MS DATA ALLIANCE

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

TOOLBOX

STRATEGIC FOCUS AREA 2

Develop a standardised MS data dictionary and support local harmonisation efforts

Tina Meißner
tina.meissner@med.uni-goettingen.de
University Medical Center Göttingen
Department of Medical Informatics

MS DATA ALLIANCE

Real-world data… **Reality Check**…

Example 1: employment

- **REGISTRY 1: Do you currently have a job or do any unpaid work outside your home?**
  - |_| Yes |_| No
- **REGISTRY 2: What is your job title?**
  - _____ (free text)
- **REGISTRY 3: What is your current employment status?**
  - |_| Working full-time
  - |_| Working part-time
  - |_| Unable to work because of sickness or disability
  - |_| Looking after home and/or family
  - |_| Student
  - |_| Retired
  - |_| Unemployed
  - |_| Doing unpaid or voluntary work

Example 2: Copaxone as DMT

COPAX
COPAXONE
COPAXONE 40
COPAXONE 40MG
COPAXONE HAUTEMENT DOSAGE
COPAXONE ORAL
COPAXONE [40]
COPAXONE(40MG)
COPAXONE-40
COPAXONE-CONFIDENCE TV-44400-CNS-40083
COPAXONE40
GLATIRAMER ACETATE

MS DATA ALLIANCE

Common Data Model - Rationale

- Every registry documents variables differently. That applies to e.g. naming conventions, level of detail or whether or not they are routinely documented at all.
- In order to facilitate collaborative research using existing, heterogeneous data sources, data needs to be harmonised.
- Harmonising data means that raw, heterogeneous data is mapped to an agreed upon, uniform representation; a common data model.

MS DATA ALLIANCE

Common Data Model - Rationale

Reduced time and costs

Through data harmonisation we aim to support

- Multiple MS collaborative research projects
- Prospective as well as retrospective data collections

by harmonising data sources once, at best.

Increased and guaranteed data quality

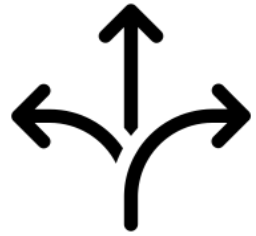
Quality checks are included in the mapping process.

Reproducable & transparent

Detailed information of strategy is documented and publicly available.

Multi-language

The local vocabulary will not be altered and is therefore still available for local use.

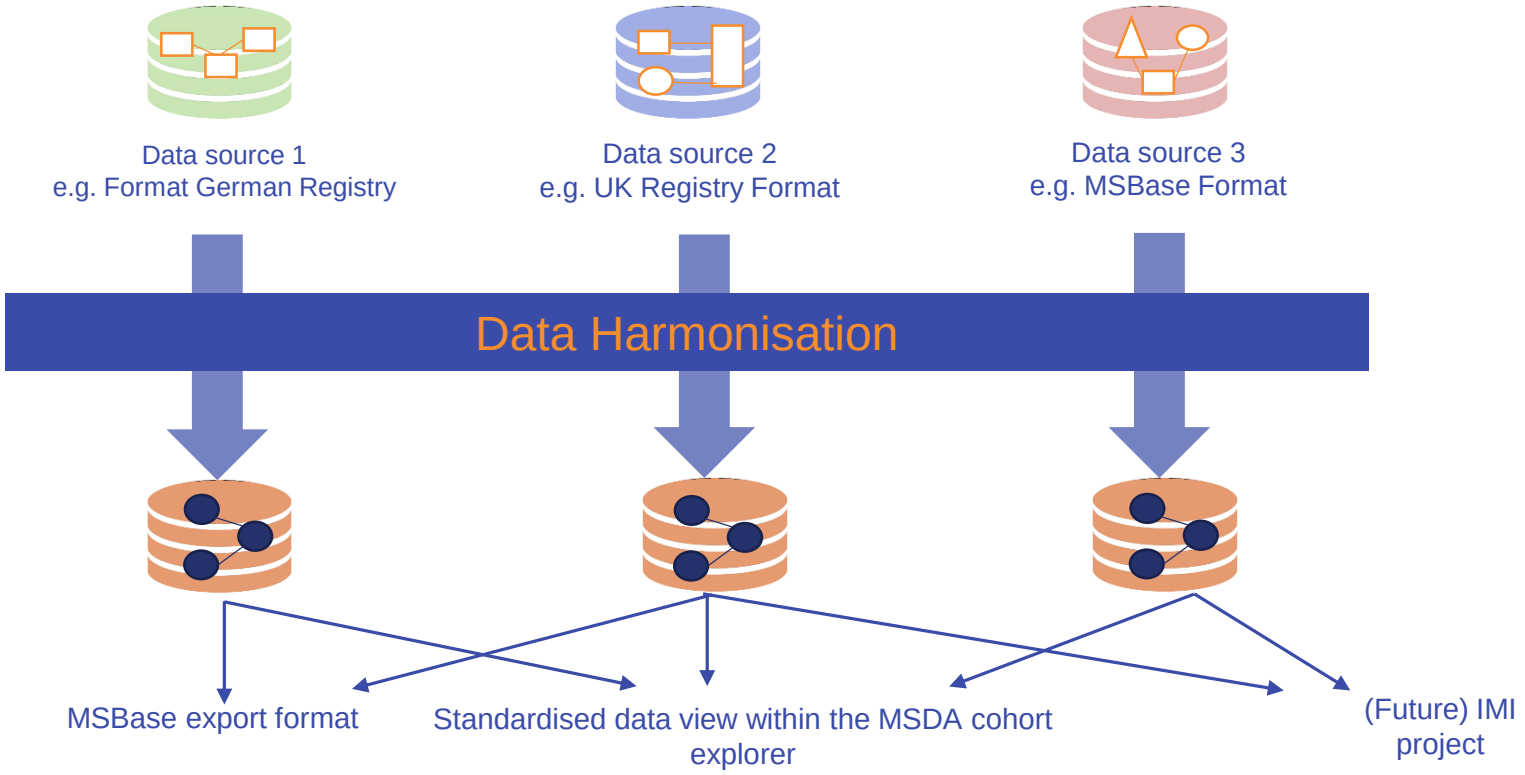


Flexible & Scalable

Common data model is adaptable and extendable along with the minimal dataset requirements.

MS DATA ALLIANCE

Schematic Representation of Data Harmonisation



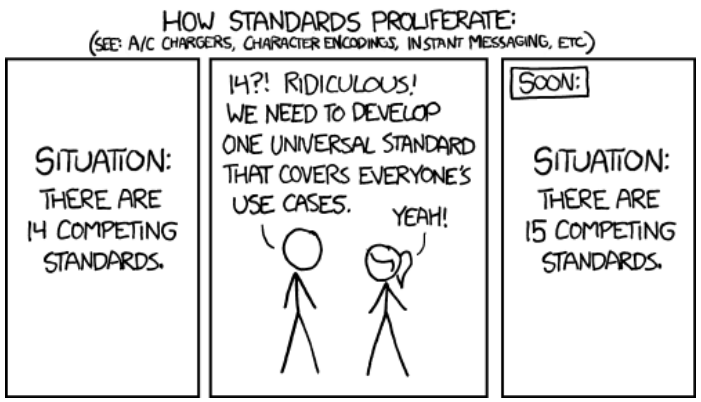
MSDA SwitchBox:

Proprietary, heterogeneous registry data

harmonised into a

common data model representation of registry data.

Common Data Model: Choosing “OMOP” – Why?



Credits: https://xkcd.com/927/

We don’t want *that*…

We chose the OMOP CDM because of the combination of following features:

- It is a mature data model for observational data (in our case registry data) in different types of databases.
- It is being used within the EMIF project which acts as the prototype for the MSDA toolbox infrastructure.
- The explicit capture of observation periods is characteristic for the OMOP CDM, enabling (long-time) follow-ups of patients. Characteristic in registries.
- The data model is extendable and adaptable to (changing) requirements.
- There is an extensive OHDSI community behind the OMOP CDM (and its tools).

MS DATA ALLIANCE

MSDA SwitchBox - What we did so far:

1. Started the documentation (MSDA SwitchBox concept) for the transformation of heterogeneous MS registry data to the OMOP Common Data Model.
2. Prepared a dictionary for the MSDA minimal dataset translation into OMOP Common Data Model.
3. Started to analyse the overall data structure of the two pilot registries:
    a) German MS registry
    b) UK MS registry.
4. Prepared a mapping dictionary of the relevant German and UK MS registry variables for the MSDA minimal dataset.
5. Started the data harmonisation within the data integration environment.

MS DATA ALLIANCE

Minimal Dataset for the Data Harmonisation

Patient Specific Information
Date of Birth
Date of Death
Gender
Feminine gender
Masculine Gender
Gender unknown
Non-binary gender

Disease Specific Information
Date of onset
Date of MS diagnosis
MS type
CIS
Relapsing remitting MS
Secondary progressive MS
Primary progressive MS
EDSS (value)
EDSS (date)
Relapses
Relapses (date)
Glucocorticoid (y/n)
treatment
No treatment

DMT (current and previous)
Drug type (DMT-MS, sympt.)
Drug (DMT-MS only)
Drug dose
Start date of Drug
End date of Drug
Reason for switch/stop/discont.

Serious Adverse Events
SAE (none/description)
treatment of SAE (y/n)
Treatment
No treatment
Date of SAE
Outcome of SAE
Adverse incident resulting in death
Adverse incident resulting in potentially permanent disabling damage
Adverse incident resulting in potentially permanent but not disabling damage
Transient abnormality unnoticed by the patient
Transient abnormality with full recovery

MS DATA ALLIANCE

Mapping dictionary

- Preparing a data dictionary that translates the MDS variables into OMOP CDM IDs/codes.
- This was a manual and lengthy step performed by looking up the MDS variables (or synonyms) in the OHDSI Search Engine ATHENA (http://athena.ohdsi.org/search-terms/terms).

Variable (Non-standard)	Concept_ID	Concept_Code (SNOMED)	Domain ID	Concept class	Concept_class_ID	Concept Class (higher)	Concept_class_id	OMOP CDM table/ comments
Patient Specific Information								PERSON
Date of Birth	4083587	184099003	Observation	Observable Entity	4181663			
Date of Death	4265167	399753006	Observation	Observable Entity	4181663			
Gender	4135376	263495000	Observation	Observable Entity	4181663			
Feminine gender	45766035	703118005	Gender	Observable Entity	4181663			
Masculine Gender	45766034	703117000	Gender	Observable Entity	4181663			
Gender unknown	4214687	394743007	Gender	Observable Entity	4181663			
Non-binary gender	36675593	772004004	Gender	Observable Entity	4181663			
Disease Specific Information								CONDITION_OCCURENCE + OBSERVATION
Date of onset	4181873	298059007	Observation	Observable Entity	4181663			"Date of onset"; not MS specific!
Date of MS diagnosis	4160852	432213005	Observation	Observable Entity	4181663			"Date of Diagnosis"; not MS specific!
MS type								
CIS	40493286	445967004	Condition	Clinical finding	441840	Demyl. Dis. Of the CNS	375801	"Clinically isolated syndrome"; not MS specific!
Relapsing remitting MS	4145049	426373005	Condition	Clinical finding (is a MS)	374919	Demyl. Dis. Of the CNS	375801	
Secondary progressive MS	4137855	425500002	Condition	Clinical finding (is a MS)	374919	Demyl. Dis. Of the CNS	375801	
Primary progressive MS	4178929	428700003	Condition	Clinical finding (is a MS)	374919	Demyl. Dis. Of the CNS	375801	
EDSS (value)	4169156	273554001	Measurement	Staging and scales	4110275	Staging and scales	4110275	Called "Kurtzke multiple sclerosis scale"
EDSS (date)	4231970	406543005	Observation	Observable Entity	4181663			"Date of visit"; not EDSS specific
Relapses (count?)	4117444	303359000	Observation	Qualifier Value	4179874			"relapse episode"; not EDSS specific; severity of a relapse
Relapses (date)	4231970	406543005	Observation	Observable Entity	4181663			"Date of visit"; not relapse specific
Glucocorticoid (y/n)	4022748	116596006	Drug	Pharma/Biol Product		Medicinal Product	35621894	
treatment	4191370	3890004	Observation	Qualifier Value	4179874			"treatment required for"
No treatment	4222632	83905007	Observation	Qualifier Value	4179874			"no treatment required for"
DMT (current and previous)								DRUG_EXPOSURE
Drug type (DMT-MS, sympt.)	4041829	229753003	Observation	Qualifier Value		Attribute	4078475	"Type of drug"
Drug (DMT-MS only)		RxNorm	Drug	Clinical Drug				DMT are single drug-specific (concerning values)
Drug dose								DMT are single drug-specific (concerning values)
Start date of Drug								DMT are single drug-specific (concerning values)
End date of Drug								DMT are single drug-specific (concerning values)
Reason for switch/stop/discont.								Reason-specific concept ID have to declared individually; stored in the drug_exposure table
Serious Adverse Events								CONDITION_OCCURENCE + OBSERVATION
SAE (none/description)	4105886	281647001	Condition	Clinical Finding	441840	Complication	433128	„Adverse reaction"; SAE-specific concept Ids have to be declared individually
	441207	62014003	Observation	Clinical Finding				„Adverse reaction to drug"
treatment of SAE (y/n)								
treatment	4191370	3890004	Observation	Qualifier Value	4179874			"treatment required for"
No treatment	4222632	83905007	Observation	Qualifier Value	4179874			"no treatment required for"
Date of SAE	4208903	439771001	Observation	Observable Entity	4181663			„Date of event"
Outcome of SAE	4181664	363788007	Observation	Observable Entity	4181663			Non-standard = CIEL (Outcome of adverse event) maps_to SNOMED „Clinical history/examination observable"
Adverse incident resulting in death	4236718	405535005	Observation	Clinical Finding	442840	Adverse incident outcome categories	4231813	Serious Adverse INCIDENTS
Adverse incident resulting in potentially permanent disabling damage	4226020	405532008	Observation	Clinical Finding	442840	Adverse incident outcome categories	4231813	Serious Adverse INCIDENTS
Adverse incident resulting in potentially permanent but not disabling damage	4236716	405531001	Observation	Clinical Finding	442840	Adverse incident outcome categories	4231813	Serious Adverse INCIDENTS
Transient abnormality unnoticed by the patient	4266810	397882007	Observation	Clinical Finding	442840	Adverse incident outcome categories	4231813	Serious Adverse INCIDENTS
Transient abnormality with full recovery	4162217	398056004	Observation	Clinical Finding	442840	Adverse incident outcome categories	4231813	Serious Adverse INCIDENTS

Table: The MSDA data dictionary for the minimal dataset within OMOP.

MS DATA ALLIANCE

Mapping dictionary

- The next step in the mapping preparation is the allocation of the registry variables to the MDS variables.
- This was done in collaboration with the two pilot registries.

Variable (Non-standard)	Formular	Item	Rückgabewert
Patient Specific Information			
Date of Birth	mnpmsfp1stammdaten	brthdtc	
Date of Death			
Gender	mnpmsfp1stammdaten	sex	
Feminine gender			2
Masculine Gender			1
Gender unknown			
Non-binary gender			
Disease Specific Information			
Date of onset	mnpmsfp1stammdaten	mhsydtc1	
Date of MS diagnosis	mnpmsfp1stammdaten	mhstdtc1	
MS type	mnpmsfp1verlauf	mslauf	
CIS			1
Relapsing remitting MS			2
Secondary progressive MS			3
Primary progressive MS			4
EDSS (value)	mnpmsfp1verlauf	edsstot	
EDSS (date)	mnpmsfp1verlauf	edssdat	
Relapses (count?)	mnpmsfp1verlauf	rel_cnt_sinc_visit	
Relapses (date)	mnpmsfp1schubereignis	schax	
Glucocorticoid (y/n)	mnpmsfp1schubereignis	steroid_oral, steroid_parenteral	
treatment	1		
No treatment	0		

Fig.: Excerpt from the Mapping dictionary of the German MS registry

MS DATA ALLIANCE

Data structures of the German and UK MS registry

- „White Rabbit“ (OHDSI tool) was used to perform an analysis scan of the data structure of the pilot registries.
- The output from that scan is a scan report that gives an overview of the tables, field variables and their characteristics.

	A	B	C	D	E	F	G	H
1	Table	Field	Type	Max lengt	N rows	N rows ch	Fraction empty	
2	casenodes_MSFP1_20190917-142811.csv	mnppid	int	4	-1	76	0	
3	casenodes_MSFP1_20190917-142811.csv	mnpctrid	int	3	-1	76	0	
4	casenodes_MSFP1_20190917-142811.csv		empty	0	-1	76	1	
5								
6	centres_MSFP1_20190917-142811.csv	mnpctrid	int	3	-1	2	0	
7	centres_MSFP1_20190917-142811.csv	mnpctrname	varchar	12	-1	2	0	
8	centres_MSFP1_20190917-142811.csv	mnpcname	varchar	11	-1	2	0,5	
9	centres_MSFP1_20190917-142811.csv		empty	0	-1	2	1	
10								
11	deactivatedcodes_MSFP1_20190917-142811.csv	formtablename	empty	0	-1	4	1	
12	deactivatedcodes_MSFP1_20190917-142811.csv	column	empty	0	-1	4	1	
13	deactivatedcodes_MSFP1_20190917-142811.csv	lookuptable	varchar	16	-1	4	0	
14	deactivatedcodes_MSFP1_20190917-142811.csv	value	int	3	-1	4	0	
15	deactivatedcodes_MSFP1_20190917-142811.csv	label	varchar	9	-1	4	0	
16	deactivatedcodes_MSFP1_20190917-142811.csv	version	varchar	12	-1	4	0	
17	deactivatedcodes_MSFP1_20190917-142811.csv	versiondate	varchar	19	-1	4	0	
18	deactivatedcodes_MSFP1_20190917-142811.csv		empty	0	-1	4	1	
19								
20	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	mnppid	int	4	-1	9	0	
21	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	mnpdocid	int	5	-1	9	0	
22	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	mnpsubdocid	int	5	-1	9	0	
23	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	fgid	int	6	-1	9	0	
24	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	position	int	1	-1	9	0	
25	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_concomitant_agent	empty	0	-1	9	1	
26	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_conco_start	empty	0	-1	9	1	
27	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_conco_end	empty	0	-1	9	1	
28	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_conco_ongoing	empty	0	-1	9	1	
29	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_conco_dose	empty	0	-1	9	1	
30	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_conco_unit	empty	0	-1	9	1	
31	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	ae_conco_indication	empty	0	-1	9	1	
32	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv	aerelconco	empty	0	-1	9	1	
33	emnpmsfp1ae_conco_med_MSFP1_20190917-142811.csv		empty	0	-1	9	1	
34								
35	emnpmsfp1ae_immunmodther_MSFP1_20190917-142811.csv	mnppid	int	4	-1	7	0	

Fig.: Excerpt from the Scan Report of the German MS registry

MS DATA ALLIANCE

Preparations for the harmonisation process (with „Rabbit in a Hat“)

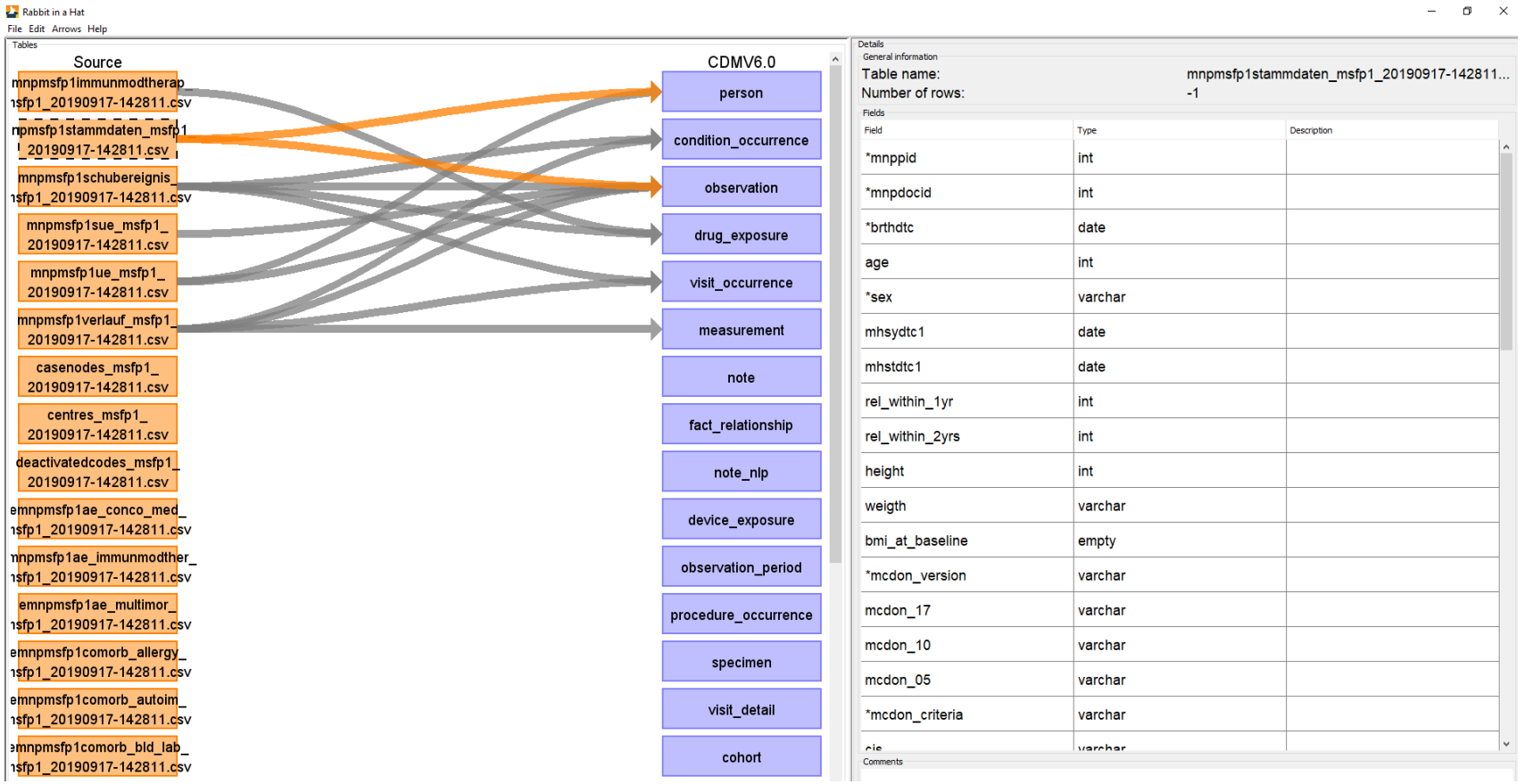


Fig.: table level mappings

Preparations for the harmonisation process (with „Rabbit in a Hat“)

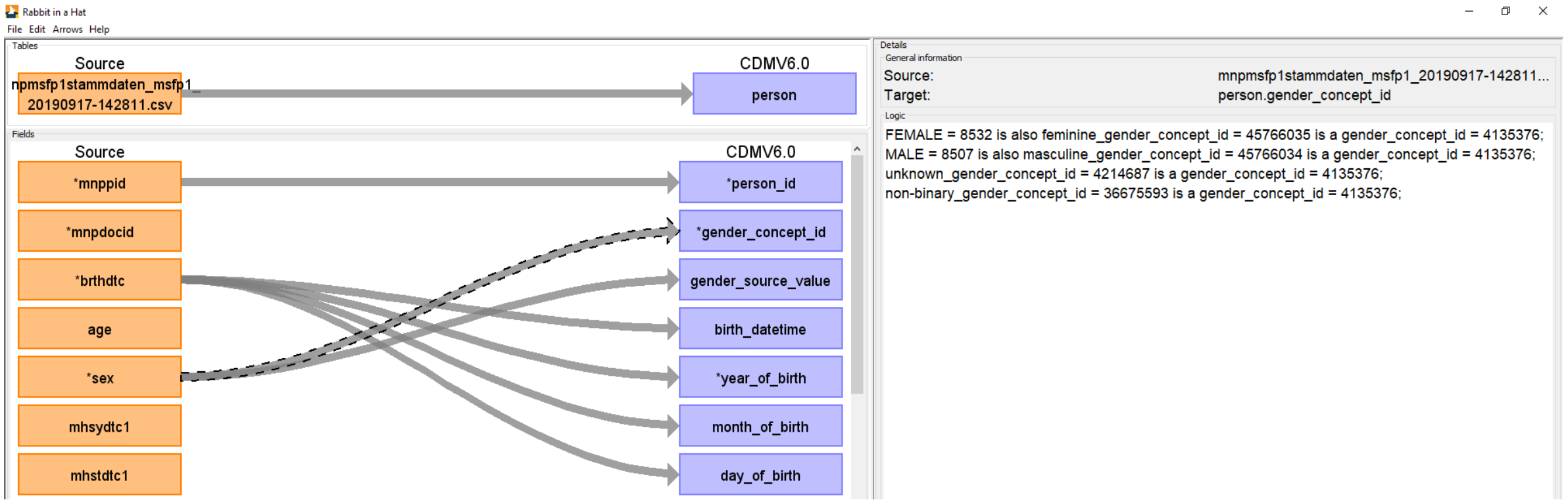
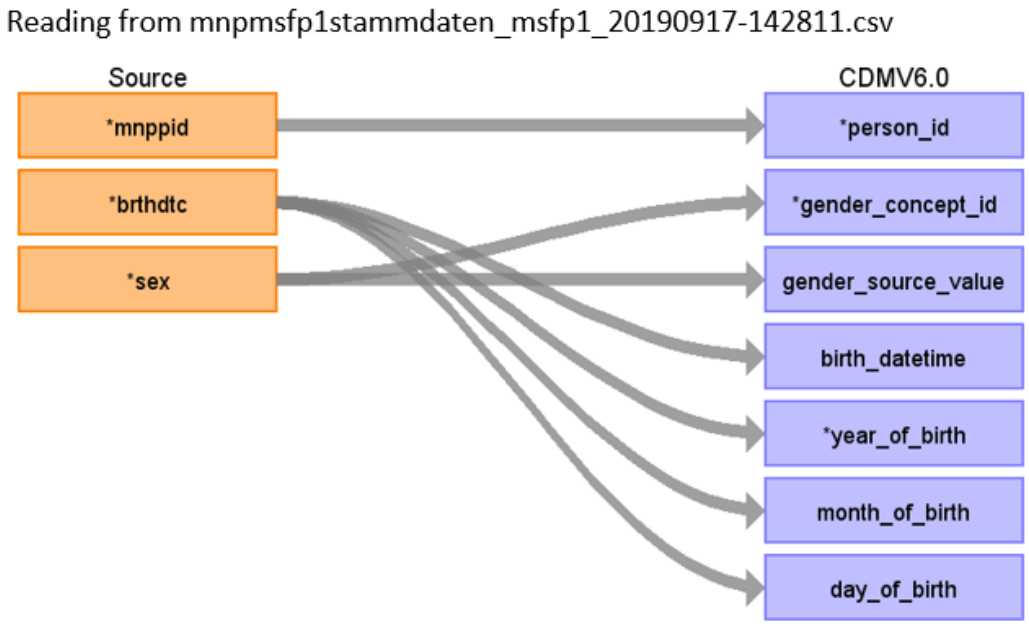


Fig.: variable level mappings

Preparations for the harmonisation process (with „Rabbit in a Hat“)



Destination Field	Source Field	Logic	Comment
person_id	mnppid		

gender_concept_id	sex	FEMALE = 8532 is also feminine_gender_concept_id = 45766035 is a gender_concept_id = 4135376;
		MALE = 8507 is also masculine_gender_concept_id = 45766034 is a gender_concept_id = 4135376;
		unknown_gender_concept_id = 4214687 is a gender_concept_id = 4135376;
		non-binary_gender_concept_id = 36675593 is a gender_concept_id = 4135376;	

gender_source_value	sex	2 = FEMALE 1 = MALE unkown_gender is n. a. non-binary_gender is n. a.	
birth_datetime	brthdtc	date_of_birth = 4083587 (concept_ID) = birth_datetime --> consists of year_of_birth and month_of_birth and day_of_birth (if applicable) here: "brthdtc" from mnpmsfp1stammdaten in	

Fig.: ETL document containing all mappings and their respective logic

MS DATA ALLIANCE

The implementation of the data harmonisation has started…

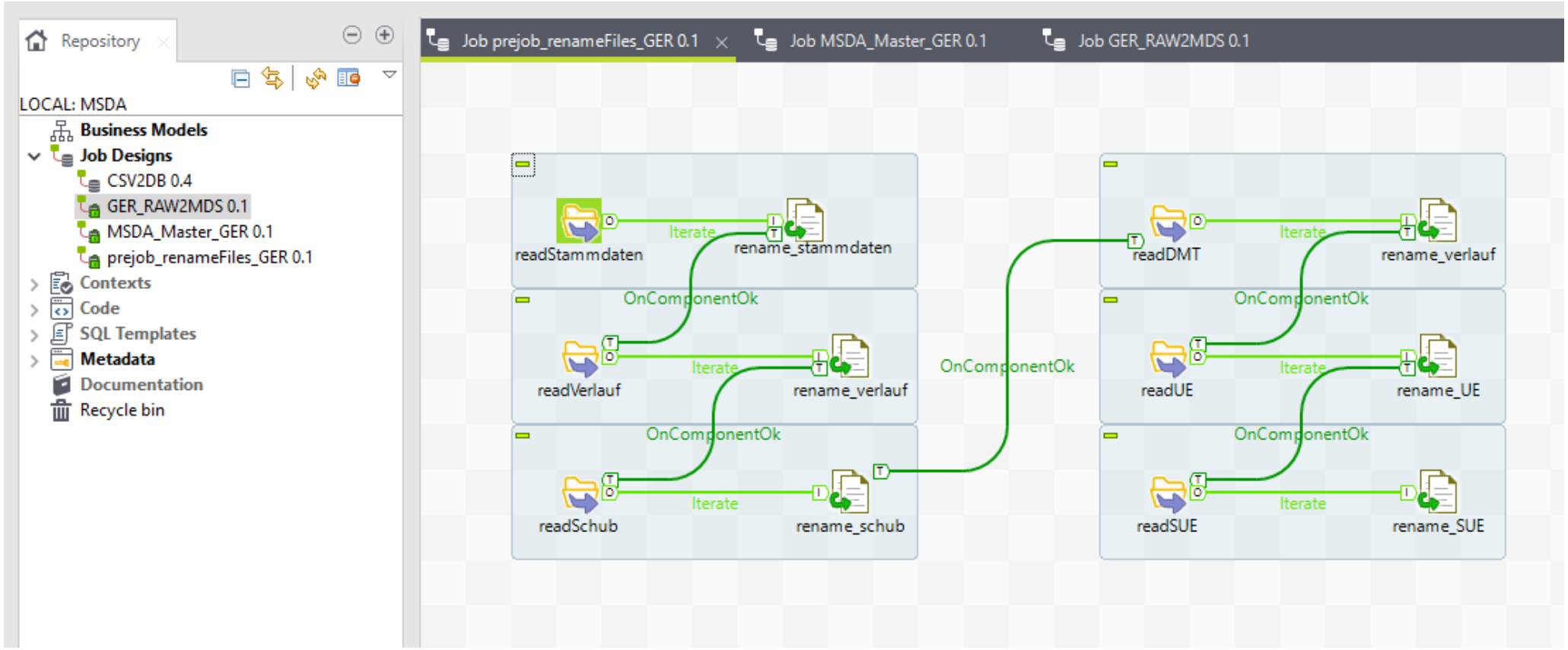


Fig.: Selecting relevant tables from the German MS registry and renaming them for future generalisation

MS DATA ALLIANCE

Next steps…

Until the end of the pilot year (beginning of 2020) we aim to:

- Prepare the mapping procedures for each OMOP CDM table for both pilot registries.
- Update the MSDA SwitchBox Concept with each step to document the mapping process.
- Implement the mapping of the raw data to OMOP CDM for at least one of the two pilot registries.
- At the end of the pilot year: Update the MSDA SwitchBox Concept in regards to its applicability to other registry mappings. Document lessons learned from the pilot year.

MS DATA ALLIANCE

And then?

In the upcoming years we‘d like to focus on:

- Mapping all interested registries/cohorts to the OMOP CDM formatted minimal dataset.
- Continuously improving the user-friendliness and decrease the workload of our mapping procedures.
- Expanding (or adapting) our minimal dataset based on the strategic input provided in the dicussions and by the feedback.

A first formal evaluation on our data harmonisation is planned during one of our stakeholder focus groups.

Tina Meißner

University Medical Center Göttingen
Department of Medical Informatics
Von-Siebold-Str. 3
37075 Göttingen

Tel.: +49 551 3961506
Mail: tina.meissner@med.uni-goettingen.de
Or add me on linkedin!