# Nobody Wants to Live in a Cold City where no Music Has Been Recorded

## Analyzing Statistics with *Explain-a-LOD*

Heiko Paulheim

Technische Universität Darmstadt
Knowledge Engineering Group
paulheim@ke.tu-darmstadt.de

**Abstract.** While it is easy to find statistics on almost every topic, coming up with an explanation about those statistics is a much more difficult task. This demo showcases the prototype tool *Explain-a-LOD*, which uses background knowledge from DBpedia for generating possible explanations for a statistic[1].

## 1 Introduction

Every year, Mercer Research publishes a ranking of the most and the least livable cities in the world. For its current version, people in 221 cities have been interviewed and asked for the perceived quality of living in their city[2].

Statistics like these are widely spread and frequently cited, e.g., in the newspapers. However, what we are typically interested in is asking: *why* are the values in a particular statistics the way they are? Looking at the Mercer example, a typical question would be: *What is it that makes Vienna (which is at the top position) more livable than, e.g., Dubai (which is on position 74)?*

In order to come up with hypotheses for answering such questions, *background knowledge* is required, and we need to find out more information about the cities. Factors that could be of interest to answering our questions could be dealing with the climate, the economy, the cultural live, the population density, and so on. Therefore, the first task is to enhance the statistics file at hand with more background information. Once this is done, tools for correlation analysis can be run on the enhanced file for finding possible hypotheses.

Compiling that background knowledge manually is a labour-intensive task, and it is prone to a priori biases – since we have an initial feeling for which information could be relevant, we are likely to include some pieces information and discard others. Thus, an automatic system for compiling the background information would be desirable. Explain-a-LOD, the tool introduced in this demo[3],

[1] This demo accompanies the paper *Generating Possible Interpretations for Statistics from Linked Open Data* [1], also included in these proceedings.

[2] http://www.mercer.com/articles/quality-of-living-survey-report-2011

[3] http://www.ke.tu-darmstadt.de/resources/explain-a-lod

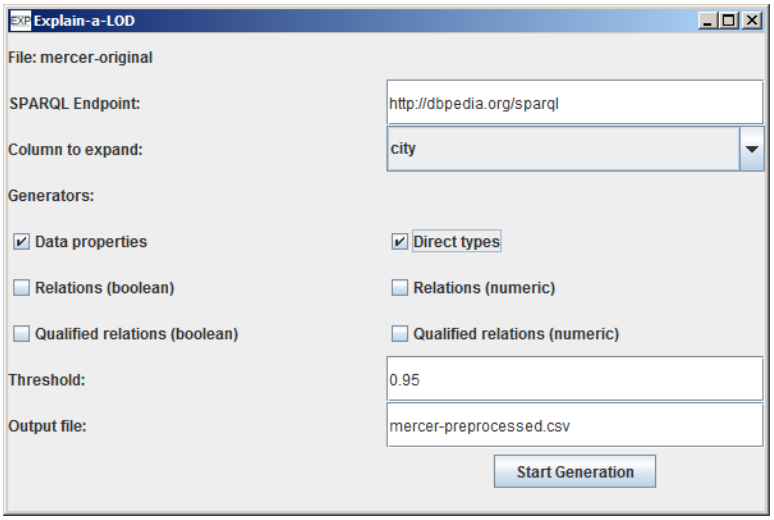


**Fig. 1.** Preprocessing a statistics file

uses data sources in Linked Open Data [2] for adding background knowledge to a statistic in a fully automated manner.

## 2 The Explain-a-LOD Workflow

Statistics are most often tables, thus, the workflow of Explain-a-LOD starts with such a table, e.g., a CSV file. The user can import such a file, specify a column name which contains the entites to gather background information for (e.g., a column with city names), and select a couple of generators and a relevance threshold for the newly generated features, as shown in Fig. 1. The preprocessed file may also be stored for later use.

Different generators are available for adding background information (see [3] for details):

- Data attributes can be added for all datatype properties. For example, a column *population* is introduced in each row of the cities statistics, which reflects the value of the `DBpedia:population` value of the respective entity.
- Direct types can be added as boolean columns. For example, the column `EuropeanCapitals` is added with value *true* for Vienna, and with value *false* for Dubai.
- Incoming and outgoing relations can be added either as boolean or numeric columns. For example, if there are any albums recorded in a city, i.e., there are incoming relations of the type `recordedIn`, the column *recordedIn_in* is filled with *true* or a positive number, with *false* or zero otherwise.
- Qualified relations may also be added, taking into account the type of the related object. For example, since Vienna is the headquarter of the organization *OPEC*, a boolean or numeric attribute *headquarter_in_Organization*

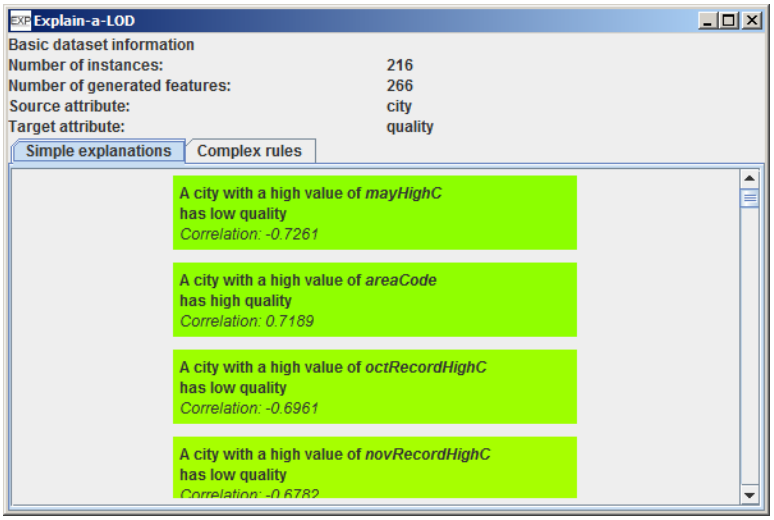


**Fig. 2.** Hypotheses generated for a statistics file

can be introduced, depicting whether the city is a `dbpedia:headquarter` of any organization, or the number of such organizations, respectively.

Once that additional data is added to the original dataset, Explain-a-LOD will start analyzing the data and try to formulate hypotheses. Two strategies are used: simple correlations are sought by analyzing the correlation coefficient between each column generated and the statistic's target value, and different rule learners are run on the dataset for formulating more complex hypotheses, using the *Weka* machine learning framework [4].

The hypotheses found are presented to the user in two lists, using color codings for the machine's confidence in those hypotheses (the correlation coefficient or the confidence of a rule, respectively), as shown in Fig. 2. The colors range from green (high confidence) to red (low confidence).

## 3 Example Hypotheses

We have created a set of hypotheses, using the different generation strategies discussed above, and have had them rated in the form a questionnaire in a user study (see [1] for details on the user study). The top-rated hypotheses were[4]:

1. Cities where many things take place have a high quality of living.
2. European capitals of culture have a high quality of living.
3. African capitals have a low quality of living.

[4] The full list of hypotheses and their ratings can be found at `http://www.ke.tu-darmstadt.de/resources/explain-a-lod/user-study`

4. Host cities of olympic summer games have a high quality of living.
5. Cities where at least 73 things are located have a high quality of living.

The first and the last hypothesis have been generated by exploiting unqualified relations, while the second, third, and fourth have been generated from direct types (e.g., *YAGO*, which is used for types in DBpedia, defines types such as *EuropeanCapitalsOfCulture* or *HostCitiesOfOlympicSummerGames*). The last hypothesis has been generated by a rule learning algorithm, which cannot only find a correlation between an attribute and the target, but also an optimal point for splitting the dataset into positive and negative examples (i.e., high and low quality cities).

While many of the hypotheses generated make sense to the users, the tool also produces some not-so well perceived hypotheses. Examples include:

1. Cities with a large latitude have a high quality of living.
2. Cities where many bands founded in 2004 originate have a high quality of living.
3. Cities where nothing has been recorded and where the maximum temperature in January does not exceed 16°C have a low quality of living.

Those examples point at challenging problems with the approach. The first hypothesis shows that the tool often cannot verbalize a hypothesis in a way that satisfies the end user. In fact, the latitude of a city is a good indicator for separating cities into cities in the first world and cities in the third world. It can be assumed that the rating for a re-formulated hypothesis like *First world cities have a high quality of living* would have been much higher, but the tool cannot detect which of the two variants will be more plausible to the user.

The second example points to a problem with DBpedia: it has a strong bias towards popular culture, especially Northern American and European popular culture. Thus, hypotheses with references to popular culture appear quite frequently, although they are in many cases not plausible. Due to that bias, that hypotheses mainly refers to the Northern American and European countries.

The third example also points to the bias problem, but also includes another challenge: at the moment, the tool is not capable of generating hypotheses that are coherent in themselves. For example, cultural life (expressed in music recorded in a city) and climate (expressed in the January temperature) may both influence the quality of living in a city, but in the users' perception, they are not interrelated. Thus, such hypotheses are ranked very low by users, although they may be quite accurate. Finding and implementing metrics for coherent rules could help remedying this problem.

## 4 Conclusion and Future Work

In this demo, we have introduced the *Explain-a-LOD*, which uses background information from Linked Open Data for enriching statistics, and which is capable of coming up with hypotheses for explaining a statistic in a fully automated way.

The tool and the hypotheses it creates have been tested with a larger number of users. In this paper, we have shown examples both for high and low ranked hypotheses, and discussed some reasons that lead to the generation of the latter.

While Explain-a-LOD is currently a prototype which can be used on various statistics datasets, there are many interesting research questions. Some of those have already been touched by the examples above: using data from the semantic web, dealing with biases, incompleteness and faultiness of data is an issue. Separating useful from useless, plausible from implausible hypotheses is also an issue which cannot be addressed trivially. Further research problems cover issues such as scalability, especially when using more complex generation strategies, or producing an intuitive verbalization and visualization of hypotheses.

Current plans of extending Explain-a-LOD include the combination of different datasets. For many statistical datasets, sources such as World Fact Book or Eurostat may be ideal candidates for generating background knowledge, while for others, such as statistics about the box office revenue of films, specialized data sets such as Linked Movie Database might be more suitable. Picking relevant datasets fully automatically for different statistics would be desirable, but requires some more in-depth research. Using table extraction mechanisms could be a way to also include tabular data from non-LOD sources.

In summary, Explain-a-LOD showcases an approach employing Linked Open Data for a use case which has not been addressed much in the past. The results prove that the approach is feasible and open up a number of interesting research questions. During the demo, the visitors will be able to try the prototype with different datasets by themselves.

**Acknowledgements**

This work was supported by the German Science Foundation (DFG) project FU 580/2 "Towards a Synthesis of Local and Global Pattern Induction (GLoc-Syn)".

## References

1. Paulheim, H.: Generating Possible Explanations for Statistics from Linked Open Data. In: 9th Extended Semantic Web Conference (ESWC). (2012)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22
3. Paulheim, H., Fürnkranz, J.: Unsupervised Feature Generation from Linked Open Data. In: International Conference on Web Intelligence, Mining, and Semantics (WIMS'12). (2012)
4. Bouckaert, R.R., Frank, E., Hall, M., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA — Experiences with a Java open-source project. Journal of Machine Learning Research **11** (September 2010) 2533–2541