Rule-Based Program and Datasets Relationship Mapping

Yizhen Chen 11/13/22024

Table of Contents

[Project Description 3](#_Toc182404526)

[Objectives 3](#_Toc182404527)

[Stakeholders 3](#_Toc182404528)

[Requirements Gathering 4](#_Toc182404529)

[Rule Development 4](#_Toc182404530)

[Program and Dataset’s Funding Source matching 5](#_Toc182404531)

[Data Fields 5](#_Toc182404532)

[Matching Rule 5](#_Toc182404533)

[Program and Dataset’s title matching 5](#_Toc182404534)

[Data Fields 5](#_Toc182404535)

[Matching Rule 5](#_Toc182404536)

[Program and Dataset’s PI matching (Yizhen Added) 6](#_Toc182404537)

[Data Fields 6](#_Toc182404538)

[Matching Rule 6](#_Toc182404539)

[Project and Dataset’s Org matching 6](#_Toc182404540)

[Data Fields 6](#_Toc182404541)

[Matching Rule 6](#_Toc182404542)

[Project and Dataset’s description matching 6](#_Toc182404543)

[Data Fields 7](#_Toc182404544)

[Matching Rule 7](#_Toc182404545)

[Grant and Dataset’s PI matching 7](#_Toc182404546)

[Data Fields 7](#_Toc182404547)

[Matching Rule 7](#_Toc182404548)

[Grant and Dataset’s Funding Source matching 7](#_Toc182404549)

[Data Fields 8](#_Toc182404550)

[Matching Rule 8](#_Toc182404551)

[Grant and Dataset’s Org matching 8](#_Toc182404552)

[Data Fields 8](#_Toc182404553)

[Matching Rule 8](#_Toc182404554)

[Grant and Dataset’s Org matching 8](#_Toc182404555)

[Data Fields 8](#_Toc182404556)

[Matching Rule 9](#_Toc182404557)

# Project Description

Develop a rule-based program to automate the mapping and relationship identification between datasets.

## Objectives

Enable accurate mapping and discovery of relationships between data entities based on predefined rules to support data integration, analysis, and curation.

## Stakeholders

Project Manager: Zack

Tech Leader: Yizhen

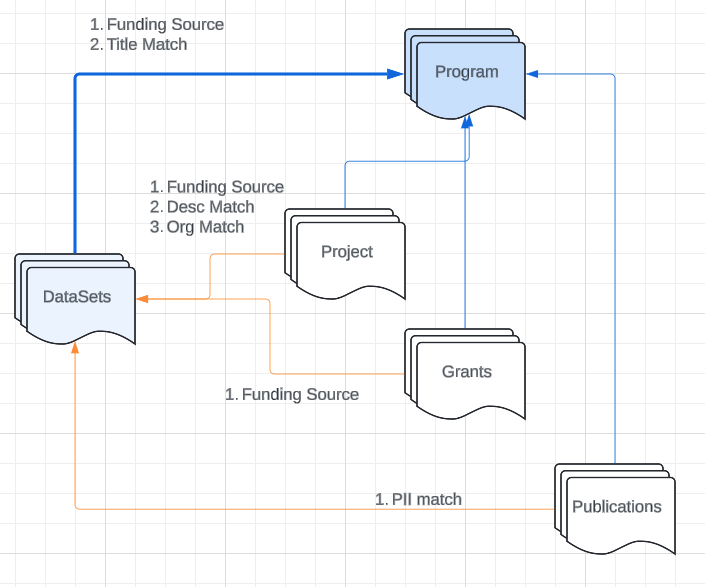
Data Curator: Ricardo

# Requirements Gathering

[2.0.0.4](https://nih.sharepoint.com/:f:/r/sites/NCI-CBIIT-FNLCCDI/Shared%20Documents/INS/2-Executing/Data%20Gathering/INS_v3.x/Data_to_Load/2.0.0.4?csf=1&web=1&e=yMnAiK)

We have Program, Project, Publication, Grant and dataset tsv file.

# Rule Development



## Program and Dataset’s Funding Source matching

Confident rate: Hight

### Data Fields

Program Fields:

Program.NOFOs

Program.Awards

DataSets Fields:

Description

funding\_source

### Matching Rule

If datasets’ fields include any item listed in Program fields (exact match, but case insensitive), then the program and dataset are considered related.

## Program and Dataset’s title matching

Confident rate: Medium

### Data Fields

Program Fields:

program\_name

program\_acronym

DataSets Fields:

Description

Title

### Matching Rule

If datasets’ fields include any item listed in Program fields (exact match, case insensitive), then the program and dataset are considered related.

## Program and Dataset’s PI matching (Yizhen Added)

Confident rate: Medium

### Data Fields

Program Fields:

Contact\_PI

DataSets Fields:

PI\_name

### Matching Rule

If datasets’ fields include any item listed in Program fields (exact match, case insensitive), then the program and dataset are considered related.

## Project and Dataset’s Org matching

Confident rate: Medium

### Data Fields

Project Fields:

project\_org\_name

DataSets Fields:

Description

### Matching Rule

If datasets’ fields include any item listed in Project fields (exact match, case insensitive), then the program and dataset are considered related.

## Project and Dataset’s description matching ML/AI

Confident rate: Medium

### Data Fields

Project Fields:

project\_abstract\_text

project\_title

DataSets Fields:

Description

Title

### Matching Rule

Sematic Matching, Similarity Matching, ML/AI

## Grant and Dataset’s PI matching

Confident rate: Medium

### Data Fields

Grant Fields:

principal\_investigators

program officers

DataSets Fields:

PI\_name

### Matching Rule

If datasets’ fields include any item listed in Grant fields (exact match, case insensitive), then the program and dataset are considered related.

## Grant and Dataset’s Funding Source matching

Confident rate: Hight

### Data Fields

Grant Fields:

grant\_opportunity\_number

DataSets Fields:

funding\_source

### Matching Rule

If datasets’ fields include any item listed in Grant fields (exact match, case insensitive), then the program and dataset are considered related.

## Grant and Dataset’s Org matching

Confident rate: Medium

### Data Fields

Grant Fields:

grant\_org\_name

DataSets Fields:

Description

### Matching Rule

If datasets’ fields include any item listed in Grant fields (exact match, case insensitive), then the program and dataset are considered related

## ~~Grant and Dataset’s Org matching~~

~~Confident rate: Medium~~

### ~~Data Fields~~

~~Project Fields:~~

~~grant\_org\_name~~

~~DataSets Fields:~~

~~Description~~

### ~~Matching Rule~~

~~If datasets’ fields include any item listed in Grant fields (exact match, case insensitive), then the program and dataset are considered related~~