# Initial Assumptions and Analysis

## System Context

At Xapien the internet is scraped for data about a subject, including a lot of web pages containing unstructured data. To process this so useful data has to be chosen and parse it into a structure that a system can understand. Xapien would like to add the ability to search the Forbes website for a subject, to do this the system must be able to understand the data on those Forbes web pages.

## System Goal

To provide a function that can parse the personal, and other useful, information about a subject from a Forbes web page. The function should take an input of the raw HTML content of a web page and return a sensible data structure containing useful information.

Examples of potential inputs have been attached and include scraped data from pages on some high-profile test cases (Elon Musk and Mark Zuckerberg for example). However, the function should be able to process other examples.

In initial development, the core functionality is only required at this time.

## Assumptions

1. The Solution is one that is going to be maintained and extended into the future.
2. In the future parsing of other sites may be required.
3. In the future parsing of other “entity types” may be required (people, companies, countries, charities…).
4. The Json produced will be used as a source of basic information about an individual.
5. The Json produced will be used as a source of financial information about an individual.

## Forbes Field Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Field | Parsed? | Justification | Always Present? | example if not | Requirement type | Justification | Schema Name |
| Name | Yes | Needed to identify the reports subject. | Yes |  | Mandatory | Needed to identify the report. | ReportSubjectName |
| Position, Company | Yes | basic information about the person. | No | Li Xiaohua and Family | Optional | The are examples of when a person doesn’t have a role or company. | CompanyPosition, CompanyName |
| Real Time net worth | No | Better data is given in the Wealth history graph. I can see the case for adding this as an extra data point though. Although it’s at a date, when the wealth chart is a month and year only, which would course issues. | Yes |  | n/a |  |  |
| net worth Change | No | A time frame for this change isn’t given so in isolation this could be misleading data. | Yes |  | n/a |  |  |
| Image | No | I can’t see a useful reason for having this as part of the schema. | No | Li Xiaohua and Family | n/a |  |  |
| About | Yes | Useful unstructured information about the person, good candidate for NLP. | Yes |  | Optional | As per assumption 2, other sites may not possess this type of data. | UnstructuredData |
| Also on Forbes | No | Not related to the subject. | Yes |  | n/a |  |  |
| Wealth history | Yes | Shows a history of the persons wealth over time up to near real time. | Yes |  | Mandatory | As per assumption 5, financial data is a primary use of this report so will be required. | Valuations |
| Forbes lists | Yes | Shows where the individual sits in relation to similar people and the world and industry. | No |  | Optional | As per assumption 2, other sites may not possess this type of data. | InfluentialRankings |
| Age | Yes | Basic information about the person. | Yes |  | Mandatory | As per assumption 4, basic information about the subject is a primary use of this report so will be required. | Age |
| Self-Made Score | No | This is a score specific to Forbes with no obvious meaning. The Forbes lists give a better indication of the information this could be used to portray. | No | Tim Cook | n/a |  |  |
| Philanthropy Score | No | This is a score specific to Forbes with no obvious meaning. The Forbes lists give a better indication of the information this could be used to portray. | No | Tim Cook | n/a |  |  |
| Children | Yes | basic information about the person. | No | Tim Cook | Optional | Despite assumption 4, not every subject has children. It is not in the place of this parser to assume the lack of data on this means the subject has no children | Children |
| Source of Wealth | Yes | basic information about the person. | Yes |  | Mandatory | As per assumption 4, basic information about the subject is a primary use of this report so will be required | SourceOfWealth |
| Residence | Yes | basic information about the person. | Yes |  | Mandatory | As per assumption 4, basic information about the subject is a primary use of this report so will be required | Residence |
| Citizenship | Yes | basic information about the person. | Yes |  | Mandatory | As per assumption 4, basic information about the subject is a primary use of this report so will be required | Citizenship |
| Education | Yes | basic information about the person. | Yes |  | Mandatory | As per assumption 4, basic information about the subject is a primary use of this report so will be required | Education |
| Marrital Status | Yes | basic information about the person. | No | Tim Cook | Optional | Despite assumption 4, not every subject has a partner. It is not in the place of this parser to assume the lack of data on this means the subject has no partner | MarritalStatus |
| Did you know | No | These seem to be more "fun facts", rather than giving information the report consumer might be interested in. | No |  | n/a |  |  |
| Related People and Companies | Yes | Gives an indication of other subjects that the report consumer might be interested in. | Yes |  | Optional | As per assumption 2, other sites may not possess this type of data | RelatedParties |
| More on Fobes | No | Not related to the subject. | Yes |  | n/a |  |  |