A Field Study on Linked and Open Data at Datahub.io

Timm Heuss[*][†], Janina Fengel[*], Bernhard Humm[*], Bettina Harriehausen-Mühlbauer[*] and Shirley Atkinson[†]
[*]University of Applied Sciences
Darmstadt, Germany
Email: {Timm.Heuss|Janina.Fengel|Bernhard.Humm|Bettina.Harriehausen}@h-da.de
[†]University of Plymouth,
Plymouth, United Kingdom
Email: {Timm.Heuss|Shirley.Atkinson}@plymouth.ac.uk

***Abstract*—We describe and conduct a study on datahub.io to explore to what, in practice, Linked and Open Data refers to. We focus on the use of formats, licenses, ages and popularity of the data. An in-depth analysis reveals information about availability, quantity, structure and vocabulary usage of the real-world RDF-based datasets contained. Results show that the most common formats are Microsoft Excel, CSV and RDF. High proportions of structured data is of tabular nature, independent from the format. The heuristics and evaluation methods developed here are released as open source and can be applied to other CKAN-based repositories and RDF-based datasets, too.**

I. Introduction

In 2006, Tim Berners-Lee introduced a 5-star rating or maturity model for data [5]. This model is used to “encourage people [. . .] along the road to good linked data” [5] and to assess “the openness and linking potential of the data” [9], p. 4. The ranking starts with Open Data, data in any format that is openly licensed (1 star), followed by structured, machine readable formats (2 stars), non-proprietary structured formats (3 stars), RDF (Resource Description Framework) or SPARQL (SPARQL Query Language) endpoints (4 stars) and ends up with the highest maturity class, which is RDF data interlinked with other data [5].

II. Motivation

As there is only a licence-based constraint for the first star, many different file formats might come into question. When building applications based on Linked and Open Data, this question of format becomes important: it requires an entirely different technology stack to integrate, assure quality, and store data from, for example, RDF - compared to the technology needed for data in MS Excel format. And while building apps based on Office formats is straight forward (thank to frameworks like Apache POI [3]), years of experience and complex infrastructure might be required to master the advanced possibilities of RDF, such as OWL (Web Ontology Language) reasoning. Furthermore, integrating different formats is challenging, e.g., if the internal structure of the data differs fundamentally. While data stored in spreadsheets might usually be of tabular nature, RDF-based data, by design, does “not necessarily consist of clearly identifiable ‘records’” [17]. RDF knowledge bases are considered to be heterogeneous, non-tabular structures, which do not resemble relational structures [18], p. 455.

Therefore, the question that needs to be answered before putting the data into use is: what are relevant formats when dealing with Linked and Open Data in practice? Under what terms of use may it be reused and processed? If the format is RDF, can it flawlessly loaded and does it require sophisticated tool support (in form of OWL reasoners)? What is the internal structure of the real-world RDF? Does it constitute homogeneous, tabular structures or is it as heterogeneous as often found in large knowledge bases? This field study reported on is conducted using the well-known data portal datahub.io, as it is commonly used as an indicator for the progress of Linked Data, e.g. upon creating the LOD Cloud diagram [22].

III. Approach

The field study is designed to consist of three parts. Firstly, the meta information about the data available at datahub.io is extracted and stored in a CSV (Comma Separated Values) file. Secondly, the CSV file is loaded, interpreted and analysed. Based on this meta data, information about formats, popularity, licences, and ages of the datasets can be acquired. Thirdly, based on the meta data, the highest rated RDF-based resources are attempted to be downloaded, imported, and analysed.

The exact analytical process is described in the following. For demonstration purposes, the third part uses excerpts of the New York Times Linked Open Data dataset “People” [7] as an example for the analysis conducted. All scripts, queries, program source code and results are published in the GitHub repositories CKANstats [15] and LODprobe [16].

A. Extraction

Like many other data portals, datahub.io is based on CKAN (Comprehensive Knowledge Archive Network), offering an open, REST-based, JSON-based API [8]. In CKAN-terminology, the actual data is published as a resource, and one or more resources are provided in units named datasets [21]. A Python script named CKANstats.py uses the CKAN-API (Version 3) [19] and extracts the meta information about the datasets registered at datahub.io [20], including: dataset names, licences, a boolean flag if the dataset is openly licensed, the dataset's page views (total + last 14 days), a resource's format statement, download URL and resource's page views (total + last 14 days).

The script stores the retrieved meta information in a CSV file [12].

B. Meta Data Analysis

In order to conduct further analysis with SQL, the first step is to load the CSV file into PostgreSQL (Version 9.4.1) using `COPY datahubio FROM 'datahubio.csv' DELIMITER ',' CSV;`. Unfortunately, values found in format column are not unified. For example, there are at least 29 different notations given for specifying a Microsoft Excel resource. Another issue with this column is the fact that in about 20% of the cases, there is no format specification at all.

To address these issues, a generic mapping table has been manually created. It assigns the various source values stated in resource to a unified format definitions. For example, `application/zip+application/vnd.ms-excel` and `microsoft excel` is combined into `Excel`.

Based on the table holding the imported data, a database view is created using this mapping table twice:

- Firstly, the format definition of datahub.io is translated via SQL-like-patterns `left outer join mptbl as a on lower(trim(resource_format)) like lower(a.expr)`.
- Secondly, for every remaining format unknown, it is attempted to join the mapping table an additional time based on the last characters of the resource URL - `left outer join mptbl as b on (a.format = 'n/a' and lower(substring(trim(resource_url) from '...$' )) like b.expr)`. So if, for example, a resource has a URL pointing at "example.com/filename.pdf", this is an indication for the file format PDF (Portable Document Format).

Both these joins produce a best-effort corrected view on the meta data extracted. Thus, further analysis is enabled based on this view, and the SQL scripts developed are documented online [15].

C. In-Depth Analysis of RDF-based Resources

The in-depth analysis is conduced for every RDF resource which has ever been visited, ergo having a `resource tracking summary total` larger than 0. At the time of the described metadata extraction this included 606 resources.

In a semi-automated process, each single resource is downloaded using wget (Version 1.15) and loaded into an empty Apache Fuseki (Version 2.0.0 2015-03-08T09:49:20, Xmx set to 14.240M) using s-put and, if that failed, using Fuseki's WWW front end. Errors during this process are logged as follows:

Not Found wget could not download the resource, either because it was no longer available or the connection timed out.

Parse Error Fuseki failed to load the downloaded resource using s-put and Fuseki Web.

TABLE I
EXCERPT OF LODPROBE ANALYSIS RESULT FOR THE NEW YORK TIMES LINKED OPEN DATA DATASET "PEOPLE".

Number of unique subjects: 9958	**Count**	**[0]**	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**
[0] cc:attribution[...]	4979	-	4979	4979	0	0	0
[1] cc:attributionURL	4979	-	-	4979	0	0	0
[2] cc:license	4979	-	-	-	0	0	0
[3] nyt:associated[...]	4979	-	-	-	-	4281	4281
[4] nyt:first_use	4281	-	-	-	-	-	4281
[5] nyt:latest_use	4281	-	-	-	-	-	-

TABLE II
EXCERPT OF THE PREVIOUS TABLE I, LOADED, MIRRORED, AND CONVERTED AS AN R MATRIX OBJECT.

Number of unique subjects: 9958	**[0]**	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**
[0] cc:attributionName	4979	4979	4979	0	0	0
[1] cc:attributionURL	4979	4979	4979	0	0	0
[2] cc:license	4979	4979	4979	0	0	0
[3] nytdata:associated[...]	0	0	0	4979	4281	4281
[4] nytdata:first_use	0	0	0	4281	4281	4281
[5] nytdata:latest_use	0	0	0	4281	4281	4281

Partly Some files of the resource were loaded, others not. Evaluation is done with the loaded files only.

Out of Memory Fuseki failed to load the resource and reported a out of memory exception or a garbage collection overhead exception.

This process is documented for all 606 resources [13]. Out of these, the resources that could successfully be loaded into Apache Fuseki are the foundation for the subsequent analysis. Thereby, the authors notice that in some cases, RDF dumps could be loaded using s-put, but not via Web front end, and in others vice versa.

The Java tool called LODprobe [16] has been specifically developed for this field study to analyse the inner structure of the RDF-based datasets. Once a dataset is entirely loaded in an empty local Fuseki dataset, LODprobe fires a number of SPARQL queries against the default graph.

As a result, quantities about several basic characteristics of the resources are extracted:

- The number of unique RDF subject identifiers.
- The number of occurrences of each RDF property the default graph contains.
- The number of co-occurrences of two RDF properties, considering every property with each other.

The result is a symmetrical matrix of co-occurrences, accompanied with individual property counts, as excerpted in Table I. For example, the value 4979 in the second row (starting with [1]) and column [2] shows that the RDF property http://creativecommons.org/ns#attributionName ([1]) co-occurs with the RDF property http://creativecommons.org/ns# attribution ([2]) in 4979 subjects.

Considering the co-occurring values of two RDF properties row- or column-wise, further insights into the structure of the resource can be gained. For the example above, the properties [0], [1], and [2] always co-occur. This holds also true for the properties [4] and [5], all co-occur as well. Both property groups seem to be part of distinct entities, as [0], [1], and

TABLE III
METRICS FOR LODPROBE RESULTS, SAMPLE MEASUREMENT RESULTS FOR THE NEW YORK TIMES LOD "PEOPLE" RESOURCE.

Metric	Sample
LODprobe analysis name	people.csv
Number of unique subjects	9958
Number of properties	20
Minimum height cluster analysis	0
Maximum height cluster analysis	1
Number of cluster groups at h=0.1	3
Number of cluster groups at h=0.2	3
Number of cluster groups at h=0.3	3
Number of cluster groups at h=0.4	3

[2] never co-occur with [4] and [5], as indicated by the zero values in the matrix. This may be concluded that this sample data contains two entities.

In such obvious cases, the result of a LODprobe analysis contains about 20x20 RDF properties with few, clearly identifiable entities. Usually, however, there are more entities, less clear co-occurrences, and / or much more properties.

In a next step, a large-scale analysis of the individual LODprobe outputs is conducted using the scriptable statistics software R. Thereby, a co-occurrence diagonal matrix from the CSV-files is loaded, converted into a numerical matrix and mirrored in a symmetric one. Individual property-counts are moved into the diagonal. Table II shows this for the chosen example. Having the LODprobe results available as R objects allows for further advanced analysis in consecutive order: calculation of the dissimilarity matrix of the LODprobe matrix, followed by a cluster analysis of the dissimilarity matrix. From this cluster analysis, a number of metrics for the structuredness of the individual resources is extracted. This metrics-based analysis is augmented by generating visualisations of the clusters detected. Follow-up analytical steps are:

1) Dissimilarity Calculation: In this step, based on the counted co-occurrences, the dissimilarity of the property-pairs in the matrix is calculated. The work showed that LODprobe results usually contain more zero than non-zero values. Therefore, non-euclidean distance metric has been applied. This field study uses the Gower Distance [10] by utilising the R function `daisy` from the `cluster` package, using `metric = "gower"`.

2) Cluster Analysis: In this step, based on their mutual (dis-)similarity calculated previously, groups or clusters of the properties are searched for. Thereby, the complete linkage method [1] is used (via R's `hclust` using `method ="complete"`), so resulting cluster analyses are usually scaled from a minimum height of 0 to a maximum height of 1.

Table III shows the metrics that are calculated for all LODprobe results for comparison purposes. They characterise a RDF resource: in the case of New York Times LOD "People" dataset, judging by the minimum and maximum height of the analysis, the dissimilarity of groups of properties in the dataset is very high - it is obvious that the RDF resource contains different entities.

Upon considering a number of cluster analyses of different RDF resources, the grouping behaviour of the clusters between the heights 0.1 and 0.4 seems to be most informative. Especially at lower heights of 0.1, 0.2 or sometimes even 0.3, properties usually seem to be clustered based on the logical entities found in the data, just before those clusters are again grouped together with other clusters. In the example above, even at lower heights, the 20 involved RDF properties constitute three groups. This is an indicator that the properties within the groups are very similar, but the groups themselves are very distinctive.

Finally, the collected metrics are compared to metrics computed for synthetically generated resources, simulating the case in which RDF data is truly heterogeneous. Thereby, property occurrence and co-occurrence counts are randomised and then normalised by the amount of actual unique subjects. This is repeated in a Monte-Carlo-like process based on two real examples, a small RDF resource with 9,958 subjects, 20x20 properties, and 1,000,000 simulations and a large one with 694,400 subjects, 222x222 properties, and 15,127 simulations.

In addition to the cluster analysis metrics, dendrograms are generated for each of the 251 LODprobe results to support the interpretation. They are generated using R's 'plot' function with the generated cluster analysis from above, without any further parameters.

IV. RESULTS

With using the described methodology, various insights on real data at datahub.io could be gained.

A. Common Formats

Considering the unified format values, the most frequently used format for data are full-featured spreadsheets such as Microsoft Excel or LibreOffice Calc documents, and CSV, including its variations like TSV [15]. Together, both tabular formats add up to almost one third of all data formats (27%), followed by RDF (11%), PDF (8%), and Images (7%).

With regard of the openness of data, the frequently used data formats are usually not openly licensed: only 21% of the spreadsheets are open, about 59% of the CSV, and roughly 14% of the PDF [15]. RDF, in contrast, is openly licensed in more than three of four cases (76%). The highest openness-percentages can be archived for the formats MARC (Machine-Readable Cataloging) (100%), GTFS (General Transit Feed Specification) (100%), and Beacon (98%).

By using the 5-star rating model [5] to classify the data, over a quarter (25%) of the data is 1-star, 6% 2-stars, 24% 3-stars, and 21% is 4 stars (or more) [15]. Excluded from this is data that is not explicitly openly licensed (about 48%). One fourth of open data formats could not be classified. The low frequencies of 2-star data can be explained by the openness criteria: Most 2 star data is in the Microsoft Excel format and not openly licensed.

B. Popular Formats

Based on existing data, a popularity measure can only be approximated using the tracked visits from the CKAN-API [20]. According to the resource tracking, in the last two weeks before the extraction, 33% of the clicks at datahub.io where on resources with an unknown format, followed by RDF (13%), RDF sample record (10%), CSV (8%), Spreadsheet (7%), SPARQL (6%) [15]. Despite the fact that PDF is quite a quite common format, resources with that format only received 2% of the tracked clicks. The most unpopular types are Links, RSS and Maps with 0,3%, 0,2% and 0,05% of the visits.

C. Licences

A clear license statement of data is important as it defines a terms of use of the referred resource(s). However, similar to the case of the original format information, the licence-field does not contain unified values [15]. Even worse, in more than 40% of the cases, there is no specification of a licence at all. In the remaining cases, properly defined licences (like Creative Commons licences) are mixed with country- or language specific licences, and insufficiently named licences like “None”, “Other (Not Open)”, or “apache”. Moreover, the boolean openness flag is set in 10,612 of the total 20,178 cases, which corresponds to 52.59%.

D. Ages

The extract contains meta information about data up to four years. Judging by the created and revision timestamp, in most cases (80%) this meta information is never updated after the dataset had been put online [15]. Of the remaining cases, more than 10% are updated within less than 50 days.

E. In-Depth RDF Analysis Results

In addition to the previous analyses, based on metadata of all data on datahub.io, the following analyses are limited to specific datasets of the type RDF and that have a popularity ranking larger than zero. At the time of extraction, this included 606 RDF resources.

1) Download and Process Results: In two-thirds of the cases, the download-URL worked and returned a server response (33% of the cases are not found). The downloaded resources, however, could only be flawlessly loaded in about 46,37% of the cases - the rest resulted primarily in parser errors (18,81%). Only a few datasets (0,83%) could not be loaded due to memory deficiencies of the test machine[1] - this includes the knowledge base DBpedia. Additional four resources[2] (0.66%) could only loaded partly, e.g., because they are split-up into several parts.

TABLE IV
AVERAGE CLUSTER ANALYSIS RESULTS OF OBSERVED DATA AT DATAHUB.IO AND OF ARTIFICIALLY SIMULATED HETEROGENEOUS DATA.

	Average at datahub.io	**Simulated Heterogenous (small)**	**(large)**
number of resources	253	1,000,0000	15,127
number of unique subjects	217,659.05 ± 994,267.50	9,958	694,400
min. height	0.00 ± 0.05	0.13 ± 0.03	0.04 ± 0.01
max. height	8.48 ± 79.17	0.58 ± 0.05	0.51 ± 0.02
Total Groups (h=0.0)	28.07 ± 33.03	20	222
Groups at h=0.1	7.66 ± 3.96	20 ± 0.5	210 ± 3.91
Groups at h=0.2	5.05 ± 2.08	16 ± 2	106 ± 13.1
Groups at h=0.3	3.76 ± 1.35	8.6 ± 1.4	20 ± 2.6
Groups at h=0.4	2.96 ± 1.00	4 ± -1	3 ± 1

2) Frequently used Properties: By summing up all individual LODprobe property counts, a big picture of the most frequently used RDF properties and vocabularies can be calculated. In total, over 373 million triples have been analysed, containing nearly 3000 different RDF properties. Unsurprisingly, basic properties are very frequent: every one out of six observed property is a rdf:type assignment, one out of 20 is a rdfs:label.

Regarding OWL, the most frequently used properties with the default namespace `http://www.w3.org/2002/07/owl` is sameAs - position 19 of the most frequently used properties with more than 4.5 million occurrences, followed by onProperty (Position 948 - 3,213 occurrences) and intersectionOf (position 1,016 - 2,038 occurrences).

3) Homogeneous Structures in RDF: Table IV shows the aggregated results of 253 LODprobe and cluster analyses for distinct resources [11]. On average, a RDF datasets contained 28 (± 33) different RDF properties, which can be grouped in only 7 (± 4) clusters at a height of 0.1 and in only 5 (± 2) clusters at a height of 0.2. Moreover, Table IV compares the measured values with two synthetically generated and simulated heterogeneous RDF datasets of different sizes. In these, there are significantly more cluster groups at lower heights, such as for h=0.1, the number of groups is (almost) identical to the total number of groups.

Accordingly, Figure 1 shows a dendrogram of the real RDF resource `southampton-ac-uk-org` that has the exact characteristics of an average resource, thus having 28 different RDF properties that form 7 groups at a height of 0.1, 5 groups at a height of 0.2, and 4 groups at a height of 0.3. The diagram gives the impression of a clearly structured resource, as many properties are already grouped together at a height of 0 and the properties seem to have quite similar co-occurrence counts.

In contrast, Figure 2 depicts a dendrogram of synthetically generated resource (simulation number 473,494 of 1,000,000) that shows the average characteristics of simulated, small heterogeneous resources - 20 properties, 20 groups at h=0.1, 16 at h=0.2, 9 at h=0.3. The lack of structure can be deduced from the high number of groups at lower heights, individually consisting of less properties (usually max. 2 properties per

[1] “dbpedia”, “library-of-congress-name-authority-file”, “semantic-xbrl”, “europeana-lod-v1” and “allie-abbreviation-and-long-form-database-in-life-science”

[2] “jiscopenbib-bl_bnb-1”, “geowordnet”, “datos-bcn-cl”, and “rkb-explorer-citeseer”

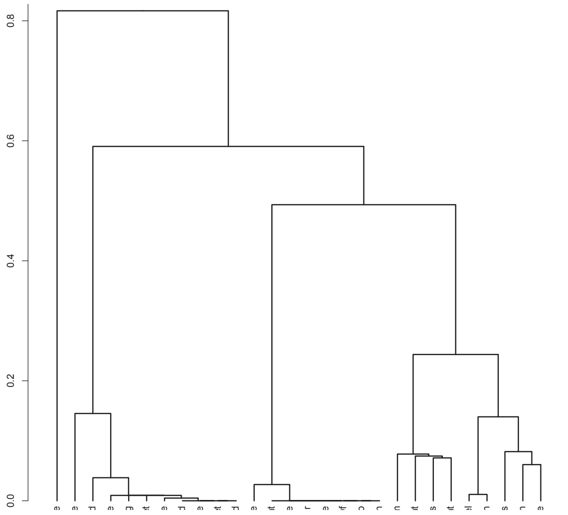

Fig. 1. A typical dendrogram of a real-world RDF resource: Many joins at lower (< 0.2) heights, indicating a high number of co-occurrences for these properties. Few joins above heights > 0.3, an high maximum height of 0.8 or above.

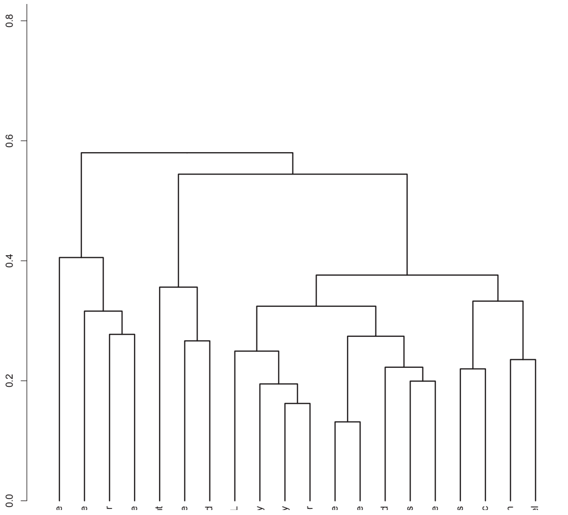

Fig. 2. A typical dendrogram of a synthetically generated heterogeneous RDF resource: Joins are evenly distributed between heights of 0.1 to 0.5, almost no joins at lower (< 0.2) heights, a low maximum height of < 0.6.

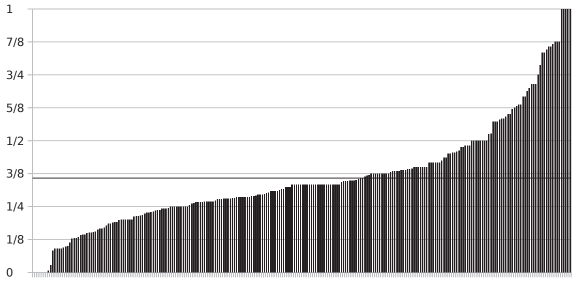

Fig. 3. Ratio of the number of properties and the number of cluster groups at h=0.1 for all analysed RDF datasets.

group at h=0.2). In addition, the lower maximum dendrogram height indicates that the properties are mutually less differentiable. As expected, the number of groups found in a cluster analysis is proportional to the number of unique properties the resource has. Resources with an RDF property count up to nine show on average 3 ± 2.4 groups at a height of 0.1. Resources with more properties, e.g. 70-79 distinct RDF properties, have 11 groups ± 2.63 at this height. This observation can be approximated by the rule of thumb “the more properties, the more groups”. Though, resources with 60 to 69 properties contain the most groups of all at a height of 0.1, 15.67 ± 4 groups.

Figure 3 visualises the observed proportionality between the number of properties and the groups they form: it compares the ratio between the number of groups at a height of 0.1 with the number of unique RDF properties a resource has in total for all involved analyses. Almost all ratios are below 4/5 and the average ratio about 3/8, while, in contrast, the simulated heterogeneous resources both have ratios of nearly 1/1.

V. Discussion

Only about half of the data on datahub.io is open, the majority of the rest bears legal uncertainty for application developers using it. Openness varies with the data formats: Excel is a very common format, but is usually not open. RDF is the third most common format, and is usually open. Accordingly, data on datahub.io is 53% Open Data, 15% Linked Data and about 11 % Linked *and* Open Data. However, of these RDF-based resources, 33% are non-existing downloads and nearly 20% are inaccessible due to parser errors. As a consequence of this distribution, RDF-only data integration approaches would not reach about 85% of the data that is out there. This fact underlines the necessity that data integration solutions support a whole spectrum of different formats.

Real data format quantities are almost evenly distributed across the scale of the 5-star rating model, despite for 2-stars data, usually represented by Microsoft Excel files which are often not openly licensed. However, the star rating in the 5-star model is proportional to the number of page views a resource gets: The more stars, the more page views. It seems that RDF is data-consumer friendly.

For a portal that conceptually only distributes URLs to resources and not the resources itself, having up-to-date meta data is essential - especially when considering the Link Rot phenomenon [6]. Unfortunately, a vast majority of meta data is never updated after it has been published. This might contribute to the 33% missing downloads mentioned above.

The cluster analysis revealed that RDF resources show clearly more characteristics of homogeneous data than of heterogeneous data. Almost all of the 251 dendrograms contain structures in form of properties that more or less exclusively co-exist with certain other properties [14].

So, even if it is not a constraint of the format, real-world RDF triples constitute somewhat differentiable entities that might fit in tabular structures as well as in graph-like ones. The cluster groups might give an approximate hint regarding the number of attributes each entity has: taking the average 28 properties and the 8 groups at h=0.1, and the 5 groups at h=0.2 into consideration, an entity of them would statistically consist of 4-6 properties.

VI. Related Work

The LODstats project [23] collects similar statistics, but is limited to Linked Data only. The original setup is also quite different: this field study is implemented on an average laptop, while [23] work with Hadoop. They evaluated more RDF resources, nevertheless, similarities are noticeable: As found here, `rdf:type` and `rdfs:label` are frequently used RDF properties. Moreover, the proportion of "problematic" resources is comparable, c.f. [24].

[2] have introduced a clustering approach to automatically partitioning an RDF dataset with the aim of size-reduction. The methodology thereby involved a bisimulation, while in this paper the counts of co-occurrences of RDF properties are used. Thus, [2] finds subgraphs, consisting of clusters of similar subjects, while the work presented here finds clusters of related subjects.

The Roomba project [4] provides a similar analysis software for CKAN like introduced here. While the present work operates on meta data to provide a big picture on all data, Roomba probes all datasets to detect the mimetype.

VII. Conclusion

This field study provides a means to perform a reality check on what the term Linked and Open Data means, in practice, for the commonly known data portal datahub.io. It shows that real Linked *and* Open Data, ergo 4+-star data, is quite rare, while ordinary Office formats like Excel are twice as common. Next to technical aspects, datasets often do not possess a clear license, and thus, clear terms of use. This leaves a huge uncertainty for developers who want to build applications based on this data. RDF data, if available, does not seem to rely on OWL properties on a broad scale. After all, large amounts of data at datahub.io seem to be of tabular nature, independent from the fact if the actual format demands it (Spreadsheets and CSVs) or or not (RDF).

As mentioned, this work has two limitations: Firstly, while the meta data analysis was based on the entire data library of datahub.io, an in-depth analysis was only conducted for RDF resources with a page-view on this portal larger than 0. Secondly, in some cases, the system used for evaluating did not have enough computing power to load certain RDF resources.

Future work could address both these restrictions, possibly with a Big Data infrastructure, as suggested in [23]. Porting involved analysis tools like LODprobe on Map-Reduce-based jobs has been found to be feasible. Accordingly, in addition, also other CKAN-based data repositories could be analysed in this matter, too.

VIII. Acknowledgements

The authors want to thank Werner Helm and Adrian Schmitt from the Competence Center Stochastics & Operations Research (CCSOR) of the University of Applied Sciences Darmstadt for consulting in the statistical foundation of this work.

References

[1] Complete-linkage clustering, Sept. 2014. Page Version ID: 625941679.

[2] A. Alzogbi and G. Lausen. Similar structures inside rdf-graphs. In C. Bizer, T. Heath, T. Berners-Lee, M. Hausenblas, and S. Auer, editors, *LDOW*, volume 996 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[3] Apache POI. Apache POI - the Java API for Microsoft Documents, 2014. Last access 2015-06-08.

[4] A. Assaf, A. Sénart, and R. Troncy. Roomba: automatic validation, correction and generation of dataset metadata, 05 2015.

[5] T. Berners-Lee. Linked Data. Webpage, June 2006.

[6] M. Ceglowski. Remembrance of Links Past (Pinboard Blog), 05 2011.

[7] Datahub.io. New York Times - Linked Open Data - People (SKOS) - the Datahub, September 2010. Last access 2015-06-02.

[8] Datahub.io. About - the Datahub, 6 2015. Last access 2015-06-11.

[9] J. Edelstein, L. Galla, C. Li-Madeo, J. Marden, A. Rhonemus, and N. Whysel. Linked Open Data for Cultural Heritage: Evolution of an Information Technology. 2013.

[10] J. C. Gower. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4):857–871, Dec. 1971.

[11] T. Heuss. CKANStats datahub.io Cluster Analyses - github.com/heussd/CKANstats/blob/master/datahub.io/clusteranalysis.csv (last accessed 2016-02-26), 2015.

[12] T. Heuss. CKANStats datahub.io Dataset Meta Data - github.com/heussd/CKANstats/blob/master/datahub.io/datahub.io.csv (last accessed 2015-06-10), 2015.

[13] T. Heuss. CKANStats datahub.io Linked Data Download Meta Data - github.com/heussd/CKANstats/blob/master/datahub.io/all_ld_rdf_res _tracking_sum.csv (last accessed 2015-06-02), 2015.

[14] T. Heuss. CKANStats datahub.io RDF Dendrograms - github.com/heussd/CKANstats/tree/master/datahub.io/png (last accessed 2015-06-09), 2015.

[15] T. Heuss. CKANStats Repository - github.com/heussd/CKANstats (last accessed 2015-06-09), 2015.

[16] T. Heuss. LODprobe Source Repository - github.com/heussd/lodprobe (last accessed 2015-06-02), 2015.

[17] A. Isaac, W. Waites, J. Young, and M. Zeng. Library linked data incubator group: Datasets, value vocabularies, and metadata element sets. Technical report, W3C Incubator Group Report, 2005.

[18] M. Morsey, J. Lehmann, S. Auer, and A.-C. N. Ngomo. DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, editors, *The Semantic Web – ISWC 2011*, number 7031 in Lecture Notes in Computer Science, pages 454–469. Springer Berlin Heidelberg, Jan. 2011.

[19] Open Knowledge Foundation. API guide — CKAN 2.4a documentation, 2013. Last access 2015-05-24.

[20] Open Knowledge Foundation. Page View Tracking — CKAN Data Management System Documentation 2.1a documentation, 2013. Last access 2015-05-24.

[21] Open Knowledge Foundation. User guide — CKAN 2.4a documentation, 2013. Last access 2015-05-24.

[22] Richard Cyganiak. lod-cloud/datahub2void - GitHub, 9 2014. Last access 2014-09-21.

[23] S. Stadtmüller, A. Harth, and M. Grobelnik. Accessing Information About Linked Data Vocabularies with vocab.cc. In J. Li, G. Qi, D. Zhao, W. Nejdl, and H.-T. Zheng, editors, *Semantic Web and Web Science*, Springer Proceedings in Complexity, pages 391–396. Springer New York, Jan. 2013.

[24] Webpage. LODStats - 9960 datasets, 6 2015. Last access 2015-06-09.