# What are the Important Properties of an Entity?
## Comparing Users and Knowledge Graph Point of View

Ahmad Assaf[1], Ghislain A. Atemezing[1], Raphaël Troncy[1] and Elena Cabrio[1,2]

[1] EURECOM, Sophia Antipolis, France. `<firstName.lastName@eurecom.fr>`
[2] INRIA, Sophia Antipolis, France. `<elena.cabrio@inria.fr>`

**Abstract.** Entities play a key role in knowledge bases in general and in the Web of Data in particular. Entities are generally described with a lot of properties, this is the case for DBpedia. It is, however, difficult to assess which ones are more "important" than others for particular tasks such as visualizing the key facts of an entity or filtering out the ones which will yield better instance matching. In this paper, we perform a reverse engineering of the Google Knowledge graph panel to find out what are the most "important" properties for an entity according to Google. We compare these results with a survey we conducted on 152 users. We finally show how we can represent and explicit this knowledge using the Fresnel vocabulary.

**Keywords:** Entities, Google Knowledge Graph, knowledge extraction

## 1 Introduction

In many knowledge bases, entities are described with numerous properties. However, not all properties have the same importance. Some properties are considered as keys for performing instance matching tasks while other properties are generally chosen for quickly providing a summary of the key facts attached to an entity. Our motivation is to provide a method enabling to select what properties should be used when depicting the summary of an entity, for example in a multimedia question answering system such as QakisMedia[3] or in a second screen application providing more information about a particular TV program[4]. Our approach consists in: (i) reverse engineering the Google Knowledge Panel by extracting the properties that Google considers as sufficiently important to show (Section 2), and (ii) analyzing users' preferences by conducting a user survey and comparing the results (Section 3). We finally show how we can explicitly represent this knowledge of preferred properties to attach to an entity using the Fresnel vocabulary before concluding (Section 4).

## 2 Reverse Engineering the Google KG Panel

Web scraping is a technique for extracting data from Web pages. We aim at capturing the properties depicted in the Google Knowledge Panel (GKP) that are

[3] `http://qakis.org/`
[4] `http://www.linkedtv.eu/demos/linkednews/`

injected in search result pages [54]. We have developed a Node.js application that queries all DBpedia concepts that have at least one instance which is `owl:sameAs` with a Freebase resource in order to increase the probability that the search engine result page (SERP) for this resource will contain a GKP. We assume in our experiments that the properties displayed for an entity are type and context dependent (country, time, query) which can affect the results. Moreover, we filter out generic concepts by excluding those who are direct subclasses of `owl:Thing` since they will trigger ambiguous queries. We obtained a list of 352 concepts[5]. For each of these concepts, we retrieve $n$ instances (in our experiment, $n$ was equal to 100 random instances). For each of these instances, we issue a search query to Google containing the instance label. Google does not serve the GKP for all user agents and we had to mimic a browser behavior by setting the $User-Agent$ to a particular browser. We use CSS selectors to check the existence of and to extract data from a GKP. An example of a query selector is *._om* (all elements with class name *_om*) which returns the property DOM element(s) for the concept described in the GKP. From our experiments, we found out that we do not always get a GKP in a SERP. If this happens, we try to disambiguate the instance by issuing a new query with the concept type attached. However, if no GKP was found again, we capture that for manual inspection later on. Listing 1 gives the high level algorithm for extracting the GKP. The full implementation can be found at `https://github.com/ahmadassaf/KBE`. We finally observe that this experiment is only valid for the English Google.com search results since GKP varies according to top level names.

**Algorithm 1** Google Knowledge Panel reverse engineering algorithm

```
INITIALIZE $equivalentClasses(DBpedia, Freebase)$ AS $vectorClasses$
Upload $vectorClasses$ for querying processing
Set $n$ AS number-of-instances-to-query
for each $conceptType \in vectorClasses$ do
    SELECT $n$ instances
    $listInstances \leftarrow$ SELECT-SPARQL($conceptType$, $n$)
    for each $instance \in listInstances$ do
        CALL http://www.google.com/search?q=$instance$
        if $knowledgePanel$ exists then
            SCRAP GOOGLE KNOWLEDGE PANEL
        else
            CALL http://www.google.com/search?q=$instance + conceptType$
            SCRAP GOOGLE KNOWLEDGE PANEL
        end if
        $gkpProperties \leftarrow$ GetData(DOM, EXIST(GKP))
    end for
    COMPUTE occurrences for each $prop \in gkpProperties$
end for
return $gkpProperties$
```

## 3 Evaluation

We conducted a user survey in order to compare what users think should be the important properties to display for a particular entity and what the GKP shows.

[5] See also the SPARQL query at `http://goo.gl/EYuGm1`

**User survey.**

We set up a survey[6] on February 25th, 2014 and for three weeks in order to collect the preferences of users in term of the properties they would like to be shown for a particular entity. We select only one representative entity for nine classes: `TennisPlayer`, `Museum`, `Politician`, `Company`, `Country`, `City`, `Film`, `SoccerClub` and `Book`. 152 participants have provided answers, 72% from academia, 20% coming from the industry and 8% having not declared their affiliation. 94% of the respondents have heard about the Semantic Web while 35% were not familiar with specific visualization tools. The detailed results[7] show the ranking of the top properties for each entity. We only keep the properties having received at least 10% votes for comparing with the properties depicted in a KGP. We observe that users do not seem to be interested in the `INSEE code` identifying a French city while they expect to see the `population` or the `points of interest` of this city.

***Comparison with the Knowledge Graphs.*** The results of the Google Knowledge Panel (GKP) extraction[8] clearly show a long tail distribution of the properties depicted by Google, with a top N properties (N being 4, 5 or 6 depending on the entity) counting for 98% of the properties shown for this type. We compare those properties with the ones revealed by the user study. Table 1 shows the agreement between the users and the choices made by Google in the GKP for the 9 classes. The highest agreement concerns the type `Museum` (66.97%) while the lowest one is for the `TennisPlayer` (20%) concept. We think properties for museums or books are more stable than for types such as person/agent which vary significantly. We acknowledge the fact that more than one instance should be tested in order to draw meaningful conclusion regarding what are the important properties for a type. With this set of 9 concepts, we are covering 301, 189 DBpedia entities that have an existence in Freebase, and for each of them, we can now empirically define the most important properties when there is an agreement between one of the biggest knowledge base (Google) and users preferences.

| **Classes** | TennisPlayer | Museum | Politician | Company | Country | City | Film | SoccerClub | Book |
|---|---|---|---|---|---|---|---|---|---|
| **Agr.** | 20% | 66.97% | 50% | 40% | 60% | 60% | 60% | 50% | 60% |

**Table 1.** Agreement on properties between users and the Knowledge Graph Panel

***Modeling the preferred properties with Fresnel.*** Fresnel[9] is a presentation vocabulary for displaying RDF data. It specifies *what* information contained in an RDF graph should be presented with the core concept `fresnel:Lens` [24]. We

[6] The survey is at `http://eSurv.org?u=entityviz`

[7] `https://github.com/ahmadassaf/KBE/blob/master/results/agreement-gkp-users.xls`

[8] `https://github.com/ahmadassaf/KBE/blob/master/results/survey.json`

[9] `http://www.w3.org/2005/04/fresnel-info/`

use the Fresnel and PROV-O ontologies[10] to explicitly represent what properties should be depicted when displaying an entity. This dataset can now be re-used as a configuration for any consuming application.

```
:tennisPlayerGKPDefaultLens rdf:type fresnel:Lens ;
  fresnel:purpose fresnel:defaultLens ;
  fresnel:classLensDomain dbpedia-owl:TennisPlayer ;
  fresnel:group :tennisPlayerGroup ;
  fresnel:showProperties (dbpedia-owl:abstract dbpedia-owl:birthDate
    dbpedia-owl:birthPlace dbpprop:height dbpprop:weight
    dbpprop:turnedpro dbpprop:siblings) ;
  prov:wasDerivedFrom
    <http://www.google.com/insidesearch/features/search/knowledge.html> .
```

**Listing 1.1.** Excerpt of a Fresnel lens in Turtle

## 4 Conclusion and Future Work

We have shown that it is possible to reveal what are the “important” properties of entities by reverse engineering the choices made by Google when creating knowledge graph panels and by comparing users preferences obtained from a user survey. Our motivation is to represent this choice explicitly, using the Fresnel vocabulary, so that any application could read this configuration file for deciding which properties of an entity is worth to visualize. This is fundamentally different from the work in [60] where the authors created a generalizable approach to open up closed knowledge bases like Google’s by means of crowd-sourcing the knowledge extraction task. We are aware that this knowledge is highly dynamic, the Google Knowledge Graph panel varies across geolocation and time. We have provided the code that enables to perform new calculation at run time and we aim to study the temporal evolution of what are important properties on a longer period. This knowledge which has been captured will be made available shortly in a SPARQL endpoint. We are also investigating the use of Mechanical Turk to perform a larger survey for the complete set of DBpedia classes.

***Acknowledgments.*** This work has been partially supported by Datalift (ANR-10-CORD-009), UCN (ANR-11-LABX-0031-01) and LinkedTV (GA 287911).

## References

1.
2. Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining RDF data with ProLOD++. In *30th IEEE International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014.
3. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.
4. J. Anja, C. Richard, and B. Chrstian. State of the lod cloud. http://lod-cloud.net/state/.

[10] http://www.w3.org/TR/prov-o/

5. Assaf Ahmad, Sénart Aline, and Troncy Raphaël. Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In *24$^{th}$ World Wide Web Conference (WWW), Demos Track*, Florence, Italy, 2015.
6. S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats - an Extensible Framework for High-performance Dataset Analytics. In *18$^{th}$ International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, Galway, Ireland, 2012.
7. S. Ben. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.
8. C. Böhm, G. Kasneci, and F. Naumann. Latent Topics in Graph-structured Data. In *21$^{st}$ ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2663–2666, Maui, Hawaii, USA, 2012.
9. C. Bohm, F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with ProLOD. In *26th International Conference on Data Engineering Workshops (ICDEW)*, 2010.
10. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *ACM International Conference on Management of Data (SIGMOD)*, 2008.
11. D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.
12. B. Christian. Evolving the Web into a Global Data Space. In *28$^{th}$ British National Conference on Advances in Databases*, 2011.
13. B. Christian, L. Jens, K. Georgi, A. Sören, B. Christian, C. Richard, and H. Sebastian. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3), 2009.
14. B. Christian, H. T, and B.-L. T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
15. B. Christoph, L. Johannes, and N. Felix. Creating voiD Descriptions for Web-scale Data. *Journal of Web Semantics*, 9(3):339–345, 2011.
16. M. Cornolti, P. Ferragina, and M. Ciaramita. A Framework for Benchmarking Entity-annotation Systems. In *22$^{nd}$ World Wide Web Conference (WWW)*, 2013.
17. R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In *5$^{th}$ European Semantic Web Conference (ESWC)*, pages 690–704, Tenerife, Spain, 2008.
18. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing Linked Datasets with the VoID Vocabulary. W3C Note, 2011. `http://www.w3.org/TR/void/`.
19. M. d'Aquin and E. Motta. Watson, More Than a Semantic Web Search Engine. *Semantic Web Journal*, 2011.
20. R. Dave. The Organization Ontology. W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-org`.
21. R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. In *7$^{th}$ European Semantic Web Conference (ESWC)*, 2010.
22. L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. In *13$^{st}$ ACM International Conference on Information and Knowledge Management (CIKM)*, 2004.
23. N. Douglas. Developing Spatial Data Infrastructures: The SDI Cookbook, 2004. `http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf`.

24. P. Emmanuel, B. Christian, K. David, and L. Ryan. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *5th International Semantic Web Conference (ISWC'06)*, pages 158–171, 2006.
25. D.-A. Ernesto, D. Lucas, S.-T. Lars, and N. Wolfgang. Real-time top-n recommendation in social streams. In *6th ACM conference on Recommender systems - RecSys*, 2012.
26. B. Eytan, R. Itamar, M. Cameron, and A. Lada. The role of social networks in information diffusion. In *12th International Conference on World Wide Web (WWW'12)*, 2012.
27. M. Fadi and E. John. Data Catalog Vocabulary (DCAT). W3C Recommendation, 2014. http://www.w3.org/TR/vocab-dcat/.
28. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. In *11th European Semantic Web Conference (ESWC)*, 2014.
29. B. Forchhammer, A. Jentzsch, and F. Naumann. LODOP - Multi-Query Optimization for Linked Data Profiling Queries. In *International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES)*, Heraklion, Greece, 2014.
30. M. Frosterus, E. Hyvönen, and J. Laitio. Creating and Publishing Semantic Metadata about Linked and Open Datasets. In *Linking Government Data*. 2011.
31. M. Frosterus, E. Hyvönen, and J. Laitio. DataFinland - A Semantic Portal for Open and Linked Datasets. In *8th Extended Semantic Web Conference (ESWC)*, pages 243–254, 2011.
32. T. Giovanni, C. Richard, C. Michele, D. Szymon, D. Renaud, and D. Stefan. Sig.ma: Live views on the Web of data. *Journal of Web Semantics*, 8(4), 2010.
33. A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data Summaries for On-demand Queries over Linked Data. In *19th World Wide Web Conference (WWW)*, 2010.
34. K. Houda, A. Ghislain, R. Giuseppe, and R. Troncy. Aggregating Social Media for Enhancing Conference Experiences. In *1st Internationl Workshop on Real-Time Analysis and Mining of Social Streams*, 2012.
35. R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data. In *9th International Semantic Web Conference (ISWC), Posters & Demos Track*, 2010.
36. C. Iván and B. Alejandro. Semantic contextualisation of social tag-based profiles and item recommendations. In *12th Internationl Conference on E-Commerce and Web Technolgoies*, 2011.
37. P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's No Money in Linked Data, 2013. http://knoesis.wright.edu/faculty/pascal/pub/nomoneylod.pdf.
38. M. James and D. E. Almasi. Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey Business Technology Office, 2001.
39. A. Jentzsch. Profiling the Web of Data. In *13th International Semantic Web Conference (ISWC), Doctoral Consortium*, Trentino, Italy, 2014.
40. T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing Linked Data Dynamics. In *10th European Semantic Web Conference (ESWC)*, 2013.
41. S. Khatchadourian and M. P. Consens. ExpLOD: Summary-based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *7th Extended Semantic Web Conference (ESWC)*, pages 272–287, Heraklion, Greece, 2010.

42. M. Konrath, T. Gottron, S. Staab, and A. Scherp. SchemEX - Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. *Journal of Web Semantics*, 16, 2012.
43. Kovács-Láng. Global Terrestrial Observing System. Technical report, GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.
44. S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain. Automatic Domain Identification for Linked Open Data. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 205–212, 2013.
45. A. Langegger and W. Woss. RDFStats - An Extensible RDF Statistics Generator and Library. In *20th International Workshop on Database and Expert Systems Application (DEXA)*, pages 79–83, 2009.
46. J. Leskovec and C. Faloutsos. Sampling from Large Graphs. In *12thth ACM International Conference on Knowledge Discovery and Data Mining (KDD'12)*, 2006.
47. H. Li. Data Profiling for Semantic Web Data. In *International Conference on Web Information Systems and Mining (WISM)*, pages 472–479, 2012.
48. E. Mäkelä. Aether - Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets. In *11th European Semantic Web Conference (ESWC), Demo Track*, Heraklion, Greece, 2014.
49. P. Marco and G. Siva. Investigating topic models for social media user recommendation. In *11th International Conference on World Wide Web (WWW'11)*, 2011.
50. B. Martin, B. Ciro, E. Ivan, F. Markus, K. Dimitris, and H. Sebastian. DataID: Towards Semantically Rich Metadata for Complex Datasets. In *10th International Conference on Semantic Systems*, 2014.
51. S. Max, B. Christian, and P. Heiko. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th International Semantic Web Conference (ISWC)*, 2014.
52. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems*, 2011.
53. H. Michael, H. Wolfgang, R. Yves, F. Lee, and A. Danny. SCOVO: Using Statistics on the Web of Data. In *ESWC*, 2009.
54. B. Mike. Deconstructing the Google Knowledge Graph. `http://www.mkbergman.com/1009/deconstructing-the-google-knowledge-graph`.
55. A. Nikolov, M. d'Aquin, and E. Motta. What Should I Link to? Identifying Relevant Sources and Classes for Data Linking. In *Joint International Semantic Technology Conference (JIST)*, 2011.
56. A. Phil and S. Gofran. Asset Description Metadata Schema (ADMS). W3C Working Group Note, 2013. `http://www.w3.org/TR/vocab-adms`.
57. D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. In *6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.
58. N. Press. Understanding Metadata. Technical report, National Information Standards Organization, 2004.
59. I. Renato and M. James. vCard Ontology - for describing People and Organizations. W3C Interest Group Note, 2014. `http://www.w3.org/TR/vcard-rdf`.

60. T. Steiner and S. Mirea. SEKI@home or Crowdsourcing an Open Knowledge Graph. In *1$^{st}$ International Workshop on Knowledge Extraction & Consolidation from Social Media (KECSM'12)*, Boston, USA, 2012.
61. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16$^{th}$ International World Wide Web Conference (WWW)*, 2007.
62. B.-L. Tim. Linked Data - Design Issues. W3C Personal Notes, 2006. `http://www.w3.org/DesignIssues/LinkedData`.
63. L. Timothy, S. Satya, and M. Deborah. PROV-O: The PROV Ontology. W3C Recommendation, 2013. `http://www.w3.org/TR/prov-o`.
64. R. Usbeck, M. Röder, A.-C. Ngonga-Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL - General Entity Annotation Benchmark Framework. In *24$^{th}$ World Wide Web Conference (WWW)*, 2015.
65. Z. Valentina and C. L. Social ranking: uncovering relevant content using tag-based recommender systems. In *2$^{nd}$ ACM conference on Recommender systems - RecSys*, 2008.
66. G. Vickery. Review of Recent Studies on PSI-use and Related Market Developments. Technical report, EC DG Information Society, 2011.
67. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment Methodologies for Linked Open Data. *Semantic Web Journal*, 2012.