Integration of Heterogeneous In-service Training Data into a Nationwide Database

LUNG-HSING KUO, HUNG-JEN YANG
National Kaohsiung Normal University
No.116, Ho-Ping 1st Rd., Kaohsiung City,
Taiwan
admi@nknucc.nknu.edu.tw, hjyang@nknucc.nknu.edu.tw
&
HSIEH-HUA YANG
Department of Health Care Administration
Orient Institute of Technology
No. 58, Sec. 2, Sihchuan Rd., Pan-Chiao City, Taipei County 22061,
Taiwan
yansnow@gmail.com
&
JUI-CHEN YU
National Science and Technology Museum
No. 720, Jiu-Ru 1st Rd., Kaohsiung City,
Taiwan
raisin@mail.nstm.gov.tw
&
LI-MIN CHEN
Department & Graduate School of Industrial Design
National Yunlin University of Science and Technology
123 University Road, Section 3, Douliou, Yunlin 64002,
Taiwan
chenlm02@yahoo.com.tw,

Abstract: The integration of heterogeneous in-service training data offers possibilities to manually and automatically draw up new information of professional human resources, which are not available when using only a single data source. Furthermore, it allows for a consistent representation and the propagation of updates from one data set to the others. However, different acquisition methods, data schemata and updating cycles of the content can lead to discrepancies in professions and expertise accuracy and correctness which hamper the combined integration. To overcome these difficulties, appropriate methods for the integration and harmonization of data from different sources and different types are needed. In this study, twenty-five databases were integrated into one national level database. More than 220,000 K--12 teachers' in-service training data were collected through out this heterogeneous databases integrating project. A unified subject-category was introduced based upon both the teaching professions and administration professions so as to absorb all twenty-five sources. The integrated feasibility was evaluated according to the efficiency and correctness.

Key-Words: Integration; Nationwide databases; In-service training data; Heterogeneous database integration

1 Introduction

A database could provide dynamic information generating if it has integrated all possible data sources. Information from the electronic audit was used to provide feedback to preceptors, modify the training curriculum, and increase accessibility[1].

1.1 Data integration – benefits, requirements, and applications

With the advance of more and more administration system and knowledge workers terminals, the number of available digital data sets is ever increasing[2, 3]. This is especially true for educational human resource management data, which are acquired by different organizations, e.g. local

public schools, local private schools, administrations like departments of education at city level, and Ministry of Education at national level. In addition to this diversity of data providers, there is a diversity of data models, data acquisitions schemes, as well as the way to categorize training subjects. All attributes have to be interpreted in order to extract explicit professional growth information[4, 5].

The diversity of available digital data sources carries the chance of integration the aim of which is to exploit the relative benefits of each. These benefits are:

- Integrated analysis with prior information: data of one data set can be inspected using also data from the other sets.
- Reference to common training: linking data sets based on common objects allows making professional reference to other district data sets.
- Mutual corrections and refinements of teaching profession: if the accuracies of the data sets are known, they can be taken into consideration to generate a new, enhanced, and expert description.
- Mutual enrichment with in-service training program: attributes and properties of one data set can be transferred or extended to the others.

Obviously, true integration is much more than just overlaying data in an in-service training information system, as it must make explicit the relations between individual objects in the different data sets.

Technically, it also means more than information fusion, if the original data sets should still be available to be used in their own right. This is a common requirement as different agencies are interested in maintaining control over the data they are responsible for and they have the knowledge to maintain the data properly.

The requirements for the realization of a true integration of teacher in-service training data sets are as follows:

- Corresponding objects of different data sets have to be identified so that they can be connected by explicit links.
- The links have to be set up by the automatic filter which takes care of disagreement.
- It should be possible to match and link arbitrary data types, for instance:
  - Date as string
  - Big5 to UTF8
  - Number as string
- Different credit units should be considered, and different units should be converted into the same unit, for instance:
  - One credit hour as eighteen hours
  - Half an hour as 0.5 hour
- The filtering and linking is to be carried out in a nationwide database environment.

Besides the general benefits of data integration, there are a lot of practical applications of integration [6-8]. Database are used in different situations, such as statistical yearbook editing, and e-learning.[9-11] One is the verification and update of data sets: in order to check the prevalence and correctness of a data set, a second or third data set can be used to check the information. The task can be extended to provide also update capability: whenever current information is available in one data set, it can be employed to update the other sets, based on the known relations between the data sets and the known link structure between the corresponding objects.

Furthermore, integration can be used to provide prior information for a dedicated analysis using one data set: for example if an in-service subject data set has to be updated, information from existing in-service subject data can be used to partition the subject and identify potential teachers for new courses. Similarly, course-joining classifications by age-group can be used as input for a more detailed inspection of dedicated training courses of interest. Also, data from one information source (e.g. Taipei) can be used to enrich data from another one (e.g. Kaohsiung). In this way, comparative data can be enriched with demographic information, which is used as an indirect professional growth in many other databases.

The remainder of the paper is structured as follows:

- The application scenario of research work and the data sets employed in our study is briefly described.
- An overview of the state of the art in data integration is presented.
- The developed architecture for database supported integration is described.
- Methods for heterogeneous data integration are presented.
- Results demonstrating the potential of the proposed solution are presented,
- Conclusions are drawn and further work is discussed.

1.2 Application scenario

The research aims were to develop concepts for the integration of heterogeneous in-service training data sets. Those different ideas had been developed and implemented in a prototype consisting of three modules: a file/database integration module, a database/database integration module, and a database module as a base for our work.

Describing and representing certain teacher's professional growth status from different districts' databases will naturally lead to different data sets. Due to the same training context, similar courses could be taken with different course providers. All the training records of a single in-service teacher could be stored in different databases belonging to the authorities in different districts. This is true for the data sets under investigation in our study. The whole picture of in-service professional growth of certain teacher required information culled from different data sets.

When overlaying the required data sets, these common objects should obviously coincide professionally. However, this is not exactly true. The reasons are manifold: First, different data acquisition methods may have been used; second and more importantly, one of the data sets may have been categorized differently and needs to be recoded.

Interestingly, those modules used methods for in-service training records alignment--one case being the alignment to a category, and the other the alignment to personal attributes. The underlying concepts and methods, however, are similar. Both modules had been implemented in the framework of a nationwide database. This database is firstly used to store all the data. Secondly, it provides mechanisms for data pre-processing that are optimized on the database. Thirdly, it models, represents and maintains the data structure into which many-to-many relationships between corresponding objects are embedded and provides interfaces to access the individual data sets and the linked partners. Finally, it controls the application processes in order to keep track of the housekeeping data.

1.3 Heterogeneous data platforms

Professional development courses of in-service teachers were provided by various institutions. There are twenty-six districts in which education administration departments are equipped with authority to keep records. Database and application language of data set of each district were listed in Table 1.

Table 1 Platform information of each district data set

District	AP Lang.	Database	OS
1.	Asp	MSSQLServer	Windows2000
2.	VS.net	MSSQLServer	Windows2003
3.	java	PostgreSQL	Centos5
4.	php	MySQL	Solaris2.8
5.	php	MySQL4.0.27	freeBSD6.3
6.	AspX	SQL2000	Windows2003
7.	Asp.net2003	SQL2000	Windows2000
8.	php4.4.6	MySQL4.12	freeBSD4.9
9.	php5.4	PostgreSQL8.1.9	freeBSD7.0
10.	Jsp	PostgreSQL	Linux
11.	php4	MySQL4.1.11	FreeBSD4.10
12.	php4	PostgreSQL	Fedoracore2
13.	php5	MySQL4.1.7	FreeBSD5.5
14.	php4	MySQL	debian
15.	ASP.net	MSSQL	Window 2005
16.	ASP.net	MSSQL	Window 2005
17.	ASP.net	MSSQL	Window 2005
18.	ASP	SQL2000	Windows2003
19.	php	MySQL	Linux
20.	ASP.net	MSSQL	Window 2005
21.	ASP.net	MSSQL	Window 2005
22.	java	DB2	UNIX
23.	ASP.net	MSSQL	Window 2005
24.	ASP.net	MSSQL	Window 2005
25.	ASP.net	MSSQL	Window 2005

26.	Asp	MSSQLServer	Windows

In those twenty-six districts, application developing languages used were:

1. ASP
2. VS.net
3. Java
4. PHP4
5. PHP5
6. ASPX
7. ASP.net

Databases used by these twenty-six district authorities were:

1. MS SQL
2. MySQL
3. PostgreSQL
4. DB2

Operating systems used by these twenty-six district authorities were:

1. Windows 2000
2. Windows 2003
3. Window 2005
4. Centos5
5. Solaris 2.8
6. UNIX
7. FreeBSD 6.3
8. FreeBSD 4.9
9. FreeBSD 4.1
10. FreeBSD 5.5
11. FreeBSD 7.0
12. debian
13. Linux
14. Fedora core 2

According to Table 1, the integration is facing quite a wide range of languages used, DBMS used, and OS used.

2. State of the art of in-service teacher training data integration

The integration of in-service training data sets presented in this paper is based upon the formal discussion of committee members.

2.1 Milestones of data integration

There were three major stages of integrating system development. The first stage was from 2003 to 2005, the foundational functions were established in those three years. In Table 2, major milestones and actions were listed. The nationwide basic database was first founded at 2003.

Table 2 The milestones and actions in the first developing stage

Year	Milestones/actions
2003	● Nationwide In-service Education Information Network Project Initiation ● Prototype database test run ● Nationwide database established
2004	● Data integration suggestion initiated in the practitioner annual meeting
2005	● Data integrating typology setup ● Prototype system in design

In Table 3, the milestones and actions were listed for years 2006 and 2007. Second data integration meeting was held. The formal regulations on district data submission was enforced by Ministry of Education based upon the general agreement of all the district authorities.

Table 3 the milestones and actions in the

second developing stage in year 2006

Year	Milestones/actions
2006	● Second Data Integration Meeting
	● Regulations on district data submission enforced
	● Data submission system implemented
	● Required fields defined
	● Data submission frequency issued by each district based on direct injection, monthly file upload, seasonal file upload, or semester file upload.
	● The function of on-line data editing after submission added
	● The function of on-line data insert after submission added
	● The function of on-line data verifying after submission added
	● The system of on-line data appending added
	● The function of school code on-line search added
	● The function of on-line data deleting after submission added
	● The function of on-line data fields referencing added
	● The "resource center of special education" added into school code data sets
	● Database integrating service survey conducted
	● Coding reference table of data fields revised

In the second developing stage, functions to supporting integration were designed and implemented. In year 2006, a work plan was first confirmed by the Second Data Integration Meeting. According to the work plan, functions to support database service were developed.

In the table 3, all the major contributions were listed for reference. In the table 4, the major contributions of year 2007 were also listed. Daily report service was provided at this stage.

Table 4 the milestones and actions in the second developing stage in year 2007

Year	Milestones/actions
2007	● Annual meetings change to half-year meetings which would be hosted by turns by each district
	● Exporting data with the file types of excel, access, pdf from integrated database added
	● The function of daily report of data submitted and sent in by e-mail added
	● The backup function of temporary data of each district added
	● A mission team to deal with submitted data initiated
	● Coding reference table of course attributes revised

By the end of this second stage, a mission team started to promote data submitting. Their works were quite welcomed by district users.

Table 5 The milestones and actions in the third developing stage

Year	Milestones/actions
2008	• The function of accepting data via post URL added • Data exchange standards including methods and definitions revised • The function of searching courses without trainee lists added • The field of total submitting member for district database verifying purpose added • The field of export status for preventing redundant download added • The on-line delete all function for the course management system added • Exporting data with the file types of XML from the integrated database added • Weekly cleaning databases of test-run systems scheduled • The Q&A of the mission team dealing with submitted data to the public posted • The prototype of the single sign on function developed

In Table 5, the milestones and actions were listed for year 2008. In this stage, the integrating of heterogeneous in-service training databases became mature.

In year 2008, the system provided service to read uploading data through web post. This function help some districts w

2.2 Attributes of data sets

Personal information and course information were two major tables of this database. The personal information table would be used for the following fields:

Name, id, birth date, school, in-service training records.

The course information table would be used for the fields of:

Course name, schedule type, credit type, date.

3. Methodology

A prototype system development method was applied when the original system was designed. The other method applied in this study was a panel discussion.

The purpose of this study was to establish a information system to integrate all twenty-six heterogeneous in-service training databases. A hypothesis was set up to test the usability of injecting data into the integrated database. Paired t-tests would be applied to verify the usability.

4. Integration architecture

In this section, the developed architecture and modeling concepts of the database integration are explained. The database architecture was designed to preprocess inputs and to store and export results of the in-service trainees' data and the in-service course integration steps.

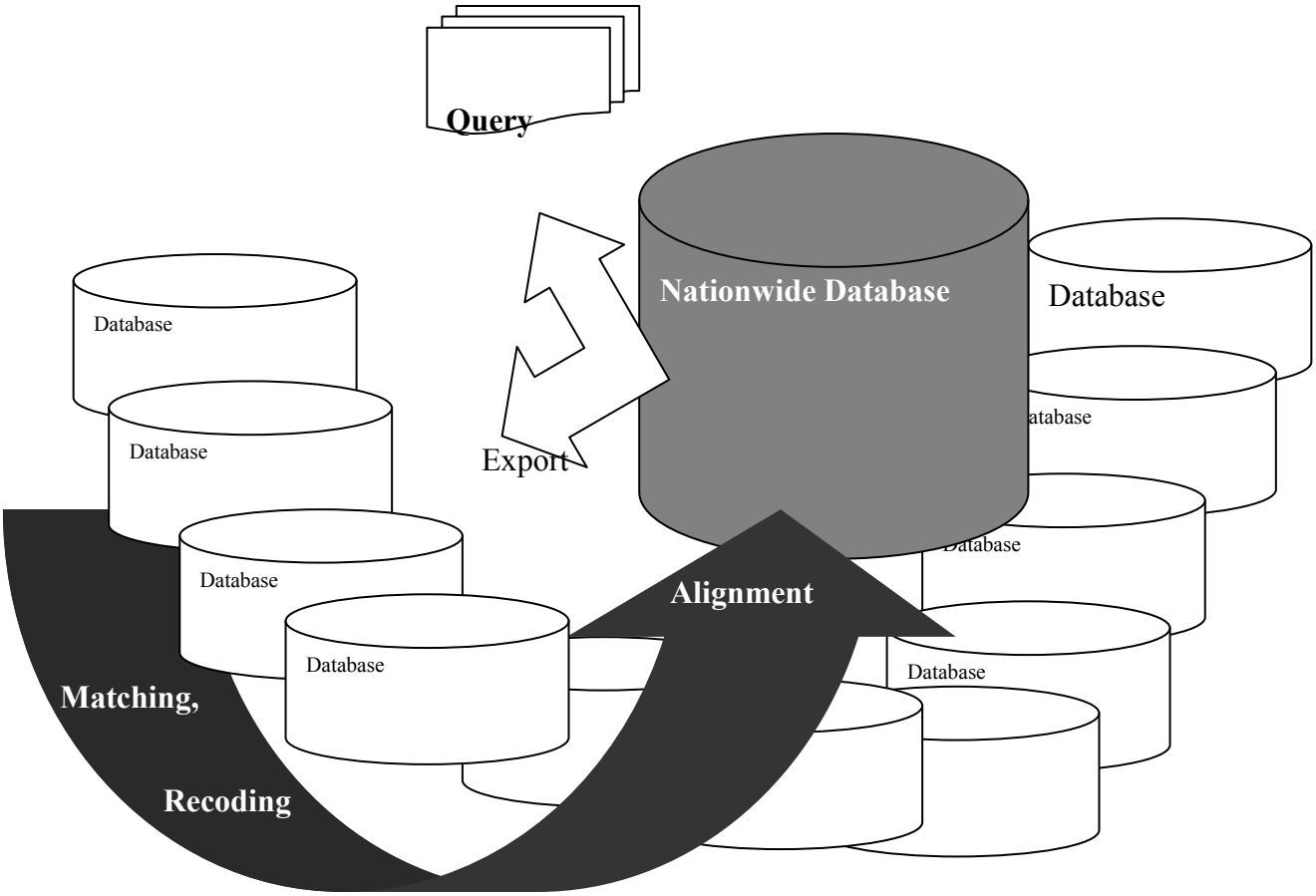

Figure 1 Architecture overview

Fig. 1 gives a simplified overview of the integration architecture with respect to the interaction between the database and the course categories matching and recoding processes.

The underlying database architecture is chosen according to the paradigm of in-service training, as it gives a close coupling and at the same time keeps district databases autonomous. Fig. 2 shows this kind of architecture.

In this way, the matching and re-coding processes are given an integrated view of the different databases via a global database schema for overall applications. Nevertheless, the access to the local individual databases is still possible. For this purpose, the known architecture of the nationwide database is extended to handle personal professional growth records, certification records, and linkage between multiple data sets. In order to join district course category systems, a mapping to harmonize the course object classes of different data sets had been defined on the schema level. For the object level, the process of matching for identifying objects was developed for data submitted based upon the structure for maintaining links between corresponding objects.

4.1. Schema adaptation

To make the structurally different data sets accessible for nationwide service, a generic but flexible export schema was designed based upon experiences with in-service teacher's professional growth data sets containing teaching and administrating objects with respect to object-relational databases. The schema contains all objects, object classes, attribute types, and attribute values, each of them in one entity consisting of two tables in the relational database management system.

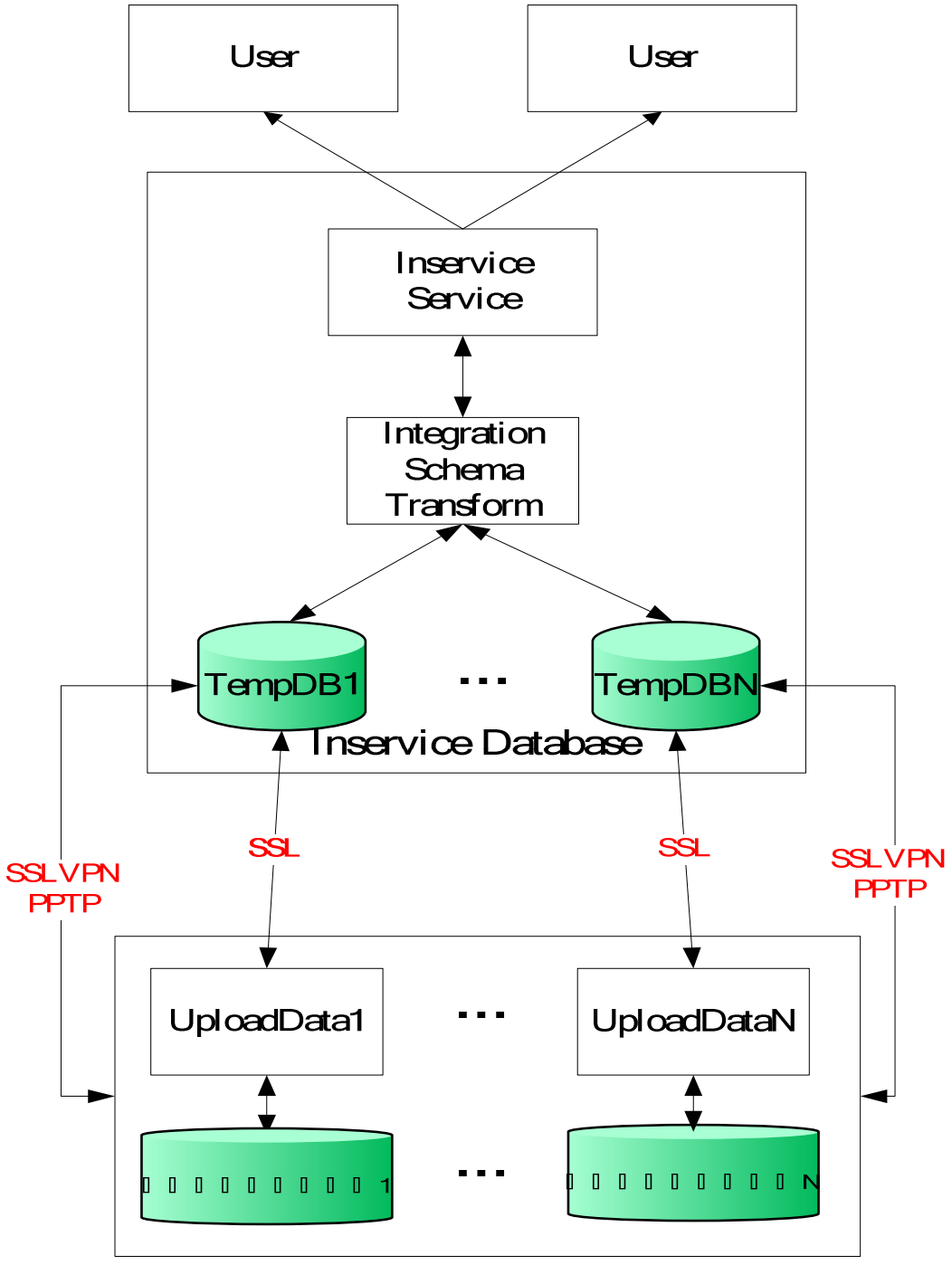

Figure 2 Architecture of a nationwide database

Figure 3 shows the schema for personal information data and course data. They have application specific attribute types and object classes according to their own representation model.

A in-service training object of entity type Personal_Object has several entries of type Personal_Attributes, namely (attribute, value)-pairs like (major, technology education). The corresponding type of the attributes or the classification of the professional objects can be

found in the collections Personal_AttributeTypes and Personal_ObjectClasses.

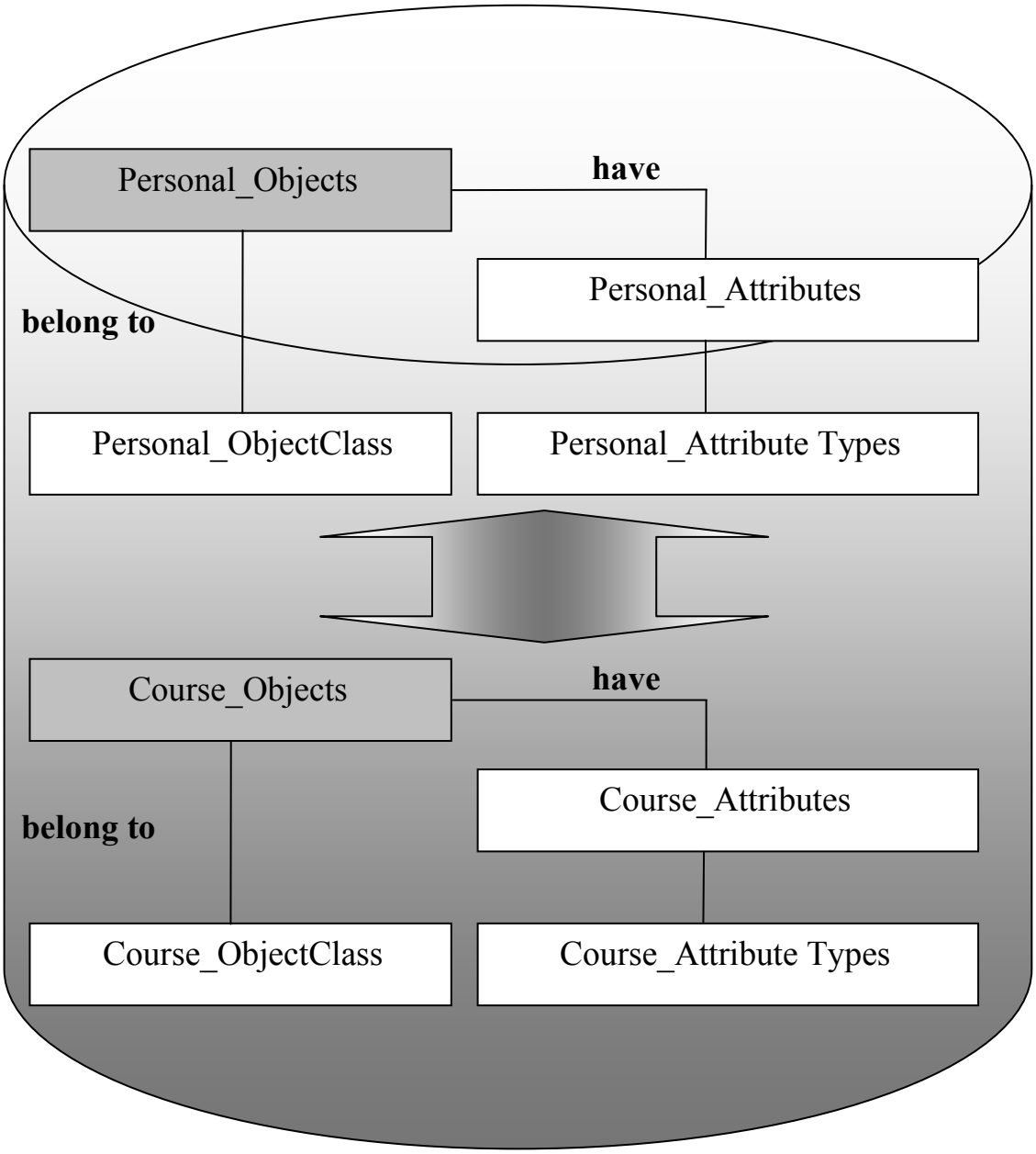

Figure 3 Schema for personal information data and course data

4.2. Course category integration

Given the structural adaptation of the different data sets, the nationwide database can be used to incorporate correspondences through re-coding. For each fields of course data sets, referenced tables were standardized and listed in Table 5.

In Table 5, major categories of each field were listed for reference purpose. They were class types, schedule types, fee types, supporters, audience, and three layers of content types.

For the field tof class types, the content was categorized into non-credit, bachelor credit, master credit, bachelor degree, master degree, intra-institutional non. These categories are listed in the table for the reference purpose.

For the field of schedule types, the content were identified as daytime of weekdays, nighttime of

weekdays, weekends, summer vacations, winter vacations, on-line. This field was designed to store the class offered time periods.

For the field of fee types, it was designed to identify how the class was paid. For the field of supporter, it was designed to identify which institution offering this class.

For the field of audience, the audience was divided into six groups. Those are high school teachers, vocational high school teachers, junior high school teachers, elementary teachers, kinder-garden teachers, special education teachers

Table 6 Standardized reference-content of course fields

Field	Major categories
Class Types	Non-credit, Bachelor credit, Master credit, Bachelor degree, Master degree, Intra-institutional
Schedule Types	Daytime of weekdays, Nighttime of weekdays, weekends, summer vacations, winter vacations, on-line
Fee Types	None, pay by oneself, pay partially
Supporter	Institution code
Audience	High school teachers, vocational high school teachers, junior high school teachers, elementary teachers, kinder-garden teachers, special education teachers
Content type	Layer0 (admin, subjects)
	Layer1
	Layer2

4.3. Integrated database

Figure 4 summaries the architecture of the integrated database. This nationwide service supports the following tasks:

1. Collecting district data contents
2. Maintaining data exchange based upon standards
3. Maintaining temporary submitted data storage and backup
4. Processing temporary submitted data and converted them into formal data

4.4. Usability of data injection

The result of paired t-tests was not significant at 0.05 level. The t value was -1.585 and the significant level was 0.133. This statistical result provide evidence against the null hypothesis: "The number of submitted data was significantly different to the number of right data". It is concluded that the usability of integrated database system is significantly well.

5 Conclusion

The purpose of this study was to identify a long term research project of constructing a nationwide heterogeneous integrated database system for in-service training data. The system usability was supported by the statistical test.

The benefits of data integration were pointed out. The application scenario was also reported. The major milestones and actions were set up and taken to persuade all the district authorities of education departments. The state of the art of in-service teacher training data integration was established based upon all the regulations.

The study findings provide information of architecture overview of the system.

The integration of heterogeneous in-service training data offers possibilities to manually and automatically draw up new information of professional human resource, which are not available when using only a single data source. Furthermore, it allows for a consistent representation and the propagation of updates from one data set to the others. However, different acquisition methods, data schemata and updating cycles of the content can lead to discrepancies in professions and expertise accuracy and correctness which hamper the combined integration. To

overcome these difficulties, appropriate methods for the integration and harmonization of data from different sources and different types are needed.

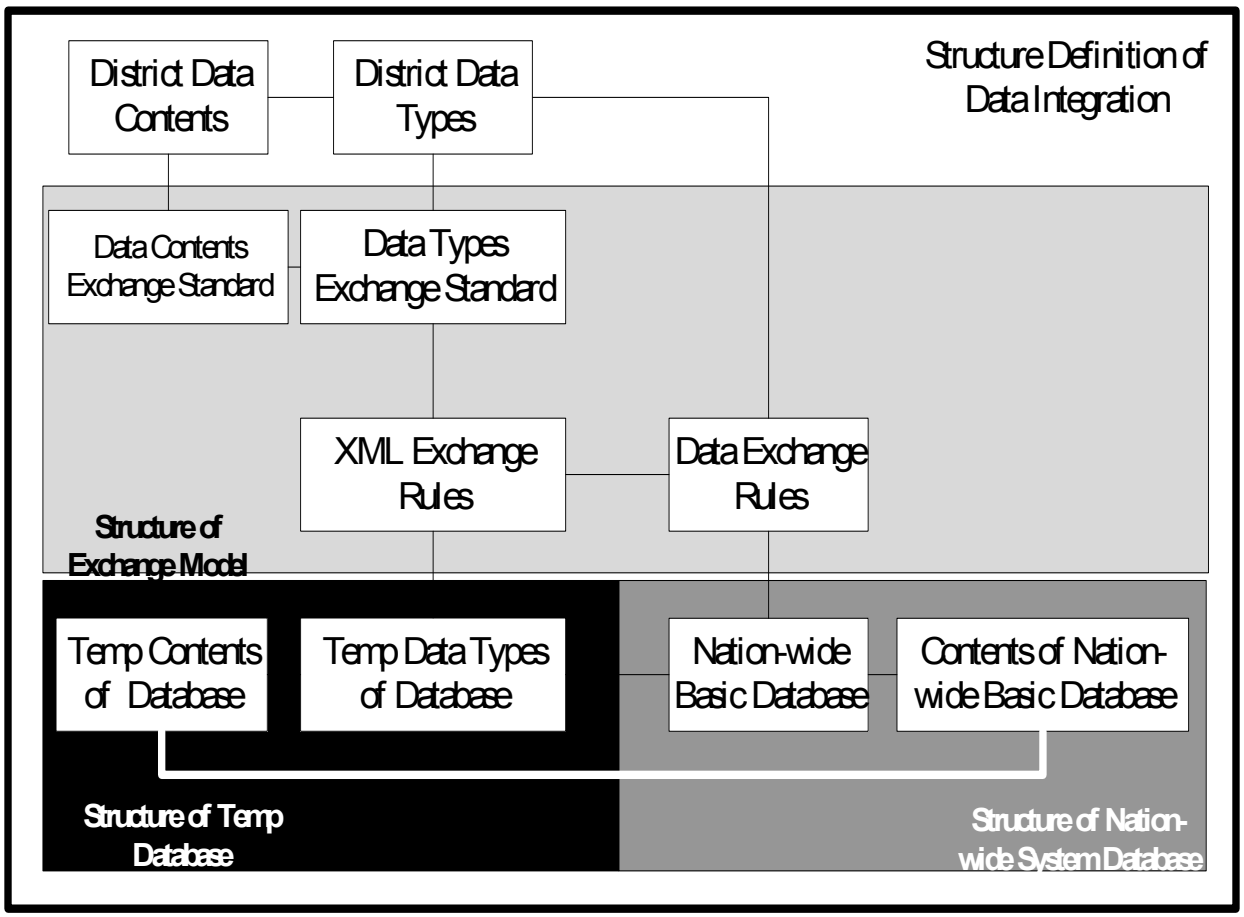

Figure 4 Structure of integrated database

In Figure 4., the structure of integrated database was illustrated. The data contents exchange standard and the data types exchange standard were setup the foundation of integrating district data. According to these two standards, XML exchange rules and data exchanges rules were established.

References:

[1] Williams, J., et al. "Use of an electronic record audit to enhance mental health training for pediatric residents". *Teach Learn Med,* 2007. **19**(4): pp. 357-61.

[2] Reisch, L.M., et al. "Training, quality assurance, and assessment of medical record abstraction in a multisite study". *Am J Epidemiol,* 2003. **157**(6): pp. 546-51.

[3] McCain, C.L. "The right mix to support electronic medical record training: classroom computer-based training and blended learning." *J Nurses Staff Dev,* 2008. **24**(4): pp. 151-4.

[4] Sujansky, W. "Heterogeneous database integration in biomedicine". J Biomed Inform, 2001. **34**(4): pp. 285-98.

[5] Wei, C.P., P.J. Hu, and O.R. Sheng. "A knowledge-based system for patient image pre-fetching in heterogeneous database environments--modeling, design, and evaluation". *IEEE Trans Inf Technol Biomed,* 2001. **5**(1): pp. 33-45.

[6] Corwin, J., et al." Dynamic tables: an architecture for managing evolving, heterogeneous biomedical data in relational database management systems". *J Am Med Inform Assoc*, 2007. **14**(1): pp. 86-93.

[7] Praz, V., V. Jagannathan, and P. Bucher. "CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature". *Nucleic Acids Res,* 2004. **32**(Database issue): pp. D542-7.

[8] Rainaldi, G., et al. "PLMItRNA, a database on the heterogeneous genetic origin of mitochondrial tRNA genes and tRNAs in photosynthetic eukaryotes" *Nucleic Acids Res*, 2003. **31**(1): pp. 436-8.

[9] Yang, H.J., Kuo, L.H., Che-Chern Lin, & Wei H.M. Integrating Databases for Compiling Statistical Yearbook of Teacher Education, April 2006, *WSEAS Transactions on Engineering Education*, **3**(4),

[10] Yang, H.H., & Yang, H.J. Investigating the Opinions of University Students on E-learning, May 2006,*WSEAS Transactions on Communications*, **5**(5)**.**

[11] Yang, H.J., Yu, J.C., & Yang H.H. A Study of On-line Problem-based Learning Framework, May 2006,*WSEAS Transactions on Computers*, **5**(5)**.**