Title: The Status of Data Management Practices: a Mixed-Method Study across German Medical Data Integration Centers

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

Quanti- tative	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 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

Overview Data Collection Items

		 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
Quali- tative	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)

Prozess Datenintegration

Adding new Participant ID 1

Participant ID

Prozess Datenintegration - Allgemeine Fragen

1

Ihr Standort * must provide value	P	Dresden Erlangen Frankfurt Freiburg Gießen Greifswald Magdeburg Mainz Mannheim Marburg	reset
Welche Vorgaben sind für die Tätigkeiten in Ihrem Datenintegratioszentrum bindend? * must provide value	Ģ	 DIZ Satzung Nutzungsordnung Datenschutzvorgabe, bitte spezifizieren Research Data Policy Standard-Arbeitsanweisungen (SOPs) Vorgaben der Herstellersysteme Weitere, bitte spezifizieren 	
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	Ģ	Strecken wie beispielhaft vorgegeben streaming klassisch per ETL andere, bitte spezifizieren	reset Expand reset
Sind für die einzelnen Verarbeitungsschritte Standardarbeitungsanweisungen (SOPs) verfügbar? * must provide value	Ģ	 nein SOPs sind in Entwicklung SOPs sind freigegeben unbekannt 	
Sind für die einzelnen Verarbeitungsschritte Rollen und Verantwortlichkeiten definiert? (Beispiele für Rollen sind z.B. Entwickler, Tester, Administrator,) * must provide value	Ģ	⊂ 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	Ģ	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	P	keine Tools 🛛 folgende Tools, bitte nennen	
Welche Provenance-Informationen werden im DIZ erhoben? * must provide value	P	 bislang keine bitte Erhebungskriterien nennen 	reset
Welche Erwartung haben Sie an die Verwendung von Provenance-Informationen?	P		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

Zugang zu den jeweiligen Quellsystemen		
Die Datenelemente sind	 nicht dokumentiert manuell annotiert (z.B. in Confluen automatisiert annotiert sonstig dokumentiert, bitte spezifiz 	
	ist versioniert	nein ja, manuell ja, über ein Versionierungstool sonstiges res
Der Code	generiert Log-Files	 nein ja, statisch ja, parametrisierbar
	ist inhaltlich dokumentiert	⊖ nein ⊖ ja res
	werden wie lange aufbewahrt?	 keine Aufbewahrung bitte angeben, wie lange res
Die Log-Files	bestehen getrennt aus Ausführungs-Logs und Fehlerprotokoll-Logs?	⊂ ja ⊂ andere Aufteilung, bitte angeben res
Daten zur Ausführunghistorie/-umgebung	den Testbetrieb.	 nein ja und manuell dokumentiert ja, automatisiert und sind kollaborativ im Zugriff sonstiges, bitte spezifizieren res
des Codes werden generiert für	den Produktivbetrieb	 nein ja und manuell dokumentiert ja, automatisiert und sind kollaborativ im Zugriff sonstiges, bitte spezifizieren

In der Ausführu	ngsphase: Datentransformation in FHI	KStandard	
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 		
	ist versioniert	 nein ja, manuell ja, über ein Versionierungstool sonstiges 	
Der Code	generiert Log-Files	○ nein ○ ja, statisch ○ ja, parametrisierbar rese	
	ist inhaltlich dokumentiert	⊂ nein ⊖ ja rese	
	werden wie lange aufbewahrt?	 keine Aufbewahrung bitte angeben, wie lange rese 	
Die Log-Files	bestehen getrennt aus Ausführungs-Logs und Fehlerprotokoll-Logs?	⊂ ja ⊂ andere Aufteilung, bitte angeben rese	
Daten zur Ausführunghistorie/-umgebung	den Testbetrieb	 nein ja und manuell dokumentiert ja, automatisiert und sind kollaborativ im Zugriff sonstiges, bitte spezifizieren rese 	
werden generiert für	den Produktivbetrieb	 nein ja und manuell dokumentiert ja, automatisiert und sind kollaborativ im Zugriff sonstiges, bitte spezifizieren 	

OMOP andere, bitte angeben		
Wie erfolgt die Befüllung des FDRs	 nicht dokumentiert manuell annotiert (z.B. in Confluence) automatisiert annotiert sonstig dokumentiert, bitte spezifizien 	
Daten zum Ursprung der einzelnen Datenelemente sind	onicht verfügbar verfügbar zusammen mit dem Datenelement reset	
	ist versioniert	 nein ja, manuell ja, über ein Versionierungstool sonstiges
Der Code	generiert Log-Files	onein ja, statisch ja, parametrisierbar rese
	ist inhaltlich dokumentiert	⊂ nein ⊖ja res
	werden wie lange aufbewahrt?	 keine Aufbewahrung bitte angeben, wie lange res
Die Log-Files	bestehen getrennt aus Ausführungs-Logs und Fehlerprotokoll-Logs?	⊂ ja ⊂ andere Aufteilung, bitte angeben res
	den Testbetrieb	 nein ja und manuell dokumentiert ja, automatisiert und sind kollaborativ im Zugriff sonstiges, bitte spezifizieren
Daten zur Ausführunghistorie/-umgebung werden generiert für	den Produktivbetrieb	res nein ja und manuell dokumentiert ja, automatisiert und sind kollaborativ im Zugriff sonstiges, bitte spezifizieren res
Velche Informationen werden in die Log-File ind diese strukuriert oder unstrukturiert? must provide value	s geschrieben?	
elche Informationen werden über die IT Inf ber die Ausführungshistorie automatisiert		Expa

Datenmanagement - Testverfahren

Sind Testverfahren während de	es Datenmanagementprozesse	s implementiert
zur Aufrechterhaltung der Datenintegrität?	onein ja reset	
	Die Testergebnisse sind	 nicht dokumentiert manuell dokumentiert automatisiert abgleitet und kollaborativ im Zugriff anders dokumentiert, bitte spezifizieren
zur Messung der Datenqualität?	⊂ nein ⊖ja	
	Die Testergebnisse sind	 nicht dokumentiert manuell dokumentiert automatisiert abgleitet und kollaborativ im Zugriff anders dokumentiert, bitte spezifizieren
Wird ein Workflow-Tool benutz * must provide value	t?	 ○ nein
Werden die vorhandene Dokur auf Nachvollziehbarkeit im DIZ * must provide value		
Gibt es ein Metadatenkonzept? * must provide value	?	 nein ja, bitte erläutern oder hochladen
		<i>•</i>
		Expand
Werden Metadaten Templates * must provide value	benutzt?	 nein ja, bitte spezifizieren welche
Werden Metadaten Tools benu * must provide value	tzt?	 ○ nein ⇒ ○ ja, bitte spezifizieren welche reset

Prozess Data Sharing - Allgemei	ne Fragen zum DIZ	
Ihr Standort * must provide value	Dresden Erlangen Frankfurt Freiburg Gießen Greifswald Magdeburg Mainz Mannheim Marburg	rese
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	o wie beispielhaft vorgegeben o andere, bitte spezifizieren	rese
Einteilung der Verarbeitungs-/Prozessschritte		
Sind für den Data Sharing Prozess SOPs (Standard- Arbeitsanweisungen) verfügbar? * must provide value	neln SOPs sind in Entwicklung SOPs sind freigegeben unbekannt	Expand
Bitte listen Sie die dazugehörigen SOPs auf:		rese
Sind für die einzelnen Data Sharing Prozesse/Schritte Rollen und Verantwortlichkeiten definiert? * must provide value	o nein ja, bitte angeben	rese
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	Expand
Welche Provenance-Tools haben Sie bereits im Einsatz? * must provide value	keine Tools 🔄 folgende Tools, bitte nennen	rese
Welche Provenance-Informationen werden im DIZ erhoben? * must provide value	🔿 bislang kelne 🔄 bitte Erhebungskriterien nennen	rese
		Expand
Welche Erwartung haben Sie an das Sammeln und die Verwendung von Provenance-Informationen? * must provide value		cxpand
		Expand

Datenmanagement während des Prozesses Data Sharing

keine Dokumentation Wie dokumentieren Sie Ihre Datenmanagement-Dokumentation über einen Datenmanagementplan/-tool Aktivitäten? gemäß SOPs * must provide value anders, bitte spezifizieren reset Zeit nicht erfasst Wieviel Zeit verwenden Sie für die Dokumentation der weniger als 8 Stunden Datenmanagement-Aktivitäten pro Studie bzw. pro zwischen 8 und 16 Stunden Anfrage? mehr als 16 Stunden * must provide value anders, bitte spezifizieren nicht zutreffend reset In welcher Form sind (Meta-) Informationen über die Machbarkeits- oder Datenanfrage vorhanden? nicht strukturiert Erläuterung: strukturiert Dazu gehören z.B. die Einschlusskriterien, Name der über ein Template/Tool strukturiert Anfragenden/Forscher, Datenmanager/Data Steward, strukturlert und maschinenlesbar Archivierungsstelle, etc. reset * must provide value

Expand

Welche organisatorischen Informationen sind über den Datenfreigabeprozess verfügbar?

	nicht verfügbar	intern verfügbar, dokumentiert	intern verfügbar auf zentraler kollaborativer Platform	nicht Teil des Prozesses, bitte spezifizieren
Anfragende/Forschende				0
Freigabe EK				o reset
Freigabe UAC				reset
Freigabe DSB				
Zustimmung der jeweiligen Data Owner				reset
Nutzungseinschränkungen				reset
Akzeptierte Datennutzungsvereinbarung/-vertrag				o
Datenausleitung (Datum)				reset O reset
Bitte listen Sie hier weitere Information zu Freigabeprozess: * must provide value	Im			

Expand

Abfragen im Datenpool basieren	nd 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		
Ist das Abfragescript inhaltlich dokumentiert? * must provide value	nein Ja, teliweise nachvoliziehbar dokumentiert Ja, komplett nachvoliziehbar dokumentiert	Expand
Ist das Abfragescript versioniert? * must provide value	nein Ja, manuell Ja, über ein Versionierungstool	reset
Erzeugt das Abfragescript Logfiles? * must provide value	nein a, als Ausführungs-Log a, als Fehlerprotokoll-Log	Tese
Welche Informationen werden in Log-Files geschrieben? * must provide value		
Warden die Zugriffe auf die unterschiedlichen Datenneele		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
	nein	Expand
Werden Daten zur Ausführungsumgebung und -historie protokolliert? * must provide value	Ja, manuell dokumentiert Ja, automatisiert und kollaborativ im Zugriff, bitte spezifizieren sonstiges, bitte spezifizieren	reset
		1636
Welche Datelformate werden für die Datenausgabephase erzeugt? * must provide value	überwiegend proprietäre Formate, bitte spezifizier offene Formate und maschinenlesbar, bitte spezifiz beides, bitte spezifizieren anders, bitte spezifizieren	
		Expand
Bleiben alle Datensets, die über das Abfragescript hinweg transformiert werden, erhalten? * must provide value	neln 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 Datenauslei	tung 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 Pr			
Wird das Ergebnisobjekt sowie dessen Artefakte auf Nachvoliziehbarkeit im DIZ hin untersucht? * must provide value	Expand neln stichprobenweise, bitte erläutern ist geplant, bitte spezifizieren reset			
	OIZ intern ohne Dokumentation			
Wie werden das Ergebnisobjekt und dessen Artefakte intern registriert? * must provide value	DIZ Intern mit Dokumentation DIZ Intern kollaborativ mit Dokumentation offen und permanent verfügbar anders, bitte spezifizieren			
Wie bzw. über welches Medium erhält der Forschende Zugriff auf die Ergebnisse (digitalen Objekte)? * must provide value	reset OUSB-Stick Internes Laufwerk Plattform, bitte spezifizieren andere, bitte spezifizieren			
	andere, bitte spezinzieren reset			
Wird eine Bestätigung vom Forschenden zum Erhalt der angefragten Daten eingefordert? * must provide value	nein ja			
- musc provide value	reset			
	reset			
Werden das Ergebnisobjekt und dessen Artefakte archiviert? * must provide value	nein Ja, bitte Metadaten oder Standard spezifizieren reset			
	TESEL			
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	onein oja, bitte erläutern oder bitte hochladen reset			
	Expand			
Bitte stellen Sie weitere (Meta-) Informationen oder das Metakonzept zur Verfügung:	⊥ <u>Upload file</u>			

📃 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
Welche Daten werden prozessiert?	strukturierte Daten (z.B. aus Tabellen)
Mehrfachnennungen sind möglich.	halb-strukturierte Daten (z.B. XML Files)
* must provide value	unstrukturierte Daten (z.B. Text, Bilder, Mails)
Aus welchen Datenquellen stammen die unstrukturierten Daten? * must provide value Sind die Datenelemente der jeweiligen Module über die gesamte Integrationspipeline annotiert?	Expand
Die Annotierung von Datenelementen bedeutet, dass die vorhandenen Variablen eines Datensatzes beschrieben sind in ihrer Struktur und	 teilweise annotiert
Bedeutung. * must provide value	komplett annotiert reset
In welcher Form liegt die Annotierung von Datenelementen vor?	nicht vorhanden
Mehrfachnennungen sind möglich.	reine textuelle Beschreibung (z.B. Confluence)
Bitte in nächstem Feld erläutern.	maschinen-lesbar und verarbeitbar (z.B. ODM)
* must provide value	🗌 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	C	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. bitte Datenelement(e) und Katalog mit Version angeb * must provide value				
			Expand	
Kodierungskatalog Intensivmedizin (therap. Ma bitte Datenelement(e) und Katalog mit Version angeb * must provide value				
Kodierungskataloge Bioproben und Artefakte * must provide value		SNOMED CT ICD-O sonstige bitte spezifizieren	Expand	
Version der Kataloge				