sehr gute Qualität für Normwertgenerierung geeignet	5	Es müssen mindestens 2 Punkte erreicht werdenfür eine Aufnahme in die Datenbank
			gute Qualität	4	
			befriedigende Qualität	3	
			ausreichende Qualität	2	
			mangelhafte Qualität	1	
			ungenügende Qualität	0	
					5 (2 von 5 Pflicht)

Auswertungsregeln

	maximale Punktzahl:	21	zur Normwertegenerierung geeignet		
	Minimale Punktzahl:	0	wird nicht in Datenbank aufgenommen		
	10*-21 Punkte		Aufnahme in Datenbank		
	2-11* Punkte		vorerst keine Aufnahme in Datenbank, Kontaktaufnahme mit Datenbereitsteller, eventuelle Verbesserung der Datenqualität möglich?		
	*11 Punkte beziehen sich auf die Summe aus den Katgeorie Q2 und Q4				

Um den Standard der Mo|Re data Qualitätsprüfung zu erhöhen, ist es vorgesehen ein „Editorial-Board“ zu etablieren, welches zusätzlich zu der internen Überprüfung eine externe Qualitätsprüfung, ein Kreuzgutachten, übernimmt. Ähnlich wie bei einem Peer-Review von wissenschaftlichen Publikationen sollen unabhängige Gutachter aus dem Fachgebiet Sportwissenschaften mit dem Schwerpunkt Motorikforschung die Eignung der eingehenden Daten beurteilen. Die Experten erhalten die unten aufgeführten Beurteilungsbögen.

Bogen 1 orientiert sich an den Qualitätskriterien nach Wang & Strong (1996) und Naumann (2007). Der Gutachter erhält zuvor die Metadaten und die Auswertung der statistischen Qualitätsprüfung (Bogen 2). Diese Informationen sollen den Gutachter beim Ausfüllen des Qualitätsbogen 1 unterstützen. Der Gutachter hat die Möglichkeit für jedes aufgeführte Qualitätskriterium 0-5 Punkte zu vergeben, der Mittelwert ergibt die Gesamtpunktzahl für das Expertenurteil. Bis zur Etablierung des Editorial-Boards wird die Qualitätsprüfung von Motorik-Experten des Instituts für Sport und Sportwissenschaft des Karlsruher Institutut für Technologie durchgeführt.

Bewertung der Qualität durch das Editorial Board in Anlehnung an Wang und Strong (1996)/ und Naumann (2007)								
		0	1	2	3	4	5	**Punktzahl**
Interne Qualitität	Glaubhaftigkeit							
	Genauigkeit							
	Objektivität							
	Reputation							
Kontextuelle Datenqualität	Mehrwert							
	Relevanz							
	Zeitnähe							
	Vollständigkeit							
	Datenmenge							
Aufnahme in die Datenbank wird empfohlen	**ja**						**Gesamt-punktzahl**	
	nein							
	nach Überarbeitung/Rücksprache							

Abb.4: Qualitätsbewertungsbogen für externe Gutachter (Bogen 1)

Interner Qualitätscheck zur Aufnahme in die Datenbank

Ausgabe: DIN A4 Seite mit diesen Infos zu jeder eingegebenen Variablen

Kriterien		Okay?	
Minimum	3		
Maximum	6		
Mittelwert	4,090909		
Standardabweichung	0,831209		
Schiefe	1,086981	1	automatisch
Kurtosis	2,285319	1	automatisch
% Werte > bzw. < 3 Stabw			automatisch

Boxplot für Ausreißerdiagnose

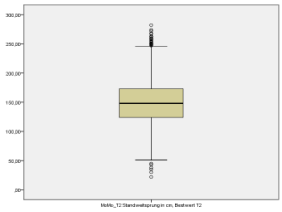

Abb.5: Ausdruck der interne, statistischen Datenqualitätsprüfung (Bogen 2)

Nach dem Upload des Datensatzes bekommt der Datenlieferant/ Creator eine Rückmeldung über den Upload des Datensatzes und den Vorgang des Prüfung/Review-Verfahren. Sobald die Qualitätsprüfung abgeschlossen ist, bekommt der Datenlieferant/ Creator entweder die Mitteilung, dass sein Datensatz in die Datenbank aufgenommen und eine DOI erhalten wird, oder aber spezielle Rückfragen notwendig sind um die Qualitätsprüfung abzuschließen. In manchen Fällen kann ein Datensatz auch wegen ungenügender Qualität vollständig zurückgewiesen werden. Wie die Abbildung 2 bereits veranschaulichte erfolgt die DOI-Vergabe für den jeweiligen Datensatz erst nach der Qualitätsprüfung und –

bewertung. Der berechnete Qualitätsindex wird dem jeweiligen Datensatz angehängt.

3. Quelle und weiterführende Literatur:

Balzert, H. Schröder, M., Schäfer, C. (2011): *Wissenschaftliches Arbeiten – Ethik, Inhalt & Form wiss. Arbeiten, Handwerkszeug, Quellen, Projektmanagement, Präsentation*. 2. Aufl. Herdecke; Witten: W3L-Verlag.

Bertelmann,R., Gebauer, P., Hasler, T., Kirchner, I., Peters-Kottig, W., Razum, M., Recker, A., Ulbricht, D., van Gasselt, S., (2014). *Einstieg ins Forschungsdatenmanagement in den Geowissenschaften*. doi: 10.2312/lis.14.01

Higgins, J.& Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0.* Zugriff am 24.10.2014 unter: http://handbook.cochrane.org.

Naumann, F. (2007). Datenqualität. *Informatik-Spektrum*, 30(1), 27-31.

Rahm, E. & Do, H.-H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* 23(4), 3–13.

Wang,R., Ziad,M. &Lee, Y.(2001). *Data quality.* Massachusetts: Kluwer.

Schendera, C. F. (2007). *Datenqualität mit SPSS*. Oldenburg. Zugriff am 24.10.2014 unter http://tocs.ulb.tu-darmstadt.de/186624018.pdf.

Wand. Y & Wang, R.(1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95.

Wang, R & Strong, D .(1996). Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.* 12(4), 5–34.