Aseguramiento de integridad de datos para el sistema de encuestas del ITSH

Assurance of data integrity for the ITSH survey system

CRUZ-LUNA, Manuel†*, LUNA-TREJO, Cupertino y URBINA-FERNÁNDEZ, José

Instituto Tecnológico Superior de Huauchinango

ID 1er Autor: *Manuel, Cruz-Luna* / **ORC ID**: 0000-0002-0640-8926, **Researcher ID Thomson**: H-8709-2018, **CVU CONACYT ID**: 368159

ID 1er Coautor: *Cupertino, Luna-Trejo* / **ORC ID**: 0000-0001-5898-8486, **Researcher ID Thomson**: I-6465-2018, **CVU CONACYT ID**: 904398

ID 2do Coautor: *José, Urbina-Fernández* / **ORC ID**: 0000-0003-1805-0973, **Researcher ID Thomson**: I-6545-2018, **CVU CONACYT ID**: 905112

Recibido: Marzo 26, 2018; Aceptado: Junio 05, 2018

Resumen

La integridad de los datos es una de las características esenciales que se deben tomar en consideración cuando se trata de bases de datos, principalmente si se trata de una base de datos ubicada en un servidor, que es alimentada de forma asíncrona desde diversos dispositivos móviles. En este caso, es un sistema de aplicación de encuestas dentro del Instituto, donde los docentes recaban información de los alumnos, la almacenan dentro de sus dispositivos móviles y posteriormente se envía al servidor para su concentración; parte de los datos que recibe el servidor, debe ser distribuida entre todos los dispositivos móviles utilizados para aplicación de encuestas, para que dichos datos sirvan como datos de entrada en la captura de información de nuevas encuestas aplicadas de manera subsecuente. En el momento de almacenar los datos en cada dispositivo móvil, se genera de manera interna una clave primaria, que pude ser diferente en otro dispositivo, aunque se trate de la misma información; al llegar estos registros al servidor, se debe identificar que se trata del mismo dato y generar una clave primaria general y se deberá hacer llegar a todos los dispositivos móviles para utilizarla en nuevas encuestas aplicadas

Integridad, Base datos, Multiplataforma

Abstract

Data integrity is one of the essential properties to take in consideration when using databases, mainly if it's about a database located into a server, which is fed in asynchronous way from several mobile devices. In this case, it's a system for survey application inside an Institute, where teachers collect information from students, they store it in theirs mobile devices and later send it to the server for their concentration; part of the data that receives the server must be distributed among all mobile devices used for survey application, so that said data serve as input data in the capture of new surveys applied subsequently. At the time of store data in each mobile device, a primary key is internally generated, which can be different from others devices, even if it's the same information; when this records arrive at the server, it must be identified it is the same data and generate a general primary key that must be sent to all mobile devices to be used in the application of new surveys

Integrity, Databases, Multiplatform

Citación: CRUZ-LUNA, Manuel, LUNA-TREJO, Cupertino y URBINA-FERNÁNDEZ, José. Aseguramiento de integridad de datos para el sistema de encuestas del ITSH. Revista de Tecnologías Computacionales. 2018. 2-6: 15-21.

*Correspondencia al Autor (Correo Electrónico: mcruzl@hotmail.com)
† Investigador contribuyendo como primer autor.

Introducción

Una de las características más importantes al momento de almacenar información dentro de una base de datos, es la integridad de dicha información, haciendo referencia principalmente a su exactitud y fiabilidad. Se tienen diversos tipos de integridades y en este trabajo se aplicará la integridad referencial, donde se apunta a las relaciones que existen entre dos o más tablas por medio de llaves primarias y foráneas [1], [3].

En este caso la base de datos se encuentra distribuida en un servidor y en diversos dispositivos móviles que es donde se generan los datos, se almacenan de manera local con sus propias llaves primarias y posteriormente se envían hacia el servidor, donde se deben generar llaves primarias globales al momento de recibir la información sin importar de que dispositivo es enviada; estas características se encuentran definidas de manera clara en la estructura y la lógica interna de la base de datos [4]. para mantener la consistencia de la base de datos se deben cumplir con las reglas de integridad que requieren los datos, sin importar las fuentes donde se genera la información [2], [6].

El esquema de esta aplicación es semejante al de una base de datos distribuida, debido a que los datos se generan de manera independiente en diversos dispositivos móviles y algunos de esos datos deben quedar disponibles para todos los demás, con el inconveniente que los dispositivos no siempre se encuentran conectados y es necesario realizar copias de la información hacia el servidor de datos en cuanto se reestablezcan las conexiones [5], [7].

Dentro de este documento se dará inicialmente un panorama del escenario donde se aplica la integridad de los datos, se hace un recuento de los tipos de integridad y la que aplica en este caso, además de los trabajos previos realizados sobre este tema, se describe la forma en que fluye la información de manera interna, se enlista la infraestructura aplicable al proyecto y los servicios que se ofrecen, el algoritmo propuesto de solución al problema existente y se muestran los resultados logrados; finalmente se hace mención de las conclusiones obtenidas.

Escenario de aplicación

En el Instituto Tecnológico Superior de Huauchinango (ITSH) se requiere que grupos de docentes apliquen encuestas entre sus alumnos como parte de un programa de tutorías, donde recaban sus datos personales, escuela de nivel medio superior de procedencia, becas obtenidas, estado de salud, datos laborales si es que trabajan e información de sus padres o tutores. Todos los datos que se generan en estas encuestas eran almacenados en hojas que utilizadas al momento de aplicar las entrevistas con los alumnos; en el momento en que un docente cambia de grupo tutorado, se debe pasar todo el expediente con el historial de cada alumno del grupo al nuevo tutor para que tenga la información necesaria de cada alumno y lo pueda guiar de manera adecuada.

En caso de que se requieran algunos datos para generar estadísticas, se deben calcular de manera manual por cada docente que tiene a su cargo el grupo tutorado, a través de la revisión física de los documentos que tiene en el expediente del grupo y obtener los resultados solicitados. Al generar el concentrado general, se obtienen valores diversos para el mismo dato y se debe llevar a cabo una depuración para obtener información que sea de utilidad.

El sistema de encuestas del Instituto debe tener la capacidad de almacenar la información, al momento de ser generada, de manera directa dentro de un dispositivo móvil por medio de una aplicación [17]; posteriormente se transferirá hacia un servidor para concentrarla y tener un solo lugar desde donde hacer las consultas que se requieran.

Algunos datos que se capturen en las encuestas deben quedar disponibles para todos los demás docentes y con esto se reduzcan los errores de captura o las diferencias de información; para esto se requiere que dichos datos sean distribuidos a las bases de datos de todos los dispositivos móviles que tengan la aplicación para realizar encuestas.

Todo este flujo de información entre bases de datos ubicadas en diferentes plataformas requiere de un control preciso de la información [13], para evitar la pérdida o duplicidad de dicha información relacionada con cada uno de los alumnos.

ISSN: 2523-6814

Integridad de la información

La integridad de la información es un término que hace referencia a la exactitud y confiabilidad de la información que se almacena dentro de una base de datos. Estos datos deben estar completos, ubicados en el lugar adecuado y con una relación estrecha y bien establecida para realizar consultas y obtener información útil con el menor esfuerzo computacional requerido y en un tiempo mínimo [15]. Algunos de los problemas de integridad de la información se generan desde el usuario, debido a que por descuido o falta de conocimiento en el manejo del equipo de cómputo, no coloca los datos de manera correcta al alimentar una base de datos, lo que da como resultado información duplicada o datos erróneos que pueden llevar a obtener datos no confiables en las consultas que se realizan a las bases de datos.

Se tienen diferentes tipos de integridad de datos, la primera conocida como integridad de la entidad, que hace referencia a que los registros de una tabla deben ser únicos, no debe haber dos registros idénticos, para esto se hace uso de llaves o claves primarias [11], [12], pudiendo estar formadas por uno o más campos; este campo o combinación de campos deben ser únicos dentro de la tabla y no pueden tener valores nulos, es decir, no pueden estar vacíos, debido a que esto limita el desempeño de la mayoría de algoritmos utilizados para perfilar y manipulas las llaves [9], [10], [16].

La siguiente integridad es conocida como referencial, en donde se debe tomar en consideración que, al relacionar dos tablas, el valor que se coloque en la tabla secundaria dentro de la llave foránea debe existir en la llave primaria de la tabla principal, para evitar que se tengan registros "huérfanos" en las tablas, generando con esto que las columnas del mismo atributo sean semánticamete equivalentes entre ellas [8].

La tercera integridad, que es de dominio, hace referencia a la validez de los datos que se encuentran en cada uno de los campos de las tablas, esto se logra estableciendo desde el inicio del diseño el tipo de campo correcto y correspondiente a los valores a almacenar [4], además de las configuraciones de restricciones aplicables a través de expresiones regulares.

La pérdida de la integridad de datos se da principalmente por errores del usuario final o procesos de control de cambios deficientes o no desarrollados de manera completa [14].

Las soluciones existentes en bases de datos móviles no son independientes del servidor, debido a que utilizan información como metadatos o funciones específicas como disparadores y marcas de tiempo [18], y en algunos otros casos se hace uso de algoritmos basados en resúmenes de los datos para identificar diferencias o semejanzas [20].

Derivado de las características del sistema de encuestas del Instituto, no será necesaria una sincronización de todos los datos de las bases de datos, debido a que no toda la información se debe tener disposible en todos los dispositivos móviles, únicamente se requiere de controlar las llaves primarias y foráneas de las tablas auxiliares (Ciudad e Institución) entre todas las bases de datos, siendo este un proceso en un solo sentido o en ambos [19].

Flujo de información

Lo datos que se van a almacenar por cada uno de los alumnos incluyen entre otros su nombre y apellidos, sexo, fecha de nacimiento, estado civil, enfermedades, pasatiempos, deporte practicado, lugar de nacimiento y nombre de la institución de procedencia entre otros.

Para cada uno de estos últimos dos datos se va a realizar una tabla de tipo catálogo con su llave primaria generada de manera automática para llevar a cabo las relaciones correspondientes hacia la tabla **Alumno**.

Para la base de datos del servidor se va a incluir un campo con la fecha de actualización de cada registro y será utilizado como referencia para la sincronización de datos hacia las demás bases de datos, la figura 1 muestra parte del diagrama de la base de datos para el servidor, donde se muestra la tabla **Alumno** y las tablas **Ciudad** e **Institucion**, haciendo referencia al lugar de nacimiento y a la institución de procedencia respectivamente.

ISSN: 2523-6814

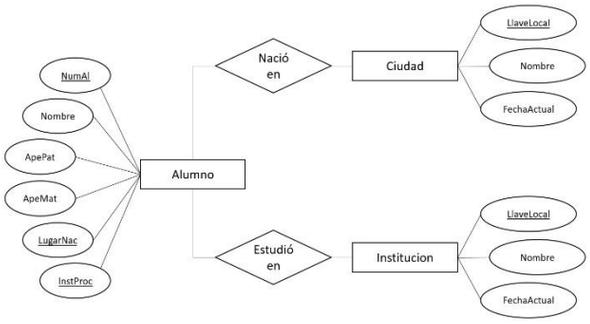

Figura 1 Diagrama parcial BD servidor
Fuente: Elaboración propia

En cuanto a la base de datos para los dispositivos móviles, contará en las tablas de catálogos con dos llaves, una de ellas será la primaria que se genera de manera local cada vez que se almacena la información de una nueva encuesta que requiere de un dato que aún no se ha almacenado; la otra llave se utilizará para hacer referencia al valor que contiene ese mismo registro, pero en la llave primaria dentro de la base de datos del servidor. Ambos valores se deben conocer dentro de cada dispositivo móvil, para que los siguientes registros que se almacenen tomando los valores ya existentes tengan forma de hacer referencia a los registros que se encuentran dentro del servidor sin la posibilidad de generar duplicidad de información o sobrecarga de operaciones innecesarias al verificar la existencia de datos que ya se han almacenado. La figura 2 muestra parte del diagrama de la base de datos para el dispositivo móvil, donde se identifican los campos para las dos llaves mencionadas.

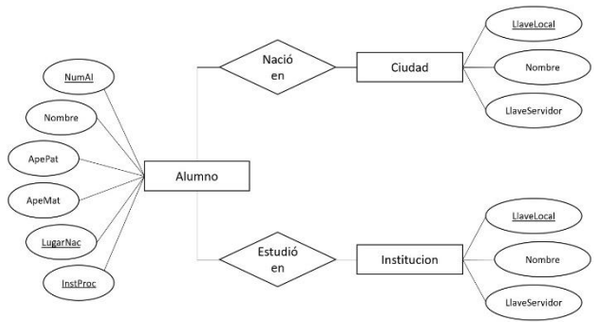

Figura 2 Diagrama parcial BD móvil
Fuente: Elaboración propia

Infraestructura utilizada

Se tiene un servidor disponible con las siguientes características:

- Sistema operativo Windows Server.
- Internet Information Services con soporte para ASP.NET
- SQL Server

Lo que nos dá como resultado los siguientes servicios:

- Servidor web. Incluye un servicio web para transferencia de información, en formato JSON, hacia y desde los dispositivos móviles.
- Servidor de datos. Contiene la base de datos donde se almacena la infomación completa de las encuestas aplicadas a los alumnos.

Dispositivos móviles con sistema operativo Android y una aplicación instalada para la captura de información de encuestas y envío de datos hacia el servidor por medio de red inalámbrica.

Todo esto hace uso de la red Institucional que se encuentra instalada dentro de las instalaciones.

Algoritmo propuesto

La descripción del siguiente procedimiento para asegurar la integridad de los datos relacionados con el lugar de nacimiento de los alumnos, también se aplica a estado civil, institución de procedencia, enfermedad, pasatiempo, deporte, comida y bebida, con sus tablas y campos respectivos.

Para el caso de la base de datos de los dispositivos móviles, contendrá también la tabla **Alumno** y las demás tablas de catálogo, con la diferencia de que estas últimas no contendrán como dato adicional la fecha de actualización, sino un campo adicional para almacenar el valor de la llave primaria con que se dio de alta en el servidor por primera vez.

La tabla 1 muestra el contenido de la tabla **Ciudad** ubicada dentro de un dispositivo móvil, donde el primer campo contiene la llave primaria que se genera dentro del mismo dispositivo móvil al momento de almacenar un nuevo registro, el segundo campo contiene el lugar de nacimiento y el tercero contendrá el valor de la llave primaria de ese lugar de nacimiento una vez que se da de alta en el servidor por medio de una actualización de datos.

ISSN: 2523-6814

La primera encuesta que se capturó en ese dispositivo móvil fue de un alumno que nació en Nuevo Necaxa, debido a esto, es el primer registro de la tabla **Ciudad**, el segundo alumno nació en Huauchinango y el tercero en Poza Rica. En el momento en que se almacenan los datos en la tabla local, el campo LlaveServidor contendrá un valor cero (0), debido a que en ese momento se desconoce su llave del servidor.

LlaveLocal	Nombre	LlaveServidor
1	Nuevo Necaxa	0
2	Huauchinango	0
3	Poza Rica	0

Tabla 1 Tabla inicial Ciudad en BD móvil
Fuente: Elaboración propia

Al enviar los datos del primer alumno, se incluye el nombre del lugar de nacimiento y el número cero como llave del servidor, cuando llegan los datos a este último equipo, se verifica esa ciudad en su tabla **Ciudad**, si ya existe, se busca la llave primaria (el valor 3) y se devuelve al dispositivo móvil para que actualice su tabla; lo mismo sucede para el segundo alumno. En caso del tercero, que no existe el nombre de la ciudad en la tabla del servidor, primero se da de alta para que se genere su llave primaria (el valor 7) y se envía al dispositivo móvil para su actualización como se muestra en la tabla 2. Este procedimiento se aplica para cada registro que requiere de un nuevo nombre de lugar de nacimiento.

LlaveLocal	Nombre	LlaveServidor
1	Nuevo Necaxa	3
2	Huauchinango	1
3	Poza Rica	7

Tabla 2 Tabla final Ciudad en BD móvil
Fuente: Elaboración propia

Por cada alumno que conteste la encuesta y sea de alguna de estas tres ciudades, se enviarán al servidor sus datos y el valor de la llave del servidor (tercer campo) para su almacenamiento directo, si necesidad de realizar la búsqueda anteriormente descrita.

Como resultado de estos movimientos de llaves, dentro del dispositivo móvil se llevará a cabo la relación entre las tablas de catálogos hacia la tabla Encuesta, por medio de las llaves primarias generadas de manera local como se muestra en la tabla 3.

NumAl	Fecha	Nombre	LugarNac
1	09/04/2018	Eugenia	1
2	09/04/2018	Claudia	2
3	09/04/2018	Gemma	3

Tabla 3 Tabla parcial Alumno en BD móvil
Fuente: Elaboración propia

Dentro del servidor, esos datos también se encontrarán almacenados, pero haciendo referencia hacia su llave primaria local como se muestra en la tabla 4.

Las tablas de catálogos, como la tabla **Ciudad** descrita anteriormente, son las que tendrán la relación entre las llaves primarias de cada dispositivo móvil y su equivalente hacia las llaves primarias del servidor; toda la información que fluya del dispositivo móvil hacia el servidor, en cualquiera de los dos sentidos, deberá pasar por la equivalencia entre llaves locales y llaves de servidor.

NumAl	Fecha	Nombre	LugarNac
…	…	…	…
27	09/04/2018	Eugenia	3
28	09/04/2018	Claudia	1
29	09/04/2018	Gemma	7

Tabla 4 Tabla parcial Alumno en BD servidor
Fuente: Elaboración propia

Otro proceso que se lleva a cabo es la actualización de las tablas de catálogos dentro de las bases de datos de los dispositivos móviles, dando esto la posibilidad a los usuarios de dichos dispositivos de contar con datos adicionales a los que van capturando para reducir el tiempo en que llenan las encuestas durante las entrevistas, al poder elegir valores en lugar de escribir cada uno de ellos.

Si algún docente desde su aplicación agrega algún alumno que haya nacido en la ciudad “Tulancingo” y actualiza la base de datos del servidor, esta ciudad junto con su llave primaria del servidor será descargada en los demás dispositivos y quedará disponible para su uso.

El motivo por el que no se agregaron todas las ciudades inicialmente en la base de datos, es porque se tendrían muchos valores que no se van a utilizar por ningún docente, lo que llevaría a contar con bases de datos muy grandes dentro de los dispositivos móviles con datos innecesarios.

ISSN: 2523-6814

Resultados obtenidos

Uno de los principales resultados que se derivan de este proyecto, es la precisión con la que se almacenan los datos dentro del sistema, teniendo la información disponible en el momento en que se genera, reduciendo al mínimo los errores en dicha información debido a que no existe una captura posterior por parte de personas diferentes a quienes aplicaron las encuestas; la veracidad de la información y su totalidad se verifican en el momento en que se obtiene por parte de los alumnos y antes de ser enviada al servidor para su concentración.

Como se muestra en el gráfico 1, el tiempo promedio para la aplicación de una encuesta utilizando papel y lápiz es de 34 minutos, utilizando dispositivos móviles se elimina el tiempo de captura posterior y en caso de que todos los datos sean nuevos (digital peor caso), se requiere en promedio de 25 minutos por encuesta.

En caso de que algunos datos ya se hayan capturado con anterioridad (digital mejor caso), el tiempo promedio es de 20 minutos por encuesta aplicada.

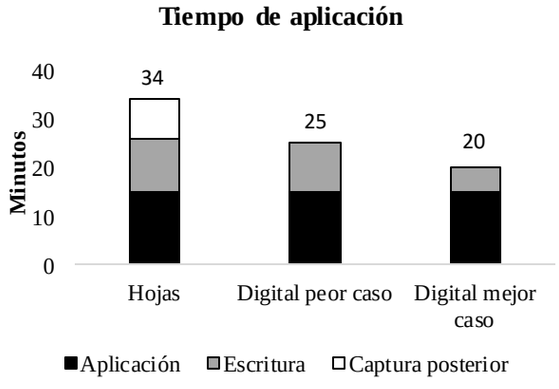

Gráfico 1 Tiempo de aplicación de encuestas
Fuente: Elaboración propia

Otro punto, no menos importante, es la reducción en el uso de papel (más de 2500 hojas ahorradas por semestre), debido a que ya no se cuenta con hojas para la realización de encuestas, se hace de manera digital por medio de un celular o Tablet de cada docente que funge como tutor de un grupo de alumnos.

Conclusiones

La integridad de la información es un tema que se debe tomar en consideración cuando se hace uso de bases de datos, principalmente si se trata de información que se encontrará distribuida en diversos equipos informáticos (computadoras y celulares entre muchos otros), que será manipulada en diferentes momentos y lugares para posteriormente ser concentrada en repositorios institucionales, desde donde se realizarán consultas para obtener información en tiempo real, con la confianza de la veracidad y totalidad de la información que se encuentra cargada en el sistema.

Además de este procedimiento, se deben tomar en consideración otros mecanismos para certificar que la información viaja de manera segura a través de todos los dispositivos de telecomunicaciones que intervienen entre el dispositivo móvil desde donde se almacena la información y el servidor de datos donde se almacenará de manera definitiva.

Referencias

[1] D-P. Pop, "Natural versus Surrogate Keys. Performance and usability", Database System Journals, 2011, pp. 55-63.

[2] P. Koutris and J. Wijsen, "Consistent query answering for primary keys", SIGMOD Records, 2016, pp. 15-22.

[3] A. H. Bahmani, M. Naghibzadeh and B. Bahmani, "Automatic databases normalization and primary key generation", 2008 Canadian Conference on Electrical and Computer Engineering, Niagara Falls, ON, 2008, pp. 000011-000016.

[4] V. Zykin, "Automatization of foreign keys construction", 2016 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, 2016, pp. 1-4.

[5] L. Zhangbing, C. Wujiang and L. Zhangbing, "A new algorithm for data consistency based on primary copy data queue control in distributed database", 2011 IEEE 3rd International Conference on Communication Software and Networks, Xi'an, 2011, pp. 207-210.

ISSN: 2523-6814

[6] A. U. Tansel, "Integrity constraints in temporal relational databases", International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004., Las Vegas, NV, USA, 2004, pp. 460-464 Vol. 2.

[7] P. Doshi and V. Raisinghani, "Review of dynamic query optimization strategies in distributed database", 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari, 2011, pp. 145-149.

[8] M. Zhang, M. Hadjieleftheriou, B. C. Ooi, C. M. Procopiuc and D. Srivastava, "Automatic discovery of attributes in relational databases", Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 2011, pp. 109-120.

[9] H. Köhler and S. Link, "Inclusion dependencies reloaded", Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015, pp. 1361-1370.

[10] M. Memari, S. Link and G. Dobbie, "SQL data profiling of foreign keys", International Conference on Conceptual Modeling, 2015, pp. 229-243.

[11] J. Motl and P. Kordík, "Foreign Key Constraint Identification and Relational Databases", ITAT 2017 Proceedings, 2017, pp. 106-111.

[12] M. Memari and S. Link, "Index Design for Partial Referential Integrity", CDMTCS Research Reports CDMTCS-482, 2015.

[13] I. Bala, S. Bishnoi, "Research Paper on Data Integrity Checking In Cloud Computing", International Journal of Enhanced Research in Management & Computer Applications, 2015, pp 55-61.

[14] E. Gelbstein, "Data integrity – Information Security's Poor Relation", ISACA JOURNAL, 2011, pp. 20-25.

[15] M. Kahng, S.B. Navathe, J.T. Stasko, D.H. Chau, "Interactive Browsing and Navigation in Relational Databases", Proceedings of the VLDB Endowment, 2016, pp. 1017-1028.

[16] H. Köhler, S. Link, X. Zhou, "Possible and Certain SQL Keys", Proceedings of the VLDB Endowment, 2015, pp. 1118-1129.

[17] P. Pocatilu, "Building Database-Powered Mobile Applications", Informatica Economică, 2012, pp 132-142.

[18] R. Singh, C. Dutta, "A Synchronization Algorithm of Mobile Database for Cloud Computing", International Journal of Application or Innovation in Engineering & Management, 2013, pp. 491-497.

[19] A.A. Imam, A. Basri, R. Ahmad, A.R. Gilal, "Data Synchronization Model for Heterogeneous Mobile Databases and Server-side Database", International Journal of Advanced Computer Science and Applications, 2018, pp. 521-531.

[20] B.S. Ramya, S.B. Koduri, M. Seetha, "A Stateful Database Synchronization Approach for Mobile Devices", International Journal of Soft Computing and Engineering, 2012, pp. 316-320.

ISSN: 2523-6814