Design and Analysis of DWH and BI in Education Domain

Manjunath T N[1], Ravindra S Hegadi[2],Umesh I M[3] ,Ravikumar G K[4]

[1]Bharathiar University, Coimbatore, Tamilnadu, INDIA

[2]Karnatak University, Dharwad, Karnataka, INDIA

[3]R V College of Engineering,Bangalore,Karnataka,INDIA

[4]Wipro Technologies, Bangalore, Karnataka, INDIA

Abstract

Data warehouse acting as a decision support systems, Data warehouses standardize the data across the organization so that there will be one view of information. Data warehouses can provide the information required by the decision makers. Developing a data warehouse for educational institute is the less focused area as educational institutes are non-profit and service oriented organizations. In present day scenario where education has been privatized and cut throat competition is prevailing, institutes needs to be more organized and need to take better decisions. Educational institute's enrollments are increasing as a result of increase in the number of branches and intake. Today, any reputed Institute's enrollments count in to thousands. The management challenges include meeting diverse student needs, increased complexity in academic processes. The complexity of these challenges requires continual improvements in operational strategies based on accurate, timely and consistent information. The cost of building a data warehouse is expensive for any educational institution as it requires data warehouse tools for building data warehouse and extracting data using data mining tools from data warehouse. The present study provides an option to build data warehouse and extract useful information using data warehousing and data mining open source tools.

Keywords: DWH, BI, ETL, Analysis, Modeling.

1. Introduction

Now a days, an educational institutes have to generate funds for their research and other operational activities as the government funding has been limited to aided institutes. Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system could address a wide range of problems by distilling data from any combination of education records maintenance system. The data mining from data warehouse can be a ready and effective system for the decision makers. A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management decisions [1].Data warehouse obtains the data from a number of operational data base systems which can be based on RDBMS or ERP package, etc. These are called data sources. The data from these sources are converted into a form suitable for data warehouse. This process is called Extraction, Transformation and Loading (ETL). In addition to the target database, there will be another data base to store the metadata, called the metadata repository. This data base contains data about data-description of source data, target data and how the source data has been modified into target data. The client software will be used to generate reports.

2. Related Work

There are some efforts in the area of data warehouse for building data warehouse for education domain. The paper by Carlo DELL'AQUILA summarizes the experience in designing and modeling an academic data warehouse. Existing facilities and databases affect the chosen data warehouse that brings them together to support decisional activities leading the whole university environment, including administrators, faculties and students. The choice to develop a dedicated system is mainly forced by the peculiar information type that defines the basic information in data warehouse widely different from institution to institution [11].In the article titled 'What academia can gain from building a data warehouse' by David Wierschem, Jeremy McMillen and Randy McBroom, they have identified the opportunities associated with developing a data warehouse in an academic environment. They begin by explaining what a data warehouse is and what its informational contents may include, relative to the academic environment. Next they address the current environment drivers that provide the opportunities for taking advantage of a data warehouse and some of the obstacles inhibiting the development of an academic data warehouse. Finally, the article provides strategies to justify developing a data warehouse for an academic institution [12].

3. Proposed Data warehouse Environment

Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system could address a wide range of problems by distilling data from any combination of education records maintenance system. The data mining from data warehouse can be a ready and effective system for the decision makers. The reputed engineering college R V College of Engineering, Bangalore, Karnataka, India, has been taken for this study.Figure-1 shows the DWH architecture of RV College where source systems are smart campus, asset management server and csv files, the information is spread across diverse platforms, data from different sources has to be taken from different sources and then consolidated to produce required report. ETL activities are performed to extract the data from heterogeneous sources and load into staging and then load the data into dimension and fact tables as per the schedules. We proceed to extract the BI report from data warehouse on demand based on requirement from the management. In an educational institute, main information required will be regarding key components of the education institute, namely students, employees and infrastructure. The purpose of this paper was to investigate current system of information delivery and proposing a better system for timely, accurate and consistent information delivery to the decision makers of the educational institute. The paper has been prepared in order to extend the usage of current available technology in decision making processes of educational institute.

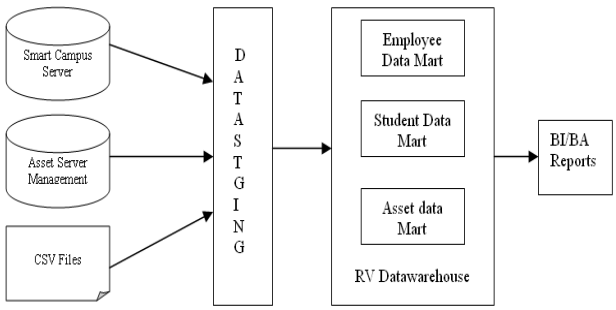

Figure 1-RV_Data warehouse architecture

3.1 Data Modeling Design

It is decided to classify the information in to three categories, Employee information, Student Information and Asset information. Hence it is decided to have three data marts namely, Student Mart, Employee Mart and Asset Mart. It is decided to have Star schema

3.1.1 Star schema

A start schema is a modeling paradigm in which the data warehouse contains a large, single central Fact Table and set of smaller Dimension tables, one for

each dimension. The fact table contains the detailed summary data. Its primary key has one key per dimension. Each dimension is a single, highly de-normalized table. Every tuple in the fact table consists of fact or subject of interest, and dimension that provide the fact. The dimension table consists of columns that correspond to the attributes of the dimension. Star Schema designed for data marts

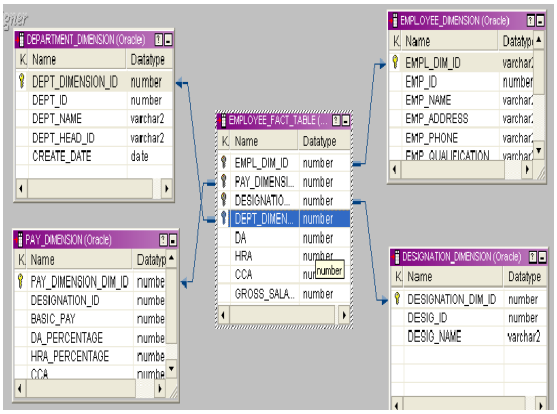

Figure 2: Employee Mart

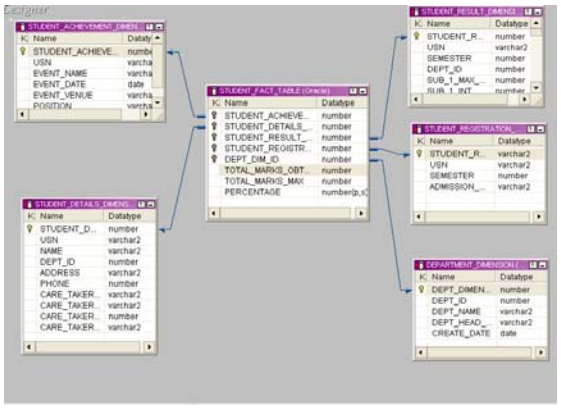

Figure 3: Student Mart

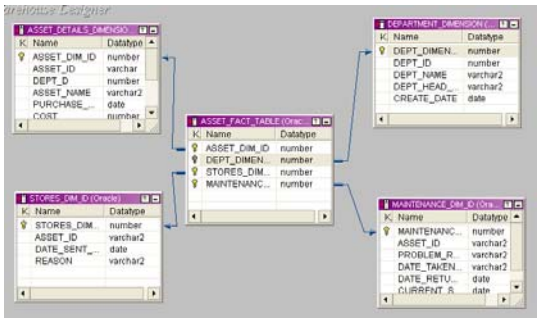

Figure 4: Asset Mart

3.1.2 ETL Activity

There are several transformations available to transform the data from source files to target tables. Expression transformation is used to transform the format of date or calculate anything if needed before loading the data into target. Filter transformation filters the records as specified in the condition part. Update strategy is used to update the records. Sequence generator transformation generates continuous values, which can be used to generate surrogate keys. Router is an advanced filter which is used to direct the output to two or more different tables based on the condition specified. Joiner transformation is used to join tables from two or more heterogeneous sources. Lookup transformation is used to check the content of one or more attributes of target table before loading the data from source.

Figure 5 shows the mapping done to load the data from department flat file to department dimension table. The details of the department such as dept_id and department_name are stored in flat file. The mapping is done to load the data from flat file to dimensional table. The mapping includes transformations like lookup, expression, filter and sequence generator. Lookup transformation is included to check whether the record with same content is already in target table or not. Checking is done on dept_id key which is unique identifier. The records are filtered using filter transformation and finally loaded to dimensional table.

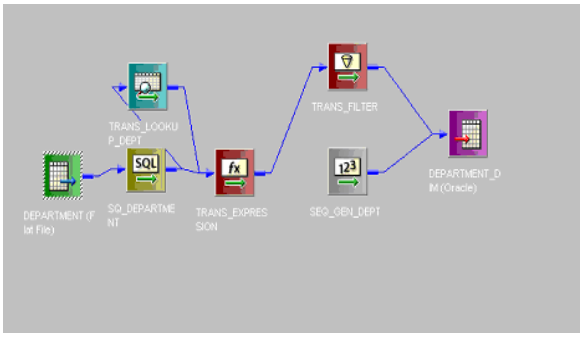

Figure 5: Mapping to load department csv to department Dim table

Figure 6 shows the mapping done to move the data from student_achievement excel file to student_achievement_dimension dimension table. In this mapping, there is no need for checking the record for duplication. The transformations used include expression and sequence generator. Expression transformation is included to convert the date format. Sequence generator is used to generate surrogate keys.

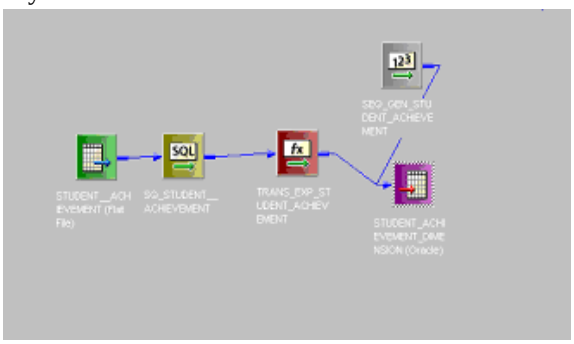

Figure 6: Mapping to load student_achievement csv file to student_achievement_dimension dim table.

Figure 7 shows the mapping done to move the data from asset_details excel file to asset_details_dim dimension table. The asset_details flat file contains the data such as asset_name, asset_id, dept_id, purchase_date and so on. Expression transformation is used to clean the data coming from source file. Date format conversion is done. Sequence generator is used to generate surrogate keys.

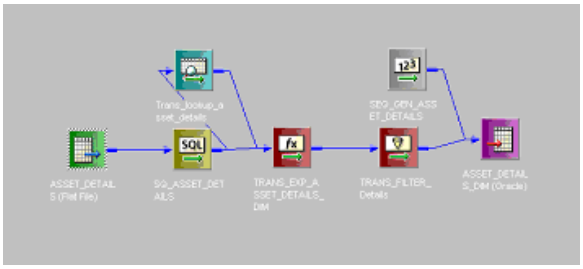

Figure 7: Mapping to load Asset_details csv file to asset_details_dim dimension table.

Data warehouse enables the decision makers with benefits listed below.

i. Phenomenal improvements in turnaround time for data access and reporting
ii. Standardizing data across the organization so that there will be one view of information.
iii. Merging data from various source systems to create a more comprehensive information source.
iv. Reduction in costs to create and distribute information and reports.
v. Encouraging and improving fact-based decision making.

4. Results

Some of the results observed after querying the data marts are documented below. The results are cross checked with the requirements specified by the different types of users. The requirements with regard to asset information were to extract the information regarding the number of assets of each type in the Institute. The three different data marts are queried using SQL query. The results returned by the queries are found accurate and meeting users demands. The sample screen shots of queries and the results are shown. Figure 8 shows the results obtained after querying asset mart. The information required was the number of equipments of each type in the institute and the amount spent on different type of assets in particular departments.

```
SQL> SELECT ASSET_TYPE,COUNT(*) FROM ASSET_DETAILS_DIM GROUP BY ASSET_TYPE;

ASSET_TYPE       COUNT(*)
---------------- ---------
 ELECTRICA            16
 EQUIPMENT            31
 FURNITURE            25
 MP MODULE             6

SQL> SELECT DEPT_ID,SUM(COST) FROM ASSET_DETAILS_DIM GROUP BY DEPT_ID;

   DEPT_ID  SUM(COST)
---------- ----------
       100     503000
       200      50100
       300      24600
       400      75400
       401     221000
```

Figure 8: Results regarding number of equipments and total amount spent

Figure 9 shows the results obtained after querying asset mart. The information required was the number of equipments whose warranty has been expired. The information can be used to count the number of assets in need of Annual Maintenance Contract.


```
SQL> SELECT ASSET_NAME,WARRANTY_EXPIRY_DATE,DEPT_ID FROM ASSET_DETAILS_DIM WHERE WARRANTY
E('29-aug-07';

ASSET_NAME                     WARRANTY_   DEPT_ID
------------------------------ --------- ---------
  PROJECTOR                    03-FEB-02       100
  OHP                          06-MAR-02       100
  COMPUTER                     06-APR-02       100
  SEATER                       03-FEB-01       100
  TABLE                        06-MAR-04       100
  COMPUTER CHAIR               06-APR-05       100
  HOD TABLE                    04-FEB-06       100
  TEA POY                      03-FEB-02       200
  BOARD                        06-MAR-03       200
```

Figure 9: Results regarding the equipments

Figure 10 shows the results after querying student mart. The needed information was the toppers in odd semester considering all the departments.

```
SQL> SELECT STUDENT_DETAILS_DIMENSION.STUDENT_NAME,STUDENT_RESULT_DIMENSION.SEMESTER,STUDI
DIMENSION.TOTAL_OBTAINED
  2    FROM STUDENT_DETAILS_DIMENSION,STUDENT_RESULT_DIMENSION
  3      WHERE STUDENT_DETAILS_DIMENSION.USN = STUDENT_RESULT_DIMENSION.USN
  4        AND STUDENT_RESULT_DIMENSION.TOTAL_OBTAINED IN (SELECT TOTAL_OBTAINED FROM STU
T_DIMENSION WHERE TOTAL_OBTAINED IN (SELECT MAX(TOTAL_OBTAINED) FROM STUDENT_RESULT_DIMEN
BY SEMESTER HAVING SEMESTER=3 OR SEMESTER=5 OR SEMESTER=7))
  5
SQL> /

STUDENT_NAME                                       SEMESTER TOTAL_OBTAINED
-------------------------------------------------- --------- --------------
SYEDA                                                     5            624
ARCHANA                                                   7            668
KUMAR ABHISHEK                                            3            704
```

Figure 10: Results regarding toppers in different semester

Figure 11 shows the results after querying student mart. The result analysis example like the number of students who have obtained different results.

```
SQL> SELECT SEMESTER,RESULT,COUNT(*) FROM STUDENT_RESULT_DIMENSION GROUP BY SEMESTER,RESUL

  SEMESTER RESUL   COUNT(*)
---------- ----- ----------
         1 FC            16
         2 FC            16
         3 FL            32
         4 FC            16
         5 FC            16
         5 FL            16
         6 FCD           16
         7 FC            32
         7 FL            16
         7 FCD           32
```

Figure 11: Results regarding result analysis

Figure 12 shows the results after querying employee mart. The information required was the number of employees in each cadre of two particular departments.

```
  1* SELECT DESIGNATION,COUNT(*) FROM EMPLOYEE_DETAILS GROUP BY DESIGNATION
SQL> /

DESIGNATION                                        COUNT(*)
-------------------------------------------------- ----------
ASSISTANT INSTRUCTOR                                        1
ASSISTANT PROFESSOR                                         4
HELPER                                                      3
INSTRUCTOR                                                  4
LECTURER                                                    6
PEON                                                        5
PROFESSOR                                                   4
PROGRAMMER                                                  2
SYSTEM ANALYST                                              1

9 rows selected.
```

Figure 12: Results regarding number of employees

Figure 13 shows the results after querying employee mart. The information required was the total salary paid to each cadre of employees in particular two departments.

```
SQL> SELECT DESIGNATION, SUM(SALARY) FROM EMPLOYEE_DETAILS GROUP BY DESIGNATIOI

DESIGNATION                                        SUM(SALARY)
-------------------------------------------------- -----------
ASSISTANT INSTRUCTOR                                      6130
ASSISTANT PROFESSOR                                      96560
HELPER                                                   15666
INSTRUCTOR                                               41776
LECTURER                                                 96240
PEON                                                     22170
PROFESSOR                                               128080
PROGRAMMER                                               18200
SYSTEM ANALYST                                           12020

9 rows selected.
```

Figure 13: Results regarding salary

6. Conclusions

Justifying a data warehouse project can be very difficult. Usually, analysis of the success of the data warehouse project is done considering the financial benefits against the investment. Since most of the educational institutes are non profit organizations and service oriented, the evaluation of the usefulness of the data warehouse can be done on the basis of its ability to meet user's requirements. The academic data which was spread all across different sources has been loaded into single platform. The decision makers can extract information regarding three main components of the institute, namely Employees, Students and Infrastructure. Employee data mart can provide the users with the information such as career growth and attrition rate. Student mart can provide the information related to the student like best outgoing

student considering his academic and non academic activities. Information regarding assets such as the investment in a particular financial year can also be accessed. In educational institute, decision makers ask "What are the expected results and benefits?" when making a data warehouse project rather than "What is the anticipated return on investment?". The data warehouse developed has met their expectations. Benefits of the present project can be more if the Institute has positive approach towards new technologies.

6. Future Scope

The scope of the study ends with building a data warehouse. Another useful concept OLAP is the main option for future enhancement. At the present stage, data required from data warehouse is to extracted by writing queries. This can be improved by having excellent front end designed for reporting purpose. Reporting tools such as Business Objects permit a user to easily to do the following tasks.

i.Place headings, titles, and explanatory information within charts, tables, and other derived figures;

ii.Add borders and shading to clarify and highlight important information and groupings;

iii.Modify font size and style to emphasize points;

iv.move, edit, or delete data, text, and graphics in final reports;

v.Produce a wide range of figures, including bar graphs, pie charts, bar and line

vi.graph combinations, multiple axis graphics, and scatter plots;

vii.export data in various formats (e.g., ASCII, Excel™);

viii.generate reports in various formats (e.g., html, PDF™, e-mail, paper); and

ix.include legends, citations, explanations, and other information.

The reports required in the format sought by university can be designed which avoids a lot of clerical work.A decision support system's reporting functions must serve a wide range of users-including novices and users with expert analytical capabilities. To accommodate this, most systems offer two primary classes of reporting tools: (1) predefined (static) reports that require little system expertise and are ideal for users with typical information needs; and (2) dynamic (ad-hoc) report-generating capabilities that require greater understanding of both the data and the querying technology, but allow users to investigate more complex questions. *Predefined reports*: Some types of data requests are quite common: How many students are enrolled this year? How many students graduated last year? What percentage of students passed in each semester in each department? Because these and many other data requests are quite common in education settings, they can be anticipated and are often preprogrammed, in predefined reports. These types of reports are especially effective tools for users who require basic and predictable information. Real time ETL refers to the software that moves data synchronously into a data warehouse with some urgency-within minutes of the execution of the business transaction. Implementation of real-time data warehouse reflects a new generation of hardware, software and techniques. Capture, Transform, and Flow (CTF) is a relatively new category of data integration tools designed to simplify the movement of real-time data across heterogeneous database technologies. The transformation functionality of CTF tools is typically basic in comparison with today's mature ETL tools, so often real time data warehouse CTF solutions involve moving data from the operational environment, lightly transforming it using the CTF tool and then staging it.

Acknowledgements

This paper is prepared through experimental results of engg_dwh of reputed institute, quieried with all possible combinations and consulted the decision makers for the usability, discussed with various SME's of datawarehous groups of various organizations in India and abroad, The authors gratefully acknowledge the time spend in this

discussions provided by Mr.Shahzad, SME, CSC USA, Mr. Parswanath Project Manager (Data Warehouse Wing). Wipro Technologies, India. Mr. Govardhan (Architect) IBM India Pvt Ltd, Mr. Arun Kumar Data Architect KPIT Cummins India,Mr.Raghavendra SME, iGATE Global solutions,India,Mrs.Radha Sarvana,Analyst,Wipro Technologies,India.

References

[1] Ralph Kimball, The Data Warehouse ETL Toolkit, Wiley India Pvt Ltd., 2006.

[2] KV.K.K Prasad, Data warehouse development Tools, Dreamtech Press, 2006.

[3] W. H. Inmon, Building the Data Warehouse. Wiley; 3rd edition March 15, 2002.

[4] Alex Berson, Data Warehousing Data Mining & OLAP, Computing Mcgraw-Hill , November 5, 1997.

[5] Arshad Khan, SAP and BW Data Warehousing, Khan Consulting and Publishing, LLC (January 1, 2005)

[6] Vivek R Gupta, "An Introduction to data warehousing",System Services Corporation.

[7] www.datawarehouse-expert.com.

[8] www.learndatamodeling.com.

[9] www.dwinfocenter.org.

[10] Carlo DELL'AQUILA,'An Academic Data Warehouse' World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA ©2007.

[11] McMillen and Randy McBroom.,'what academia can gain from building a data warehouse' No.1,pp.41-46.

[12] Channah F. Naiman, Aris M. Ouksel "A Classification of Semantic Conflicts in Heterogeneous Database Systems", Journal of Organizational Computing, Vol. 5, 1995.

[13] John Hess, "Dealing With Missing Values In The Data Warehouse" A Report of Stonebridge Technologies, Inc-1998.

[14] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in DataWarehouse Systems",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (1) , 2010, 477-485.

[15] Jaideep Srivastava, Ping-Yao Chen,"Warehouse Creation-A Potential Roadblock to Data Warehousing", IEEE Transactions on Knowledge and Data Engineering January/February 1999 (Vol. 11, No. 1) pp.118-126.

[16] Amit Rudra,Emilie Yeo (1999) "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences – 1999.

[17] Jesús Bisbal et all,"Legacy Information Systems: Issues and Directions", IEEE Software September/ October 1999.

[18] Scott W. Ambler "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", Practice Leader-2001.

Authors Profile

Manjunath T.N. received his Bachelor's degree in computer Science and Engineering from SJCIT, Karnataka, India during the year 2001 and M. Tech in computer Science and Engineering from Jawaharlal Nehru National College of Engineering, Shimoga, Karnataka, India during the year 2004. Currently pursing Ph.D degree in Bharathiar University, Coimbatore, Tamilnadu. He is having total 10 years of experience which includes Industry and academics. His areas of interests are Data Warehouse & Business Intelligence, multimedia and Databases. He has published and presented papers in journals, international and national level conferences.

Dr.Ravindra S Hegadi received his Master of Computer Applications (MCA) & M.Phil and Doctorate of Philosophy Ph.D. in year 2007 in computer science from Gurbarga University, Karnataka; He is having 15 years of Experience. He has visited overseas to various universities as SME.His area of interests are Image Mining, Image Processing and Databases and business intelligence. He has published and presented papers in journals, international and national level conferences.

Umesh.I.M. received his Master of Science (MSc) & M.Phil in year 2007 in computer science from Bharathidasan University, Tamilnadu,He is working in RV College of Engineering,Bangalore,Karnataka,India.He is having 10 years of Experience.His area of interests are Image Mining, Image Processing and Databases and business intelligence. He has published and presented papers in journals, international and national level conferences.

Ravikumar G.K. received his Bachelor's degree from SIT, Tumkur (Bangalore University) during the year 1996 and M. Tech in Systems Analysis and Computer Application from Karnataka Regional Engineering College Surthakal (NITK) during the year 2000. He is working as a Project Manager in Wipro Technologies, Bangalore for Data warehousing projects. He had worked with IGATE Global solutions Bangalore and also has worked with SJBIT as Prof and HOD of Dept of CSE and ISE having around 14 years of Professional experienced which includes Software Industry and teaching experience. His area of interests are Data Warehouse & Business Intelligence, multimedia and Databases. He has published and presented papers in journals, international and national level conferences.