# Data Model - brainstorming

This is a brainstorming document; I'll simply jot down what I think could be valuable for our data model.

### Ressources

- Guide to data modelling with MongoDB

- UML Data Modelling
  Our example will be simpler I guess.

### My thoughts

A t first I thought we should get the data from different sources, transform them and only after the transformation store them in the database.
Maybe I was wrong here and we could benefit if we store the data from each source first in the database, transform and save again afterwards. I think we are mainly interested in the uniform data model (which includes all your different sources) but maybe in the future we change our minds about this uniform structure and in this case we would have
no raw data from the sources and would be forced to get all the data again (in some caes this may be not possible at all, who knows).

So we could make a seperate collection for all data sources.

So far we have 3 data sources, I try to add new ones before our meeting on Wedenesday, I promis at least one new source ;)

### Data Sources

**Muse.com**

- **job_title**: name of the job, should not be NULL
- **min_salary**: value can be :"NOT_FOUND" which equals NULL
- **max_salary**: value can be :"NOT_FOUND" which equals NULL
- **currency**: value can be :"NOT_FOUND" which equals NULL
- **skills**: value can be [] => empty list which equals NULL
- **publication_date**: when the offer was published, should not be NULL
- **id**: every data I save contains an id, this id is distinct for the data source
- **location**: list that contains somethimes multiple entries, one of them should be city, country, field has to be checked, I have discovered so far something like this:

```
"location": [
{
"name": "Flexible / Remote"
},
{
"name": "London, United Kingdom"
}
]
```

- **experience**: level of exp required, have not seen yet but I guess could be NULL
- **html_link** : link to full job offer description , could valuable if we get data from multiple sources to identify doubles (our sources are often services that query data sources, for example okjobs queries LinkIn), should not be NULL
- **type**: describes tpye as "external" (sample data has no other values), could be NULL imho
- **company_name**: should not be NULL
- **company_id**: muse company id => not useful for us !?
- **source**: redundant (see html_link above) !?
- **created** : I would say publication_date is enough !?