Understanding Data and their Environment

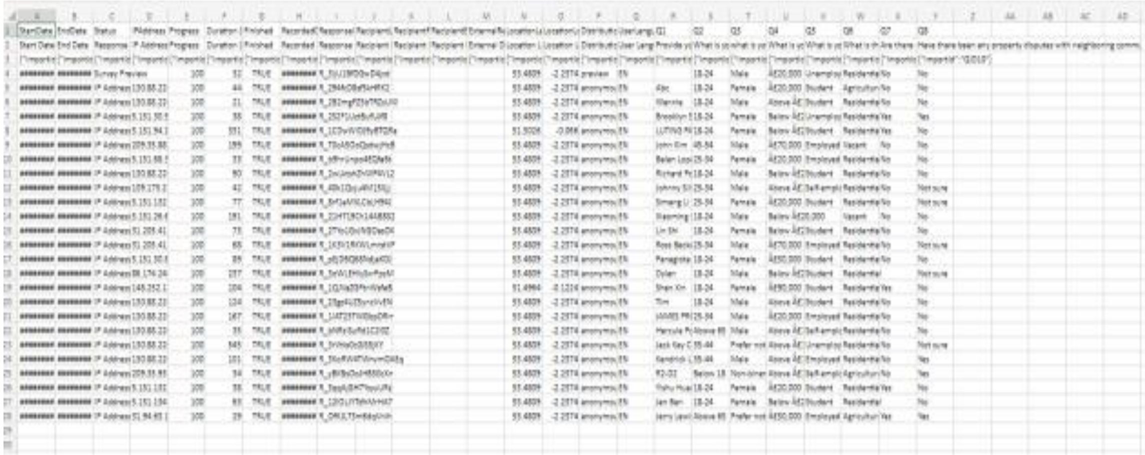
PROVENANCE REPORT
ID-*11349153*

DATA71011

ABSTRACT

Provenance plays an important role in the modern world of endless data. Very often we work with multiple data without knowing about its origin. It can lead to serious issues in critical projects where the source is crucial. That's where provenance is useful- it helps in tracking the flow of data from its origin thus ensuring transparency, consistency, completeness, and authenticity. We get the answers for which, when, and whom at every level of the workflow.

DATASET USED



Land Registry Dataset

DECIDING ENTITIES,ACTIVITIES,AGENTS FOR THE GRAPH

Before getting into how I chose entity, agent & and activity for this task, first let us investigate what those terms exactly mean.

Agent-One who is responsible for the process involving entities.

Entity-Data sources & outputs influenced by activities and agents.

Activity-Process or task being done by agents.

Accordingly, as per our final objective of the model and activities involved, I have chosen the following

- Agents- Alice, Bob, Charlie, and Mike who are members of our team(coral) and team Beige. Also professor was involved in this so we consider them agents.

- Entity- Topic,survey_questions, both the datasets used, indicator results and data,valid license holder column,income column, metadata and its issues.

- Activity- Analysis, creating survey questions, discussion about metadata, creating both the datasets, deriving metadata from datasets, Addressing loopholes in metadata.These all can be considered as activities/processes used in the model.

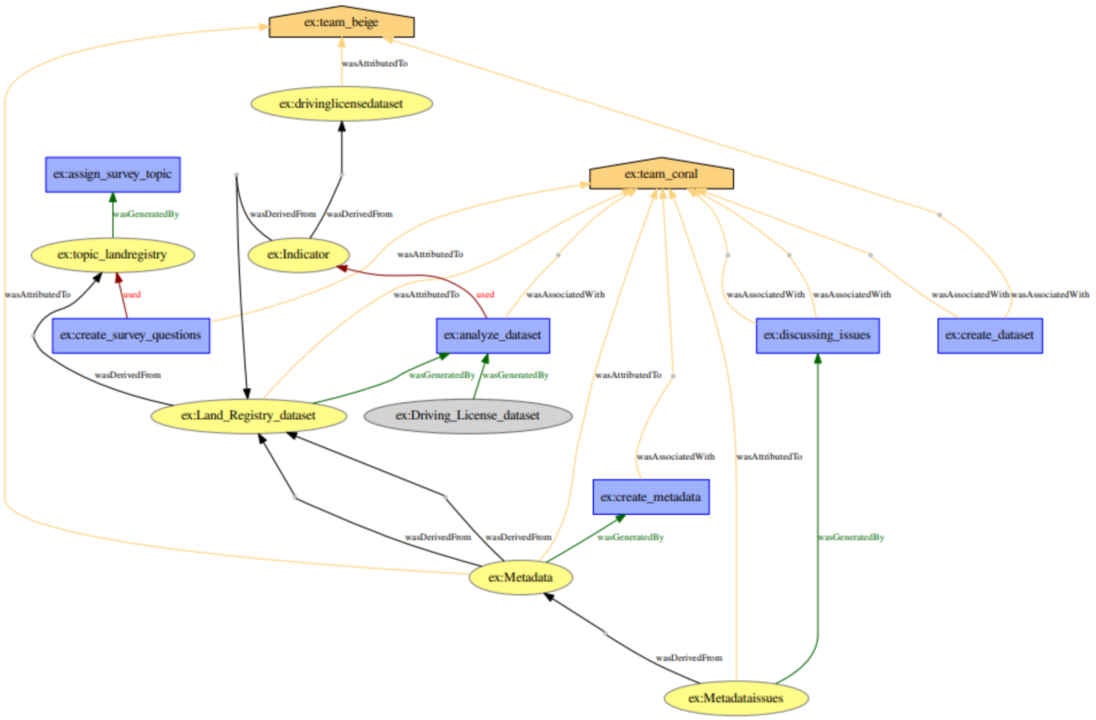
- Identifiers and types -I have kept identifiers which describes clearly the agents,activities and entities used in graph as per their definitions.

I have not included agents like "qualtrics" which was used for survey and then detailed activities about how we calculated the indicator.

PROVENANCE GRAPH

Lower Granularity (GENERIC)

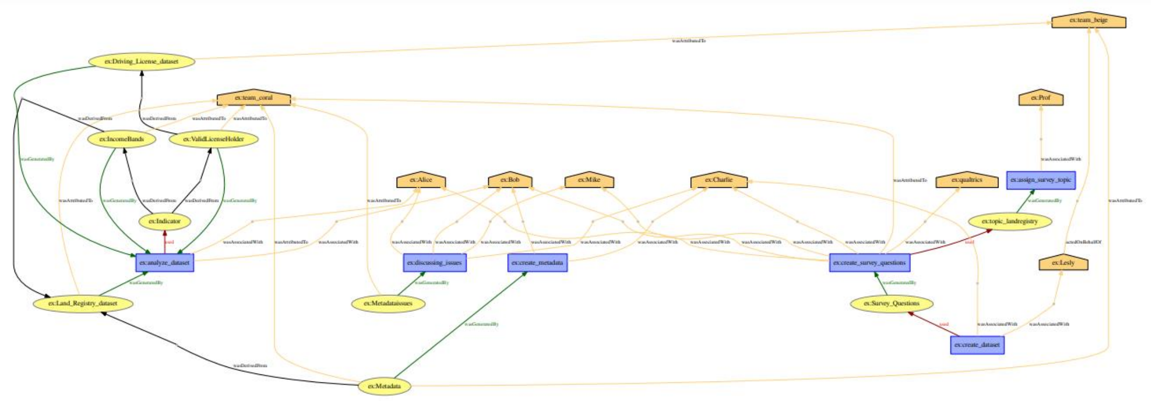
In such cases, we consider things on a broad level avoiding the inner complexities of a problem. In short, it's not detail oriented. Similarly, if we wanted a lower-granularity model then we wouldn't have gone into the details like how we are calculating indicators, cleaning, and checking of erroneous data, who did which part. Instead, we would have focused more on the final outcomes. Here I didn't show anything about indicator functions or agent roles.



Lower Granularity Model

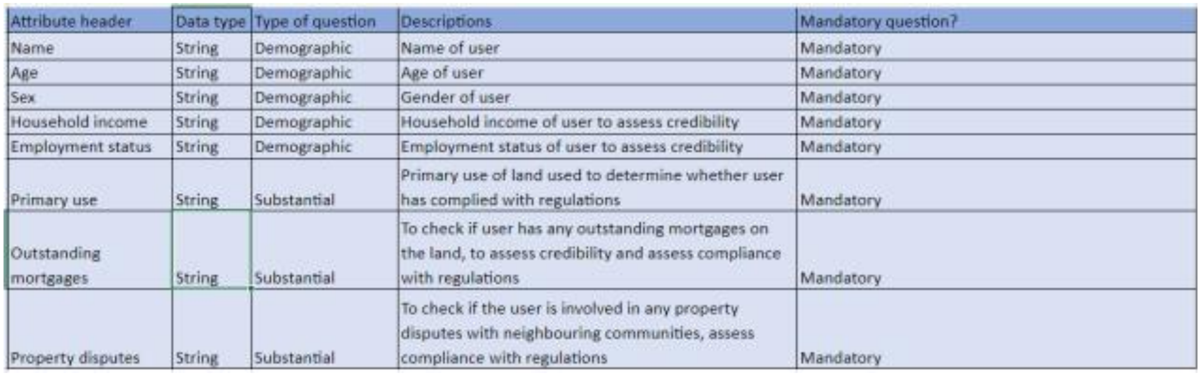
High Granularity (DETAILED)

This approach is more detail oriented. We can get a complete overview of the dataflow including who was responsible for which part of the activity. Sometimes timestamps are also provided for authenticity. We have created a high granularity model here.



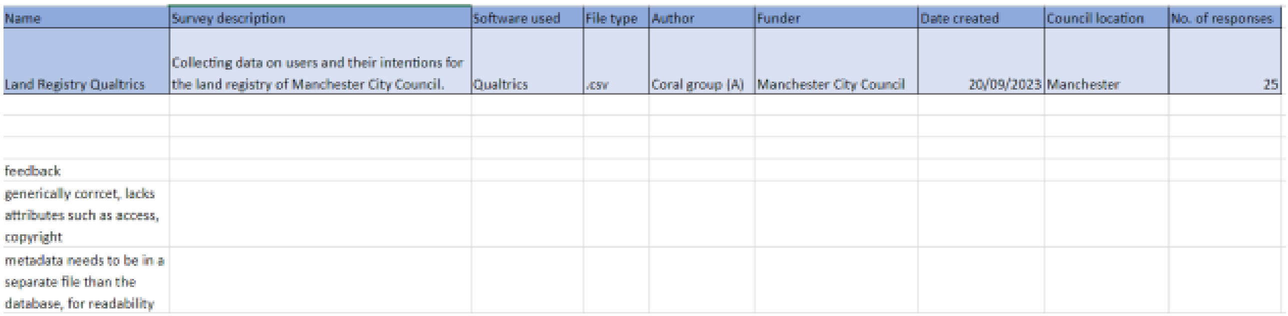
High Granularity Model

METADATA



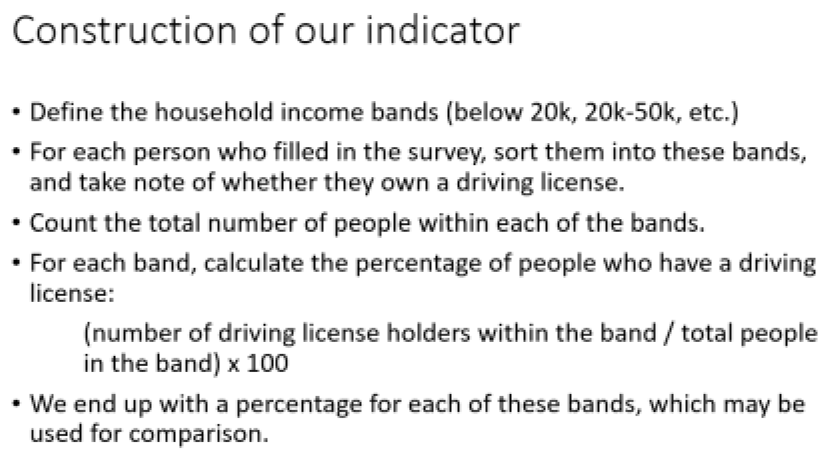
Attribute header	Data type	Type of question	Descriptions	Mandatory question?
Name	String	Demographic	Name of user	Mandatory
Age	String	Demographic	Age of user	Mandatory
Sex	String	Demographic	Gender of user	Mandatory
Household income	String	Demographic	Household income of user to assess credibility	Mandatory
Employment status	String	Demographic	Employment status of user to assess credibility	Mandatory
Primary use	String	Substantial	Primary use of land used to determine whether user has complied with regulations	Mandatory
Outstanding mortgages	String	Substantial	To check if user has any outstanding mortgages on the land, to assess credibility and assess compliance with regulations	Mandatory
Property disputes	String	Substantial	To check if the user is involved in any property disputes with neighbouring communities, assess compliance with regulations	Mandatory

Issues in Metadata:-



Name	Survey description	Software used	File type	Author	Funder	Date created	Council location	No. of responses
Land Registry Qualtrics	Collecting data on users and their intentions for the land registry of Manchester City Council.	Qualtrics	.csv	Coral group (A)	Manchester City Council	20/09/2023	Manchester	25
feedback								
generically corrcet, lacks attributes such as access, copyright								
metadata needs to be in a separate file than the database, for readability								

INDICATOR(PERCENTAGE OF PEOPLE HAVING DRIVING LICENSE UNDER EACH INCOME GROUP)



Construction of our indicator

- Define the household income bands (below 20k, 20k-50k, etc.)
- For each person who filled in the survey, sort them into these bands, and take note of whether they own a driving license.
- Count the total number of people within each of the bands.
- For each band, calculate the percentage of people who have a driving license:

  (number of driving license holders within the band / total people in the band) x 100
- We end up with a percentage for each of these bands, which may be used for comparison.

CHALLENGES

Provenance overall is easy to understand. Basically, we need to backtrack the flow from final outcome to data source involving agents, entities and activities. Logically it's simple to think as a flow but while trying to document that as provn file things become a bit hard because of the syntax and since it’s a niche concept, there aren’t many resources available online for support. Thankfully we were given some websites to follow from university which helped. Also while trying to make the model highly granular it becomes difficult to describe the small, complex relations among entities and agents.Linking the minor entities with agents and activities was hard keeping in mind the data flow, responsibility view and process view of the model remains intact.

REFERENCES

- *https://s11.no/2020/prov/provn-cheat-sheet/*
- *https://s11.no/2020/prov/validating-and-visualising-prov/*