



jHound – Large-Scale Profiling of Open JSON Data

Motivation and Aims

- JSON offers a flexible way for storing arbitrary data
- jHound delivers an overview on large JSON documents collections
- Users can gain insights into the usage of JSON
- jHound collects and computes statistics of JSON documents (data type distribution, property presence, nested object level with most data, and more) to detect interesting patterns

Profiling Stages

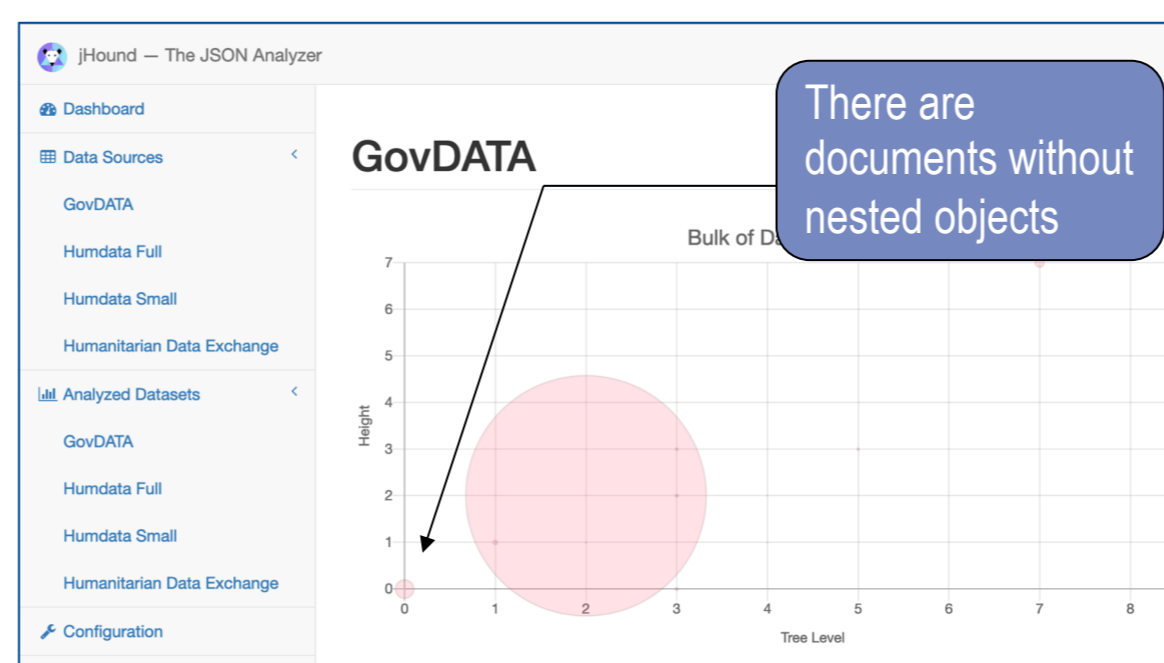
Inspection

UUID Prefix	Req. Propls.	Sum Propls.	Str	Num	Int	Bool	Arr	Obj	Null	Simple	Only Arr	Obj
6d5acfee	0	5	21	84	0	0	1	21	0	0	0	1
8df91f89	0	786	0	0	0	0	0	131	0	0	0	1
9c...	150	0	0	0	0	0	0	0	0	0	0	1
69...	1600	0	0	0	0	0	0	0	0	0	0	1
d2...	175	0	0	0	0	0	0	0	0	0	0	1
a7...	256	0	0	0	0	0	0	0	0	3	1	1
55138d0e	10	2	6	258	0	0	131	6	0	127	3	1
3326d3ab	4	7	375	1050	0	0	601	450	0	375	225	1

Only 5 different properties in the document

Document has 1 array and 21 objects

Analysis



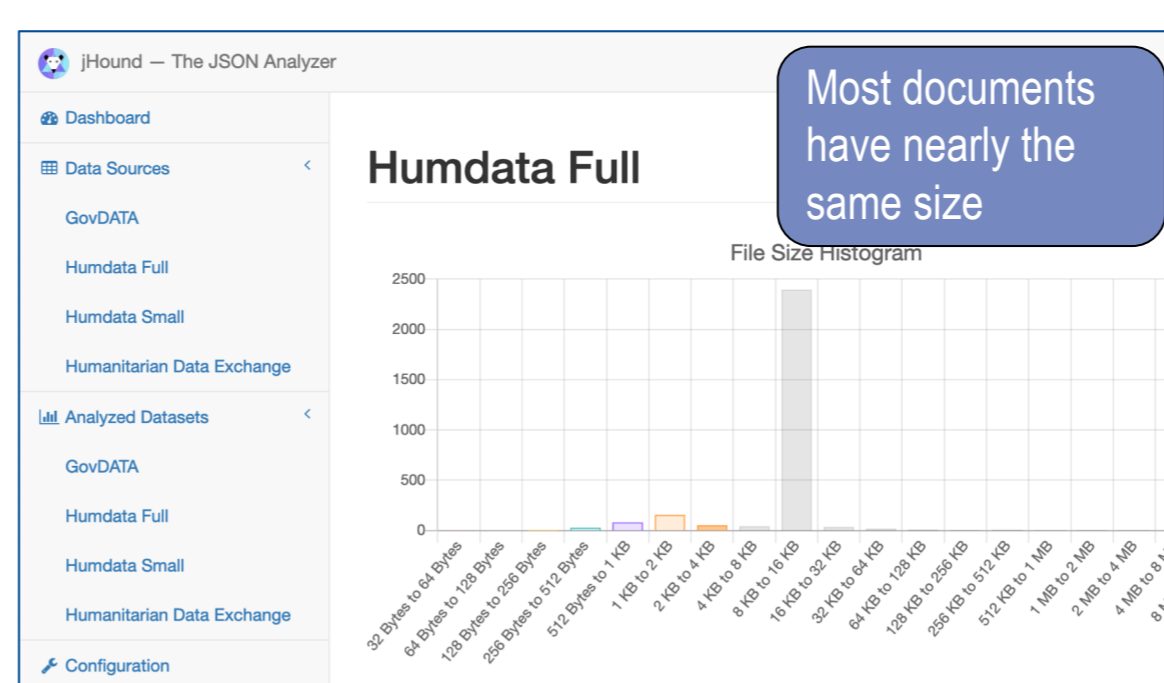
Detection

```
{
  "jahr": "1999",
  "pkw_besatz": 444.7,
  "durchschnittsalter_kfz": 6.1,
  "kfz_besatz": 583.6,
  "durchschnittsalter_privater_halter": 44.4
},
{
  "jahr": "2006",
  "pkw_besatz": 443.5,
  "durchschnittsalter_kfz": 7.6,
  "kfz_besatz": 586.0,
  "durchschnittsalter_privater_halter": 47.0
}
```

Detects documents with a relational-like structure

UUID Prefix	Req. Propls.	Sum Propls.	Str	Num	Int
a2716c01	6	0	413	3	0
ce4afd89	6	0	413	3	0
0041124f	6	0	413	3	0
4c201fb1	6	0	413	3	0
ca524790	6	0	361	3	0
b0a9b8c1	6	0	413	3	0
db09200a	6	0	413	3	0

Lots of JSON documents with very similar metrics



```
{
  "page": 1,
  "pages": 1,
  "per_page": "10000",
  "total": 59
},
{
  "indicator": {
    "id": "SP_POP_TOTL",
    "value": "Population, total"
  }
},
{
  "page": 1,
  "pages": 1,
  "per_page": "10000",
  "total": 59
},
{
  "indicator": {
    "id": "SP_POP_DPND_YG",
    "value": "Age dependency ratio, young (% of working-age population)"
  }
}
```

Helps to detect documents with a similar structure

Analysis Workflow

jHound's Workflow consists of multiple stages and is designed for crawling, downloading and analyzing Open JSON Data from CKAN-compatible repositories. A webfrontend guides through the workflow stages

