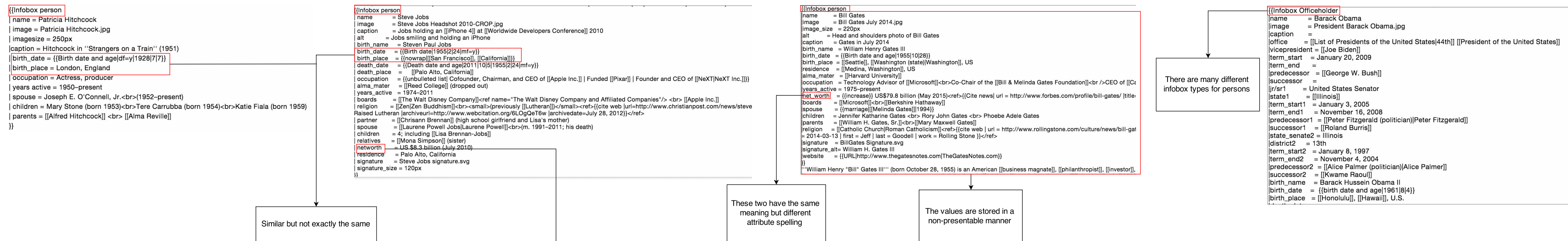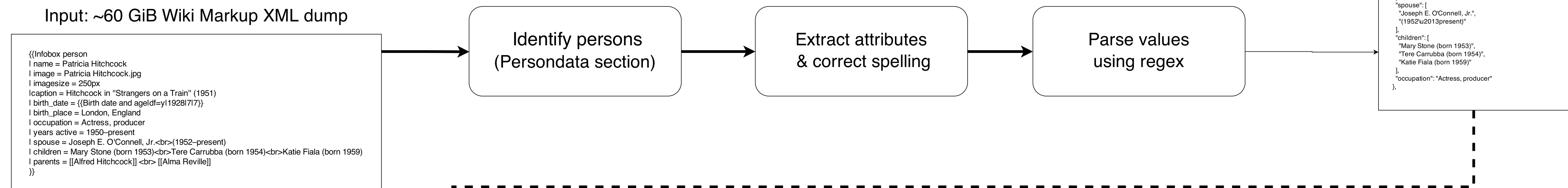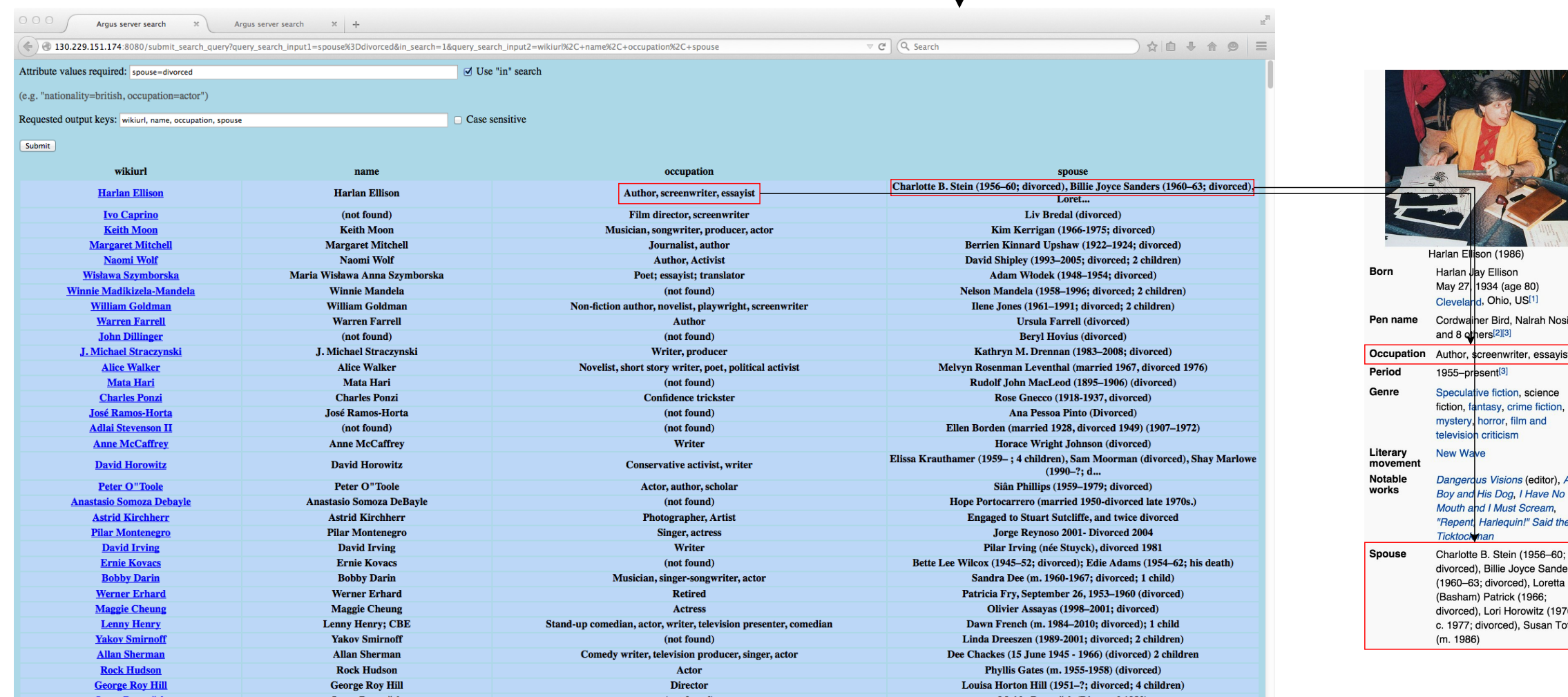# Collecting Data about People from Wikipedia

The objective of the project was to retrieve as much information as possible regarding as many people as possible from Wikipedia and produce a JSON file containing all the information in a standardised form.
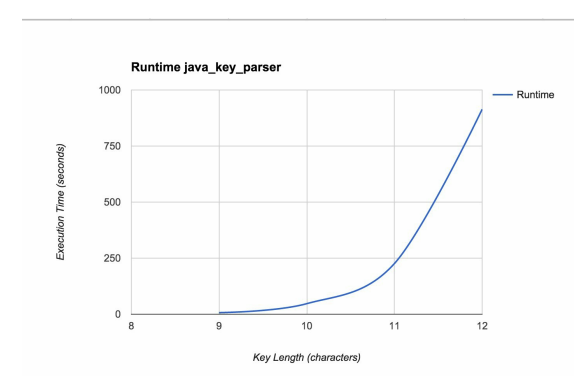

Similar but not exactly the same

These two have the same meaning but different attribute spelling

The values are stored in a non-presentable manner

There are many different infobox types for persons

Our approach:

Output: JSON

Input: ~60 GiB Wiki Markup XML dump

```
{{Infobox person
| name = Patricia Hitchcock
| image = Patricia Hitchcock.jpg
| imagesize = 250px
| caption = Hitchcock in "Strangers on a Train" (1951)
| birth_date = London, England
| birth_place = London, England
| occupation = Actress, producer
| years active = 1950–present
| spouse = Joseph E. O'Connell, Jr.<br>(1952–present)
| children = Mary Stone (born 1953)<br>Tere Carrubba (born 1954)<br>Katie Fiala (born 1959)
| parents = [[Alfred Hitchcock]] <br> [[Alma Reville]]
}}
```

Identify persons (Persondata section) → Extract attributes & correct spelling → Parse values using regex
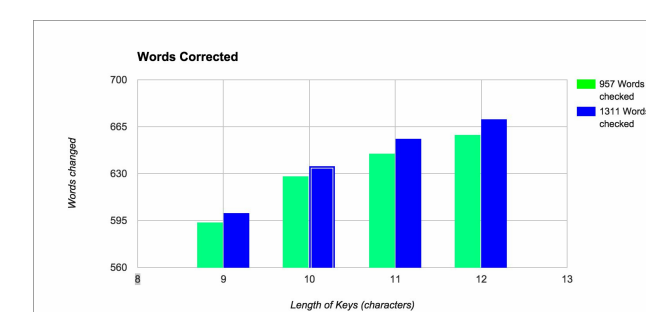
Application: A Searchable Database

Statistics - Spell Checker

The execution time rises with the length of keys we compare.

The amount of words which are corrected by our algorithm increases with the length of the keys. It also increases with the amount of words compared but less significantly.

The Average Wikipedian

(from sample)