Towards a Vocabulary for Incorporating Predictive Models into the Linked Data Web

Evangelos Kalampokis[1,2], Areti Karamanou[1,2], Efthimios Tambouris[1,2], and Konstantinos Tarabanis[1,2]

[1] Information Systems Lab, University of Macedonia,
Egnatia 156, 54006 Thessaloniki, Greece
[2] Informatics and Telematics Institute, Centre for Research & Technology - Hellas
6th km Xarilaou - Thermi, 57001, Thessaloniki
{ekal,akarm,tambouris,kat}@uom.gr

Abstract. Predictive modeling reflects the process of using data and statistical or data mining methods for predicting new observations. The predictive models that are created out of this process could be reused in different applications in the same sense that open data is reused. Towards this end, a few standards have been proposed in order to enable transfer of predictive models across platforms and applications. In this paper we suggest the need for incorporating predictive models into the Linked Data Web. Towards this end, we propose an RDF Schema vocabulary that will enable the creation of predictive models descriptions adhering to the Linked Data principles. The incorporation of these descriptions into the Linked Data Web could create new potentials beyond cross-platform model reuse. In particular, it will enable (a) easy discovery and reuse of appropriate models at a Web Scale and (b) creation of more accurate models exploiting connections of models to other models, datasets and other resources on the Web.

Keywords: Linked data, statistical data, predictive analytics, vocabulary, RDF, predictive model, interoperability

1 Introduction

In the context of quantitative empirical modeling, the term *predictive analytics* refers to the building and assessment of a model aimed at making empirical predictions using data and statistical or data mining methods [1]. In general, the goal of predictive models is to predict the output value (Y) for new observations given their input values (X). The inputs are often called the predictors, and more classically the independent variables while the outputs are called the response, or classically the dependent variables. Examples of predictive models consider the prediction of stock market volatility from Yahoo! Finance message board [2], movies success from weblog content [3], influenza-like illnesses from Google search queries [4], product sales from Amazon reviews [5] and levels of rainfall from Twitter posts [6].

These predictive models can be reused by different applications, in the same way that open data is reused [7, 8]. Towards this end, standardization activities such as the Predictive Model Markup Language (PMML)[3] has been suggested. This XML–based language enables importing and exporting developed models as components in other processes and systems using XML files. However, the discovery of an appropriate model for a task at hand is at the moment a time consuming activity that requires a lot of manual effort involving searching in scientific articles, contacting researchers or professionals and exchanging files.

In addition, predictive models incorporate knowledge about a domain or a problem area. For example, a model could include variables that effect economic development. However, usually more than one models can be created about a specific problem using different data and statistical or data mining methods. These models provide fragmented views on a specific problem. Moreover, these views could be either complementary or controversial. As a result the capability of connecting these different views could enhance the understanding of a problem and could facilitate the building of more accurate models.

At the same time, the adoption of the Linked Data principles and technologies [9] has promised to enhance the analysis of *statistical data* at a Web scale. For example, Linked Data could facilitate performing data analytics on top of combined statistical datasets that were previously closed in disparate sources and can now be linked in order to provide unexpected and unexplored insights into different domains and problem areas [10]. Moreover, linking statistical data to the Linked Data Web could enable the enrichment of a particular dataset and thus the extraction of interesting and previously hidden insights related to particular events [11].

In this paper we suggest that the incorporation of predictive models into the Linked Data Web could enable new potentials beyond the reuse of models across different platforms. In particular, this could enable the discovery of predictive models at a Web scale in an easy and effective manner. For example, it will make possible queries such as "On which data mining method the most accurate model that predicts influenza-like illnesses from Google queries is based on?" or "What predictor variables should a model aiming at predicting unemployment include?". Moreover, in this paper we propose an RDF Schema vocabulary, named the Linked Statistical Models *(limo)* vocabulary, that will enable the incorporation of descriptions of predictive models into the Linked Data Web and establish links to other resources such as datasets, other models, academic articles and studies.

The remaining of the paper is organized as follows. In section 2 we describe the motivation behind the incorporation of predictive models descriptions into the Linked Data Web. In section 3 we present related work regarding (a) existing endeavors for describing predictive models and (b) widely used RDF vocabularies in the area of statistics. Section 4 presents the Linked Statistical Models *(limo)* vocabulary. Finally, in section 5 a number of use cases are presented while in section 6 conclusions are drawn along with future work.

[3] http://www.dmg.org

2 Motivation

Different models could present controversial results in the same problem area and for the same variables depending on the statistical methods and/or the data that have been employed. For example, Chiricos [12] reviewed 68 studies about the relationship between crime and the unemployment rate and he found that only less than half of these studies have found positive significant effects of the unemployment on crime rates. In addition, Kalampokis et al. [13] reviewed 52 empirical predictive models that employ predictors related to Social Media. They identified that the predictive power of a model is directly related to the predictors, the statistical method, the datasets and the evaluation method that have been selected. Thus, in order to better understand a problem we need to be able to discover and analyze various models that share common characteristics.

In addition, statistical models that have been developed based on a specific dataset can indeed be reused in another case. For example, a model developed for predicting sales based on data from Company X could be efficiently reused with data from Company Z. Moreover, a model predicting sales using a specific data mining method can be reused as a baseline for another model that uses a different method.

Publishing descriptions of statistical models on the Web following the Linked Data principles could have the following benefits:

1. Discovery of variables that a predictive relationship between them have been suggested by an empirical model. For example, it will be possible to discover that X number of models show a predictive relationship between product sales and advertising budget while Z number of models show a negative or no relationship between them.
2. Discovery of all predictor variables that are connected to product sales through successful empirical predictive models.
3. Discovery of statistical or data mining methods that have been used to identify relationships between variables. For example, most of the models that are able to accurately predict product sales from advertising budget have used linear regression methods.
4. Discovery of datasets that have been used to identify predictive relationships between variables. For example, models that show a strong predictive relationship between product sales and advertising budget have employed data from the U.S. in the period between 1975 and 2004.
5. Discovery of a specific predictive model that shows a relationship between variables based on aspects such as its creator, the affiliation of the creator, the journal that the results have been published in, etc.
6. Discovery of new datasets in order to reuse existing models. For example, identification of datasets in Europe from the last ten years in order to reuse a predictive model produced with data from the U.S.
7. Discovery of predictive models that could be used as baseline models in building new more accurate predictive models.

These benefits will be achieved only if a vocabulary to model predictive models as RDF will be specified and Linked Data descriptions of predictive models will be published at a wide range. The scoping of this paper focuses on the former.

3 Related Work

The Predictive Model Markup Language (PMML) is an XML standard that represents and describes data mining and statistical models, as well as some of the operations required for cleaning and transforming data prior to modeling [14]. PMML aims to provide enough infrastructure for an application to be able to produce a model and another application to consume by reading the PMML XML data file [15]. PMML has the following general structure:

- The *Header* contains information about the application generated the model including a time stamp.
- The *Mining Build Task* contains vendor specific information about how the model was built.
- The *Data Dictionary* contains details about the variables, called *Data Fields* that participate in the model. These can be though of representing the actual data used to develop the model including information such as the name, the type of data (e.g. string, numeric) and how it is used (e.g. is it a continuous numeric value, a categorical value, etc.).
- The *Transformation Dictionary* describes how to manipulate the data fields from the data dictionary into variables that exist within the PMML definition. This includes normalization, discretization, value mapping etc.
- The *Model* that contains model-specific features according to the model types (e.g. association rules, clustering, general regression, support vector machines, and neural networks). For instance, the *NeuralNetwork* element includes the *activationFunction* attribute that specifies the activation function to be used by the network neurons when processing incoming data. Furthermore, it contains elements that are common to all model types such as *Outputs* that define the different types of results (e.g. *predictedValue*, *standardError*, *probability*, *residual*) that can be generated by a model and *Mining Schema* that defines what to do in case any of the data fields defined in the DataDictionary element are missing or contain invalid or outlier values.

This structure presents only top-level elements. PMML is a very rich language that specifies a very big number of both elements and attributes that are related to data setup, data pre-processing and model representation. All these elements aim at enabling model reuse across heterogeneous platforms and environments for all the major statistical and data mining techniques. We should, however, note that the first version of *limo* that is presented in this paper does not intend to cover all the details required for importing and executing a predictive model into an actual platform.

In addition, a number of widely used RDF Schema vocabularies are closely related to statistics. The DDI-RDF vocabulary [16] focuses on raw record-level datasets and describes their structure, while the RDF Data Cube vocabulary [17] aims at multidimensional aggregated data and provides for the description of both the structure and the actual data of a dataset.

The DDI-RDF contains the `disco:aggregation` property that indicates that a `qb:DataSet` was derived by aggregating a record-level dataset. Moreover, the DDI-RDF vocabulary uses the class `disco:Variable` which provides a definition of the column in a rectangular data file and thus enables understanding of the content of a dataset. In addition, it defines the `disco:LogicalDataSet` class and subclass of `dcat:Dataset` to provide a description of the content of a data set. `disco:LogicalDataSet` is associated with `disco:DataFile`, subclass of `dcat:Distribution` as well as of `dctype:Dataset`, that actually represents the physical subsistence of the data set. A `disco:LogicalDataset` is organized into a set of instances of `disco:Variable`.

In RDF Data Cube, the `qb:DataSet` represents the resource of the entire data set, a data set that corresponds to the defined structure of the RDF Data Cube. The data sets are allowed to be organized in several slices. The structure of `qb:DataSet` or of a slice of the actual data is defined by the class `qb:DataStructureDefinition`. `qb:DataStructureDefinition` associates to the qb:component property in order to specify the component(s) of the datasets structure. The `qb:ComponentProperty` is the super class property of the properties that represent dimensions, measures and attributes namely `qb:DimensionProperty`, `qb:MeasureProperty` and `qb:AttributeProperty` respectively.

Finally, based on these modeling endeavors a number of open statistical datasets published by important international organizations such as the OECD, the World Bank and the IMF have been transformed to Linked Data by third parties [18, 19]. Towards this end, a number of tools have been developed. For example, Capadisli et al. [18] created a tool for transforming statistical data from SDMX-ML format to Linked Data while Salas et al. [20] from CSV and OLAP databases.

4 The *limo* Vocabulary

In this section we present the RDF Linked Statistical Models *(limo)* vocabulary that allows for the description of statistical and data mining models in the RDF model and thus enables the incorporation of these models on the Linked Data Web and linking to others resources such as datasets, organizations, people and articles.

In general, predictive analytics comprise predictive models designed for predicting new (or future) observations or scenarios as well as methods for evaluating the predictive power of a model [21]. The outcome value for a new set of observation could be continuous (or quantitative) or categorical (or qualitative). In the former case the problem is ofter referred to as a regression problem while in the latter a classification problem. Predictive power refers to an empirical

models ability to predict new observations accurately. In contrast, explanatory power refers to the strength of association indicated by a statistical model.The predictive power of a model should be tasted based on out-of-sample data (e.g. cross-validation or a holdout sample) and with adequate predictive measures (e.g. RMSE, MAPE, PRESS etc.). A popular method to obtain out-of-sample data is to initially partition the data randomly, using one part (the training set) to fit the empirical model, and the other (the holdout set) to assess the model's predictive accuracy.

The vocabulary's main classes are depicted in Fig. 1. Classes and properties from existing widely used vocabularies were reused whenever possible.

- `limo:Model` is the actual predictive model that is described by the vocabulary. The model has the following attributes:
  - dct:title which is a name given to describe the model.
  - dct:description for a descriptive comment about the model and its goals.
  - dct:issued which defines the actual data that the model has been created.
  - limo:modelType which describe the main categories of models that can be developed, namely classification, regression, clustering and dimensionReduction.
  - limo:spatial is an attribute that describe the spatial dimension of the model. The spatial dimension of the model is derived from the actual data that have been employed. For example, a model could have `limo:spatial` U.S. in the case the data used for the development of the model comes from the U.S.
  - limo:temporal is an attribute that describe the time period that the model covers. The time period of the model reflects the period that is described in the actual data that have been used for the development of the model.

  `limo:Model` is connected through `limo:data` property to a multi-dimensional data set i.e. a `qb:DataSet`. This dataset contains the actual data that have been used for the development of the model. As a result, the temporal and spatial dimension of the model could be also extracted from this dataset. In predictive analytics we have three different types of data, namely evaluation, validation and training data. So, *limo* includes three different sub-properties of the `limo:data` property, one for each of these three types of data.

  `limo:Model` is also connected through `limo:rawData` property to a `dctype:Dataset`. This dataset includes the raw data that have been used in the process of building the model. For example, this dataset could be a dump of raw tweets or a `dcat:Dataset` which thereafter was analyzed in order to produce the actual data employed be the model.

  Moreover, the `limo:Model` can be connected to a different `limo:Model` through the `limo:baseline` property which explicitly denotes that the predictive power of a model has been evaluated against the power of another model.

  The `limo:Model` can be also published in a scientific article or report. Hence we have included the `limo:publishedIn` property to express this relationship.

Finally, `limo:Model` is connected to a `foaf:Agent` through the `dct:creator` property. This property denotes the person or organization that actually builds the model.

- `limo:Variable` represents the variables that are included in the predictive model. The Variable class includes the following attributes
  - The dct:title denotes the actual name of the variable.
  - The dct:description enables the inclusion of a small text in order to describe what the variable is about.
  - The limo:variableType attributes denotes whether the variable is continuous, categorial or ordinal.
  - The limo:usageType denotes whether the variable is the response of the model or one of the predictors.

  In addition, `limo:Variable` is categorized using the limo:theme property which connects the Variable to a `skos:Concept`
- `limo:Method` describes the statistical or data mining method used for creating the model. We assume that this class uses a set of predefined concepts such as linear regression, logistic regression, markov models, support vector machine, random forests, neural networks etc. As a result, we assume that `limo:Method` is subclass of `skos:Concept`.
- `limo:Power` describes the predictive power of the model. The predictive power has the following attributes:
  - limo:evaluationMethod is used to infer the predictive power of a model. The evaluation methods include out-of-sample evaluation with statistics such as Predicted Residual Sums of Squares, Root Mean Square Error or cross-validation techniques.
  - limo:outcome is the actual value that the evaluation method produces.
- `limo:File` describes a file that can be imported in a particular platform such as R or SAS and execute the model. This could also be a PMML-XML file.

We should note that in this preliminary version of the vocabulary the execution of the model is possible through a PMML XML file. In the next version we aim at providing a more detailed description of the model in order to enable the execution of a model through its *limo* description. Full documentation of the *limo* vocabulary is available online[4].

5 Using *limo*

In this section we present how *limo* vocabulary can be used in order (a) to describe a predictive model and (b) to enable the discovery of predictive models that address some requirements.

Below we present the *limo* description of the predictive model developed by Ginsberg et al. and presented in [4]. This model aims at predicting influenza-like illness (ILI) physician visits from ILI-related queries. The models employs

[4] http://purl.org/limo-ontology/limo

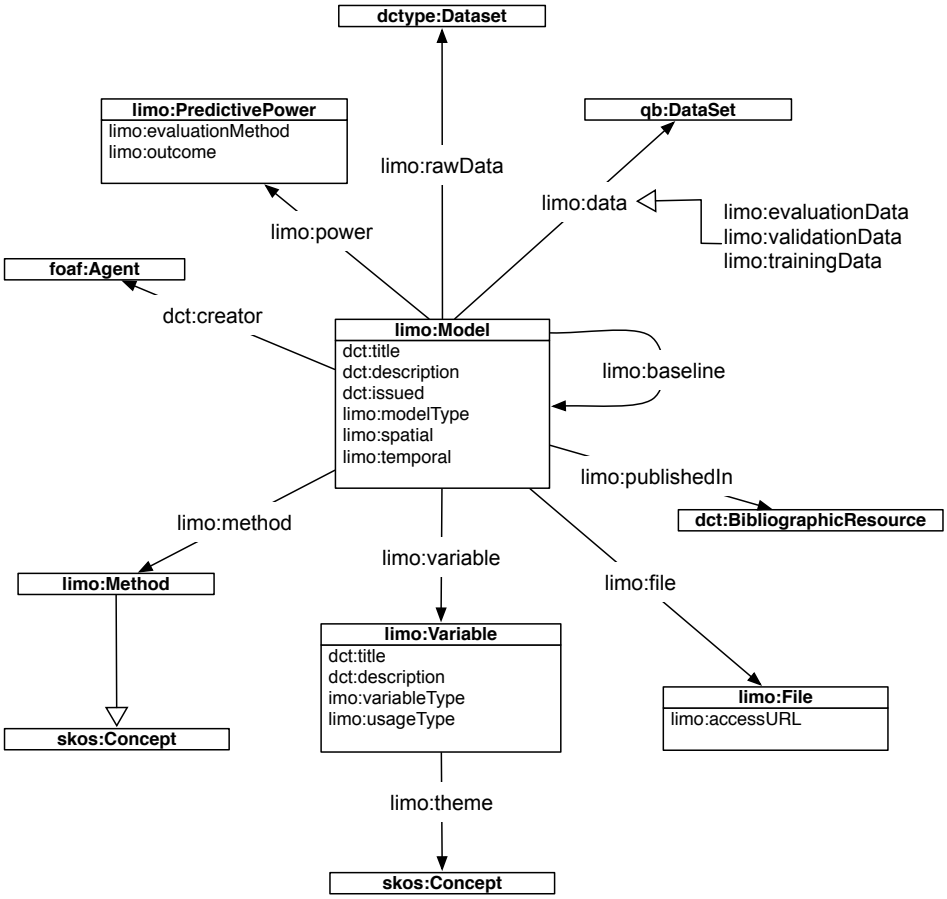

Fig. 1. The Linked Statistical Models vocabulary

a linear regression method as well as data from Google and the US Centers for Disease Control and Prevention. The data is about nine regions of the United States between 2003 and 2008. The model was assessed using cross validation against out-of-sample data partitions and they obtained a mean correlation of 0.97.

Description of the predictive model presented in [4] with limo

```
eg:DDCILImodel a limo:Model;
    dct:title "CDC-ILI model"@en;
    limo:spatial [rdf:type dbpedia:United_States];
    limo:temporal
        [a dc:terms PeriodOfTime;
        limo:startDate "2003-09-28"^^xsd:date;
        limo:endDate "2008-05-11"^^xsd:date;];
    limo:modelType eg:regression;
    limo:variable eg:resp;
    limo:variable eg:pred;
    limo:method eg:linearregression;
    limo:power eg:CDCILIpower;
    limo:file eg:CDCILIfile;
    limo:rawData eg:CDCILIdataset;
    limo:evaluationData eg:CDCILIevaluationdata;
    limo:validationData eg:CDCILIvalidationdata;
    limo:trainingData eg:CDCILItrainingdata;
    dct:creator eg:ginsberg, eg:mohebbi, eg:patel, eg:brammer,
    eg:smolinski, eg:brilliant;
eg:resp a limo:Variable;
    limo:variableType eg:continuous;
    dct:description "Percentage of physician visits in which a
    patient presents with influenza-like symptoms in a region"@en;
    limo:usageType eg:response;
    limo:theme eg:ILIphysvisits.
eg:pred a limo:Variable;
    limo:variableType eg:continuous;
    dct:description "Probability that a random search query
    submitted from a region is ILI-related"@en;
    limo:usageType eg:predictor;
    limo:theme eg:ILIrandquery.
eg:CDCILIpower a limo:Power;
    limo:evaluationMethod eg:crossvalidation;
    limo:outcome 0.97.
eg:CDCILIdataset a dctype:DataSet;
    dct:resource <http://www.cdc.gov/flu/weekly>.
```

In addition, *limo* will enable the performance of queries across distributed description of predictive models. For example below we present a query answering

the question "How many models exist that show relationship between the percentage of influenza-related physician visits and the probability that a random search query submitted from a region is influenza-related?".

A query for identifying models that predict influenza-like illnesses from search query data

```
SELECT (count( ?model ) as ?nmodels)
WHERE {
         {
            ?model limo:variable ?variable1;
                 limo:variable ?variable2.
            ?variable1 limo:usageType eg:response;
                 limo:theme eg:ILIphysvisits;
            ?variable2 limo:usageType eg:predictor;
                 limo:theme eg:ILIrandquery;
         } UNION
         {
            ?model limo:variable ?variable1;
                 limo:variable ?variable2.
            ?variable1 limo:usageType eg:predictor;
                 limo:theme eg: LIphysvisits.
            ?variable2 limo:usageType eg:response;
                 limo:theme eg: ILIrandquery.
         }

       }
```

Moreover, a query based on *limo* could unveil the variables that are predictors of influenza-related physician visits through empirical model(s) constructed by data regarding the U.S. The identification of these variables could enhance the process of building predictive model for influenza illnesses.

A query for identifying predictors of inluenza-like illnesses

```
SELECT ?variable
WHERE {
         ?model limo:variable ?variable1.
              limo:variable ?variable2.
              limo:spatial ?sp1.
         ?variable1 limo:usageType eg:predictor.
         ?variable2 limo:usageType eg:response;
              limo:theme eg:ILIphysvisits.
         ?sp1 rdf:type dbpedia:United_States.
       }
```

6 Conclusions

Predictive analytics refer to the process of building a model that enables the prediction of new observations using data and statistical or data mining methods. Predictive models are very important in businesses, academia and governments as they can predict values such as sales and identify patterns regarding e.g. profitable customers or the behavior of citizens. These models can be indeed reused across platforms and in different cases. Although, standards for transferring models across different platforms have been proposed, at the moment it is difficult to discover an appropriate model for a task at hand at a Web scale.

In this paper we suggested that descriptions of predictive models should be incorporated into the Linked Data Web and we proposed an RDF Scheme vocabulary towards this end. We described the main classes of the vocabulary and we presented an example of how the vocabulary can be used in order to describe a predictive model. We also demonstrated how the vocabulary can be used in order to facilitate the discovery of predictive models.

We believe that the adoption of the vocabulary could create new potentials beyond cross-platforms reuse of models. In particular, the vocabulary will enable (a) easy discovery and reuse of appropriate models at a Web Scale and (b) creation of more accurate models exploiting connections of models to other models, datasets and other resources on the Web.

Future work includes further evaluation of the vocabulary by describing a larger number of predictive models and by incorporating a linked data set into the linked data cloud. This will enable the execution of more complex queries and the evaluation of the vocabulary in real world settings. In addition, the possibility of extending *limo* with execution capabilities will be considered. This includes enriching *limo* with classes and attributes that will allow for importing RDF data into popular open platforms such as R and executing the actual model.

Acknowledgments

The work presented in this paper was partially carried out in the course of the *Linked2Safety*[5] project, which is funded by the European Commission within the 7th Framework Programme under grand agreement No. 288328.

References

1. Shmueli, G.: To Explain or to Predict? Statistical Science, 25(3), 289–310 (2010)
2. Antweiler, W., Frank, M.Z.: Is all that talk just noise? the information content of internet stock message boards. Journal of Finance, 59(3),1259–1294 (2004)
3. Mishne, G., Glance, N.: Predicting Movie Sales from Blogger Sentiment. In American Association for Artificial Intelligence 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (2006)

[5] http://www.linked2safety-project.eu/

4. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012–4 (2009)
5. Ghose, A., Ipeirotis, P.G.: Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498–1512 (2011)
6. Lampos, V., Cristianini, N.: Nowcasting Events from the Social Web with Statistical Learning. ACM Transactions on Intelligent Systems and Technology, 3(4) (2012)
7. Grossman, R., Mazzucco, M.: DataSpace: a data Web for the exploratory analysis and mining of data. Computing in Science and Engineering, 4(4), 44–51 (2002)
8. Grossman, R. L., Hornick, M. F., Meyer, G.: Data mining standards initiatives. Communications of the ACM, 45(8), 59–61 (2002)
9. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 122 (2009)
10. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked Open Government Data Analytics. In: Wimmer, M.A., Janssen, M., Scholl, H.J. (eds.) EGOV 2013. LNCS, vol. 8074, pp. 99–110. IFIP International Federation for Information Processing (2013)
11. Paulheim, H.: Generating Possible Interpretations for Statistics from Linked Open Data. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 560–574. Springer, Heidelberg (2012)
12. Chiricos, T.: Rates of Crime and Unemployment: An Analysis of Aggregate Research Evidence, Social Problem 34, 187–212 (1987)
13. Kalampokis, E., Tambouris, E., Tarabanis, K.: Understanding the Predictive Power of Social Media. Internet Research, 23(5) (2013)
14. Wettschereck, D. and Muller, S. (2001) Exchanging Data Mining Models with the Predictive Modelling Markup Language. International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning
15. Pechter, R.: What's PMML and What's New in PMML 4.0?. ACM SIGKDD Explorations Newsletter, 11(1), 19–25 (2009)
16. Bosch, T., Cyganiak, R., Gregory, A., Wackerow, J.: DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. In: LDOW2013, May 14, 2013, Rio de Janeiro, Brazil (2013)
17. W3C, The RDF Data Cube Vocabulary. W3C Working Draft (2013), http://www.w3.org/TR/vocab-data-cube/
18. Capadisli, S., Auer, S., Ngonga Ngomo, A.-C.: Linked SDMX Data: Path to high fidelity Statistical Linked Data for OECD, BFS, FAO, and ECB. Semantic Web (2013)
19. Capadisli, S.: Statistical Linked Dataspaces. Master's thesis, National University of Ireland (2012), http://csarven.ca/statistical-linked-dataspaces
20. Salas, P. E. R., Martin, M., Mota, F. M. D., Auer, S., Breitman, K., Casanova, M. A.: Publishing Statistical Data on the Web. In: IEEE Sixth International Conference on Semantic Computing (ICSC), pp. 285–292. IEEE Press, New York (2012)
21. Shmueli, G., Koppius, O.R.: Predictive Analytics in Information Systems Research. MIS Quarterly 35(3), 553–572 (2010)