Overview of NHS England Dedicate Work

v14/04/2023

Contents

Generic guidance for developing data lineage

Section overview

1. Data lineage development method
2. The metadata model
3. Assets
4. Relationships
    - Technology assets
    - Logical attributes and columns
    - Parties
    - Rules, mapping specifications and field mapping
    - Rules, mapping specifications and crosswalks
5. Lineage examples using the Civil Registration Deaths data set

Data lineage development method

The quality of secondary uses data can be understood in terms of flow through the data pipeline and the technological context in which it sits at any point in time. Data goes through various stages associated with landing the data, preparing the data and making it available for dissemination, which mean that the disseminated data may be different to the submitted data. In addition, the data will work its way through different aspects of the technology estate which contributes to understanding of context and further transformations the data may have gone through. A firm understanding gained through clearly documented metadata at each stage of the process is key to understanding the voracity of the disseminated data.

Development of data lineage for any given data set can be undertaken in several stages:

1. Identify key teams/contacts in the submission to dissemination pipeline and wider system
2. Acquire key specification and schema information
3. Establish where key versions of the data existed within the technical infrastructure
4. Identify where key transformations to the data took place
5. Establish additional dependencies such as use of specifications to develop the pipeline
6. Analyse the Central Metastore model and make amendments as required to effectively store the relevant metadata (with clear consideration of other unrelated requirements), including development of outstanding technical infrastructure
7. Ingest the relevant schemas and specifications, including associated rules, code lists and mapping
8. Create pertinent example lineage diagrams to illustrate the flow of data through the pipeline and provide context to the stages involved

The Metadata Model

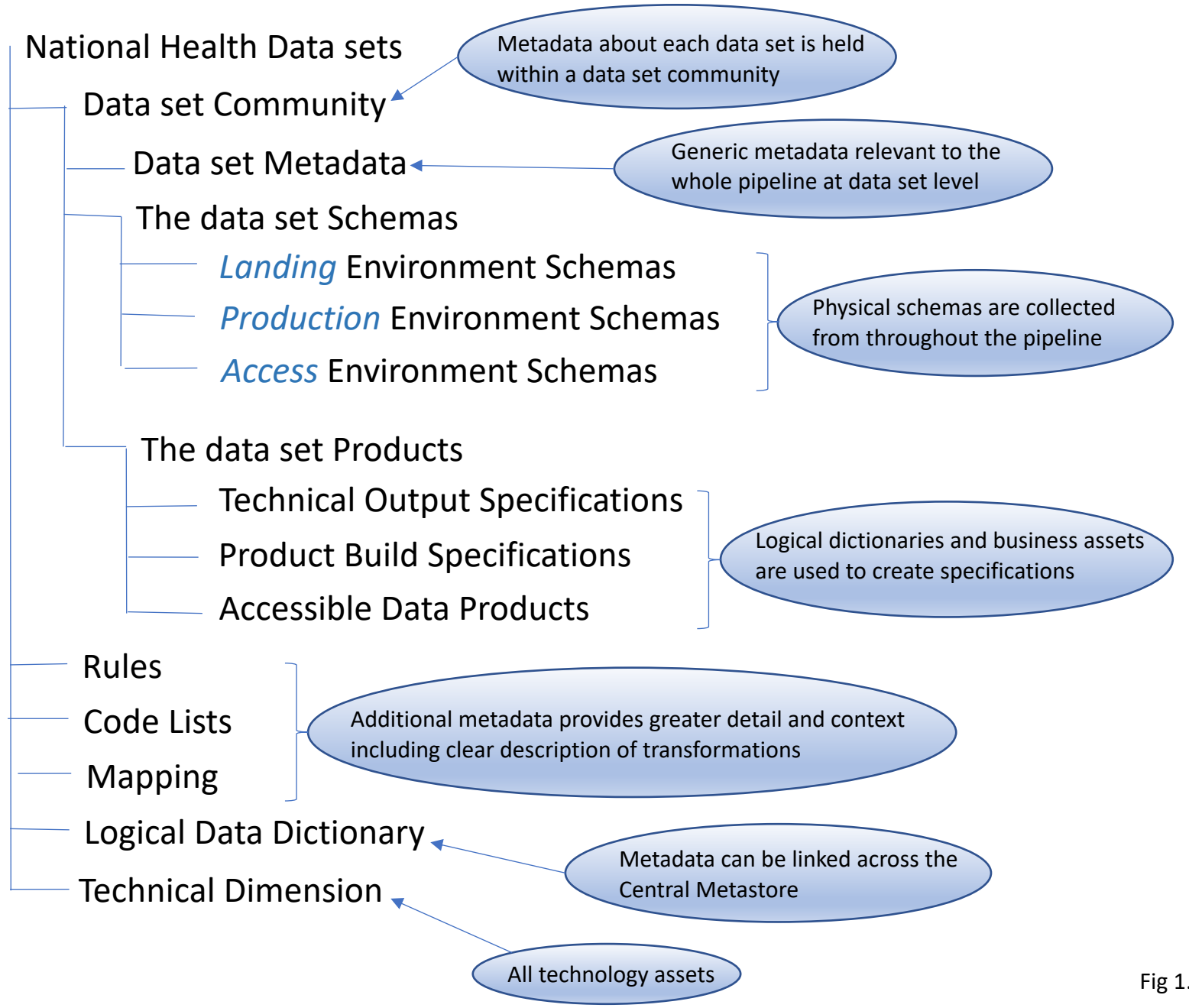


Fig 1.0

Developing a Lineage Template - Assets

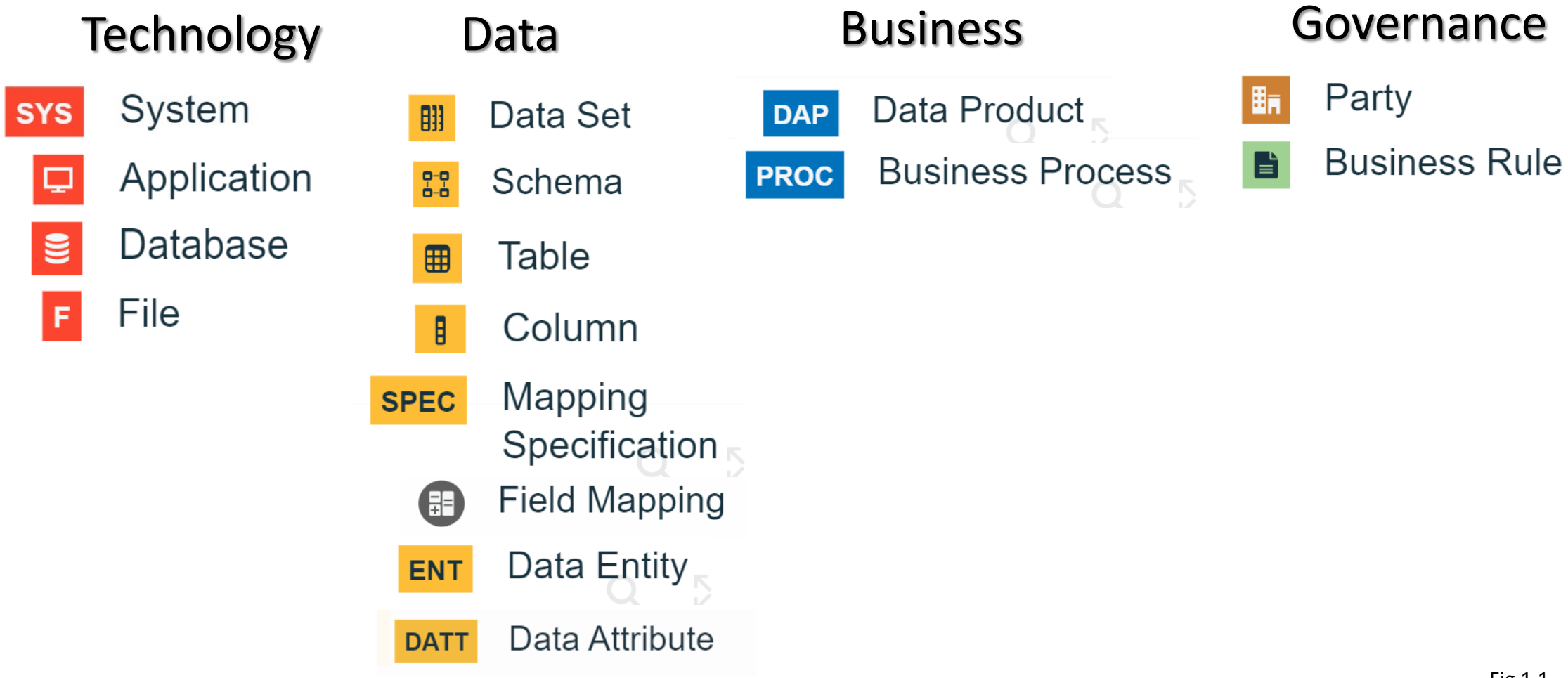


Fig 1.1

Technology assets

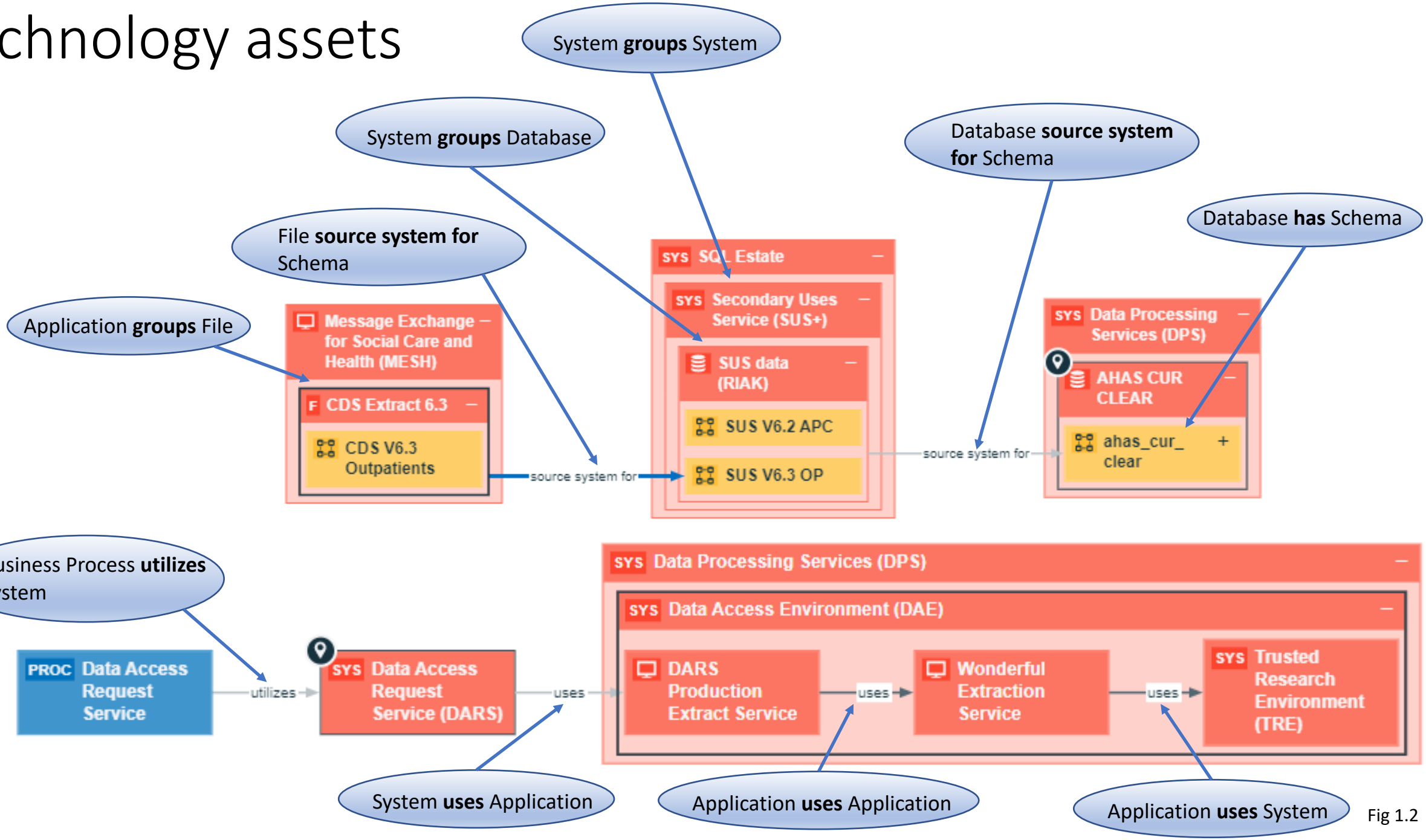


Fig 1.2

Logical attributes and columns

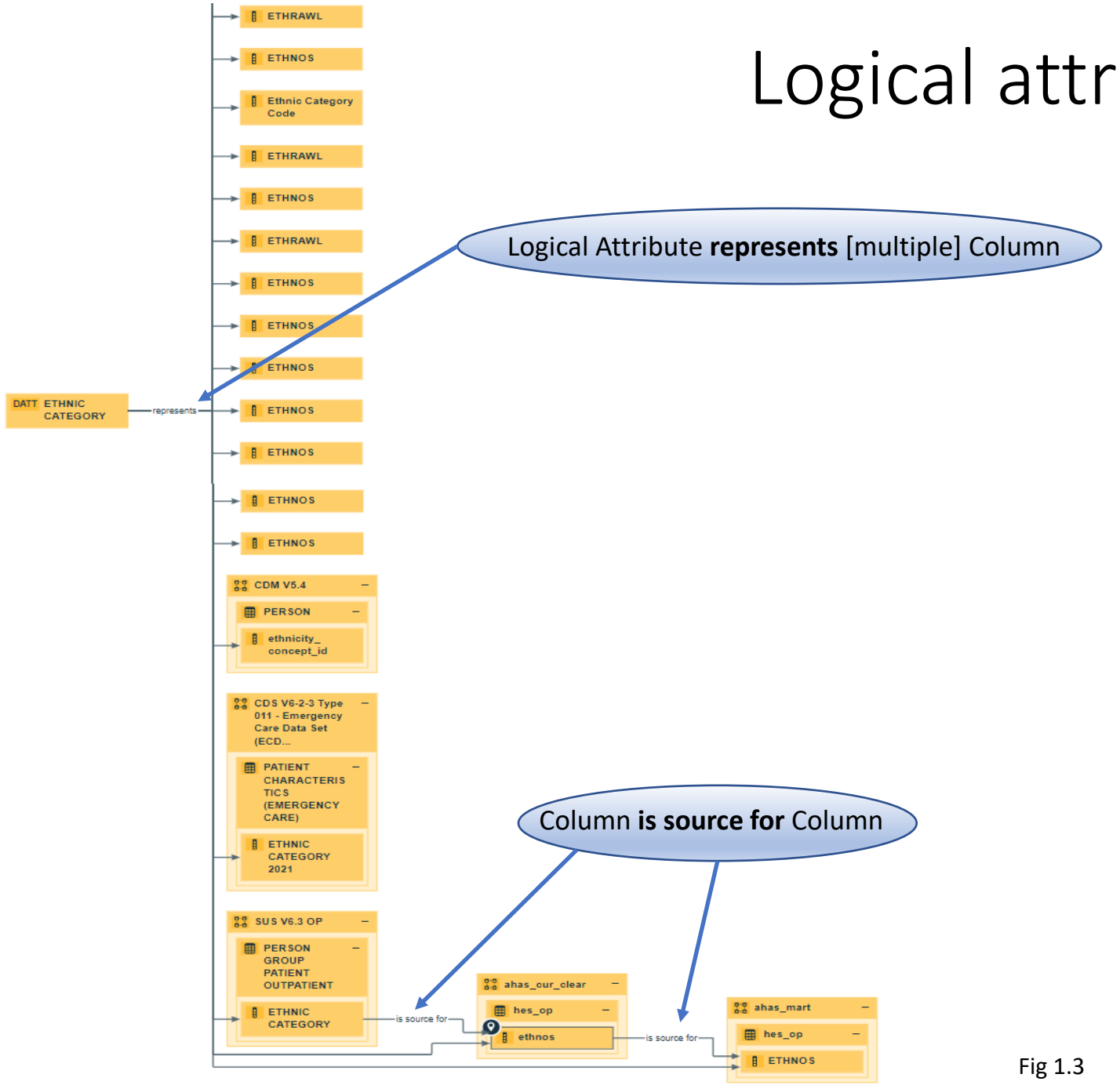
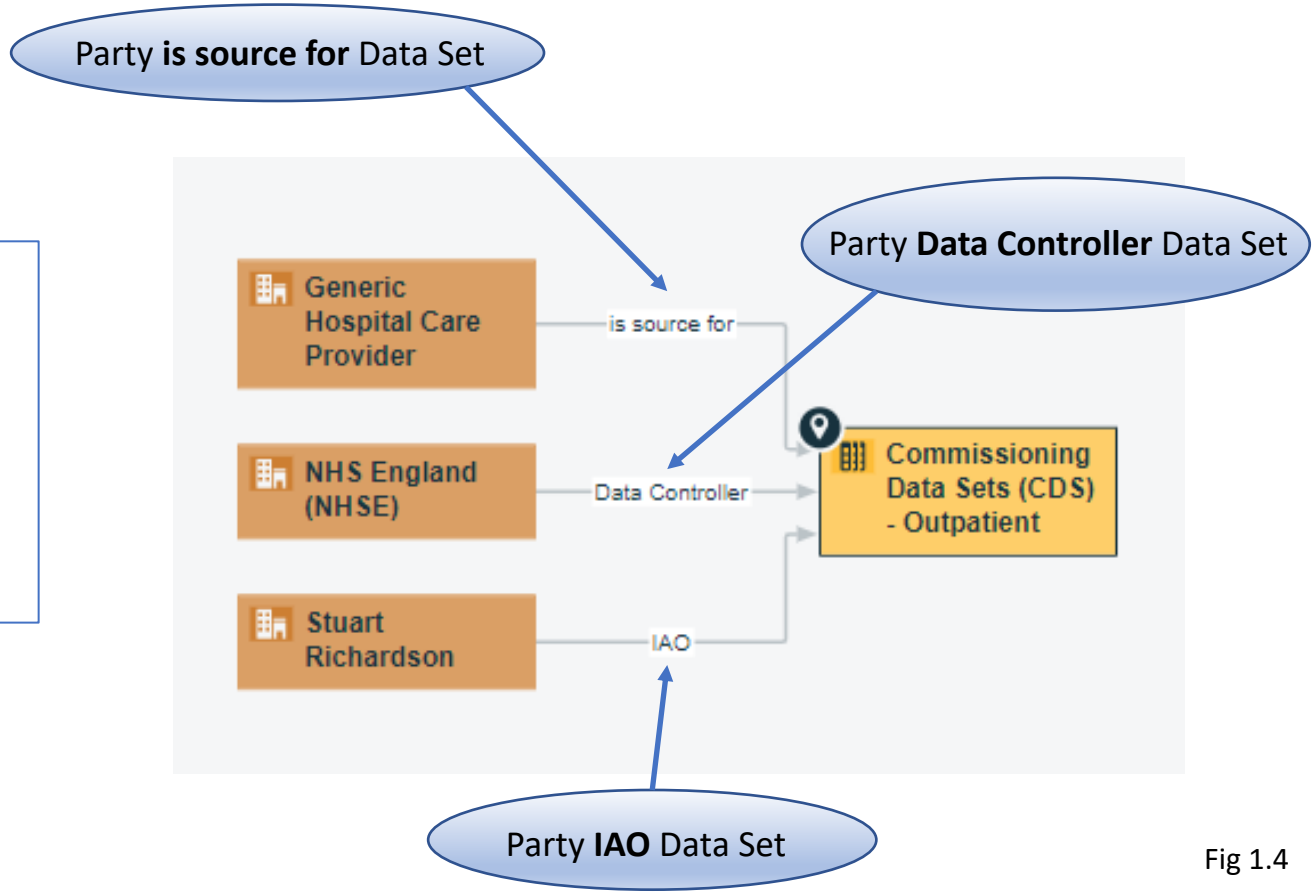


Fig 1.3

Parties

Parties describe key individuals or organisations. A party is likely to relate to more than one data set. Using a relationship to make the link allows a single party asset to be reused.



The Data Set contains key high level metadata that should be relevant to the associated data wherever it appears in the pipeline. It also groups the logical attributes that describe any associated physical columns.

Fig 1.4

Rules, mapping specifications and field mapping

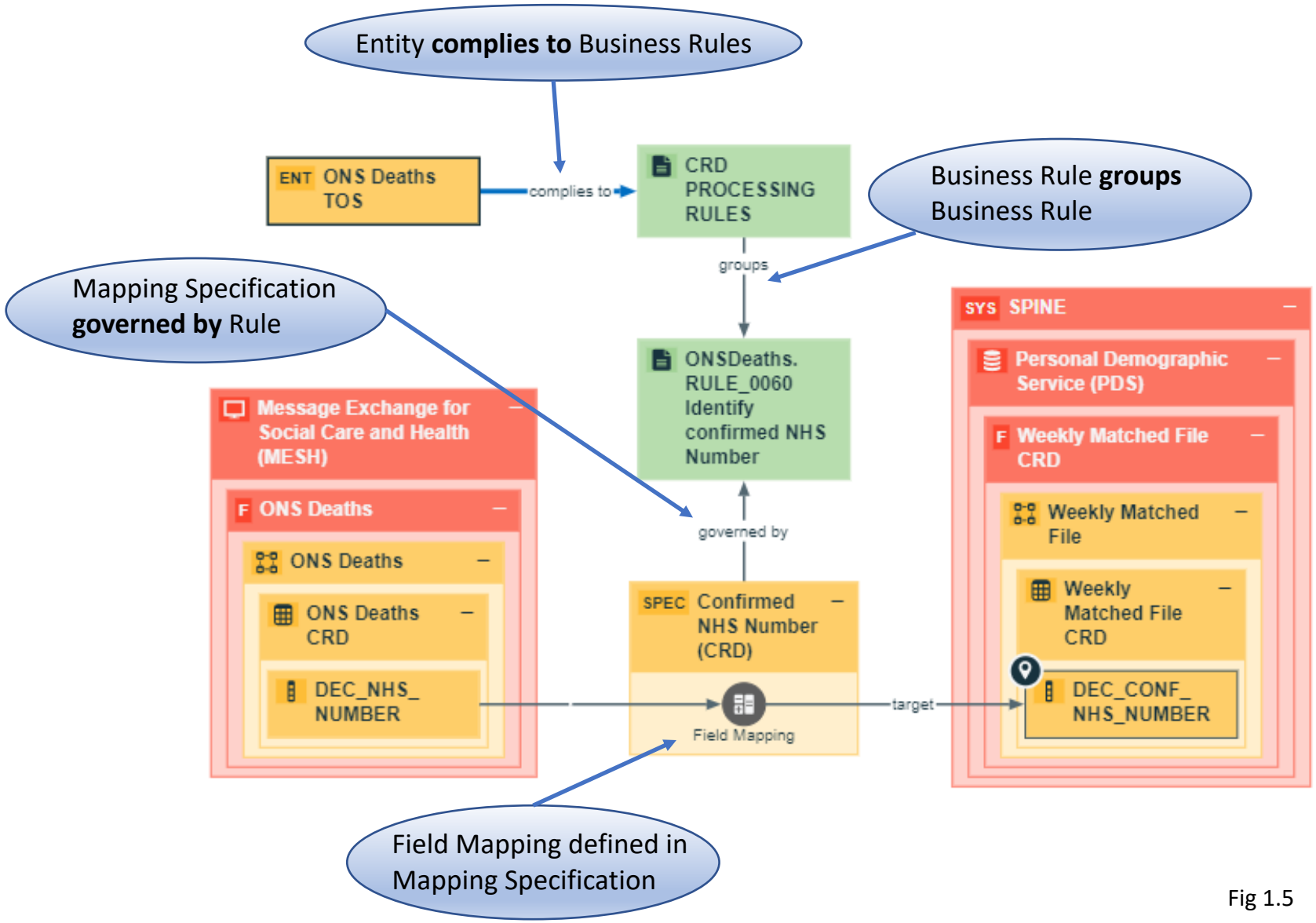


Fig 1.5

Rules, Mapping Specifications and Crosswalks

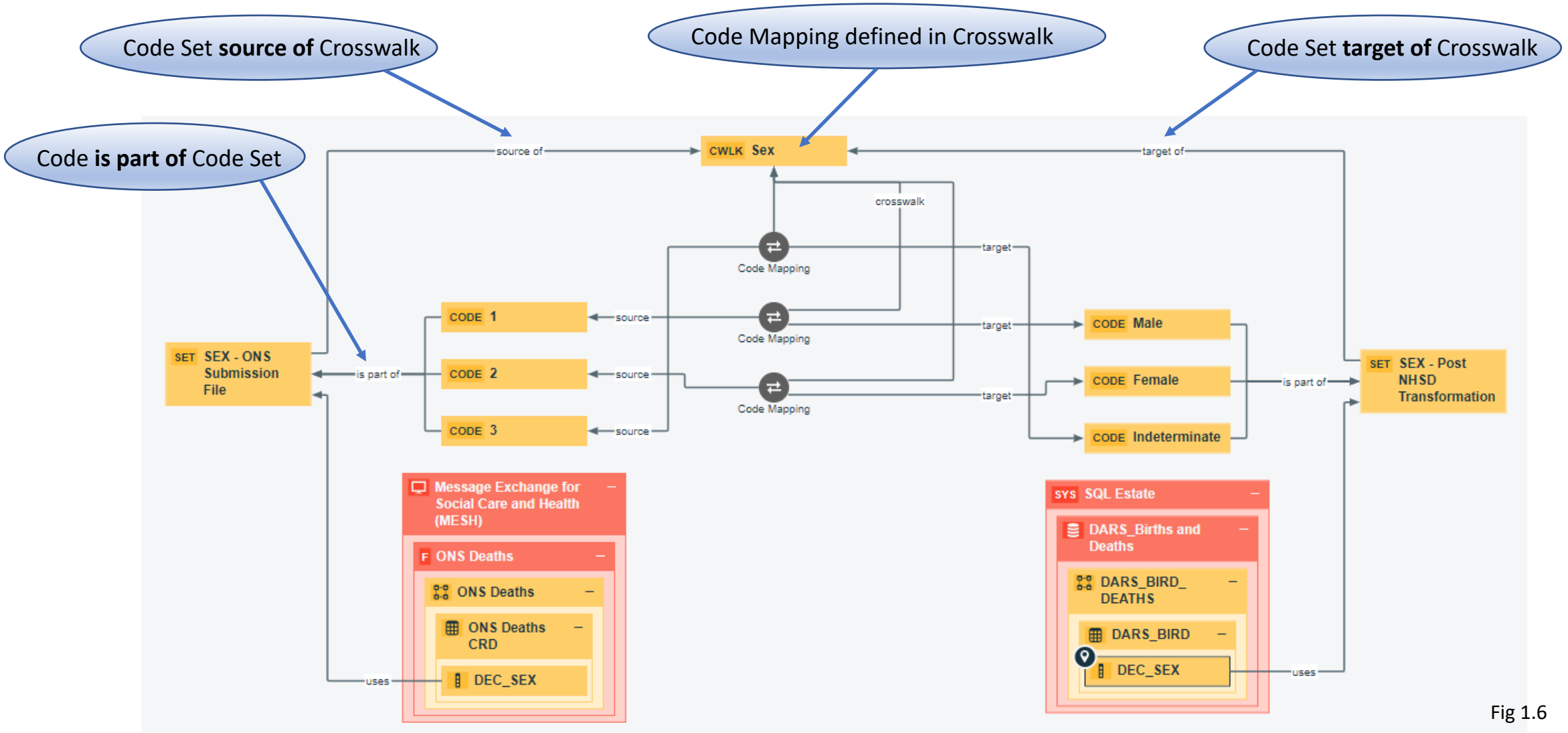


Fig 1.6

Guidance for developing data lineage using Collibra

Section overview

1. Stage 1: Technology asset domain
2. Stage 2: Creating a data model
3. Stage 3: Rules and mapping specifications

Stage 1: Technology Asset domain

This includes creating the various technical components of the overall process flow involved in deriving the data lineage for the data set at the highest level of granularity.

The major Asset Types used for the data set e.g. HES APC

a. System
b. Application
c. Database

Steps -

1. Create a Technology Asset Domain. *(Need to be done only once for all data sets together.)*
2. Once created, individual technology assets can be created by clicking on the **+** sign on the top right corner of the Collibra tool and selecting the type of Asset to be created.
3. Once the asset is created, add in the details like description, location, version, relation etc.
4. If an asset is a part of another technology asset, assign the relation **"is grouped by Technology Asset"**
5. If an asset contains any schema, use relation "**has schema**"
6. If any asset is source to a different Data Asset, use relation **"source system for Data Asset"**
7. Explore the "Add Characteristics" to add more relations if any

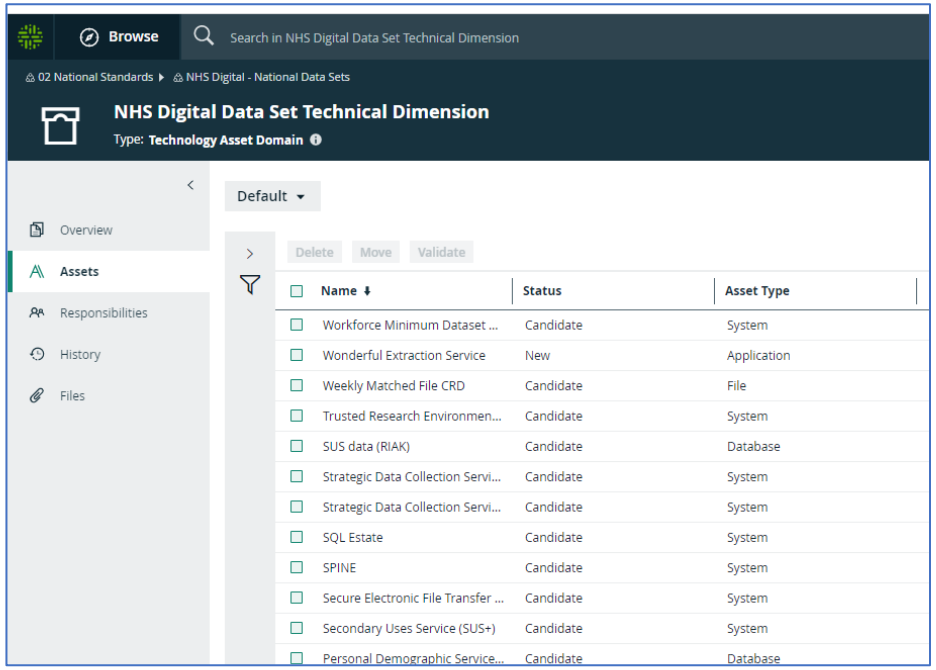


Fig 2.0

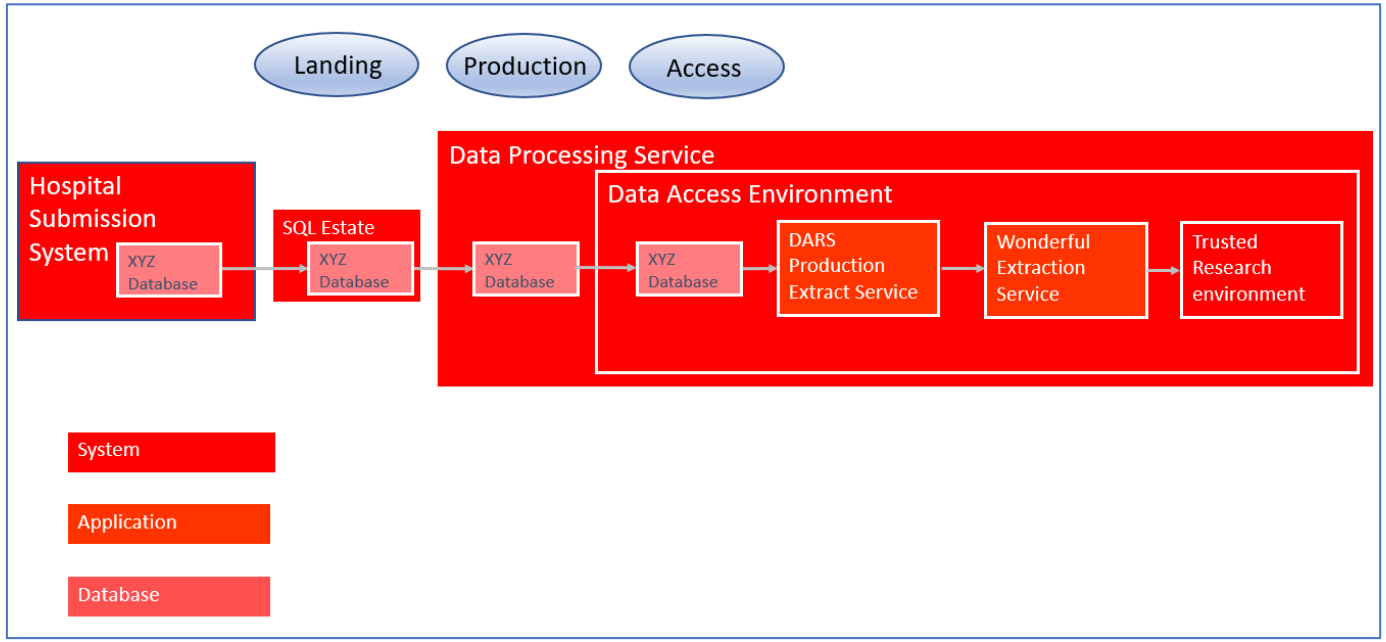


Fig 2.1

Stage 2: Creating a Data Model

1. Use specification and physical metadata to create multiple models to describe the data journey
   - Submission/Technical Output Specifications are modelled as Logical Data Dictionaries using Entities and Attributes
   - Platform Schemas are modelled as Physical Data Dictionaries using Schemas, Tables and Columns
   - Product descriptions such as the Data Access Request Service Product Specification found in the Civil Registration Deaths Accessible Data domain in the screenshot below are modelled as Business Assets using Data Products, Data Domains (not to be confused with 'Domain types') and Data Concepts.
2. Start creating the individual assets based on the technical specifications document received for every individual data model. These can also be created by the **+** sign. But this is not a feasible option when there are 200 or more columns to be created. It might also not capture all the required properties/relations of all the assets
3. If so, use the Entity Attribute template, Schema Table Column template or Product Domain Concept template as appropriate at the following link.
   https://nhsdigital.collibra.com/asset/f95da9ff-f7e6-4387-8081-7e75b8a29b82?tabbar=Files
4. Add more vertical columns to the spreadsheets to add more characteristics/properties to your fields. The templates will also help in assigning relationships across tables and columns

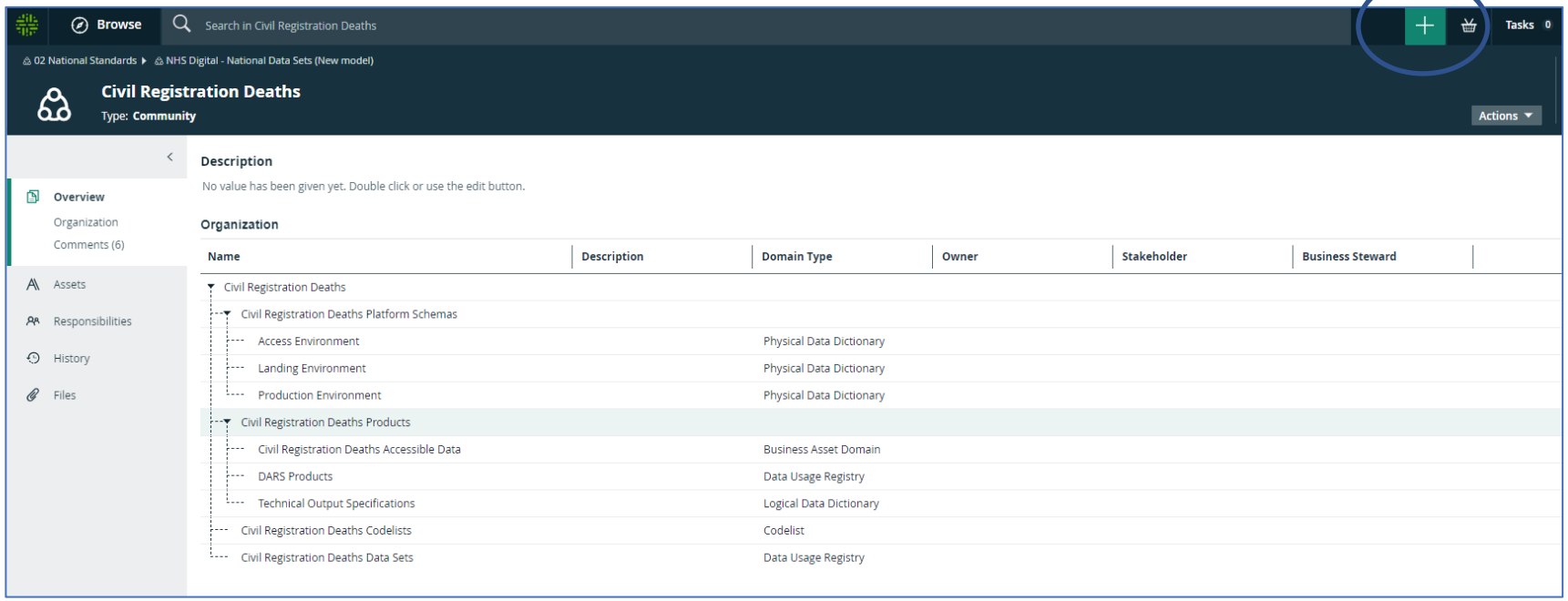


Fig 2.2

Stage 3: Rules and Mapping Specifications

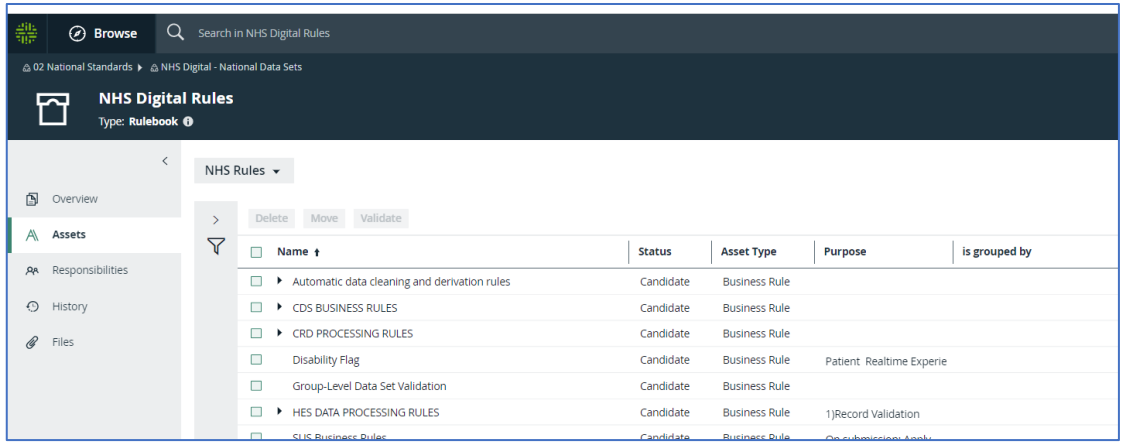


Fig 2.3

Creating Rulebooks for the data sets. These are basically the Business Rules and Data Quality rules (Used for various purposes like validation, derivation, quality checks etc.) applied at the various stages of data flow.

This is a 2 step process –

a) Creating placeholders for the rules. Creating actual rules with details.
b) Mapping the rules to the associated fields. Meaning which rules are applicable to which columns of a particular table.

Steps -

1. Use Rules Ingest Template to create the rules hierarchy and define the association with mapping specifications. There can be nested rules as well. (Note – The "**applies to**" fields can be used to apply the rules to specific assets like schemas, tables etc.) https://nhsdigital.collibra.com/asset/f95da9ff-f7e6-4387-8081-7e75b8a29b82?tabbar=Files
2. Relationships to other specifications such as the Submission Technical Output Specification can be defined manually from the associated asset.
3. Use Field Mapping Ingest Template to create the field mapping.

Cont'd . . .

Stage 3: Rules and Mapping Specifications (cont'd)

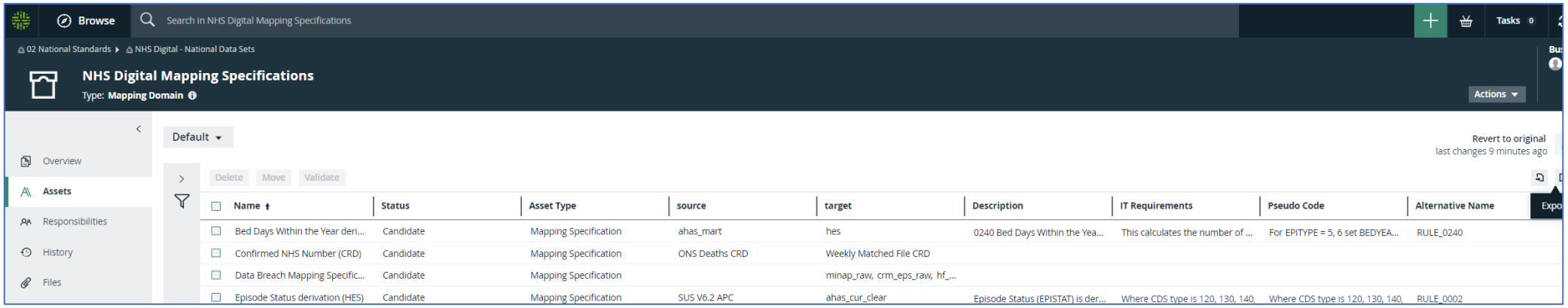


Fig 2.4

Mapping specs can be created to show a flow/relation/mapping between multiple components/assets with complex relations.

Example – There can be some transformation logic/rules applied during the transition or flow of data between the SUS and HES MART databases. Mapping specifications can be created to show the relationship between the data structures along with the transformations.

The following template creates Mapping specs, including pseudo code , source and target data structure information https://nhsdigital.collibra.com/asset/f95da9ff-f7e6-4387-8081-7e75b8a29b82?tabbar=Files

****Be Careful****

Updating detailed information for columns/tables etc. via template imports can cause unnecessary updates/deletes/additions. Keep a close check on the options selected while importing the templates and do a test import.

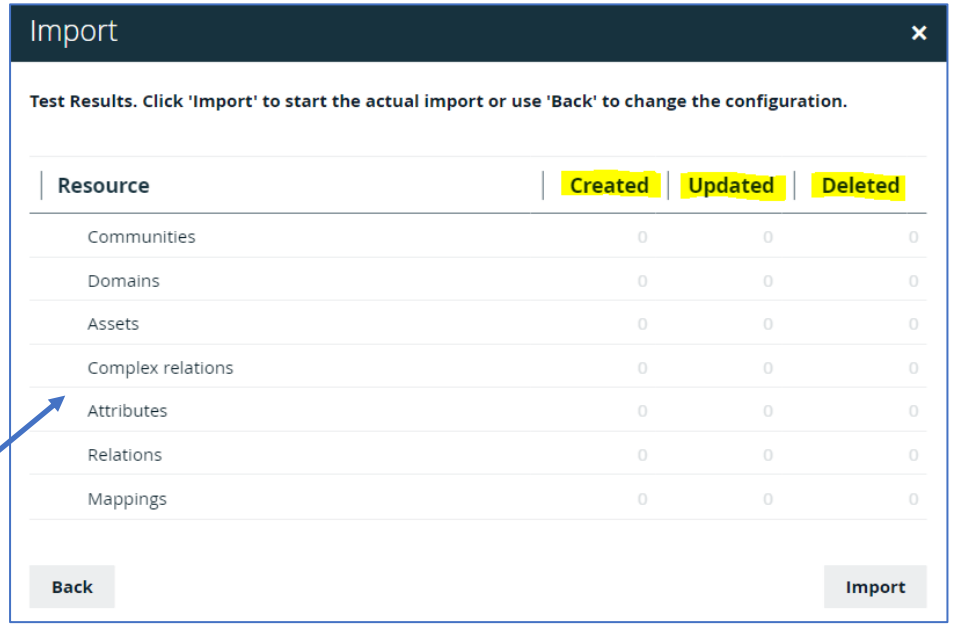


Fig 2.5

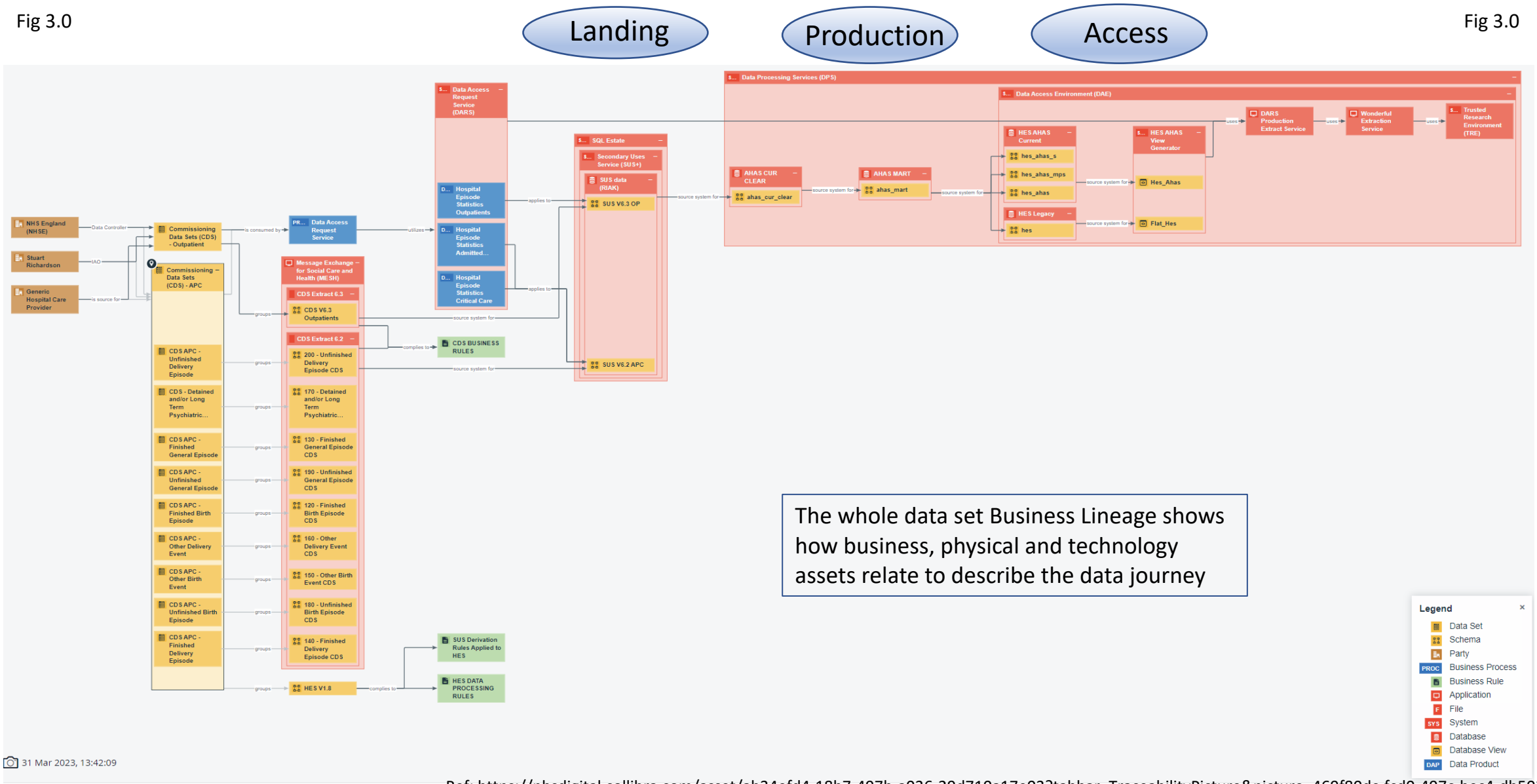
Lineage examples using the Hospital Episode Statistics data set

Section Overview

1. HES APC, OP and CC whole data set Business Lineage summary view
2. HES APC, OP and CC whole data set Business Lineage partially expanded view
3. HES APC field level lineage – Example Derivation
4. HES OP field level lineage – Example Validation
5. HES CC field level lineage – Example Removal

HES APC, OP and CC whole data set Business Lineage summary view

Fig 3.0

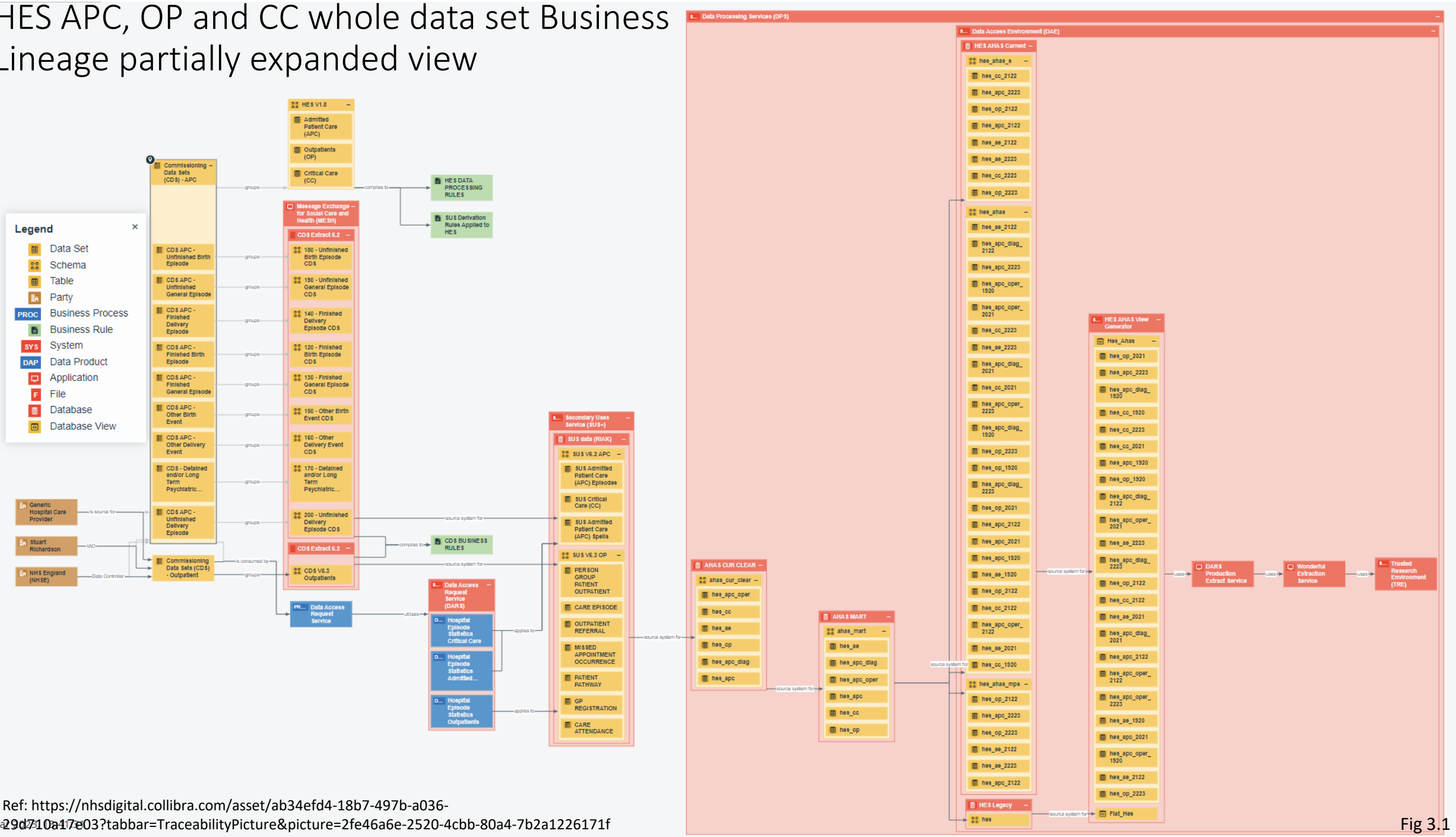


The whole data set Business Lineage shows how business, physical and technology assets relate to describe the data journey

Fig 3.0

Ref: https://nhsdigital.collibra.com/asset/ab34efd4-18b7-497b-a036-29d710a17e03?tabbar=TraceabilityPicture&picture=469f80dc-fcd0-497c-bec4-db50310c9b1b

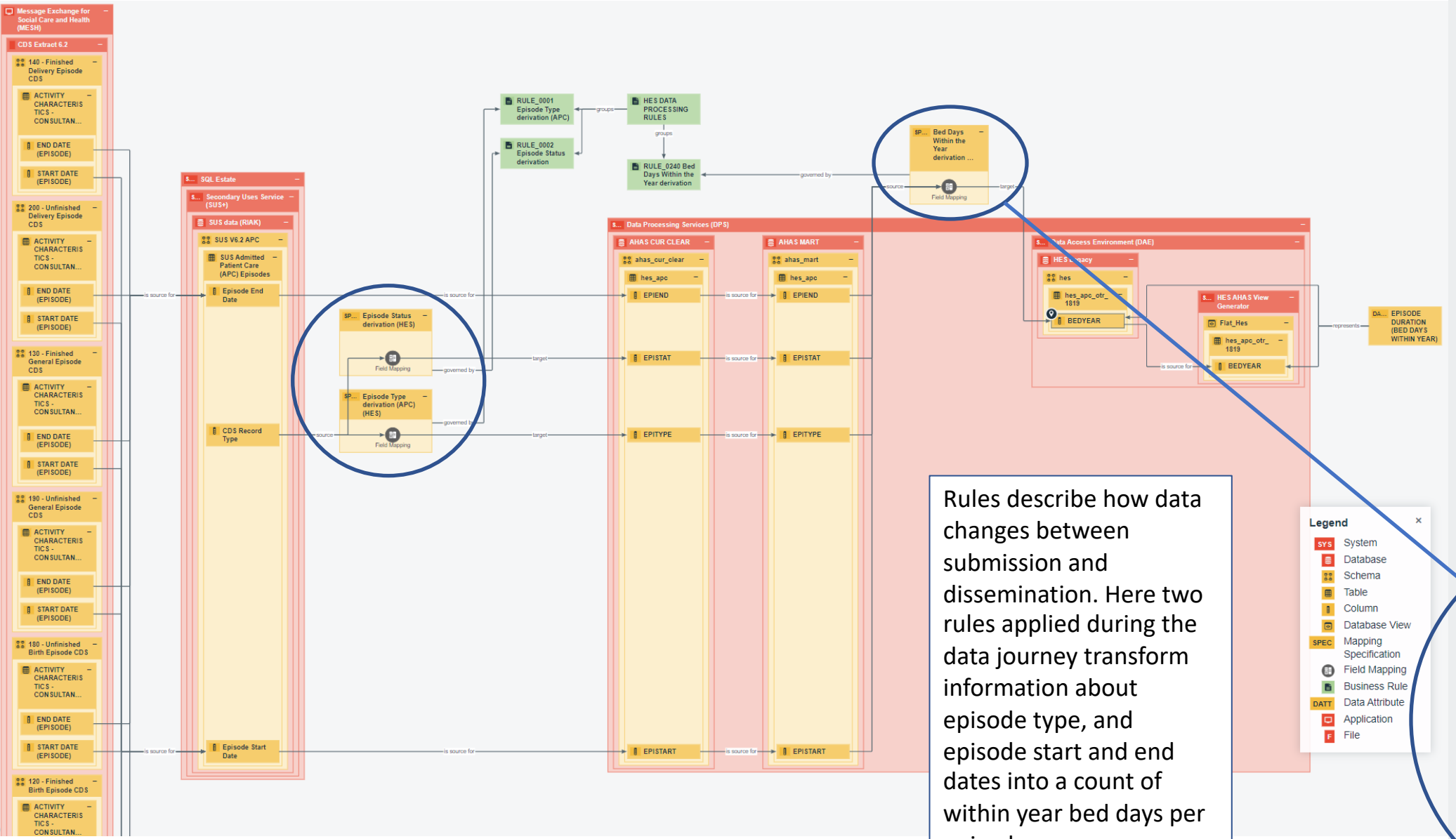
HES APC, OP and CC whole data set Business Lineage partially expanded view



Ref: https://nhsdigital.collibra.com/asset/ab34efd4-18b7-497b-a036-29d710a17e03?tabbar=TraceabilityPicture&picture=2fe46a6e-2520-4cbb-80a4-7b2a1226171f

Fig 3.1

HES APC field level lineage – Example Derivation



Rules describe how data changes between submission and dissemination. Here two rules applied during the data journey transform information about episode type, and episode start and end dates into a count of within year bed days per episode.

SPEC Bed Days Within the Year derivation (HES)

Candidate

Created on 2 Aug 2022, 15:43:00

Last modified on 20 Mar 2023, 10:07:38

Description

0240 Bed Days Within the Year derivation

IT Requirements

This calculates the number of bed days that an episode used during the current data year. It is calculated by subtracting the Episode Start Date (EPISTART) from the Episode End Date (EPIEND). If the EPISTART is prior to the beginning of the data year, then the EPISTART is replaced by the HES year start date. BEDYEAR is not valid for record types that are 'other maternity events'.

source Data Structure

ahas_mart

target Data Structure

hes

Pseudo Code

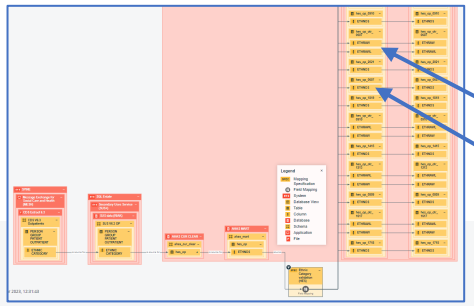
For EPITYPE = 5, 6 set BEDYEAR = null For EPITYPE <= 4 If EPISTART is null or not valid then set BEDYEAR = null If EPISTART is valid and < year start date, set startdate = year start date, else startdate = EPISTART If EPIEND is valid and > period end date, set enddate = period end date, else set enddate = EPIEND If EPISTAT = 3 (ie the record is finished) If EPIEND is null or not valid or if enddate < startdate then set BEDYEAR = null else BEDYEAR = enddate – startdate If EPISTAT = 1, 9 (unfinished) if EPISTART <= period end date then BEDYEAR = period end date + 1 – startdate (in days)

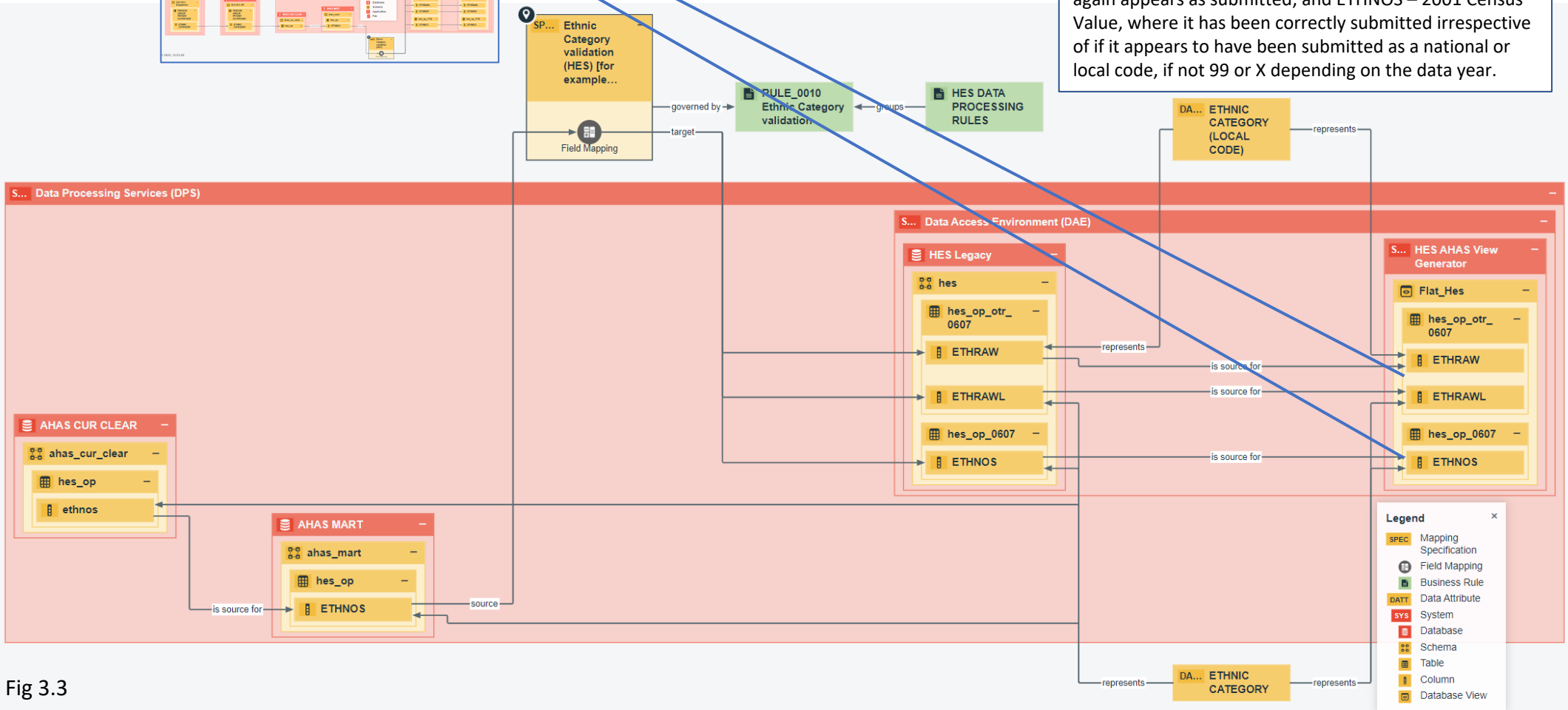
Fig 3.2

Ref: https://nhsdigital.collibra.com/asset/6c9dcf6a-c4f5-4003-a1ca-575d3f0bf438?tabbar=TraceabilityPicture&picture=18f4a20d-bde7-4544-a582-47a3df76a81b

HES OP field level lineage – Example Validation

In this example the submitted Ethnic Category data is flowed through the pipeline as: ETHRAW – data that should be submitted as one of the recognised 2001 census codes but appears as submitted, even if incorrect; ETHRAWL – an additional local code that can be submitted as part of the original Ethnic Category Data, again appears as submitted; and ETHNOS – 2001 Census Value, where it has been correctly submitted irrespective of if it appears to have been submitted as a national or local code, if not 99 or X depending on the data year.





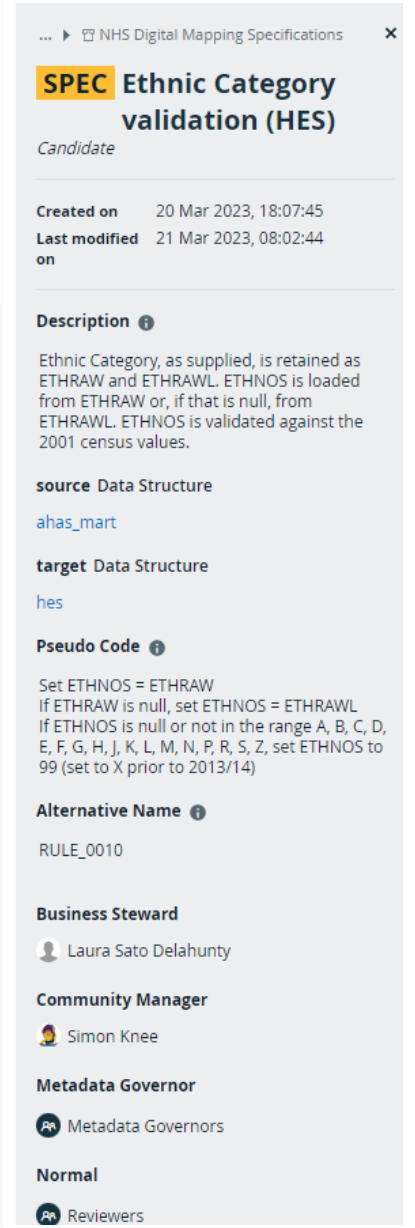


Fig 3.3

Ref: https://nhsdigital.collibra.com/asset/d1903569-d023-49b0-b1b7-639ca48d0590?tabbar=TraceabilityPicture&picture=759138b6-e198-4f25-8560-1ea1e4ea8485
https://nhsdigital.collibra.com/asset/fc6a05a2-0bb2-4523-aa8f-6966bc8bff0f?tabbar=TraceabilityPicture&picture=fab61d08-1b70-4ecc-8239-27507bcaebf6

HES CC – Example Removal

The rule here requires removal of orphaned HES CC records i.e. where they do not link to an supporting APC record

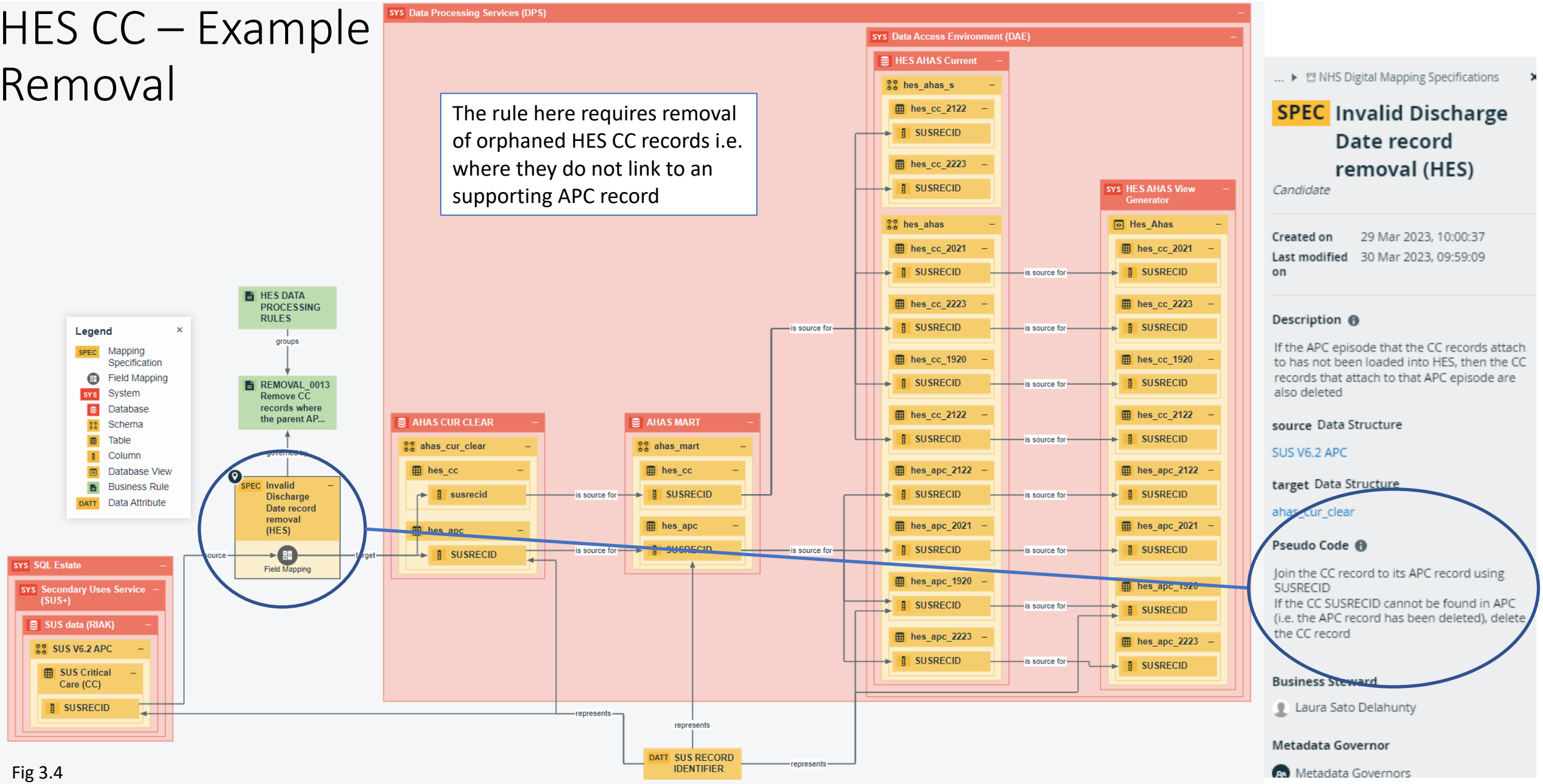


Fig 3.4

NHS Digital Mapping Specifications

SPEC Invalid Discharge Date record removal (HES)

Candidate

Created on 29 Mar 2023, 10:00:37
Last modified on 30 Mar 2023, 09:59:09

Description

If the APC episode that the CC records attach to has not been loaded into HES, then the CC records that attach to that APC episode are also deleted

source Data Structure

SUS V6.2 APC

target Data Structure

ahas_cur_clear

Pseudo Code

Join the CC record to its APC record using SUSRECID
If the CC SUSRECID cannot be found in APC (i.e. the APC record has been deleted), delete the CC record

Business Steward

Laura Sato Delahunty

Metadata Governor

Metadata Governors

Ref: https://nhsdigital.collibra.com/asset/10b3b7d3-c079-4cd4-ad1e-b4385cff4df4?tabbar=TraceabilityPicture&picture=e4c1fba2-eb40-44a4-8ebe-eee71c62a459

Lineage examples using the Civil Registration Deaths data set

Section Overview

1. CRD whole data set Business Lineage view
2. CRD field level lineage – Example Derivation
3. CRD field level lineage – Example Code List

CRD whole data set Business Lineage view

Landing

Production

Access

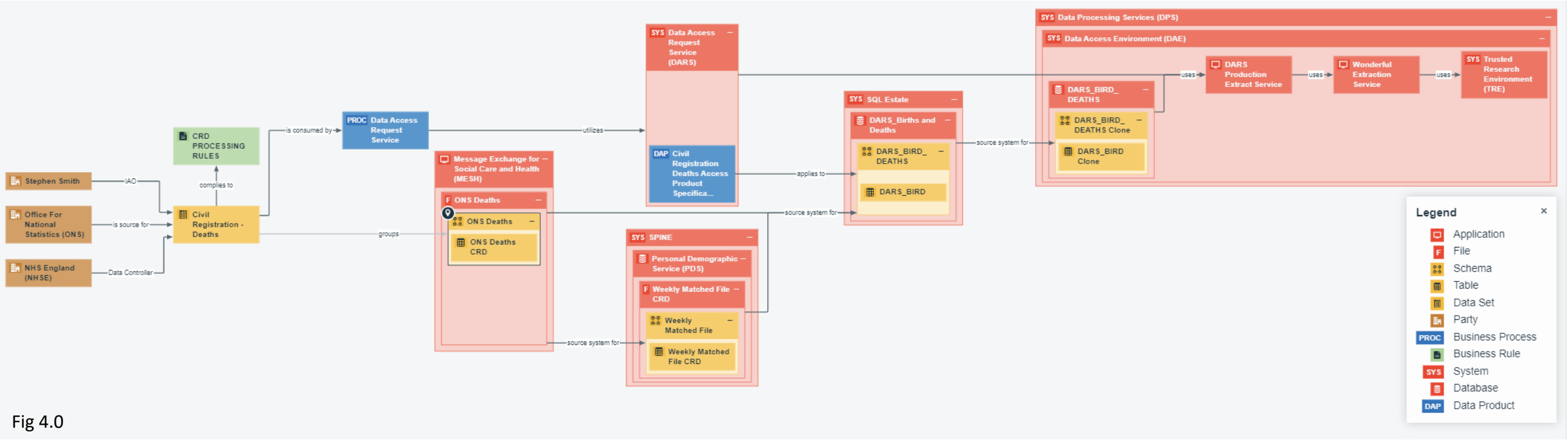


Fig 4.0

Ref: https://nhsdigital.collibra.com/asset/e035bad6-c023-4732-8e52-3dae35bc9579?tabbar=TraceabilityPicture&picture=92a8d7ae-0a89-4dcf-9954-61b523859004

CRD field level lineage – Example Derivation

The submitted NHS Number is made available to the Personal Demographic Service to confirm it validity. Two versions of the NHS Number continue through the pipeline, the submitted and the subsequently verified. In most cases these should be the same, but could vary if the submitted NHS number cannot be validated.

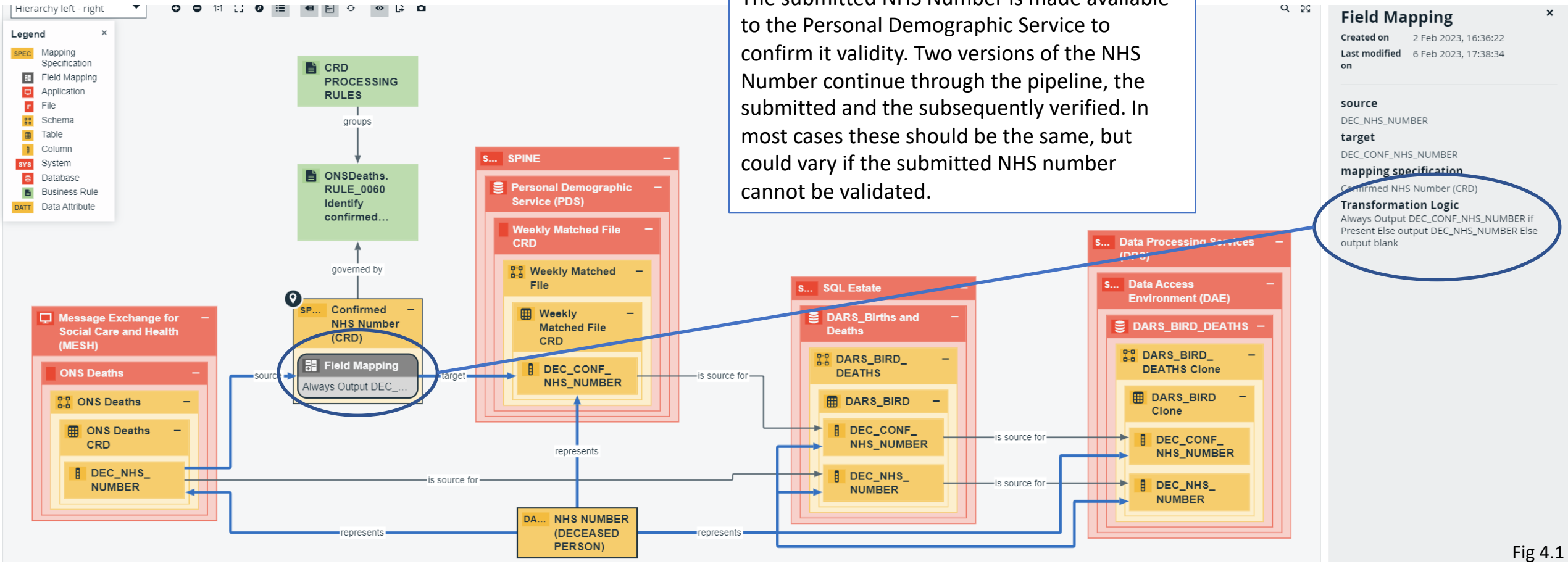


Fig 4.1

Ref: https://nhsdigital.collibra.com/asset/0f710bb4-9853-42ff-ad1b-4ad98b23709e?tabbar=TraceabilityPicture&picture=3f370d2f-5dc0-41a2-8140-790242a347e4

CRD field level lineage – Example Code List

Complex mapping (developed in the Central metastore as a 'Crosswalk') is used to present 'human readible' code descriptions to the user. The illustration shows a relatively simple 1:1 relationship between three codes, and the equivalent three descriptions. However this method can also be used to record greater complexity. For example: two codes in submitted data of 'Unstated' and 'Other', might map to a single 'Other' code in the disseminated NHSE data.

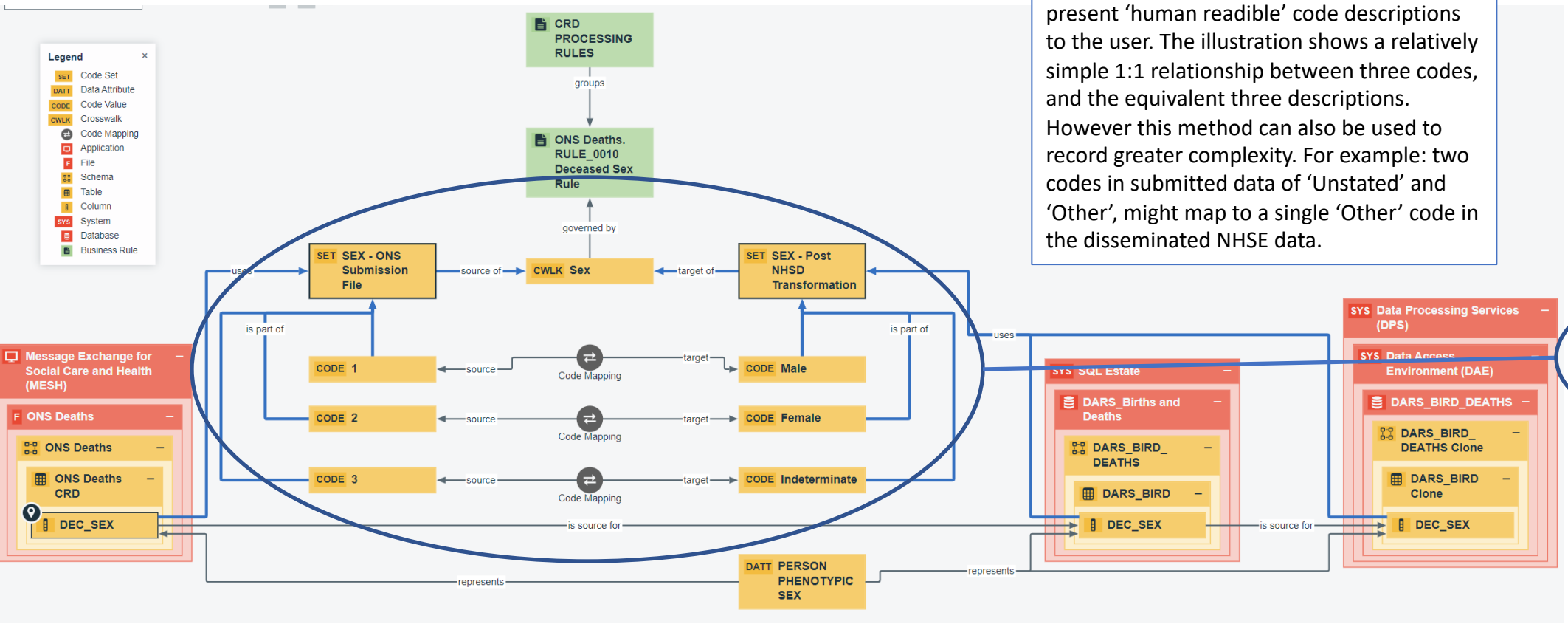


Fig 4.2

Ref: https://nhsdigital.collibra.com/asset/73df8b08-b38e-4dcc-811f-b547e544293b?tabbar=TraceabilityPicture&picture=cd891424-5d52-4482-a393-f100cacef932