Proceso de Explotación de Información para Analítica Académica en FaCENA-UNNE

Information Mining Process for Academic Analytics in FaCENA-UNNE

Mariano E.A.Lopez, Gladys N.Dapozo, Emanuel A. Irrazabal, Cristina L.Greiner
Facultad de Ciencias Exactas y Naturales y Agrimensura (UNNE)
Corrientes, Argentina
m_villa@hotmail.com, gndapozo@exa.unne.edu.ar, emanuelirrazabal@gmail.com, cgreiner@exa.unne.edu.ar

***Resumen*—Se presenta un proceso de explotación de información adaptado del modelo propuesto por Vanrell, el cual combina las etapas de la metodología CRISP-DM con actividades de la metodología COMPETISOFT, con el objetivo de aplicar un abordaje ingenieril para el desarrollo de una propuesta tecnológica como soporte para la implementación de Analítica Académica en la FaCENA-UNNE.**

***Abstract*—A process for information mining adapted from the model proposed by Vanrell is presented, which combines activities of CRISP-DM methodology with the ones of COMPETISOFT methodology. The objective of this work seeks for an engineering approach for the development of a technological proposal for supporting the implementation of Academic Analytics in the FaCENA-UNNE.**

***Palabras Clave*—Modelo de proceso, explotación de información, analítica académica.**

***Index Terms*—Process model, information mining, academic analytics.**

I. INTRODUCCIÓN

La fabricación de software es uno de los sectores que mantienen un crecimiento constante y representan una de las principales actividades económicas tanto para los países desarrollados como los que se encuentran en vías de desarrollo [1] [2]. Y junto con el desarrollo de software es cada vez más importante la explotación de los datos obtenidos en el marco de los sistemas de información. Así como en el desarrollo de software, en la disciplina de explotación de información se está trabajando en mejorar las metodologías de desarrollo de proyectos de explotación de información, especialmente en el área académica [3].

Existen diversas metodologías, como por ejemplo, CRISP-DM [4], el estándar de facto de la industria actualmente, aunque debido a los avances continuos de la tecnología las empresas utilizan cada vez más sus propias metodologías [5]. La segunda metodología más utilizada es SEMMA [6]. Otra de las metodologías utilizadas es P3TQ, compuesta por dos modelos: el modelo de negocio y el modelo de explotación de información [7]. Como fortalezas de estas metodologías se pueden señalar: [a] la identificación de problemas de inteligencia de negocio, [b] la caracterización parcialmente abstracta de los mismos, [c] la identificación de las relaciones entre las técnicas de explotación de información y las variables que modelan los problemas de inteligencia de negocio, y [d] el planteo parcial de los procesos a desarrollar. En tanto, entre sus debilidades se encuentran: [a] se centran fuertemente en las técnicas de explotación de información y en la tipificación de los datos, [b] no determinan cómo las variables vinculadas a los datos modelan el negocio, y [c] no identifican cuáles son los procesos de explotación de información, ni el modelo asociado, que, a partir de aplicar las técnicas al conjunto de valores de las variables, permiten obtener una solución para cada problema de inteligencia de negocio.

Existen otros trabajos, como el de [7] que buscan desarrollar un modelo de explotación de información para PYMEs. En especial, en [8] se realizó la comparación entre los tres modelos principales de explotación de información y el modelo de mejora de procesos de desarrollo software orientados a PYMES: COMPETISOFT [9] y se propone un modelo de procesos para proyectos de explotación de información con base en la fusión del modelo COMPETISOFT y el propuesto por CRISP-DM

Asimismo, la explotación de la información académica es cada vez más necesaria. La analítica académica [10] combina los datos institucionales, el análisis estadístico y los modelos predictivos permitiendo la exploración de datos para identificar informaciones nuevas y útiles para atender las expectativas y necesidades estratégicas de las organizaciones de educación superior [11].

Academic analytics o Analítica académica es un nuevo campo surgido en la educación superior como consecuencia de las prácticas de minería de datos y la utilización de herramientas de inteligencia de negocios. Puede referirse ampliamente a las prácticas de toma de decisiones basadas en datos para fines operativos a nivel de universidad, pero también puede ser aplicado a las dificultades del proceso de enseñanza y aprendizaje de los estudiantes. Por ejemplo predecir la probabilidad de abandono de los estudiantes o el tiempo de finalización de los estudios, aunque en la actualidad, el énfasis está puesto en "inteligencia procesable", información que puede ser entregada con tiempo suficiente para hacer una diferencia en el rendimiento académico [3]. Vinculado con este concepto, el término analítica de aprendizaje (del inglés Analytics Learning) se refiere a la medición, recopilación, análisis y presentación de informes de datos sobre alumnos y sus contextos, a los efectos de comprender y optimizar el aprendizaje en donde ocurre [12].

En este trabajo se presenta un proceso de explotación de información, basado en herramientas de Inteligencia de Negocio, especialmente orientado a la Analítica Académica, tomando como referencia el modelo de proceso de desarrollo

de un proyecto de explotación de información propuesto por Vanrell [8], con el objetivo de aplicar un abordaje ingenieril para dotar al proceso de desarrollo de: objetividad, sistematicidad, racionalidad, generalidad y fiabilidad. A su vez se exponen las lecciones aprendidas al momento de utilizar el proceso de explotación en el marco de una institución universitaria.

II. METODOLOGÍA

La metodología CRISP-DM [4] consta de 4 niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general (comprensión del negocio) hasta los más específicos (plan de implementación). El modelo de proceso COMPETISOFT [9] incluye una guía de buenas prácticas para el desarrollo de software en 3 categorías: Alta Dirección, Gerencia y Operación. Esta última, se subdivide en Administración de un proyecto específico, Desarrollo y Mantenimiento. Basado en estas dos metodologías, en [8] se definen dos procesos diferenciados, uno vinculado a la administración de proyectos de explotación de información y otro vinculado con el desarrollo de los mismos.

En este trabajo se aplicó el modelo de proceso de desarrollo de un proyecto de explotación de información propuesto por Vanrell, adaptado al caso de estudio particular que consiste en un proceso de explotación de información para analítica académica basado en datos del sistema de gestión de alumnos SIU Guarani de la FaCENA-UNNE, tomando en esta primera etapa los datos de la carrera Licenciatura en Sistemas de Información correspondiente al plan vigente (2010-2016).

A continuación, se describen las etapas y actividades resultantes:

1. Entendimiento del negocio

Tarea: Determinar las metas del proyecto de explotación de información

Actividades:

1. Trasladar los interrogantes del negocio a metas del proyecto de explotación de información.
2. Especificar el o los tipos de problemas de explotación de información.

Técnicas: Informe indicando una traducción de los interrogantes a metas de negocio y los tipos de problemas de explotación de información.

Salidas:

1. Meta: Diseñar e implementar un modelo multidimensional basado en un conjunto de indicadores para el seguimiento académico de los estudiantes de la carrera Licenciatura en Sistemas de Información (LSI) de la Facultad de Ciencias Exactas, Naturales y Agrimensura de la Universidad Nacional del Nordeste (FaCENA-UNNE) plan 2009.
2. Glosario. Incluye los términos propios utilizados en la gestión académica.
3. Informe del monitoreo de los procesos académicos que se llevan a cabo en la carrera LSI.
4. Requerimientos: El modelo debe contemplar las etapas de **Ingreso**, **Cursado** y **Finalización** para el seguimiento del desempeño estudiantil, asociadas a los "procesos de negocio" que se utilizan para la gestión académica, detallados en la Tabla I.

Tabla I. REQUERIMIENTOS DEL NEGOCIO

Etapa	Proceso de negocio de soporte	Requerimiento
Ingreso	Inscripción a carrera	Cantidad de aspirantes a carrera
		Cantidad de aspirantes aceptados a carrera
		Cantidad de aspirantes rechazados a carrera
		Cantidad de ingresantes a carrera
Cursado	Resultados del cursado	Cantidad de alumnos por materia
		Cantidad de alumnos recursantes por materia
		Cantidad de alumnos regulares por materia
		Cantidad de alumnos promocionados por materia
		Cantidad de alumnos libres por materia
		Cantidad de alumnos insuficientes por materia
		Cantidad de alumnos que abandonaron la materia
	Exámenes finales	Cantidad de alumnos inscriptos a examen final por materia
		Cantidad de alumnos ausentes en examen final por materia
		Cantidad de alumnos aprobados en examen final por materia
		Cantidad de alumnos desaprobados en examen final por materia
		Cantidad de alumnos que obtienen la categoría AP "Aprobó Práctico" en examen final por materia
	Gestión de equivalencias	Cantidad de equivalencias otorgadas por materia
Finalización	Habilitación para la emisión de títulos	Cantidad de egresados con título de pre-grado
		Cantidad de años requeridos por el alumno para la obtención del título de pre-grado
		Cantidad de egresados con título de grado
		Cantidad de años requeridos por el alumno para la obtención del título de grado

2. Entendimiento de los datos

Tarea: Reunir los datos iniciales.

Actividades:

1. Identificar cual es la información necesaria.
2. Chequear si toda la información necesaria se encuentra actualmente disponible.
3. Especificar el criterio de selección.
4. Seleccionar tablas o archivos de interés.
5. Seleccionar datos dentro de las tablas o archivos.
6. Definir el período de historia a ser utilizado.
7. Describir cómo se deben extraer los datos.

Técnicas: Reuniones con el personal técnico del SIU Guaraní, utilización de guías para la selección de los indicadores de rendimiento académico.

Salida:

1. Los datos serán proporcionados desde una copia local del sistema transaccional SG que posee la facultad (Informix 11.50).
2. Se contemplarán los años académicos desde 2010 hasta 2015, correspondientes al plan de estudio vigente.
3. Selección de indicadores de rendimiento académico, basados en el Sistema Integral de Información sobre

la Educación Superior en América Latina (INFOACES) [11] y el Anuario de Estadísticas Universitarias de la Secretaria de Política Universitarias de la República Argentina [12], se han definido los indicadores que se detallan en las Tablas II y III.

Fuente: Anuario de Estadísticas Universitarias de la Secretaria de Política Universitarias de la República Argentina "SPU".

Tarea: Describir los datos.
Actividades:

1. Identificar los datos y métodos de captura.
2. Acceder a las fuentes de datos.
3. Reportar tablas y sus relaciones.
4. Chequear el volumen de los datos y complejidad.
5. Chequear los tipos de atributos.
6. Chequear los rangos de valores de los atributos.
7. Analizar la correlación de atributos.
8. Entender el significado de cada atributo y valor de atributo en términos del negocio.
9. Evaluar si el significado del atributo es usado consistentemente.

Técnicas:

1. Utilización de la herramienta (Aqua Studio) del tipo DBMS (Database Management System) para el análisis de atributos a través de lenguaje SQL (Structured Query Language).
2. Diagramas Entidad Relación generados por el DBMS para documentar tablas y relaciones.
3. Entrevistas estructuradas al personal técnico del SIU Guaraní.

Salida:

1. Selección de 25 tablas del SG, en base a los requerimientos (ver TABLA I) e indicadores de rendimiento académico (ver TABLA II).
2. Análisis del dominio y tipo de dato de cada atributo de las tablas seleccionadas.
3. Generación de un diagrama de Entidad Relación por cada proceso de negocio seleccionado.
4. Análisis de la normativa académica, para comprender cómo son implementadas en el sistema SG. Por ejemplo: códigos de regularidad de los alumnos.

Tarea: Explorar los datos.
Actividades:

1. Analizar las propiedades de los atributos interesantes en detalle.
2. Considerar y evaluar información y resultados en los reportes de descripción de datos.
3. Formular hipótesis e identificar acciones.

Técnicas: Análisis de las propiedades de los atributos, subpoblaciones, hipótesis y sus transformaciones a metas.

Fuente: Sistema Integral de Información sobre la Educación Superior en América Latina "INFOACES".

Salida: Se identificó la falta del atributo "plan" en las claves primarias de las tablas *sga_alumno* y *sga_carrera* (ver Fig. 1), por ende, se requirió incorporar la tabla *sga_cambios_plan* y realizar una unión para así descartar toda acción (inscripción a materia, inscripción a examen final, otorgación de equivalencias, reinscripción, etc.) relacionada a planes no vigentes.

Tabla II. INDICADORES DE RENDIMIENTO ACADÉMICO - INFOACES

Indicador	Cálculo
Número total de estudiantes matriculados	$I_j = V_{1j} + V_{2j}$ (1) V_{1j} = Número de aspirantes en el sector de estudios j. V_{2j} = Número de estudiantes reinscriptos en el sector de estudios j.
Porcentaje de estudiantes no pertenecientes a la región en que se ubica la IES	$I_j = 100 * \left(\frac{V_{4j}}{V_{5j}}\right)$ (2) V_{4j} = Número de estudiantes de una carrera j no pertenecientes a la región. V_{5j} = Número total de estudiantes de la carrera j.
Tasa de matrícula femenina	$I_j = 100 * \left(\frac{V_{6j}}{V_{7j}}\right)$ (3) V_{6j} = Número de estudiantes de sexo femenino de una carrera j. V_{7j} = Número total de estudiantes de la carrera j.
Tasa de abandono inicial de la titulación	$I_j = 100 * \left(\frac{V_{8j}}{V_{9j}}\right)$ (4) V_{8j} = Número de estudiantes que han iniciado los cursos de carrera j en el año n y que no están matriculados en ella en el año n+1 ni en el año n+2. V_{9j} = Número de estudiantes que han iniciado los cursos de una carrera j de una IES en el año n.
Tasa de eficiencia en la graduación de la titulación	$I_j = 100 * \left(\frac{V_{12j}}{V_{13j}}\right)$ (5) V_{12j} = Número de estudiantes de una carrera j que logran finalizarla. V_{13j} = Número de estudiantes ingresados en la carrera j.

Tabla III. INDICADORES DE RENDIMIENTO ACADÉMICO - SPU

Indicador	Cálculo
Número total de aspirantes a carrera	$I_j = V_{1j} + V_{2j}$ (6) V_{1j} = Número de aspirantes aceptados por el sector de estudios de la carrera j. V_{2j} = Número de estudiantes rechazados por el sector de estudios de la carrera j.
Número total de ingresantes a carrera	$I_j = V_{1j}$ (7) V_1j= Número de aspirantes aceptados que aprueban la primera materia disciplinar del plan de estudios de la carrera j.
Número total de inscriptos a materia	$I_{jk} = V_{1jk}$ (8) V_{1jk} = Número de alumnos de la carrera j inscriptos a la materia k.
Promedio de meses requeridos por los alumnos para aprobación de las materias desde su regularidad	$I_{jk} = \frac{V_{1jk}}{v2_{jk}}$ (9) V_{1jk} = Sumatorio de los meses requeridos por los n alumnos de la carrera j en aprobar la materia k desde su regularidad. $v2_{jk}$ = Cantidad de alumnos de la carrera j que aprobaron el examen final de la materia k.
Promedio de años requeridos para el egreso	$I_{jk} = \frac{V_{1jk}}{v2_{jk}}$ (10) V_{1jk} = Sumatorio de los años requeridos por los n alumnos de la carrera j en finalizar sus estudios de nivel k. $v2_{jk}$ = Cantidad de egresados de la carrera j de nivel k.

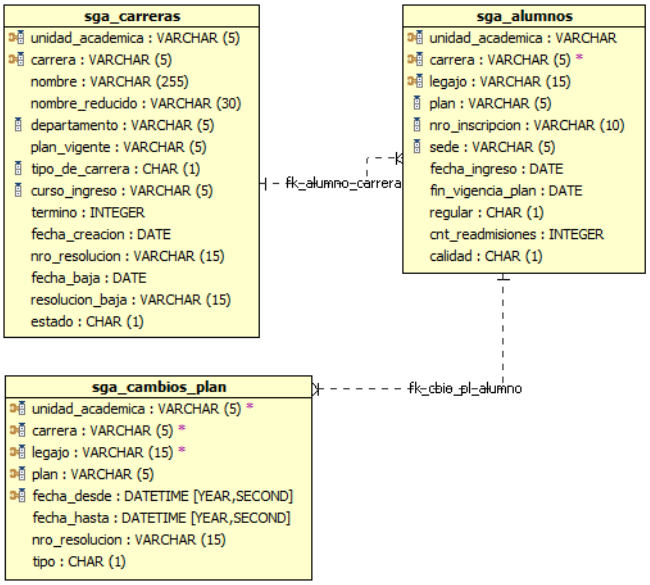

Fig. 1. Diagrama Entidad-Relación alumnos reducido

Tarea: Verificar la calidad de los datos.
Actividades:

1. Identificar valores especiales y catalogar su significado.
2. Chequear claves.
3. Identificar atributos faltantes y campos en blanco.
4. Dar significado a la falta de datos.
5. Chequear consistencias y redundancias entre las diferentes fuentes.

Técnicas: Identificación de atributos faltantes, campos en blanco, valores afectados a través de consultas SQL.

Salida: Se detectó que algunos de los atributos de la tabla *sga_personas* (ver Fig. 2), como *fecha_nacimiento, sexo, anio_egreso_sec*, admiten valores nulos.

3. Preparación de los datos

Tarea: Seleccionar los datos.
Actividades:

1. Reunir los datos producidos en la fase de preparación de datos que serán utilizados para modelar la mayor parte del trabajo de análisis del proyecto.
2. Seleccionar diferentes subconjuntos de datos (datasets) para satisfacer los requerimientos.
3. Chequear las técnicas disponibles para el muestreo de datos.

Técnicas: Consultas con SQL.
Salida:

1. Generación de consultas SQL para obtener los datasets necesarios para satisfacer los requerimientos de **Ingreso**, **Cursado** y **Finalización** de los alumnos de la carrera LSI a partir de las tablas seleccionadas del sistema SG.
2. El chequeo de los datasets se llevó a cabo tomando aleatoriamente actas de exámenes finales y cursados de algunas materias de las que se disponían de planillas de control.

Tarea: Limpiar los datos.
Actividades:

1. Corregir, remover o ignorar los *ruidos* (información indeseada, no representativa).
2. Decidir cómo considerar los valores especiales y su significado.

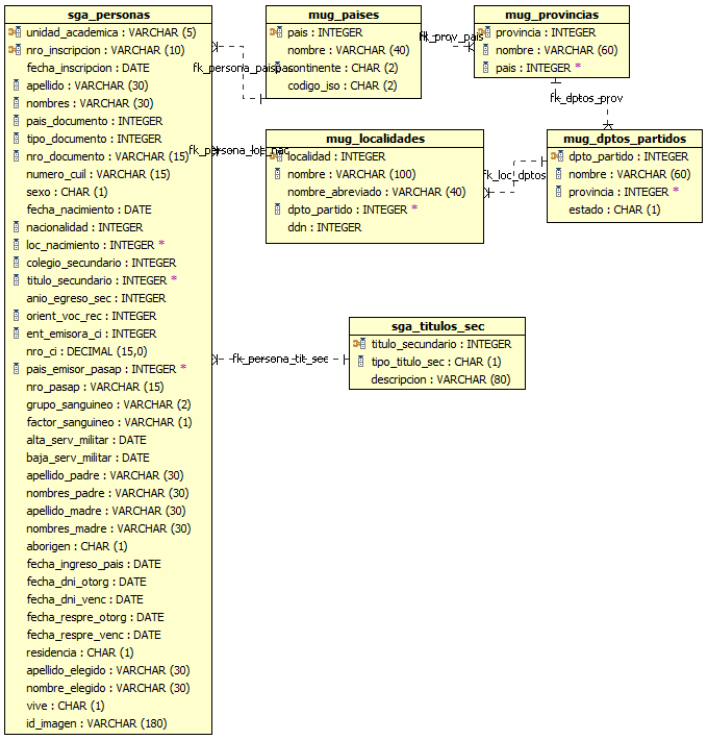

Fig. 2. Diagrama Entidad-Relación Personas Reducido

Técnicas: Corrección de valores con SQL y la herramienta de ETL Pentaho Spoon.
Salida:

1. Los campos de la tabla *sga_personas* con poca aparición de valores nulos fueron incluidos de igual manera. Aquellos con alta aparición de valores nulos, como ser el atributo de *obra_social,* fueron descartados.
2. El atributo *turno_examen*, del tipo varchar, fue normalizado utilizando expresiones regulares "RegEx". En 1 de la Fig. 3 se observa el resultado de la consulta SQL solicitando el dominio de valores del atributo *turno_examen* (28 registros). En 2 la expresión regular junto a reemplazos en la cadena de texto utilizando Pentaho, y en 3 el atributo normalizado (10 registros).
3. Se corrigieron las materias que tenían código de cuatrimestre erróneo (como consecuencia del último cambio de plan de estudio).

Tarea: Integrar los datos.
Actividades:

1. Especificar los pasos de transformaciones necesarias.
2. Realizar los pasos de transformación.
   Integrar las fuentes y almacenar los resultados.

Técnicas: Utilización de la herramienta de ETL Pentaho Spoon, consultas SQL, programación en Java script y los motores de base de datos Informix y SQL Server.
Salida:

1. Construcción de 9 sub-procesos de ETL (Carrera, Materia, Fecha, Alumno, Historia Académica, Egreso, Reinscripciones, Aspirantes y Rendimiento Académico) utilizando Spoon.
2. Cada sub-proceso incorpora las tareas:
   a. Extracción de los datos de las distintas tablas sistema del SG a través de SQL. Cabe aclarar la necesidad de incorporar el driver JDBC (Java Database Connectivity) propio del motor de base de datos "Informix" a la herramienta para poder

establecer la comunicación entre Spoon y el sistema SG.

b. Corrección y limpieza a los datos anteriormente mencionados.
c. Cambio de tipo de datos, por ejemplo, el legajo del alumno de varchar a integer, y renombramiento de los campos.
d. Aplicación de filtros, por ejemplo, descartar los registros con año académico menor a la fecha del cambio al plan vigente.
e. Denormalización de atributos para la generación de atributos de medidas (Measures). En la Fig. 4 se observa la creación de atributos de medida a partir del atributo situación del aspirante *situacion_asp*. Como resultado de la denormalización se obtuvieron 8 nuevos campos, con la posibilidad de almacenar solamente valores 0 o 1 para cuantificar el atributo del hecho.
f. Integración de los datos en un almacén de datos sobre SQL Server 2008.

A modo ilustrativo, en la Fig. 5 se observa el sub-proceso ETL para el proceso de Reinscripciones.

3. Construcción de un script de ejecución de tareas desarrollado con Spoon (ver Fig. 6), cuyo propósito es ejecutar los 9 sub-procesos de ETL mencionados anteriormente, de manera secuencial.

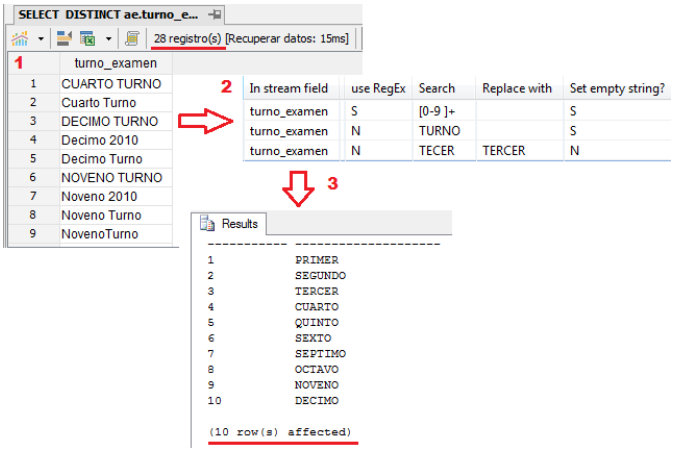

Fig. 3. Normalización del atributo turno_examen

4. Modelado

Tarea: Seleccionar la técnica de modelado.
Actividades:

1. Decidir sobre la técnica apropiada para ejercitar teniendo en mente la herramienta seleccionada.

Técnicas: Informe conteniendo los criterios de selección de la técnica a utilizar.
Salida: Se decidió hacer uso de las cualidades que brindan los almacenes de datos (Data Warehouse-DW) y herramientas de inteligencia de negocios (Business Intelligence-BI) para la aplicación de estadística descriptiva:

1. Generar un entorno virtual para el despliegue de todas las herramientas necesarias (motores de base de datos, ETL, BI), aprovechando las propiedades de la virtualización. La arquitectura técnica puede ser observada en la Fig. 7.
2. Desarrollar e implementar un almacén de datos histórico de los alumnos de la carrera LSI plan vigente de la FaCENA-UNNE, bajo la metodología de DW/BI Life Cycle (ver Fig. 8).
3. Utilizar herramientas de BI para la creación de cubos de procesamiento analítico en línea "On-Line Analytical Processing OLAP" y reportes para la explotación de información.

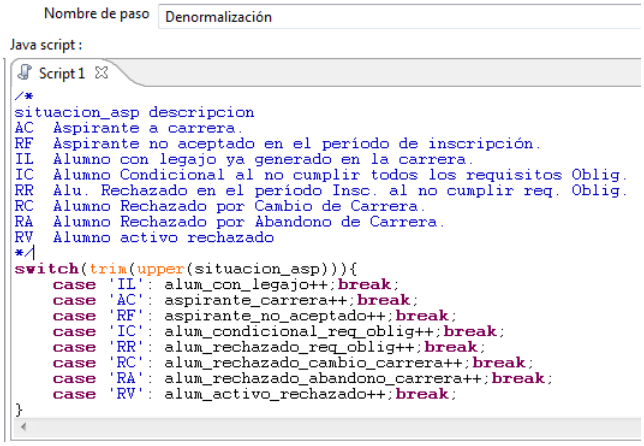

Fig. 4. Creación de atributos de Measure para Aspirantes

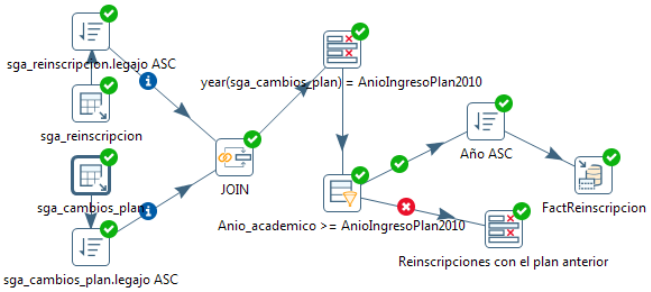

Fig. 5. Sub-Proceso ETL Reinscripciones

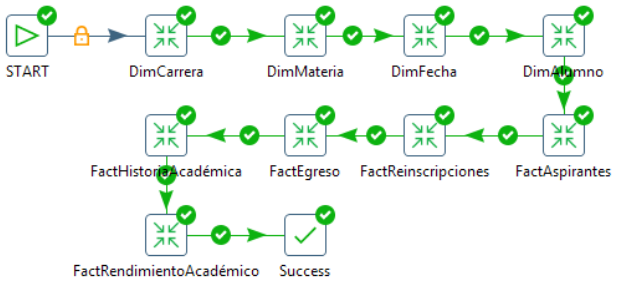

Fig. 6. Script de ejecución de tareas

Tarea: Construir el modelo.
Actividades:

1. Ejecutar la técnica seleccionada en el conjunto de datos de entrada para producir el modelo.
2. Post procesamiento de los resultados de explotación de información.

Fig. 7. Arquitectura Técnica

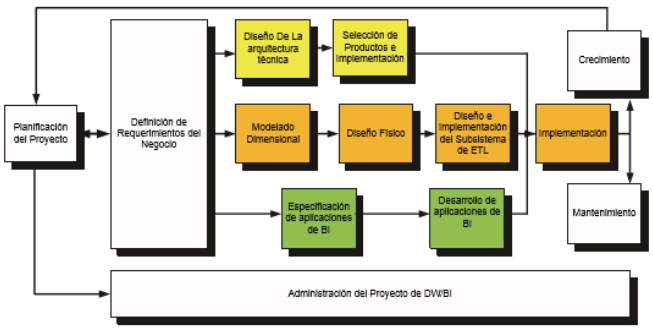

Fig. 8. Metodología DW/BI Life Cicle

Técnicas: Creación de cubos OLAP a partir de DW.
Salida:

1. A partir del modelo dimensional generado con 5 tablas de hechos: Historia Académica, Rendimiento Académico, Aspirante, Reinscripción y Egreso; y 4 tablas de dimensiones: Materia, Alumno, Fecha y Carrera (ver Fig. 9), se ejecutó el script de tareas ETL detallado en la Fig. 6, obteniendo como resultado el almacén de datos desplegado sobre SQL Server, alimentado con información histórica de los alumnos de la carrera LSI plan 2009 de FaCENA - UNNE proveniente del sistema SG.
2. Utilizando el almacén de datos como repositorio de datos se construyeron 5 cubos OLAP, uno por cada tabla de hechos, con SQL Server Analysis Services. Al ser una solución OLAP, se cuenta con las características de autonomía del usuario y alta flexibilidad en las consultas, permitiendo al usuario realizar combinaciones de **n** atributos en las filas, **m** atributos en las columnas y **z** campos de medidas en las celdas, donde **n** y **m** estarán relacionadas con la cantidad de atributos en las dimensiones involucradas y **z** con la cantidad de medidas en la tabla de hechos, como se observa en la Fig. 10. Asimismo, la utilización de operaciones OLAP como ser: DRILL, ROLL, PIVOT, SLICE y DICE.

Fig. 9. Modelo Dimensional

5. Evaluación

Tarea: Evaluar resultados**.**
Actividades:

1. Entender los resultados del proyecto de explotación de información en relación a la meta establecida, a fin de evaluar la correcta interpretación e integración de los datos y el modelado.
2. Chequear la confiabilidad de los resultados mediante:
   a. validación con un caso de estudio particular
   b. comparación con resultados obtenidos con la modalidad de consulta actualmente utilizada
3. Verificar la satisfacción de los requisitos definidos en la primera etapa.
4. Chequear el modelo contra la base de conocimiento dada para ver si la información descubierta es novedosa y útil
5. Determinar si la solución propuesta es amigable para los usuarios consumidores, aunque no cuenten con experiencia en el área de análisis de datos.
6. Analizar si existen nuevos objetivos de negocio que deban ser evaluados más tarde en el proyecto o en nuevos proyectos

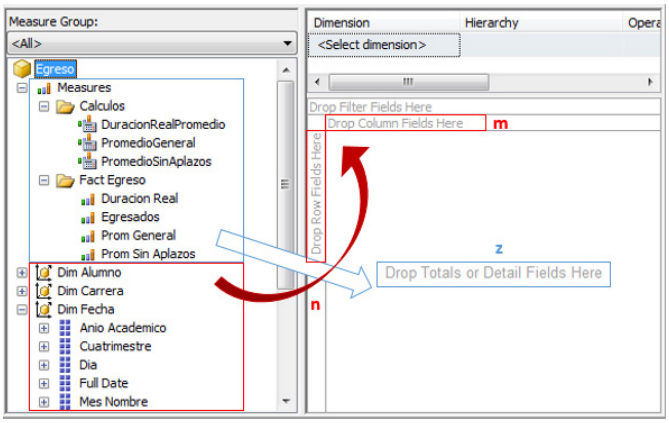

Fig. 10. Autonomía del usuario en OLAP

Técnicas: Reunión de expertos de FaCENA-UNNE (secretarios académicos, directores de carrera, personal del área de Sistemas, personal del área Acreditación) para comprobar los resultados ofrecidos por los cubos comparando con los valores obtenidos a través del sistema transaccional y los recopilados en distintas instancias de evaluación del avance de los alumnos.
Salida: Evaluación de los resultados de explotación de información respecto a las metas establecidas.

6. Entrega

Tarea: Producir un reporte final.
Actividades:

7. Identificar los reportes necesarios.
8. Elaborar reportes.
9. Llevar a cabo la presentación.

Técnicas: Informe conteniendo un detalle de los principales indicadores en las 3 categorías predefinidas: Ingreso, Cursado y Finalización
Salida: Reporte final

III. RESULTADOS

Como resultado del desarrollo del proyecto de explotación de información, se obtuvieron 5 cubos OLAP (ver Fig. 11) para la carrera Licenciatura en Sistemas de Información de la FaCENA-UNNE, contemplando las etapas Ingreso, Cursado y Finalización.

Los cubos permiten la generación de consultas del tipo ad-hoc con autonomía del usuario bajo cualquier combinación de atributos, como se detalla en la Fig. 10.

En la Fig. 12 se observa la salida resultante de la consulta de cantidad de alumnos que aprobaron el cursado (alumnos que obtuvieron la condición de Regular o Promoción) para las materias del primer año primer cuatrimestre de la carrera LSI "Algebra" y "Algoritmos y Estructuras de Datos I (AED1)", utilizando el cubo "HistoriaAcademica" mediante la operación DRILL-DOWN sobre año académico de la materia AED1 con el propósito de observar la información con mayor detalle.

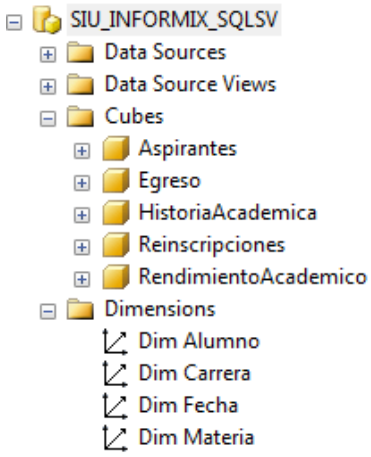

Fig. 11. Cubos OLAP

Nombre	Anio Academico	Cursado Alumnos	CursadoAprobados	CursadoAprobadosPorcentual
⊞ Algebra		2230	650	29.00%
⊟ Algoritmos y Estructuras de Datos I	2010	327	136	42.00%
	2011	370	147	40.00%
	2012	318	145	46.00%
	2013	317	126	40.00%
	2014	264	103	39.00%
	2015	272	122	45.00%
	Total	1868	779	42.00%
Total		4098	1429	35.00%

Fig. 12. Cantidad de alumnos aprobados DRILL-DOWN AED i por año académico

También se ha diseñado una serie de informes estándar de formato predefinido y acceso web a partir de los cubos, que proporcionan a los usuarios un conjunto básico de información. Algunos ejemplos de informes son:

- Análisis de obtención de títulos: Cantidad de egresados y promedio de años requeridos para el egreso, por carrera, segmentados por año académico y sexo.
- Análisis del desgranamiento de alumnos: Cantidad de alumnos aspirantes, reinscriptos, ingresantes, egresados, como así también la matrícula y la tasa de eficiencia en la graduación de la titulación, por carrera y por año académico.
- Análisis del desempeño de los alumnos por año en el plan de estudios: Detalla la cantidad de alumnos inscriptos al cursado, alumnos aprobados y alumnos desaprobados al finalizar el cursado, por carrera y por año en el plan de estudio, periodo lectivo, materia y año académico.

Los ejemplos se muestran en distintos navegadores como demostración de su funcionamiento en cualquiera de ellos. En la Fig. 13 se observa el informe "Análisis de obtención de títulos" desde el navegador web Chrome. En el mismo se visualiza que la cantidad de egresados al 29/12/2015 para la carrera LSI plan 2009 de FaCENA–UNNE era:

- Título intermedio "Analista Programador Universitario": 26 (21 de ellos de sexo masculino y 5 sexo femenino) con promedio de 5 años para la obtención del título.
- Título de grado "Licenciado en Sistemas de Información": 1 (sexo masculino) con promedio de 6 años para la obtención del título.

En la Fig. 14 se observa el informe "Análisis del desgranamiento de alumnos" desde el navegador web Edge. En el mismo se visualiza que la carrera LSI plan 2009 de FaCENA–UNNE cuenta con un promedio por año de: 287 aspirantes, 431 reinscriptos, 718 matricula, 316 ingresantes, 5 egresados (incluye ambas titulaciones) y 1.18% de tasa de eficiencia en la graduación de la titulación.

En la Fig. 15 se observa el informe "Análisis del desempeño de los alumnos por año en el plan de estudios" desde el navegador web FireFox.

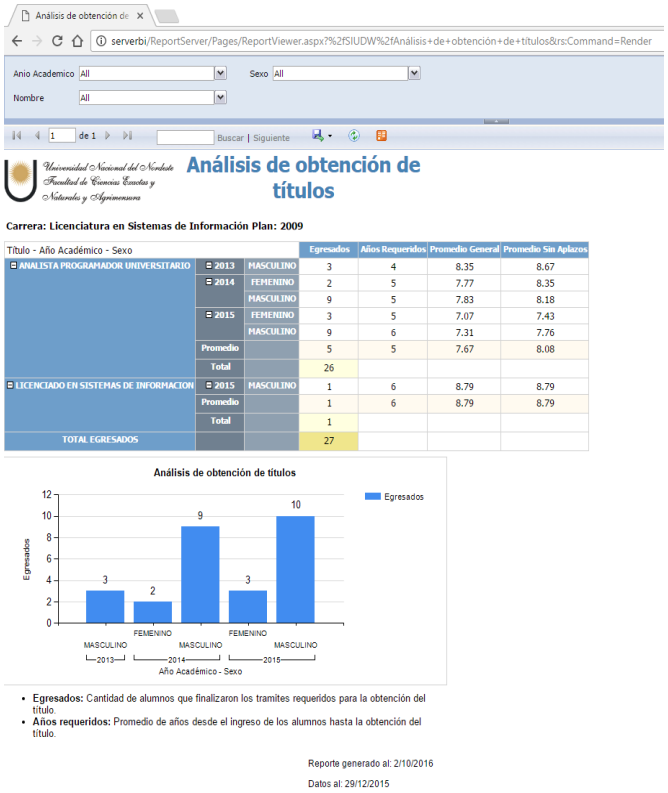

Fig. 13. Análisis de obtención de títulos

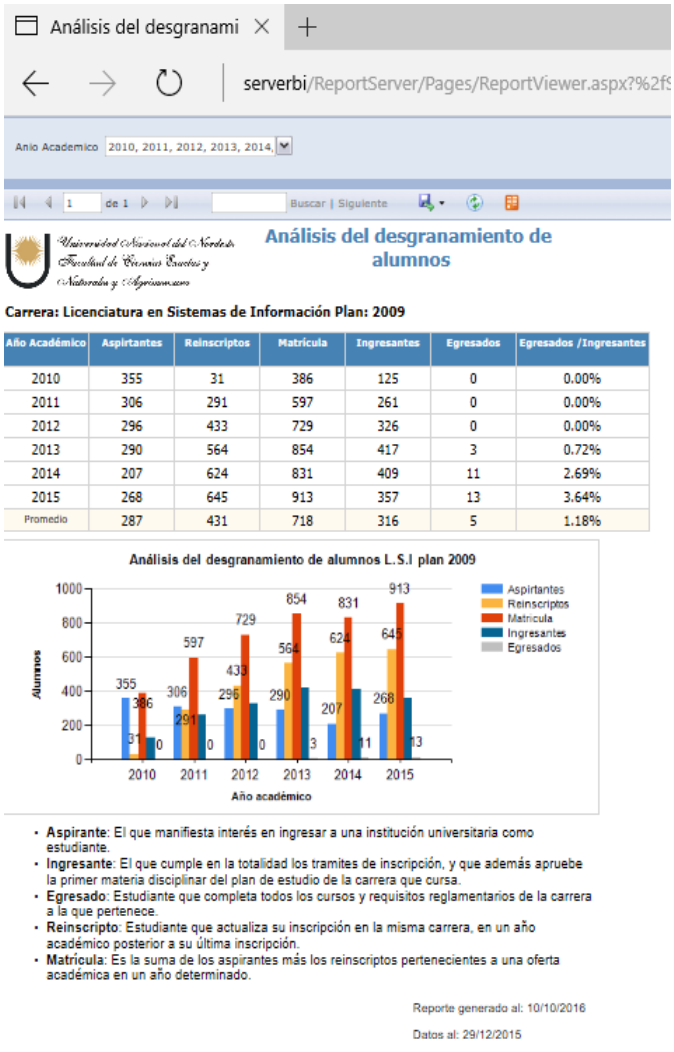

Fig. 14. Análisis del desgranamiento de alumnos

En el mismo se visualiza que la carrera LSI plan 2009 de FaCENA–UNNE cuenta con un promedio de alumnos por materia de:

- 1° año: 276 alumnos con 36.47% de aprobados
- 2° año: 100 alumnos con 57.32% de aprobados
- 3° año: 55 alumnos con 66.23% de aprobados
- 4° año: 28 alumnos con 74.75% de aprobados
- 5° año: 9 alumnos con 76.11% de aprobados

IV. LECCIONES APRENDIDAS

Como lecciones aprendidas, considerando la validación del proceso de explotación aplicado al caso de estudio particular, la gestión académica de los alumnos de la Licenciatura en Sistemas de Información de la UNNE, cabe destacar:

La etapa 1 que se enfoca en el entendimiento del negocio y la etapa 2 que abarca el entendimiento de los datos (sistemas, tablas, relaciones, atributos, dominio, claves primarias, redundancia, normalización para corrección y denormalización para atributos de medida), son las que demandaron mayor esfuerzo. Su importancia es tal que condiciona el modelo y requiere de varias iteraciones hasta lograr definir los requerimientos del proyecto.

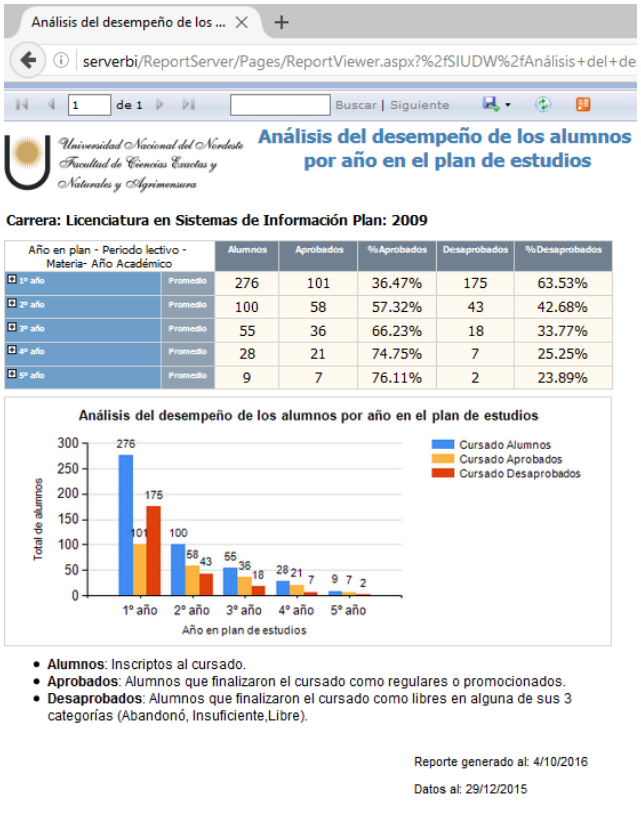

Fig. 15. Análisis del desempeño de los alumnos por año en el plan de estudios

En la etapa 3, en la tarea de integración de datos, la definición de sub-sistemas de extracción, transformación y carga fueron las actividades más importantes de este trabajo. Requirió comprender los procesos de negocio de la institución y de cómo la información de estos procesos se registra en el sistema de gestión de alumnos. Se realizó un análisis minucioso de los datos verificando: tipo, dominio, veracidad y exactitud de los mismos, y la coherencia de los procesos de negocio. En función del resultado de este análisis se definieron mecanismos de limpieza y corrección de los datos. Cabe señalar que un mal análisis o diseño en un sub-sistema de ETL puede llevar al fracaso del proyecto dado que cualquier información proveniente del mismo será errónea. Asimismo, la disponibilidad de una herramienta que permita la verificación de la calidad de los datos y su manipulación y adecuación, es clave para este proceso. Utilizar Pentaho Spoon fue de gran ayuda para la construcción del sistema, dado que brinda una interfaz drag & drop que permite visualizar los datos y realizar cambios de manera fácil y rápida.

La etapa 4 de Modelado, referida en particular al diseño e implementación del almacén de datos, tuvo como principal dificultad el cambio de enfoque del esquema OLTP a un esquema multidimensional destinado al análisis. Este último exige adquirir otro nivel de abstracción saliendo de la arraigada visión del modelo entidad-relación. Costó llevar a la práctica los conceptos de dimensión, hecho, medida.

Asimismo, el diseño e implementación de aplicaciones BI permitió la explotación del almacén de datos a través de la construcción de cubos OLAP. Las soluciones OLAP permiten pasar de datos en reposo a la explotación de información, brindando la posibilidad de combinaciones n-dimensionales, operadores especiales de análisis de datos como ser generalización, especialización y sesgo, entre otras. Cabe destacar la potencialidad del paquete SQL Server, el cual permite acceder a los cubos de forma local o remota desde Management Studio o Excel, ofreciendo interfaces drag & drop para visualización de los datos y generación de consultas, siendo este un punto clave para la autonomía de los usuarios.

Las soluciones de BI son hechas a medida, cada negocio posee su propia lógica, requiere concentrarse casi el 90% del tiempo que demanda el proyecto en los sub-procesos de ETL. Pero al finalizar se contará con una solución que permita realizar la explotación de información con pocos clics, sin la necesidad de escribir consultas complejas y sin necesidad de que los usuarios consumidores cuenten con experiencia en el área de análisis de datos.

V. CONCLUSIONES

Se obtuvo un proceso detallado para el desarrollo de un proyecto de explotación de información orientado a brindar información para realizar analítica académica que puede ser adoptado por cualquier institución de educación superior en nuestro país dado que contempla las problemáticas comunes en las universidades argentinas.

La información académica de los alumnos de la Licenciatura en Sistemas de Información, disponible en el sistema SIU Guaraní, se utilizó para validar el proceso. La solución tecnológica, formada por un DW y cubos OLAP resulta apropiada para brindar información para un conjunto de indicadores que brindan una visión amplia del avance de los estudiantes en la mencionada carrera, considerando las etapas de Ingreso, Cursado y Finalización, permitiendo monitorear y detectar dificultades y anticipar acciones para mejorar el avance de los alumnos.

Además, se contempla que la información disponible permitirá extender la analítica para obtener información acerca de otros fenómenos que afectan la terminalidad de los estudios, como el desgranamiento, abandono y prolongación de la duración real de la carrera, así como otras cuestiones vinculadas más estrechamente con el proceso de enseñanza y aprendizaje. Por ejemplo, la relación entre distintas metodologías de enseñanza y los resultados académicos que se obtienen.

A futuro, se contempla ampliar el almacén de datos con la información de todas las carreras de la Facultad de Ciencias Exactas y Naturales y Agrimensura de la UNNE, proveyendo

un mecanismo ágil para el seguimiento académico de las mismas.

RECONOCIMIENTOS

El presente trabajo se ha realizado en el marco del proyecto F010-2013 "Métodos y herramientas para la calidad del software", financiado por la Secretaría de Ciencia y Técnica de la Universidad Nacional del Nordeste (SECYT-UNNE) y como parte del plan de beca de pregrado del alumno Mariano Lopez, otorgada por la SECYT-UNNE para el periodo 2016-2017.

REFERENCIAS

[1] M. Piattini y J. Garzás, Fábricas de software: experiencias, tecnologías y organización, Paracuellos de Jarama: RA-MA, 2010.

[2] F. Mundial, «The Global Information Technology Report,» 2016.

[3] P. Baepler y C. J. Murdoch , «Academic Analytics and Data Mining in Higher Education,» International Journal for the Scholarship of Teaching and Learning, vol. 4, nº 2, 2010.

[4] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz y R. Wirth, The CRISP-DM process model, 1999, p. 310.

[5] G. Piatetsky y K. Polls, «KDNuggets,» 2014. [En línea]. Available: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[6] SAS Institute Inc, «From Data to Business Advantage: Data Mining, SEMMA Methodology and the SAS System (White Paper),» 1997.

[7] R. García Martínez, R. Lelli, H. Merlino, L. Cornachia, D. Rodriguez, P. Pytel y H. Arboleya, «Ingeniería de Proyectos de Explotación de Información para PYMES,» Workshop de Investigadores en Ciencias de la Computación, vol. XIII, 2011.

[8] J. Vanrell, Un Modelo de Procesos para Proyectos de Explotación de Información, Tesis de Maestría en Ingeniería en Sistemas de Información, Buenos Aires, 2012.

[9] H. Oktaba, M. Piattini, F. Pino, M. Orozco y C. Alquicira, COMPETISOFT: Mejora de Procesos Software para Pequeñas y Medianas Empresas y Proyectos, Madrid: Alfaomega Ra-Ma, 2009.

[10] J. Campbell y D. Oblinger, Academic analytics. Educause Quarterly, 2007, pp. 1-20.

[11] A. Rodriguez Almeida y S. da Silva Camargo, «Academic Analytics: Aplicando técnicas de Business Intelligence sobre datos de performance académica en enseñanza superior.,» Interfaces Científicas - Exatas e Tecnológicas. ISSN ELETRÔNICO - 2359-4942, vol. 1, nº 2, pp. 35-46, 2015.

[12] G. Siemens y D. Gasevic, «Learning and Knowledge Analytics,» Educational Technology & Society, nº 15, 2012.

[13] Sistema Integral de Información sobre la Educación Superior en América Latina "INFOACES", Sistema Basico de Indicadores para la Educacion Superior de America Latina, Valencia: Universidad Politécnica de Valencia, 2012.

[14] Secretaría de Políticas Universitaria, Anuario de Estadísticas Universitarias - Argentina, Buenos Aires, 2013.

M. Lopez. Licenciado en Sistemas de Información en la Universidad Nacional del Nordeste (UNNE). Becario de investigación de pregrado (SECYT-UNNE).

G. Dapozo. Docente-investigadora UNNE. Magister en Informática y Computación (UNNE). Directora del proyecto F010-2013 "Métodos y herramientas para la calidad del software" (SECYT-UNNE).

E. Irrazabal. Docente-investigador UNNE. Doctor en Sistemas de Información (Universidad Rey Juan Carlos-España). Integrante del proyecto F010-2013 "Métodos y herramientas para la calidad del software" (SECYT-UNNE).

C. Greiner. Docente-investigadora UNNE Magister en Informática y Computación (UNNE). Codirectora del proyecto F010-2013 "Métodos y herramientas para la calidad del software" (SECYT-UNNE).