Exploring Student Predictive Model that Relies on Institutional Databases and Open Data Instead of Traditional Questionnaires

Farhana Sarker
University of Southampton
United Kingdom
fs5g09@ecs.soton.ac.uk

Thanassis Tiropanis
University of Southampton
United Kingdom
tt2@ecs.soton.ac.uk

Hugh C Davis
University of Southampton
United Kingdom
hcd@ecs.soton.ac.uk

ABSTRACT
Research in student retention and progression to completion is traditionally survey-based, where researchers collect data through questionnaires and interviewing students. The major issues with survey-based study are the potentially low response rates and cost. Nevertheless, a large number of datasets that could inform the questions that students are explicitly asked in surveys is commonly available in the external open datasets. This paper describes a new student predictive model for student progression that relies on the data available in institutional internal databases and external open data, without the need for surveys. The results of empirical study for undergraduate students in their first year of study shows that this model can perform as well as or even out-perform traditional survey-based ones.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous

Keywords
Student retention, progression to completion, institutional internal/external data sources, survey/questionnaire data, predictive models, Categorical Principal Component Analysis, Logistic regression, Linked data.

1. INTRODUCTION
Student retention and progression to completion is one of the key issues to be addressed by higher education institutions around the world [1]. Increasing student retention is a long-term goal in all academic institutions. The consequences of student dropout are significant for students, academic staff and administrative staff. Since one of the criteria for government funding in tertiary education in the UK is the level of retention rate, both academic staff and administrative staff are under pressure to come up with strategies that could increase retention rates. The first year of study is recognized as a key stage, as during this period a new student is most likely to dropout from Higher Education Institutions (HEI) [2-5]. Yorke noted about one third of students [4] and Thomas et. al, noticed about 77% of students [2] withdraw from their courses during their first year. The Higher Education Statistics Agency (HESA[1]) published that the rate of non-continuation rate in the UK higher education after one year of study varied from 7.9 to 9.5 between 2001/02 and 2009/10. The disproportionate number of students who leave higher education is a major problem and is the focus for retention studies. A number of theoretical models have been developed on student retention from many years. The first and most commonly used model in the student retention literature is Tinto's model [6-8], where the likelihood of a student withdrawing from higher education is seen as being determined by individual attributes, familial attributes, prior qualifications, social integration, academic integration, individual commitment, institutional commitment and external family and societal factors. Research on factors related to student retention has traditionally relied on surveying a student cohort and following them for a specified period of time to determine whether they ultimately dropped out or whether they continued their education. Using this design, researchers have worked to validate theoretical models of student retention including Tinto's widely employed model of student integration [9][10].

Although it has been successfully used to-date, survey based research may be too burdensome to sustain, as individual institutions may not have the capacity to construct and administer a similar instrument to study their unique retention situation. Even if an institution is capable of fielding a one-time retention survey, repeated administrations over time may be too oppressive. Moreover, another major limitation of survey-based test is low participation rate, which may often compromise the precision of the output. Thus it is key for enrollment professionals and researchers to have sufficient means of evaluating the trends in the circumstances of student retention at their institution in order to develop or adjust support programs accordingly. Data-informed decision-making helps higher education institutions know whether they are achieving their missions [11]. Institutions routinely collect a broad array of information on their students' backgrounds and academic progress. Also in the UK, the Higher Education Statistics Agency (HESA[2]), the Higher Education Funding Council for England (HEFCE[3]), the Office for National Statistics (ONS[4]), and Unistates[5] routinely publish some open datasets; which could be used to develop student predictive model in the place of questionnaire-based predictive models that have been used to-date.

The combination of datasets from internal institutional databases and external data sources presents certain challenges. Although a large amount of data is available, data is frequently maintained in different locations, in different formats and often with different identifiers. Data aggregation also presents organizational challenges related to the ownership and use of the data [12]. Linked data technologies are

[1] http://www.hesa.ac.uk/index.php?option=com_content&task=view&id=2064&Itemid=141

[2] http://www.hesa.ac.uk/
[3] http://www.hefce.ac.uk/
[4] http://www.ons.gov.uk/ons/index.html
[5] http://unistats.direct.gov.uk/

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

considered to be well suited for data integration. Linked data is interlinked RDF data that enables users to retrieve quality information from different data sources[6]. In this study, we examine the sufficiency of existing linked data standards and datasets in supporting student retention, progression and completion.

In section 2, we define the methodology of this study, in section 3, we explain the experimentation and results of the study; in section 4, we discuss the findings of the study and section 5 proposes a linked data infrastructure to develop new predictive model, while the last section presents the conclusion and future work.

2. METHODOLOGY

The purpose of this study is to explore a new predictive model that relies on data commonly available in institutional internal and external (or open) data sources instead of questionnaires used in the traditional student predictive models. We developed the predictive models with the variables used by Pascarella et al. in their study of first year student retention [9] based on Tinto's theory of integration where they used a set of questionnaires called Institutional Integration Scale (IIS) developed by Pascarella and Terenzini [10] to measure various dimensions identified by Tinto as corresponding to the likelihood of persistence, which is being traditionally used in retention literature for many years.

2.1 Data and Data sources

In this study we considered 3 types of variables a) variables from institutional internal data sources (IDS), b) variables from traditional questionnaires/ institutional integration scale (IIS) and c) variables from institutional external data sources/ open data sources (EDS). Table 1 provides the list of all the variables used in this study with their sources. In this study we used National Student Survey (NSS) result published in Unistates website (as institutional external data source/ open data source) to replace the IIS variables which measures student's academic and intellectual development, faculty student interaction and faculty concern for student development. Every year the NSS conducted to measure students' satisfaction in different dimensions of their study subjects in their institutions such as satisfaction in teaching and learning, assessment and feedback, academic support, organization and management, learning resources and personal development. As IIS were also used to measure different dimensions of student satisfaction and integration, we include total 16 questionnaire items (see in table 2) from NSS which are related to student faculty interaction, faculty concern for student development, students' development and about their course among the 22 common questionnaires for all subjects as a replacement of the IIS questionnaire.

Unistats does not publish individual student data. NSS measures students' satisfaction on their program of study in a 5 points scale (Definitely Disagree, Moderately Disagree, Neither Agree nor Disagree, Moderately Agree, Definitely Agree). The website publishes the percentages of respondents in each scale for an individual course. We considered the actual value (for % Agree) for those 16 questions for 2010-2011 academic year's published result for the university of Southampton to include in our study to develop the predictive model.

2.2 Design of empirical study

As we did not have direct database permission from the institution, all the data (questionnaires and institutional internal database items) for this study were collected through online questionnaires. In the first stage of this study all students who enrolled in the academic year 2010/2011 were asked to complete an online questionnaire, which was designed to collect survey based data as well as the data which are available in the institutional internal databases such as in the admission database, students' academic performance dataset. We have had ethics approval from the university to conduct the questionnaire session. Total number of participants in this study was 149. The respondents' subsequent academic outcome status was determined based on 2 criteria: a) the students who failed to progress according to their academic year or semester that means if a student enrolled in October 2010 then they were expected to be in their second year second semester at the time of the questionnaire conducted but if they are behind their expected year and semester of their study then they were identified as "at Risk" students and b) the students who got less than 50% marks in their first year or in their first year's first semester exam are also identified as "at Risk" students.

Apart from these data, we used NSS data to replace some traditional questionnaire data to develop student predictive model. The traditional questionnaire, the Institutional Integration Scale (IIS), which is traditionally used in retention study for many years [9][13][14] includes questionnaires about student's academic and intellectual development, academic support and satisfaction on teaching and learning. NSS also have some similar measurements of questionnaire. We explored whether we could replace those traditional questionnaires about student's academic and intellectual development, academic support and satisfaction on teaching and learning with the NSS questionnaires. Table 2 presents the questionnaire items for IIS and NSS.

Table 1. List of variables and sources

Variable	Variable source
Gender, Ethnicity, A Level tariff points, Accommodation Type, First year's first semester marks, Source of tuition fee, Study field	IDS
Parents' have HE qualification	IDS
Student's working status in their first year of study.	Questionnaire item
Peer Group interaction (7 items/variables)	Questionnaire (IIS)
Student-Faculty interaction (5 items/variables)	Questionnaire (IIS)
Faculty Concern For Student Development and Teaching (5 items/variables)	Questionnaire (IIS)
Academic and Intellectual Development (7 items/variables)	Questionnaire (IIS)
Goal Commitment I	Questionnaire (IIS)
Institutional Commitment I	Questionnaire (IIS)
Goal Commitment II	Questionnaire (IIS)
Institutional Commitment II (2 items/variables)	Questionnaire (IIS)
Intention	Questionnaire (IIS)
The teaching on my course (4 items/variables)	EDS (Unistates)
Assessment and feedback (5 items/variables)	EDS (Unistates)
Academic support (3 items/variables)	EDS (Unistates)
Personal development (3 items/variables)	EDS (Unistates)
Overall satisfaction with the quality of the course	EDS (Unistates)

*IDS: Institutional Internal Data sources, EDS: Institutional External Data sources.

[6]http://www.w3.org/DesignIssues/LinkedData.html

In this study we developed three predictive models (model 1, model 2 and model 3), where the first model (model 1) includes all the independent variables considered by Pascarella et. al. to develop the predictive model to find out probable withdrawal students in their first year of study, which is a survey-based model [14]. Second model (model 2) includes only the variables from model 1, which are commonly available in the institutional internal databases to see how the model performs with only the available data in the institutions. Finally model 3 includes all the variables from model 2 and includes new variables from external data source as the replacement of the traditional questionnaire items/variables from model 1.

2.3 Data Analysis

The objective of data analysis was to establish:

i. whether it is possible to have a valid predictive model by omitting questions that are not available in institutional datasets.
ii. whether it is possible to have a valid and precise predictive model by replacing questions that would be asked in surveys by related data found in the open data cloud.

For the above objectives, an analysis of the contribution of a number of variables to the predictive model was necessary. Categorical Principal Component analysis (CATPCA) and logistic regression were used in this study. The goal of PCA is to reduce an original set of variables into a smaller set of uncorrelated components that represent most of the information found in the original variables. CATPCA is an optimal scaling method belonging to the nonlinear multivariate analysis techniques. It is the nonlinear equivalent of PCA: it aims at the same goals of traditional PCA, but it is suited for variables of mixed measurement level that may not be linearly related to each other [15]. In this study CATPCA was applied to avoid multicollenearity problem. Moreover it was applied to extract factors (F) as well as to discover the factors structure, which are significantly correlated with the student outcome status. We followed Kaiser's rule to retain the factors for the further analysis, if the analysis has more than 30 input variables, factors with eigenvalues greater than 1 are

Table 2. List of Traditional questionnaire (IIS) and NSS questionnaire

Traditional Questionnaire (IIS)	NSS questionnaire
Since coming to this university, I have made close personal relationship with other students.	Staff are good at explaining things.
The student friendships I have developed at the university have been personally satisfying.	Staff have made the subject interesting.
My interpersonal relationships with other students have had a positive influence on my personal growth, attitudes, and values.	Staff are enthusiastic about what they are teaching.
My interpersonal relationships with other students have had a positive influence on my intellectual growth and interest in ideas.	The course is intellectually stimulating.
It has been difficult for me to meet and make friends with other students.	The criteria used in marking have been clear in advance.
Few of the students I know would be willing to listen to me and help me if I had a personal problem.	Assessment arrangements and marking have been fair.
Most students at this university have values and attitudes different from my own.	Feedback on my work has been prompt.
My non-classroom interactions with faculty have had a positive influence on my personal growth, values and attitudes.	I have received detailed comments on my work.
My non-classroom interactions with faculty have had a positive influence on my intellectual growth and interest in ideas.	Feedback on my work has helped me clarify things I did not understand.
My non-classroom interactions with faculty have had a positive influence on my career goals and aspirations.	I have received sufficient advice and support with my studies.
Since coming to this university, I have developed a close, personal relationship with at least one faculty member.	I have been able to contact staff when I needed to.
I am satisfied with the opportunities to meet and interact informally with faculty members.	Good advice was available when I needed to make study choices.
Few of the faculty members I have had contact with are generally interested in students.	The course has helped me present myself with confidence.
Few of the faculty members I have had contact with are generally outstanding and superior teachers.	My communication skills have improved.
Few of the faculty members I have had contact with are willing to spend time outside of class to discuss issues of interest and importance to students.	As a result of the course, I feel confident in tackling unfamiliar problems.
Most of the faculty members I have had contact with are interested in helping students grow in more than just academic areas.	Overall, I am satisfied with the quality of the course.
Most faculty members I have had contact with are genuinely interested in teaching.	
I am satisfied with the extent of my intellectual development since enrolling in this university.	
My academic experience has had a positive influence on my intellectual growth and interest in ideas.	
I am satisfied with my academic experience at this university.	
Few of my courses this year have been intellectually stimulating.	
My interest in ideas and intellectual matters has increased since coming to this university.	
I am more likely to attend a cultural event now than I was before coming to this university.	
Your choice of this institution was?	
My academic performance has met my expectation.	
It is important for me to graduate from this university.	
I am confident that I have made the right decision in choosing to attend this university.	
Getting good result is not important to me.	
What is the highest expected academic degree?	
It is likely that I will register at this university next year.	

Table 3: Component structure of the above significant components/factors

F5	F8	F11	F12
Most of the faculty members I have had contact with are interested in helping students grow in more than just academic areas.	Intention	My academic performance has met my expectation.	First Year 1st Semester mark
I am satisfied with the opportunities to meet and interact informally with faculty members.	It is important for me to graduate from this university.	First Year 1st Semester mark	A level points
Most faculty members I have had contact with are genuinely interested in teaching.			
Few of the faculty members I have had contact with are generally interested in students.			
My interest in ideas and intellectual matters has increased since coming to this university.			

* The highest loading variables put first in the table and the lowest loading variables are in the last of the table.

normally retained while it is recommended to retain factors (F) with eigenvalues greater than 0.7 with input variables less than 30 input variables [16]. Also the variable factor loadings which were smaller than 0.4 were ignored, that is if a variable's loading on a factor was found to be smaller than 0.4, it did not come towards the factor.

To further optimize factor loadings, the varimax rotation algorithm with Kaiser normalization was applied to the resulting factor matrix. The varimax rotation is the most popular of all rotation algorithms and aims to produce a few high valued loadings and many low-valued loadings so that the number of variables per factor is minimal with each variable having a maximum loading with regards to that factor [17]. To enable further analysis with the data set using factors rather than variables, factor scores were saved in the data set using the Anderson-Rubin method as recommended by [16]. This method ensures that there are no correlations between factor scores. In the next step a correlation test was applied on retained factor scores and the students' outcome status (at Risk, Not at Risk) looking for relationship between factors and students' outcome status. Finally, logistic regression was applied to develop the predictive models with the significant factors only. Logistic regression analysis is used when the dependent variables are categorical, rather than continuous. We used binary logistic regression, as our dependent variable has two categories (at Risk and Not at Risk). Repeated hold-out method was applied to validate the predictive models. The cases (dataset) were randomly divided into two sets, where training set containing 70% of the cases and the test set containing the rest 30% of the cases. With this method, the predictive model can be made reliable by repeating the training and testing process through randomly partitioning the dataset, and average the accuracy rate of all repetition to produce the overall accuracy rate [18]. IBM SPSS Statistics (version 20) was used for the data analysis.

Table 4. Component structure of the above significant components/factors

F3	F5
A level points	Parents Higher Education
First Year 1st Semester mark	

▪ The highest loading variables put first in the table and the lowest loading variables are in the last of the table.

3. EXPERIMENTATION AND RESULTS

The purpose of this study is to explore a new predictive model that relies on data commonly available in institutional internal and external data sources instead of questionnaires. In this study we developed three predictive models (model 1, model 2 and model 3).

For model 1, a total of 39 variables were used in CATPCA. Following the approach stated in the data analysis section a total of 13 factors were retained for model 1. A correlation test was applied between these 13 variables and the students' outcome status and found only 4 factors (5, 8, 11 and 12) are significantly correlated with the students' outcome status. The factors, which are significantly correlated, are summarized with their associated input variables in table 3. Factor 5 composed with five input variables and factor 8, 11 and 12 composed with 2 input variables each (see in table 3). Highest loading variables put first in the table. Student predictive model was developed with these four significant factors using binary logistic regression and the total accuracy of the model achieved 88.86%. Utilising the same procedure we develop model 2. We found only 8 variables are available in the institutional internal database among all of the 39 variables in model 1. A total of 5 factors were retained for model 2 and two of them were significantly correlated with the students' outcome status. Table 4 presents the factors/components structure of these two significant factors. The total accuracy for model 2 was achieved 84.94%. Model 3 includes only the database item from model 1 as well as data from external data source to replace questionnaire data from model 1. Total 24 input variables were considered to develop model 3 (8 database items and 16 NSS questionnaire items). Total 8 factors were retained and among these 8 factors 3 factors were found significantly correlated with the student output status. Table 5 presents the components structure of these 3 significant factors. The total accuracy of model 3 was achieved 89.20%.

The summary of these three models presented in Table 6 with total number of input variables, total number of factors retained after applying CATPCA based on Kaiser's rule, number of significant factors which significantly correlate with the student outcome status, sensitivity of the model, specificity of the model and the overall model accuracy. The result indicates that institutional internal and external database origin predictive model performs best in predicting "at Risk" students among these three models, with the total accuracy 89.20% while the second best performing predictive model was survey-based predictive model with the total accuracy 88.86%. The predictive accuracy of these two models is quite comparable while the predictive accuracy of the model with only institutional internal database items was 84.94%.

4. DISCUSSION

Research on retention typically relies on surveying of student perceptions in relation to the factors believed to theoretically influence persistence decisions. However, this resources-intensive methodology is not always feasible for retention research at individual institutions. Caison in 2007 compares traditional survey based retention research methodology with an alternative approach that relies on data commonly available in institutional student database [13]. His result confirms that only the variables available from the institutional databases are not sufficient to build a good

Table 5. Component structure of the above significant components/factors

F2	F6	F8
Staffs have made the subject interesting.	A level points	Parents Higher Education
Staffs are enthusiastic about what they are teaching.	First Year 1st Semester mark	
Field of Study		
Gender		
I have received detailed comments on my work.		

* The highest loading variables put first in the table and the lowest loading variables are in the last of the table.

Table 6. Model summary

	Model 1	Model 2	Model 3
Number of Input variables	39	8	24
Number of Factors retained after CATPCA	13	5	8
Number of significant factors	4	2	3
Sensitivity of the model (%)	84.33	61.33	89.33
Specificity of the model (%)	89.57	87.46	89.63
Total model accuracy (%)	88.86	84.94	89.20

performing predictive model, it requires more additional data to perform better. Also the same stands for the IIS based model. In our study we also found that the model based on only institutional internal data sources performs the lowest with model accuracy 84.94% while, replacing the questionnaire data from external data source provided 89.20% of model accuracy and with traditional questionnaire the predictive accuracy of the model was 88.86%.

The result of this study strongly supports the use of institutional internal and external data sources to conduct institution specific retention and progression to completion research to identify at risk students to arrange intervention programs for them. These findings offer important validation for institutional researchers looking to utilize the considerable amount of data, which they routinely collect, and which are available in institutional external data sources. The findings of this study do not weaken the results of the model developed using traditional questionnaires; rather, this study offers researchers new approach to utilize in retention studies. This expanded toolkit for retention research offers the possibility for more research in diverse settings which given resource constraints, would not have otherwise been possible. This study lays the groundwork for this effort. Moreover, this study supports the prospects of linked data technologies in institutional research to support student retention and progression to completion as the potential data are spread out in different institutional internal and external data sources.

5. TOWARDS AN LINKED DATA INFRASTRUCTURE IN BUILDING NEW PREDICTIVE MODEL

As data are in different data sources and the biggest challenges and opportunities lie in connecting these disparate datasets to create a new single set of data for analysis, and to develop the new predictive model. Combining data into a common location is inhibited by different technology standards, lack of unique identifiers, and organizational challenges to the ownership and use of the data [12]. In recent years, there has been an increasing interest in linked data in higher education [19-21]. Linked data is well suited for data integration while data is in different formats in different data sources. Therefore, we motivate to develop an linked data infrastructure to support student retention, progression to completion by integrating related data from disparate data sources (internal or external) and analyzing the new set of linked data to provide the new student predictive model. Figure 1 depicts our vision in developing student predictive model through integrating data from disparate data sources.

We considered the following requirements are important for the infrastructure.

- The ability to convert raw data to RDF.
- The ability to perform SPARQL query over different RDF sources.
- The ability to join multiple SPARQL query results into a single dataset.
- The ability to develop the predictive model or to generate an excel file of the final data set to use in any other software to further analysis of the data.

The infrastructure consists of four components to fulfill the above requirements.

RDF generator: As most of the datasets of interest are not yet in linked data format, we developed a number of scripts, which is able to automatically convert the datasets (.csv) into RDF triples. Besides, there are many existing tools to convert data into RDF, as needed, such as Grinder[7], google-refine[8].

SPARQL engine: We can connect to different SPARQL endpoints. It only supports sending SPARQL queries via HTTP requests (i.e. sending queries to SPARQL endpoints) and accepts query results via HTTP as well.

Aggregator: It supports to join multiple SPARQL query results into a single dataset based on a common identifier. For example, there are two query results R1 and R2. R1 and R2 have common identifier student ID. Then, based on this common id the "Aggregator" joins these two datasets into a single dataset. Hence R3=R1UR2.

Model generator or Excel file generator: After combining multiple RDF sets into a single RDF set, the next step is to develop student predictive model based on the single dataset aggregated from different data sources. Also this have the ability to generate an excel file of the final RDF set to use in any custom written or any available software, such as R statistics, SPSS, Rapid Miner. So that, anyone can develop the predictive model based on this dataset.

Though linked data is efficient in integrating data from different data sources, we still are not getting the full benefit from it. We noticed that the major issue in integrating data from multiple data sources is the lack of standardization in the data as most of the interested data are in 2 star (.xls) or 3 star (.csv) format[9]. This would much improve when data providers would decide to publish their data in linked data format using standardized vocabularies and ontologies. Also data integration will be easier while data provider will make their data available via a SPARQL endpoint. Moreover, RDF data integration is done by loading all data into a single repository and querying the merged data locally. This is not feasible for legal or technical reasons. Possible technical reason is that local copies are not up-to-date. In context of statistical methods SPARQL is still at the early stage. Different frameworks and tools, which are using SPARQL, have already implemented aggregate functions like MAX, MIN, AVG or SUM. Some of these extensions found recognition and are planned to be included in the next revision of the language, SPARQL 1.1[10].

[7] https://github.com/cgutteridge/Grinder
[8] http://code.google.com/p/google-refine/
[9] http://www.w3.org/DesignIssues/LinkedData.html
[10] http://www.w3.org/TR/sparql11-query/

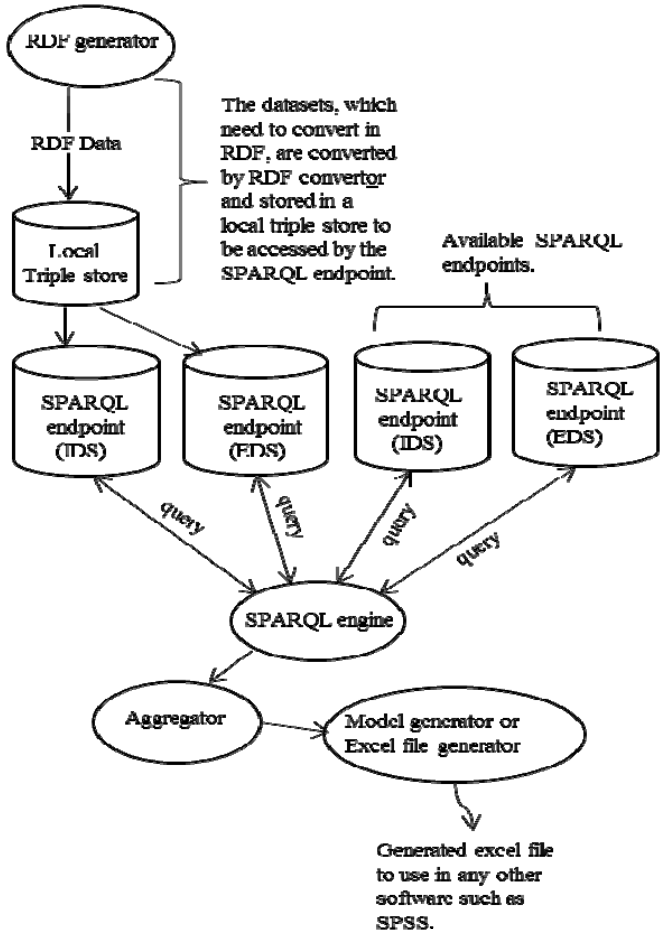

Figure 1. Linked data infrastructure to develop student predictive model.

More complex statistical methods are still missing in the current plans. An overview of proposed and implemented extensions can be found at the corresponding page in the W3C-Wiki[11].

6. CONCLUSIONS AND FUTURE WORK

In this study we propose ways to build student predictive models through integrating data from institutional internal databases and external open data sources without having to rely on questionnaires that have been essential in existing predictive models. The result of this study shows that model based on institutional internal databases and external open data sources performs best among the three predictive models with the highest model accuracy 89.20%. The model based on survey/questionnaire performs second best with the model accuracy 88.96% and the model accuracy based on only institutional internal database was 84.94%. This study underlines the importance of linked open data sources in developing predictive models to support student retention and progression to completion instead of questionnaires. We propose a linked data infrastructure to develop such predictive models to support student retention and progression to completion. In this study we used SPSS to create our predictive model, in future we will develop our own predictive model using the infrastructure. Also in a next study, we will explore model to predict students' marks based on internal databases and external, open data sources.

[11] http://esw.w3.org/SPARQL/Extensions

7. REFERENCES

[1] Sarker, F., Davis, H. & Tiropanis, T. 2010. A Review of Higher Education Challenges and Data Infrastructure Responses. International Conference for Education Research and Innovation (ICERI2010), Madrid, Spain.

[2] Thomas, M., et al. 1996. Student Withdrawal from Higher Education, *Educational Management and Administration,* 24(2), 207-221.

[3] Tinto, V. 1998. Colleges as communities: taking research on student persistence seriously, *Review of Higher Education,* 21(2), 167-177.

[4] Yorke, M. 1999. Leaving Early: Undergraduate Non-completion in Higher Education, London: Falmer Press.

[5] Harvey, L., et al. 2006. The first year experience: a literature review for the Higher Education Academy, York: Higher Education Academy.

[6] Tinto, V. 1975. Dropout from Higher Education: A TheoreticalSynthesis of Recent Research, *Review of Educational Research,* 45(1), pp. 89-125.

[7] Tinto, V. 1987. Leaving Early: Rethinking the Causes and Cures of Student Attrition, Chicago,, IL: The University of Chicago Press.

[8] Tinto, V. 1993. Leaving College: rethinking the Causes and Cures of Student Attrition, Chicago, 2nd edition, IL: The University of Chicago Press.

[9] Pascarella, E. T., Duby, Paul B., Iverson, Barbara K., 1983. A Test and Reconceptualization of a Theoretical Model of College Withdrawal in a Commuter Institution Setting, *Sociology of Education,* 56(2), 88-100.

[10] Pascarella, E. T. & Terenzini, P. T. 1980. Predicting freshman persistence and voluntary dropout decisions from a theoretical model, *Journal of Higher Education,* 51(1), 60–75.

[11] Celeste Schwartz, Mary Lou Barron & Mauger, A. J. 2010. Using Technology to Impact Student Retention at Montgomery County Community College, EDUCAUSE Quarterly (EQ) Magazine, 33(4).

[12] Arnold, K. E. 2010. Signals: Applying Academic Analytics, EDUCAUSE Quarterly (EQ) Magazine, 33(1).

[13] Caison, A. L. 2007. Analysis of Institutionally Specific Retention Research: A Comparison Between Survey and Institutional Database Methods, *Research in Higher Education,* 48(4).

[14] Herzog, S. 2005. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen, *Research in Higher Education,* 46(8), 883-928.

[15] Linting, M., Meulman, J. J., Groenen, P. J. F. & Van der Kooij, A. J. 2007. Nonlinear Principal Components Analysis: Introduction and Application., *Psychological Methods,* 12(3), 336–358.

[16] Field, A. *Discovering Statistics Using SPSS,* 3rd edition, Sage, 2009.

[17] Abdi, H. 2003. Factor Rotations in Factor Analyses. In M. Lewis-Beck, A. Bryman, and T. , *Futing, editors, Encyclopedia of Social Sciences Research Methods. Sage, Thousand Oaks, CA*.

[18] Duda, R. O., Hart, P. E., & Stork, D. G. *Pattern Classification* (2nd edition). Wiley-Interscience, 2001.

[19] Tiropanis, T., Davis, H., Millard, D. & Weal, M. 2009. Semantic Technologies for Learning and Teaching in the Web 2.0 Era, *Intelligent Systems, IEEE,* 24(6), 49-53.

[20] Tiropanis, T., Davis, H., Millard, D. & Weal, M. 2009. Semantic Technologies for Learning and Teaching in the Web 2.0 era - A survey *WebSci'09: Society On-Line.* Athens, Greece.

[21] Tiropanis, T., Davis, H., Millard, D., Weal, M., White, S. & Wills, G. 2009. Semantic Technologies in Learning and Teaching (SemTech) Report, JISC Technical Report.