Informationen zur Ergebnisdatenübermittlung an die NAKO Transferstelle

Version: 2 vom 21.01.2021
Autoren: Gunthard Stübs

1. Einleitung

Die Nutzungsordnung des NAKO e.V. legt fest (§ 10), dass Ergebnisdaten inkl. Dokumentation an die Transferstelle der NAKO zu übermitteln sind, damit diese in die Forschungsdatenbank der NAKO integriert werden. Es ist notwendig, dass vor der Integration der eigentlichen Ergebnisdaten die Metadaten (beschreibenden Daten) an das Integrationszentrum übermittelt werden.

Dieses Dokument beschreibt alleine die Rückübermittlung von alphanummerischen Ergebnisdaten und nicht von komplexen Ergebnisdaten wie bearbeitete oder annotierte DICOM-Bilder.

1.1. Unterstützte Datentypen

Aktuell können Daten mit den folgenden Datentypen an das Integrationszentrum übermittelt werden:

Ganzzahlen	2 (short), 4 (integer) oder 8 Bytes (long)
Fließkommazahlen	4 (float) oder 8 Byte (double), decimal
Kategoriale Werte	Wahrheitswerte (boolean), Auswahllisten (options)
Datums- und Zeitangaben	Datum (date), Uhrzeit (time), Datum und Uhrzeit (datetime)
Text	Variable Zeichenketten bis 256 Zeichen (varchar), Text (bis 65.500 Zeichen oder bis 4 GiB Zeichen)
Geometrische Formen	Speicherung als Liste von Koordinaten im eigenen Koordinatensystem als Text

Abbildung 1 Aktuell unterstützte Datentypen

2. Metadaten

Die Metadaten sind die Summe der Beschreibungen der einzelnen Datenelemente (Variablen). Die Metadaten werden im Data Dictionary des Integrationszentrums der NAKO gespeichert und im Datenverzeichnis der Webanwendung TransferHub der Transferstelle der NAKO angezeigt. Die Datenelemente sollten eine baumartige Gliederung aufweisen, um sie im Datenverzeichnis zu gruppieren und deren Auffinden zu erleichtern (s. Abb. 1).

Ein Datenelement (Variable) beschreibt die Daten, die in einer einzelnen Spalte der Ergebnisdaten an das Integrationszentrum übermittelt werden (Absatz 2.1). Für die qualifizierten Missings gibt es im Absatz 2.2 gesonderte Hinweise. Informationen zur Gliederung bzw. Gruppierung der Datenelemente gibt es in Absatz 2.3. Die Formate zur Speicherung der Metadaten im Excel- oder CSV-Format sind in den Absätzen 0 resp. 2.5 beschrieben.

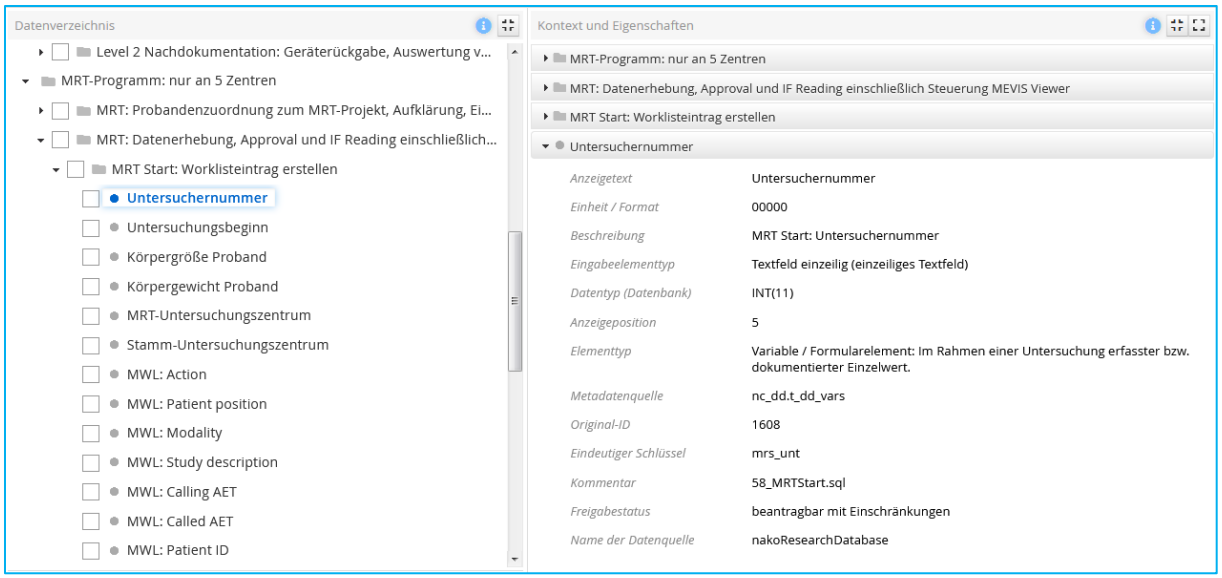

Abbildung 2 Darstellung von Datenelementen im Datenverzeichnis des TransferHub

2.1. Beschreibung der Datenelemente

Für jedes Datenelement müssen mindestens für die verpflichtenden Attribute (Absatz 2.1.1) Werte angegeben werden. Das Data Dictionary der NAKO unterstützt aktuell noch keine Mehrsprachigkeit. Falls die Metadaten in Deutsch und Englisch verfügbar sind, sollte die englische Sprache bevorzugt werden.

2.1.1. Verpflichtende Attribute

***Name** – Kurzreferenz für das Datenelement*

Das Attribut „Name“ darf maximal 20 Zeichen lang sein und muss im gesamten NAKO Data Dictionary eindeutig sein. Der Name darf nur aus kleinen ASCII-Buchstaben (a-z, keine Umlaute etc.), Ziffern und Unterstrichen (underscore) bestehen. Um die Eindeutigkeit sicher zu stellen, sollte der Name mit „u“, der Antragsnummer und einem Unterstrich beginnen z.B. „u551_heart_dis_1“

***Titel** – natürlich-sprachlicher Bezeichner für das Datenelement*

Das Attribut „Titel“ ist ein aussagekräftiger Bezeichner (label) des Datenelements und wird z.B. im Baum des Datenverzeichnisses angezeigt. Der Titel sollte aus den üblichen alphanummerischen und wenn notwendig Satzzeichen bestehen. Der Titel sollte auch unter Einbeziehung der Gliederung (siehe 2.3) gewählt werden. Der Titel sollte die Länge von 60 Zeichen nicht überschreiten. Sofern mehr Informationen vorhanden sind, können diese in der optionalen Beschreibung stehen.

***Datentyp** – der Datentyp des Datenelements*

Das Attribut „Datentyp“ bezeichnet den SQL-Datentyp für die Speicherung der Daten im Datenbankmanagementsystem (DBMS; in der NAKO MariaDB). Grundsätzlich können hier alle im SQL-Standard erlaubten Datentypen angegeben werden. Wenn es einfacher ist, können insbesondere in der Excel-Tabelle (siehe 0) auch die natürlich-sprachlichen Typen *"Ganzzahl", "Fließkommazahl", "Auswahlliste", "Ja+Nein", "Text max 200 Zeichen", "Text max 50 Zeichen", "Text max 20 Zeichen", "Datum", "Uhrzeit", "Datum+Uhrzeit" und "Lange Ganzzahl"* verwendet werden. (Diese können in der Excel-Datei ausgewählt werden.)

***Skalenniveau** – das Skalennvieau des Datenelements*

Das „Skalenniveau“ des Datenelements ist entweder *"metrisch", "nominal", "ordinal", "hierarchisch" oder "Text".* (Diese können in der Excel-Datei ebenfalls ausgewählt werden.)

2.1.2. Optionale Attribute

***Beschreibung** – eine ausführlichere Beschreibung des Datenelements*

Das Attribut „Beschreibung“ des Datenelements kann alle weiteren spezifischen Informationen zum Datenelement selbst enthalten, z.B. Erläuterungen zur Methodik der Generierung und Bedeutung der Werte, Hinweise zur Nutzung u.a.

***Einheit** – die Maßeinheit für die Werte des Datenelements*

Das Attribut „Einheit“ enthält die einheitliche SI-Maßeinheit für die im Allgemeinen metrischen Werte des Datenelements in Textform, z.B. „cm“, „mm“, „cm^3“.

***Optionen** – die Auswahloptionen bei Auswahllisten oder Missing-Kodierungen*

Das Attribut „Optionen“ enthält die Auswahloptionen von Auswahllisten mit oder ohne Missing-Kodierungen oder nur die Kodierungen von qualifizierten Missings (siehe Abs. 2.2). Eine Option besteht aus einer ganzzahligen Kodierung und einem Text als Wert, die durch „=“ voneinander getrennt sind, z.B. „1 = Anzeichen für Herzinfarkt vorhanden“. Für kategoriale Datenelemente enthalten die „Optionen“ eine Liste an Auswahloptionen, wobei die gültigen Auswahloptionen mit positiven Ganzzahlen und die Missings mit negativen Ganzzahlen zu kodieren sind. Für die Datenelemente, die nur kodierte Missings enthalten, können die Kodierungen beliebige Ganzzahlen sein. Im Excel-Format (s. Abs. 0) werden die einzelnen Auswahloptionen durch Zeilenumbrüche getrennt, im CSV-Format (s. Abs. 2.5) durch „|“ (senkrechter Strich).

2.2. Behandlung von qualifizierten Missings

2.2.1. Kodierung von qualifizierten Missings

Bei der Übermittlung von Ergebnisdaten soll es keine unqualifizierten Missings von Werten geben. D.h. sofern bei einem Datenelement nicht für jeden Teilnehmer ein Wert generiert wurde, es also fehlende Werte gibt, so müssen die erlaubten Gründe dafür im Voraus als qualifizierte Missings in den Metadaten kodiert werden. Die qualifizierten Missings sind als Ganzzahlen zu kodieren (siehe auch „Optionen“ in Abs. 2.1.2).

2.2.2. Definition von Missing-Kodierungen

Aktuell gibt es keine allseits abgestimmten und festgelegten Missings hinsichtlich der Kategorisierung und Bezeichnung. Trotzdem wäre es sinnvoll, wenn nicht jeder Nutzer das Rad neu erfinden muss. Folgende Missings werden daher aktuell empfohlen, wenn diese nicht ausreichen, können diese eigenständig ergänzt werden.

Kodierung	Bezeichnung	Erläuterung
-1	Fehlende Rohdaten	Die Rohdaten liegen für den TN nicht vor, z.B. ist eine MRT-Sequenz nicht gemessen, oder eine Blutprobe nicht bereitgestellt.
-2	Unzureichende Qualität	Die Qualität der Rohdaten reicht insgesamt nicht für den Beginn der Wertermittlung aus, z.B. wenn MRT-Bilder unscharf sind oder eine Serumprobe voller Erythrozyten oder degradiert ist.
-3	Unzureichende Quantität	Die Menge der Rohdaten reicht nicht aus, z.B. reicht das Probenvolumen nicht für eine Messung aus.
-4	Wertermittlung nicht erfolgreich	Konkret dieser Wert konnte nicht ermittelt werden d.h. die Messung nicht erfolgreich, z.B. weil im MRT ein Organ nicht segmentiert werden konnte oder die Serummessung kein Ergebnis geliefert hat.

Abbildung 3 Empfohlene Missings und Kodierungen, die individuell erweiter werden können

2.2.3. Ausblick auf die Speicherung von Werten und Missings

Da die ganzzahligen Missing-Kodierungen sowieso nur in Wertespalten mit dem Typ Ganzzahl gespeichert werden könnten dort aber auch mit den erlaubten Werten kollidieren könnten, sollen die Missing-Kodierungen in einem weiteren Datenelement / Wertespalte gespeichert werden.

Wenn also die Werte eines Datenelements Missings aufweisen, muss in den Metadaten ein zusätzliches Datenelement für die entsprechenden Missings definiert werden. Der „Name“ dieses Datenelements für die Missings sollte sich am „Namen“ des Datenelements mit den Werten anlehnen, z.B. mit einem Suffix „_m“.

Wenn also die Ergebnisdaten in das Rückgabeformat übertragen werden, dann sollten in der Wertespalte die Werte stehen, bei den Missings (d.h. wo keine Werte enthalten sind) wird statt des Werts „null“ eingetragen oder der Wert leer gelassen. In den Fällen, wo für einen Teilnehmer/Pseudonym in der Wertespalte „null“ eingetragen oder leer gelassen wird, wird dann in der Missing-Spalte die Missing-Kodierung eingetragen. Wenn in der Wertespalte hingegen Werte eingetragen werden, wird bei den Missings ebenfalls „null“ eingetragen (siehe Beispiel) oder der Wert leer gelassen.

Teilnehmer	**u555_heart_dis_1**	**u555_heart_dis_1_m**
1234567891	157.30	null
1234456784	145.87	null
4561379163	null	-4
7983197646	162.17	null

Abbildung 4 Beispieltabelle für die Speicherung von Werten und Missings

2.3. Gliederung der Datenelemente

Die Gliederung der Datenelemente dient der Gruppierung zum Zweck des leichteren Auffindens von Datenelementen im Datenverzeichnis und dem besseren Verständnis der Daten. Das Ziel der NAKO ist es, dass spätere Antragsteller die bereits durchgeführten Auswertungen von komplexen Daten oder Bioproben nicht nochmal wiederholen müssen sondern stattdessen im Datenverzeichnis auswählen und nach einer Sperrfrist nachnutzen können.

Im Datenverzeichnis definiert die Gliederung den Baum, in den die Datenelemente einsortiert sind. Falls z.B. aus einer MRT-Sequenz jeweils mehrere Dimensionen aus der Leber und der Milz gemessen werden, ist es sinnvoll die beiden Organe unterschiedlich in die Gliederung einzusortieren. z.B.

+ Ergebnisdaten aus Nutzungsantrag NAKO-555

|------ + Vermessung der Leber

|---------- + Lebervolumen

|---------- + Länge der Leber dorsal-ventral

|------ + Vermessung der Milz

|---------- + Milzvolumen

|---------- + Länge der Milz dorsal-ventral

2.3.1. Gliederung bis 1000 Datenelemente

Wenn von einem Nutzungsantrag weniger als 1000 Datenelemente zurückgeliefert werden, dann stehen maximal zwei Ebenen zur Gliederung zur Verfügung: „Abschnitt 1“ und „Abschnitt 2“. Der „Abschnitt 2“ ist immer einem konkreten „Abschnitt 1“ untergeordnet.

Titel

Jeder Abschnitt hat als verpflichtendes Attribut den „Titel“ als natürlich-sprachlichen Bezeichner für den Abschnitt. Der „Titel“ sollte aus den üblichen alphanummerischen und wenn notwendig Satzzeichen bestehen. Der „Titel“ sollte die Länge von 60 Zeichen nicht überschreiten. Sofern mehr Informationen vorhanden sind, können diese in der optionalen Beschreibung stehen.

Beschreibung

Jeder Abschnitt hat als optionales Attribut die „Beschreibung“, die alle weiteren spezifischen Informationen zum Abschnitt selbst enthalten, z.B. Erläuterungen zur Methodik oder allgemeine Hinweise zu allen Datenelementen im Abschnitt.

Bei wenigen zurückgelieferten Datenelementen, müssen nicht beide Abschnitte verwendet werden, es reicht dann „Abschnitt 1“.

2.3.2. Gliederung für mehr als 1000 Datenelemente

Wenn von einem Nutzungsantrag fast 1000 oder mehr Datenelemente zurückgeliefert werden, dann sollte eine zusätzliche dritte Gliederung genutzt werden: „Abschnitt 0“. Die Abschnitte 1 und 2 sollten weiterhin genutzt werden, wobei ein „Abschnitt 1“ immer einem konkreten „Abschnitt 0“ untergeordnet ist. Ein „Abschnitt 0“ sollte mehrere hundert bis unter 1000 Datenelemente enthalten. z.B.

```
+ Abschnitt 0
|---------- + Abschnitt 1
|---------- + Abschnitt 2
|---------- + Datenelement 1
```

2.4. Excel-Format für die Metadaten

Wenn die Datenelemente händisch beschrieben werden, sollten diese in die begleitend bereitgestellt Excel-Tabelle eingetragen werden. Ein Datenelement wird in einer Zeile beschrieben und jede Zeile muss ein Datenelement enthalten (d.h. es darf keine Zeilen geben, die nur Abschnitte enthalten). Siehe auch das Beispiel in Absatz 2.7, Seite 7).

2.4.1. Definition der Gliederung in Abschnitte

Bei wenigen Datenelementen muss die Gliederung durch Abschnitte nicht genutzt werden. Bei vielen Datenelementen stehen zuerst die „Abschnitt 1“ und bei sehr vielen Datenelementen auch die „Abschnitt 2“ zur Verfügung. Ein (neuer) Abschnitt wird bei dem ersten Datenelement in einem Abschnitt eingetragen (d.h. in derselben Zeile). Bei weiteren Datenelementen im selben Abschnitt können die Zellen in „Abschnitt 1“ und „Abschnitt 2“ leer gelassen werden. Es kann keine neue Beschreibung eingetragen werden, wenn sich nicht auch der Titel ändert. Wenn z.B. ein „Abschnitt 2“ enden soll, ohne dass ein neuer „Abschnitt 2“ beginnt, muss bei „Abschnitt 2.Titel“ und „Abschnitt 2.Beschreibung“ ein Minuszeichen („-„) eingetragen werden.

2.4.2. Definition der Datenelemente

Die Bedeutung und die formalen Anforderungen an die Attribute wurden in Absatz 2.1 beschrieben. Die verpflichtenden Attribute „Name“, „Titel“, „Datentyp“ und „Skalenniveau“ sind in der Excel-Tabelle mit roten Köpfen markiert. Die erlaubten Werte für „Datentyp“ und „Skalenniveau“ sind als Auswahlliste hinterlegt (rechts unten neben der Zelle den kleinen Pfeil nach unten anklicken). In dem Feld „Optionen“ (siehe auch Abs. 2.1.2) müssen die einzelnen Auswahloptionen durch Zeilenumbrüche in der Eingabezelle voneinander getrennt werden. (In Excel kann man durch „Alt + Enter“ einen Zeilenumbruch in der Zelle einfügen.)

Die Datenelemente für die Missings sollten unmittelbar auf die Datenelemente für die Werte folgen.

2.5. CSV-Format für die Metadaten

Wenn eine große Anzahl Datenelemente beschrieben werden sollen, die tw. auch computergeneriert sind, ist es möglich, die generierten Metadaten in einer CSV-Datei abzuspeichern.

2.5.1. Definition der CSV-Inhalte

Die CSV-Datei muss in der ersten Zeile die folgenden verpflichtenden und kann die optionalen Spalten (Attribute) enthalten.

„Abschnitt 0.Titel“
„Abschnitt 0.Beschreibung“
„Abschnitt 1.Titel“
„Abschnitt 1.Beschreibung“
„Abschnitt 2.Titel“
„Abschnitt 2.Beschreibung“
„Name“
„Titel“
„Beschreibung“
„Datentyp“
„Skalenniveau“
„Einheit“
„Optionen“

Die Inhalte bzw. erlaubten Werte dieser Spalten sind in Abs. 2.1 beschrieben. Es müssen immer Werte für „Name“, „Titel“, „Datentyp“ und „Skalenniveau“ angegeben sein. Ein Datenelement ist immer in einer Zeile zu beschreiben (siehe auch Abs. 2.4). Zeilenumbrüche sind nicht erlaubt. Die Auswahloptionen werden in der Spalte „Optionen“ nicht durch Zeilenumbrüche, sondern durch senkrechte Striche „|“ getrennt.

2.5.2. Definition von CSV-Formatvorgaben

Die folgenden Vorgaben sind für die CSV-Datei einzuhalten:

1. Das Trennzeichen ist das Semikolon „;“.
2. Eine Zeile ist mit einem Zeilenumbruch („\n“) zu beenden.
3. Jede Zeile muss gleich viele Spalten und damit Semikolons „ ;“ enthalten
4. Zeichenketten müssen an ihrem Anfang und Ende in obere Anführungszeichen ,“‘ eingeschlossen sein.
5. Andere Zeichen als ASCII-Zeichen, z.B. Umlaute, sind als UTF-8 zu kodieren.
6. Wenn in einem Feld nicht auf Zeilenumbrüche verzichtet werden können, müssen diese durch „\n“ maskiert werden.

2.6. Übermittlung und Prüfung der Metadaten

Die Metadaten können unter Angabe der Antragsnummer z. B. „Metadaten für NAKO-551“ als E-Mail mit Datei-Anhang an gunthard.stuebs@uni-greifswald.de geschickt werden. Die Excel-Datei soll auch so angehängt und nicht in CSV konvertiert werden.

2.7. Beispiel für Metadaten in einer Excel-Tabelle

	A	B	C	D	E	F	G	H	I	J	K
1	Abschnitt 1.Titel	Abschnitt 1.Beschreibung	Abschnitt 2.Titel	Abschnitt 2.Beschreibung	Name	Titel	Beschreibung	Datentyp	Skalenniveau	Einheit	Optionen
2	Physiognomie	Die Physiognomie aus den Neurosequenzen	Schädel	Ausmaße des Schädels	u551_skull_le_re	Schädelbreite	Ausmaße des Schädels rechts-links	Ganzzahl	metrisch	mm	
3					u551_skull_le_re	Schädelbreite Mssings	Missings für Ausmaße des Schädels vorne-hinten	Ganzzahl	nominal		-1 = Fehlende Rohdaten -2 = Unzureichende Qualität -4= Wertermittlung nicht erfolgreich
4					u551_skull_fro_ba	Schädeltiefe	Ausmaße des Schädels vorne-hinten	Ganzzahl	metrisch	mm	
5					u551_skull_fro_ba_m	Schädeltiefe Missings	Missings für Ausmaße des Schädels vorne-hinten	Ganzzahl	nominal		-1 = Fehlende Rohdaten -2 = Unzureichende Qualität -4= Wertermittlung nicht erfolgreich
6			-	-	u551_skull_volume	Hirnvolumen	Volumen des Gehirns	Fließkommazahl	metrisch	mm^3	
7					u551_skull_volume_m	Hirnvolumen Missings	Missings für Volumen des Gehirns	Ganzzahl	nominal		-1 = Fehlende Rohdaten -2 = Unzureichende Qualität -4= Wertermittlung nicht erfolgreich
8			Ohren	Ausmaße der Ohren	u551_ear_length	Ohrenlänge	Länge der Ohren	Ganzzahl	metrisch	mm	
9					u551_ear_length_m	Ohrenlänge Missings	Missings der Länge der Ohren	Ganzzahl	nominal		-1 = Fehlende Rohdaten -2 = Unzureichende Qualität -4= Wertermittlung nicht erfolgreich
10	Hirnmasse	Kenngrößen zur Hirnmasse	Graue Hirmasse	-	u551_brain_gra_vol	Volumen der grauen Hirnmasse	Messung der grauen Hirnmasse mit Schwellwert 121 nach der Methode von White und House et al.	Fließkommazahl	metrisch	mm^3	
11					u551_brain_gra_vol_m	Volumen der grauen Hirnmasse Missings	Missings zur Messung der grauen Hirnmasse	Ganzzahl	nominal		-1 = Fehlende Rohdaten -2 = Unzureichende Qualität -4= Wertermittlung nicht erfolgreich
12					u551_brain_lesion	Sind Läsionen sichtbar?	Sind Läsionen in der grauen Hirnmasse sichtbar?	Ganzzahl	nominal		1 = Ja 2 = Nein -1 = Fehlende Rohdaten -2 = Unzureichende Qualität -4= Wertermittlung nicht erfolgreich
13	Haarlänge	Länger der Haare	-	-	....	....	...				
14											

Abbildung 5 Beispieltabelle für die Metadaten

3. Ergebnisdaten

Die Ergebnisdaten sind alle personenbezogenen Endergebnisse, die entweder für die eigenen Auswertungen oder für andere Antragsteller relevant sein können. Nach der Übermittlung und Prüfung der Metadaten können die Ergebnisdaten nach den im Folgenden beschrieben Vorgaben als CSV-Datei generiert und gespeichert werden.

3.1. Definition der Inhalte

Für die Inhalte der Ergebnisdaten und deren Speicherung in einer Tabelle gelten die folgenden Vorgaben:

1. Die erste Zeile ist die Zeile mit den Spaltenköpfen.
2. Der erste Spaltenkopf lautet „Proband“.
3. Alle weiteren Spaltenköpfe entsprechen den „Namen“ der Datenelemente.
4. Auf ein Spaltenkopf mit dem Datenelement für die Werte sollte, falls vorhanden, das Datenelement für seine Missings folgen.
5. Die Werte für einen Teilnehmer werden in eine Zeile geschrieben.
6. In das erste Feld wird das Pseudonym des Teilnehmers geschrieben.
7. In die weiteren Felder werden die Werte und die Missings geschrieben.

3.2. Definition von Formatvorgaben

Zusätzlich zu den Vorgaben in Abs. 2.5.2 gibt es die folgenden Vorgaben:

1. Bei Fließkommazahlen sollen die Kommastellen mittels eines Dezimalpunktes getrennt werden (englische Notation). Es dürfen keine „Tausendertrennzeichen“ (z.B. Komma) verwendet werden.
2. Für Datumsangaben sind nur die Formate „yyyy-MM-dd“ (empfohlen), „yyyy/MM/dd“, „dd-MM-yyyy“ oder „dd/MM/yyyy“ erlaubt.
3. Für Uhrzeiten ist nur das Format „HH-mm-ss“ als 24-Stunden-Angabe erlaubt.
4. Für Datum und Uhrzeit sind durch Leerzeichen getrennte Kombinationen von Datum und Uhrzeit erlaubt.

3.3. Übermittlung der Ergebnisdaten

Nach der Prüfung der Metadaten wird dem Antragsteller eine Url des Integrationszentrums Greifswald und ein Token zum verschlüsselten Upload der Ergebnisdaten mitgeteilt.