**Correspondence Tables old IDs and new IDs**

1. **New IDs with multiple old IDs**
   1. *Unimodal cases Builtwith:*

Assign modal website to the new ID.

* 1. *Multimodal cases Builtwith (No clear mode)*:
     1. **Best website identified**: Compute similarity score between each pair of panjiva raw names and the websites for a given new ID. Identify what is the most similar website for each old ID of a given new ID. If all old IDs share the same most similar website, then we assign this ‘best website’ to the new ID. If all old IDs do not share the same most similar website, we use the approach in 1.2.2.

For these cases, the correspondence tables will have multiple observations for a given new ID, what will differ is the old ID and the website will be the same across all of them.

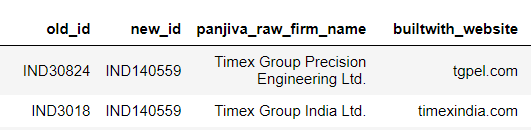
* + 1. **Best website not identified**:
       1. Compute similarity between websites for the given new ID. If the similarity between websites is high, they are likely different websites of the same firm. In such cases, we would like to define the technology (from the Builtwith dataset) for the given new ID based on all the websites, not just one. For instance, if the e-commerce technology is used on at least one of the websites during a month, but not on the others, then we can simply state that the e-commerce technology for this given new ID is in use, as it was adopted by one of the websites.

In these cases, the correspondence table will have multiple observations for a given new ID, what will differ is the old ID and also the website (because we are going to use all of them to create a group of technology).

* + - 1. If the similarity between website is low, we need to manually check these cases. Manually checking we can identify two cases:
  + Among the old IDs for the given new ID, there is one old ID that has a correct website, and the other websites are not correct. In such cases, we assign to the new ID the correct website.

For these cases, the correspondence tables will have multiple observations for a given new ID, what will differ is the old ID and the website will be the same across all of them when there is a correct website.

* + Each old ID has a correct website, i.e., the old IDs with the same new ID are totally different firms, meaning that the error comes from the algorithm that assigns the IDs. For example, Timex Group Precision Engineering Ltd is a manufacturing Engineering company, while Timex Group India is a brand of wrist watches. These cases need to be dropped as they represent different firms or send to Ana for inspection and potential new IDs assignment.



Drop for now… wait Ana’s corrections

* Each old ID has an incorrect website, let duplicate rows for now and these cases will be dropped when running the algorithm of similarity score between websites and names.
  1. *Only-one-matched:*

The “only-one-matched” cases refer to situations in the correspondence tables where new  
IDs are associated with multiple old IDs, but only one of these old IDs is matched to a  
BuiltWith website. This implies that for a given new ID, only for one of the old IDs it was successfully retrieved the BuiltWith website. For the multiple old IDs with the similar new ID, this could happen for the following reasons:

* + Because the website for the given old ID was not found in the Builtwith dataset.
  + Because the website for the given old ID (and its panjiva raw name) was not identified in the Google Places API.

For these cases, we assign to the new ID the website of the only old ID for which it was possible to match the website in the Builtwith dataset.

For these cases, the correspondence tables will have multiple observations for a given new ID, what will differ is the old ID and the website will be the same across all of them when and will be the website of the only old ID that was matched to the Builwith dataset.

1. **Old IDs with multiple new IDs**

These are cases where in the correspondence tables the old IDs have multiple new IDs. Then, we will assign the website of the old ID to the multiple new IDs it represents.

BUT THIS IS NOT CORRECT AT ALL….

1. **Old IDs and new IDs with 1:1 correspondence**