Chiffon Report

Data Modelling Approach:

In our approach to data provenance modeling, we began by specifying the outcome our readers are interested in—metadata—as the ultimate target of the process. From this endpoint, we traced backward to identify key entities, activities, and agents that played a role in producing this output.

To clarify the flow of information, we selected core documents as entities, focusing on points where there are significant transformations in the data. Entities were assigned an identifier that reflects their type and purpose, ensuring that their roles within the model are easily understandable. Activities were defined as major actions reflecting key data transformations or decision points. These activities capture essential steps in the process, each labeled to convey its specific function within the provenance model. Lastly, we identified agents as contributors in the research process, including roles such as researchers and participants to track contributions through these agents.

We believe identifiers should be clear for the reader to easily identify each resource. Rather than naming datasets as dataset1, dataset2, dataset3, we used more descriptive names like chiffonDataset, navyDataset, and cyanDataset. In PROV-N, types indicate the category or nature of a resource, helping clarify the resource's purpose and function. Depending on whether there have been appropriate types defined in standard vocabularies such as schema.org, we used

those types to improve interoperability, if not then we chose to define our own types in a custom namespace.

Modelling Decisions:

We excluded the other team's processes that we do not have detailed knowledge of. For instance, we only included the dataset and metadata schema from other teams because those are the entities we have access to. We left out the activities, entities, and agents used to generate their data because including them implies assumptions about their internal processes. We only focus on the processes that we directly perform to ensure the integrity of our provenance.

If we had modelled the provenance for lower granularity or for a broader audience, we would exclude details related to the surveying process and leave out the information about the specific timestamps.

We would incorporate provenance information from other teams by inquiring about their internal procedures if we were to model the provenance with higher granularity. Also, we would include information about the program versions and the techniques we used for data cleaning and analysis. This would enable other users to replicate our findings exactly and achieve a better understanding of our data.

Writing provenance is not easy. The first challenge is learning the syntax of PROV-N and ensuring that the outcome complies with the standards. The second challenge lies in defining the dependencies and interactions between different resources in complex provenance processes, especially given the limited number of examples online that are similar to our assignment.

One disagreement centered around defining the goal of the provenance: should it focus on the dataset or metadata schema? While we understand that, in the real world, the provenance of the dataset is often more critical, for this assignment, our audience might be more interested in the provenance of our metadata schema. Another point of disagreement was whether URL to the entities should be included in prefixes or as an attribute. We found examples using the former approach and chose to follow this standard.

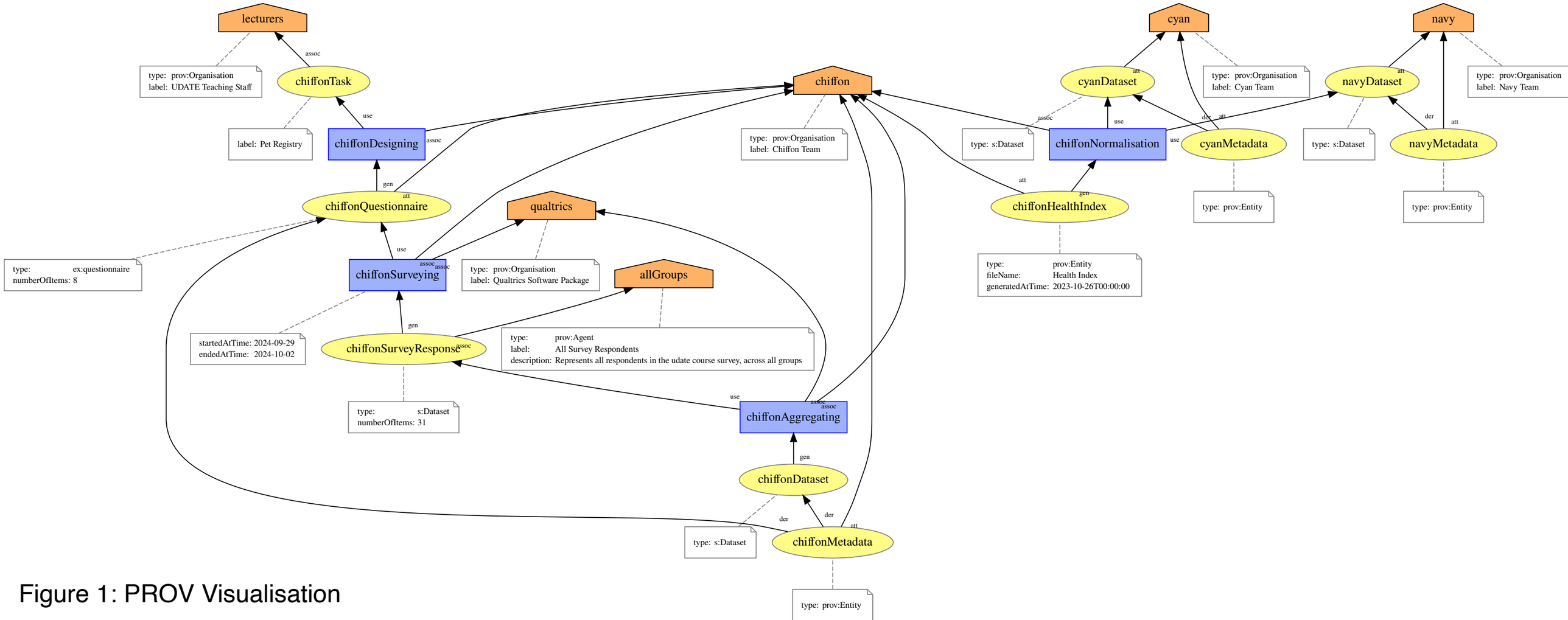

Figure 1: PROV Visualisation