

Title: The Status of data management practices throughout the Data Life Cycle: a Mixed-Method Study across MIRACUM Data Integration Centers

Overview Data Collection Items (English language) and original questionnaires (German language)

Overview Data Collection Items

Quantitative	general items	- participating Miracum Sites - general items related to DIC regulations, SOPs and associated responsibilities - processes description and division - provenance, self assessment, expectation, tools in use, collection
	items in data integration process	- artifacts and their properties from data management activities during data extraction and data transformation phase regarding - data elements - versioning - logfiles - execution history and environment - testing procedures during data management activities - traceability examination of data management activities - existence of metadata concept - usage of metadata template and/or tools
	items in data sharing process	- availability of documentation of resources and activities for data sharing process - availability of organizational information related to data release process: researcher, ethics committee, UAC, data protection officer, consent of data owner, accepted data use agreement, data delivery - items related to querying the data pool based on the data request in order to deliver a result object - generation of querying script - documentation of script available - log files available - content of log files - logging of access to data pools - logging of information about execution environment and history - validation available - format of delivery of result object - storage of data request artifacts - items related to the result object and data delivery - compliance given to the usage regulations - traceability examination of result object and artifacts

		- performed - internal registration of result object and artifacts performed - access medium to result object for researcher - acknowledgement of receipt obtained from researcher - archiving of result object performed - evaluation of result object for compliance to FAIR criteria - metadata concept available for data sharing process
	data interoperability via using terminology services: data element groups from MII core data set and their catalogs status from source system until FHIR profile	- diagnosis, procedure, laboratory finding, medication, intensive care parameter (physiological measurements and therapeutic measures), biospecimens - used coding catalogs and version per data element group and status of FHIR mapping for selected data elements - additional need for metadata
Qualitative	list of comment fields:	- description to obligatory demands for DIC - description DI processes data integration/data sharing - provenance tools in use - defined provenance criteria collected - expectation on provenance - Data Integration process - code properties (versioning) and retention period - logfiles (execution/error protocol) - metadata tools/templates - workflow tools - transparency/traceability aspect revealed in DIC - measurement of data integrity and data quality - content of logs - information automatically collected about IT infrastructure and execution history - FDR repositories used - data generated on execution history and execution environment - Data Sharing process - infos to data sharing process - validation information - content of log-files - formats produced for researchers - permanent availability of script - result object for researcher

		- storage of data request artefacts - compliance given to the usage regulations - internal registration of result object and artefacts performed - access medium to result object for researcher - metadata concept - Status data elements/catalogues - sources of not structured data - status of annotation of data elements
--	--	--

Original questionnaires (German language)

 Adding new Participant ID 1

Participant ID1

Prozess Datenintegration - Allgemeine Fragen

Ihr Standort

* must provide value

☐ Dresden
☐ Erlangen
☐ Frankfurt
☐ Freiburg
☐ Gießen
☐ Greifswald
☐ Magdeburg
☐ Mainz
☐ Mannheim
☐ Marburg

reset

Welche Vorgaben sind für die Tätigkeiten in Ihrem Datenintegrationszentrum bindend?

* must provide value

☐ DIZ Satzung
☐ Nutzungsordnung
☐ Datenschutzvorgabe, bitte spezifizieren
☐ Research Data Policy
☐ Standard-Arbeitsanweisungen (SOPs)
☐ Vorgaben der Herstellersysteme
☐ Weitere, bitte spezifizieren

reset

In welche groben Verarbeitungsschritte (Strecken) ist Ihre Datenintegrationspipeline unterteilt?

Erläuterung:

Bitte geben Sie an, in welche elementaren Teilabschnitte der Prozess der Datenextraktion aus den Quellsystemen bis hin zum Erreichen des FHIR Standards und der Überprüfung in ein Forschungsdatenrepositorium untergliedert ist, wie z.B. Quellsystem - Staging - DWH - FHIR - Forschungsdatenrepositorium.

* must provide value

☐ Strecken wie beispielhaft vorgegeben

☐ streaming
☐ klassisch per ETL

☐ andere, bitte spezifizieren

reset

Expand

reset

Sind für die einzelnen Verarbeitungsschritte Standardarbeitsanweisungen (SOPs) verfügbar?

* must provide value

☐ nein
☐ SOPs sind in Entwicklung
☐ SOPs sind freigegeben
☐ unbekannt

reset

Sind für die einzelnen Verarbeitungsschritte Rollen und Verantwortlichkeiten definiert?

(Beispiele für Rollen sind z.B. Entwickler, Tester, Administrator, ...)

* must provide value

☐ nein
☐ ja, bitte Rollen angeben

reset

Wie schätzen Sie die Expertise Ihres DIZ zur Erstellung und Verarbeitung von Provenance-Informationen während des Datenintegrationsprozesses ein.

* must provide value

Starter Level

Fortgeschrittenes Level

Experten Level

Change the slider above to set a response

reset

Welche Provenance-Tools haben Sie bereits im Einsatz?

* must provide value

☐ keine Tools
 ☐ folgende Tools, bitte nennen

reset

Welche Provenance-Informationen werden im DIZ erhoben?

* must provide value

☐ bislang keine
☐ bitte Erhebungskriterien nennen

reset

Welche Erwartung haben Sie an die Verwendung von Provenance-Informationen?

Expand

Datenmanagement während der Datenextraktions- und Datentransformationsphase

Wie dokumentieren Sie Ihre Datenmanagement-Aktivitäten?

In der Planungsphase		
☐ keine Dokumentation ☐ Dokumentation über einen Datenmanagementplan/-tool ☐ anders, bitte spezifizieren		
In der Ausführungsphase: Datenextraktion aus den unterschiedlichen Quell- Systemen		
Zugang zu den jeweiligen Quellsystemen		
Die Datenelemente sind ...	☐ nicht dokumentiert ☐ manuell annotiert (z.B. in Confluence) ☐ automatisiert annotiert ☐ sonstig dokumentiert, bitte spezifizieren	
Der Code ...	... ist versioniert	☐ nein ☐ ja, manuell ☐ ja, über ein Versionierungstool ☐ sonstiges
	... generiert Log-Files	☐ nein ☐ ja, statisch ☐ ja, parametrisierbar
	... ist inhaltlich dokumentiert	☐ nein ☐ ja
Die Log-Files ...	... werden wie lange aufbewahrt?	☐ keine Aufbewahrung ☐ bitte angeben, wie lange
	... bestehen getrennt aus Ausführungs-Logs und Fehlerprotokoll-Logs?	☐ ja ☐ andere Aufteilung, bitte angeben
Daten zur Ausführungshistorie/-umgebung des Codes werden generiert für ...	... den Testbetrieb.	☐ nein ☐ ja und manuell dokumentiert ☐ ja, automatisiert und sind kollaborativ im Zugriff ☐ sonstiges, bitte spezifizieren
	... den Produktivbetrieb	☐ nein ☐ ja und manuell dokumentiert ☐ ja, automatisiert und sind kollaborativ im Zugriff ☐ sonstiges, bitte spezifizieren

In der Ausführungsphase: Datentransformation in FHIR Standard		
Die Datenelemente sind ...	☐ nicht dokumentiert ☐ manuell annotiert (z.B. in Confluence) ☐ automatisiert annotiert ☐ sonstig dokumentiert, bitte spezifizieren	
Daten zum Ursprung der einzelnen Datenelemente sind ...	☐ nicht verfügbar ☐ verfügbar zusammen mit dem Datenelement	
Der Code ...	... ist versioniert	☐ nein ☐ ja, manuell ☐ ja, über ein Versionierungstool ☐ sonstiges
	... generiert Log-Files	☐ nein ☐ ja, statisch ☐ ja, parametrisierbar
	... ist inhaltlich dokumentiert	☐ nein ☐ ja
Die Log-Files ...	... werden wie lange aufbewahrt?	☐ keine Aufbewahrung ☐ bitte angeben, wie lange
	... bestehen getrennt aus Ausführungs-Logs und Fehlerprotokoll-Logs?	☐ ja ☐ andere Aufteilung, bitte angeben
Daten zur Ausführungshistorie/-umgebung werden generiert für	... den Testbetrieb	☐ nein ☐ ja und manuell dokumentiert ☐ ja, automatisiert und sind kollaborativ im Zugriff ☐ sonstiges, bitte spezifizieren
	... den Produktivbetrieb	☐ nein ☐ ja und manuell dokumentiert ☐ ja, automatisiert und sind kollaborativ im Zugriff ☐ sonstiges, bitte spezifizieren

In der Ausführungsphase: Datenintegration in Forschungsdatenrepositorium (FDR)		
☐ i2b2 ☐ OMOP ☐ andere, bitte angeben 		
Wie erfolgt die Befüllung des FDRs	☐ nicht dokumentiert ☐ manuell annotiert (z.B. in Confluence) ☐ automatisiert annotiert ☐ sonstig dokumentiert, bitte spezifizieren	
Daten zum Ursprung der einzelnen Datenelemente sind...	☐ nicht verfügbar ☐ verfügbar zusammen mit dem Datenelement	☐ reset
Der Code ...	... ist versioniert	☐ nein ☐ ja, manuell ☐ ja, über ein Versionierungstool ☐ sonstiges
	... generiert Log-Files	☐ nein ☐ ja, statisch ☐ ja, parametrisierbar
	... ist inhaltlich dokumentiert	☐ nein ☐ ja
Die Log-Files ...	... werden wie lange aufbewahrt?	☐ keine Aufbewahrung ☐ bitte angeben, wie lange
	... bestehen getrennt aus Ausführungs-Logs und Fehlerprotokoll-Logs?	☐ ja ☐ andere Aufteilung, bitte angeben
Daten zur Ausführungshistorie/-umgebung werden generiert für ...	... den Testbetrieb	☐ nein ☐ ja und manuell dokumentiert ☐ ja, automatisiert und sind kollaborativ im Zugriff ☐ sonstiges, bitte spezifizieren
	... den Produktivbetrieb	☐ nein ☐ ja und manuell dokumentiert ☐ ja, automatisiert und sind kollaborativ im Zugriff ☐ sonstiges, bitte spezifizieren
Welche Informationen werden in die Log-Files geschrieben? Sind diese strukturiert oder unstrukturiert? * must provide value		
Welche Informationen werden über die IT Infrastruktur bzw. über die Ausführungshistorie automatisiert aufgenommen? * must provide value		

Datenmanagement - Testverfahren

Sind Testverfahren während des Datenmanagementprozesses implementiert ...

... zur Aufrechterhaltung der Datenintegrität?	☐ nein ☐ ja reset	
	Die Testergebnisse sind ...	☐ nicht dokumentiert ☐ manuell dokumentiert ☐ automatisiert abgeleitet und kollaborativ im Zugriff ☐ anders dokumentiert, bitte spezifizieren reset
... zur Messung der Datenqualität?	☐ nein ☐ ja reset	
	Die Testergebnisse sind ...	☐ nicht dokumentiert ☐ manuell dokumentiert ☐ automatisiert abgeleitet und kollaborativ im Zugriff ☐ anders dokumentiert, bitte spezifizieren reset

Wird ein Workflow-Tool benutzt?

* must provide value

- ☐ nein
☐ ja, bitte spezifizieren

[reset](#)

Werden die vorhandene Dokumentation sowie deren Artefakte auf Nachvollziehbarkeit im DIZ hin untersucht?

* must provide value

- ☐ nein
☐ ja, bitte Methode und Ergebnisse erläutern
☐ ist geplant, bitte spezifizieren

[reset](#)

Gibt es ein Metadatenkonzept?

* must provide value

- ☐ nein
☐ ja, bitte erläutern oder hochladen

[reset](#)



[Expand](#)

Werden Metadaten Templates benutzt?

* must provide value

- ☐ nein
☐ ja, bitte spezifizieren welche

[reset](#)

Werden Metadaten Tools benutzt?

* must provide value

- ☐ nein
☐ ja, bitte spezifizieren welche

[reset](#)



Prozess Data Sharing - Allgemeine Fragen zum DIZ

Ihr Standort

* must provide value

- ☐ Dresden
- ☐ Erlangen
- ☐ Frankfurt
- ☐ Freiburg
- ☐ Gießen
- ☐ Greifswald
- ☐ Magdeburg
- ☐ Mainz
- ☐ Mannheim
- ☐ Marburg

reset

Welche Vorgaben sind für die Tätigkeiten in Ihrem Datenintegrationszentrum bindend?

* must provide value

- ☐ DIZ Satzung
- ☐ Nutzungsordnung, bitte spezifizieren
- ☐ Datenschutzvorgabe, bitte spezifizieren
- ☐ Research Data Policy
- ☐ Standard-Arbeitsanweisungen (SOPs)
- ☐ Weitere, bitte spezifizieren

In welche groben Verarbeitungsschritte ist Ihr Data Sharing Prozesses unterteilt?

z.B. Anfrage- & Vertrag - Datenanalyse - Datenbereitstellung - Archivierung

* must provide value

- ☐ wie beispielhaft vorgegeben
- ☐ andere, bitte spezifizieren

reset

Einteilung der Verarbeitungs-/Prozessschritte

Expand

Sind für den Data Sharing Prozess SOPs (Standard-Arbeitsanweisungen) verfügbar?

* must provide value

- ☐ nein
- ☐ SOPs sind in Entwicklung
- ☐ SOPs sind freigegeben
- ☐ unbekannt

reset

Bitte listen Sie die dazugehörigen SOPs auf:

Sind für die einzelnen Data Sharing Prozesse/Schritte Rollen und Verantwortlichkeiten definiert?

* must provide value

- ☐ nein
- ☐ ja, bitte angeben

reset

Expand

Wie schätzen Sie die Erfahrung Ihres DIZ zur Erstellung und Verarbeitung von Provenance-Informationen während der Data Sharing Phase ein?

* must provide value

Starter Level Fortgeschrittenes Level Experten Level

Change the slider above to set a response

reset

Welche Provenance-Tools haben Sie bereits im Einsatz?

* must provide value

- ☐ keine Tools
- ☐ folgende Tools, bitte nennen

reset

Welche Provenance-Informationen werden im DIZ erhoben?

* must provide value

- ☐ bislang keine
- ☐ bitte Erhebungskriterien nennen

reset

Expand

Welche Erwartung haben Sie an das Sammeln und die Verwendung von Provenance-Informationen?

* must provide value

Expand

Datenmanagement während des Prozesses Data Sharing

Wie dokumentieren Sie Ihre Datenmanagement-Aktivitäten?

* must provide value

- ☐ keine Dokumentation
- ☐ Dokumentation über einen Datenmanagementplan/-tool
- ☐ gemäß SOPs
- ☐ anders, bitte spezifizieren

reset

Wieviel Zeit verwenden Sie für die Dokumentation der Datenmanagement-Aktivitäten pro Studie bzw. pro Anfrage?

* must provide value

- ☐ Zeit nicht erfasst
- ☐ weniger als 8 Stunden
- ☐ zwischen 8 und 16 Stunden
- ☐ mehr als 16 Stunden
- ☐ anders, bitte spezifizieren
- ☐ nicht zutreffend

reset

In welcher Form sind (Meta-) Informationen über die Machbarkeits- oder Datenanfrage vorhanden?

Erläuterung:

Dazu gehören z.B. die Einschlusskriterien, Name der Anfragenden/Forscher, Datenmanager/Data Steward, Archivierungsstelle, etc.

* must provide value

- ☐ nicht strukturiert
- ☐ strukturiert
- ☐ über ein Template/Tool strukturiert
- ☐ strukturiert und maschinenlesbar

reset

Expand

Welche organisatorischen Informationen sind über den Datenfreigabeprozess verfügbar?

	nicht verfügbar	intern verfügbar, dokumentiert	intern verfügbar auf zentraler kollaborativer Plattform	nicht Teil des Prozesses, bitte spezifizieren	
Anfragende/Forschende	☐	☐	☐	☐	reset
Freigabe EK	☐	☐	☐	☐	reset
Freigabe UAC	☐	☐	☐	☐	reset
Freigabe DSB	☐	☐	☐	☐	reset
Zustimmung der jeweiligen Data Owner	☐	☐	☐	☐	reset
Nutzungseinschränkungen	☐	☐	☐	☐	reset
Akzeptierte Datennutzungsvereinbarung/-vertrag	☐	☐	☐	☐	reset
Datenausleitung (Datum)	☐	☐	☐	☐	reset

Bitte listen Sie hier weitere Information zum Freigabeprozess:

* must provide value

Expand

Abfragen im Datenpool basierend auf der Datenanfrage

Wird das Abfragescript validiert?

* must provide value

- ☐ nein oder nicht dokumentiert
- ☐ manuell dokumentiert (unstrukturiert)
- ☐ manuell dokumentiert (strukturiert)
- ☐ automatisiert getestet und kollaborativ

falls validiert, bitte angeben wie validiert:

z.B. gemäß Vier-Augen-Prinzip

Expand

Ist das Abfragescript inhaltlich dokumentiert?

* must provide value

- ☐ nein
- ☐ ja, teilweise nachvollziehbar dokumentiert
- ☐ ja, komplett nachvollziehbar dokumentiert

reset

Ist das Abfragescript versioniert?

* must provide value

- ☐ nein
- ☐ ja, manuell
- ☐ ja, über ein Versionierungstool

reset

Erzeugt das Abfragescript Logfiles?

* must provide value

- ☐ nein
- ☐ ja, als Ausführungs-Log
- ☐ ja, als Fehlerprotokoll-Log

Welche Informationen werden in Log-Files geschrieben?

* must provide value

Expand

Werden die Zugriffe auf die unterschiedlichen Datenpools geloggt?

z.B. über einen Audit-Trail

* must provide value

- ☐ nein
- ☐ ja, bitte erfasste Elemente angeben

reset

Expand

Werden Daten zur Ausführungsumgebung und -historie protokolliert?

* must provide value

- ☐ nein
- ☐ ja, manuell dokumentiert
- ☐ ja, automatisiert und kollaborativ im Zugriff, bitte spezifizieren
- ☐ sonstiges, bitte spezifizieren

reset

Expand

Welche Dateiformate werden für die Datenausgabephase erzeugt?

* must provide value

- ☐ überwiegend proprietäre Formate, bitte spezifizieren
- ☐ offene Formate und maschinenlesbar, bitte spezifizieren
- ☐ beides, bitte spezifizieren
- ☐ anders, bitte spezifizieren

Expand

Beiben alle Datensets, die über das Abfragescript hinweg transformiert werden, erhalten?

* must provide value

- ☐ nein
- ☐ verfügbar auf Anfrage (Zugriffsbeschränkung)
- ☐ Intern zentral verfügbar ohne Metadaten
- ☐ Intern zentral verfügbar mit Metadaten
- ☐ anders, bitte spezifizieren (auch Teilschritte der Aufbewahrung)

reset

Weitere Information zur Erhaltung von Datensets

Expand

Ergebnisobjekte

Unter Ergebnisobjekt ist der erzeugte Datensatz zwecks Datenausleitung zu verstehen.

Wird das Ergebnisobjekt vor der Ausgabe an den Forschenden auf Einhaltung der Nutzungsordnung nochmals geprüft und das Ergebnis dokumentiert?

- ☐ nein
☐ Ja, bitte Elemente der Prüfung beschreiben

[reset](#)

[Expand](#)

Wird das Ergebnisobjekt sowie dessen Artefakte auf Nachvollziehbarkeit im DIZ hin untersucht?

* must provide value

- ☐ nein
☐ stichprobenweise, bitte erläutern
☐ Ist geplant, bitte spezifizieren

[reset](#)

[Expand](#)

Wie werden das Ergebnisobjekt und dessen Artefakte intern registriert?

* must provide value

- ☐ DIZ Intern ohne Dokumentation
☐ DIZ Intern mit Dokumentation
☐ DIZ Intern kollaborativ mit Dokumentation
☐ offen und permanent verfügbar
☐ anders, bitte spezifizieren

[reset](#)

Wie bzw. über welches Medium erhält der Forschende Zugriff auf die Ergebnisse (digitalen Objekte)?

* must provide value

- ☐ eMail
☐ USB-Stick
☐ Internes Laufwerk
☐ Plattform, bitte spezifizieren
☐ andere, bitte spezifizieren

[reset](#)

Wird eine Bestätigung vom Forschenden zum Erhalt der angefragten Daten eingefordert?

* must provide value

- ☐ nein
☐ Ja

[reset](#)

[reset](#)

Werden das Ergebnisobjekt und dessen Artefakte archiviert?

* must provide value

- ☐ nein
☐ Ja, bitte Metadaten oder Standard spezifizieren

[reset](#)

Standard für die Auswahl von Metadaten bei der Archivierung:

* must provide value

[Expand](#)

Inwieweit liegen Ergebnisse für die Evaluierung der digitalen Objekte auf die Erfüllung der FAIR-Prinzipien vor?

* must provide value

- ☐ keine Untersuchung bzw. nicht dokumentiert
☐ stichprobenweise, bitte Erfüllungsgrad beschreiben
☐ komplette Evaluierung, bitte Erfüllungsgrad beschreiben

[reset](#)

[Expand](#)

Gibt es ein Metadatenkonzept?

* must provide value

- ☐ nein
☐ Ja, bitte erläutern oder bitte hochladen

[reset](#)

[Expand](#)

Bitte stellen Sie weitere (Meta-) Informationen oder das Metakonzept zur Verfügung:

[Upload file](#)

Teilstatus Datenintegration

 Adding new Participant ID 1

Participant ID

1

Ihr Standort

* must provide value

- ☐ Dresden
- ☐ Erlangen
- ☐ Frankfurt
- ☐ Freiburg
- ☐ Gießen
-  ☐ Greifswald
- ☐ Magdeburg
- ☐ Mainz
- ☐ Mannheim
- ☐ Marburg

[reset](#)

Welche Daten werden prozessiert?

Mehrfachnennungen sind möglich.

* must provide value

- ☐ strukturierte Daten (z.B. aus Tabellen)
-  ☐ halb-strukturierte Daten (z.B. XML Files)
- ☐ unstrukturierte Daten (z.B. Text, Bilder, Mails)

Aus welchen Datenquellen stammen die unstrukturierten Daten?

* must provide value



[Expand](#)

Sind die Datenelemente der jeweiligen Module über die gesamte Integrationspipeline annotiert?

Die Annotierung von Datenelementen bedeutet, dass die vorhandenen Variablen eines Datensatzes beschrieben sind in ihrer Struktur und Bedeutung.

* must provide value

- ☐ überhaupt nicht annotiert
-  ☐ teilweise annotiert
- ☐ komplett annotiert

[reset](#)

In welcher Form liegt die Annotierung von Datenelementen vor?

Mehrfachnennungen sind möglich.

Bitte in nächstem Feld erläutern.

* must provide value

- ☐ nicht vorhanden
-  ☐ reine textuelle Beschreibung (z.B. Confluence)
- ☐ maschinen-lesbar und verarbeitbar (z.B. ODM)
- ☐ anders, bitte unten erläutern

Bitte nennen Sie Details zur Umsetzung der Annotierung:



[Expand](#)

Matrix - Vorhandensein von Kodierungskatalogen

Bitte geben Sie an, welche Kodierungen und Kataloge in welchen Systemabschnitten vorliegen.

	Quellsystem	Lokales Data Mart	FHIR-Profil
Diagnose	☐	☐	☐
Prozedur	☐	☐	☐
Laborbefund	☐	☐	☐
Medikation	☐	☐	☐
Intensivmedizin (Physiolog. Messwerte)	☐	☐	☐
Intensivmedizin (Therap. Massnahmen)	☐	☐	☐
Bioproben	☐	☐	☐
Kodierungskatalog Diagnose * must provide value	☐ ICD ☐ SNOMED CT ☐ Alpha-ID ☐ Orpha ☐ sonstige bitte spezifizieren		
Kodierungskatalog Prozeduren * must provide value	☐ OPS ☐ sonstige bitte spezifizieren		
Kodierungskatalog Laborbefund und Artefakte * must provide value	☐ LOINC ☐ UCUM ☐ sonstige bitte spezifizieren		
Kodierungskatalog Medikation und Artefakte * must provide value	☐ ATC ☐ ASK ☐ sonstige bitte spezifizieren		
Kodierungskatalog Intensivmedizin (physiolog. Messwerte) bitte Datenelement(e) und Katalog mit Version angeben * must provide value			
	Expand		
Kodierungskatalog Intensivmedizin (therap. Maßnahmen) bitte Datenelement(e) und Katalog mit Version angeben * must provide value			
	Expand		
Kodierungskataloge Bioproben und Artefakte * must provide value	☐ SNOMED CT ☐ ICD-O ☐ sonstige bitte spezifizieren		
Version der Kataloge			