Propuesta de Artefactos para el Subproceso de Gestión del Modelo de Proceso de Proyectos de Explotación de Información (G-MoProPEI)

Sebastian Martins[1,2], Patricia Pesado[3], Ramón García-Martínez[2]

[1] Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata
[2] Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información. Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús
[3] Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC
smartins089@gmail.com, ppesado@lidi.info.unlp.edu.ar, rgm1960@yahoo.com

Resumen. Se ha señalado que los modelos de procesos existentes presentan carencias al momento de abordar un proyecto de explotación de información. Como respuesta a dicha necesidad, se definió un modelo de proceso (MoProPEI) el cual permita sistematizar el desarrollo de un proyecto, considerando aquellos aspectos de gestión y control requeridos para esta disciplina. En este trabajo se presenta una visión detallada del subproceso de gestión con detalle del conjunto de formalismos utilizados para representar artefactos de entradas y salidas para cada una de las actividades. Se ilustran en una prueba de concepto perteneciente al área de educación.

Palabras Clave. Ingeniería de Explotación de Información. Modelo de proceso. Gestión de Proyectos. Artefactos de Gestión. Minería de Datos.

1 Introducción

El termino minería de datos está fuertemente ligado al concepto de base de datos y se remonta a la definición de algoritmos de búsqueda de patrones de conocimiento en grandes bases de datos [1]. Sin embargo, hoy existen líneas de investigación en campos tales como: minería de textos [2], minería de imágenes [3], minería de patrones en flujos de información [4], minería en la web [5], entre otras. En este contexto, en la línea de investigación en la que se enmarca este artículo, se conviene utilizar el término "explotación de información" como referencia genérica a cualquiera de los tipos de minería precitados. Con base en que la Ingeniería de Software ha sido definida en el SWEBOK [6] como: "la aplicación de un enfoque sistemático, disciplinado y cuantificable al desarrollo, operación y mantenimiento de software, y el estudio de estos enfoques, es decir, la aplicación de la ingeniería al software"; se conviene en definir a la Ingeniería de Explotación de Información [7] como la aplicación de un enfoque sistemático, disciplinado y cuantificable al desarrollo de proyectos de explotación de información, y el estudio de este enfoque, es decir, la aplicación de la ingeniería a la explotación de información.

Se ha señalado [8] que los modelos de procesos existentes presentan carencias al momento de abordar un proyecto de explotación de información. Por ello, se definió un modelo de proceso (MoProPEI) que mediante la visión de procesos (fases, tareas, técnicas de representación y procedimientos de ejecución de la tarea) permita sistematizar del desarrollo de un proyecto, considerando aquellos aspectos de gestión y control requeridos en todo proceso productivo.

En [9], se ha realizado una comparación con los principales modelos de procesos utilizados en la industria, centrándose en los aspectos de gestión, resaltando las ventajas introducidas por el proceso propuesto.

En este trabajo se presenta una descripción de las problemáticas abordadas (sección 2), se presenta una visión detallada del subproceso de gestión, presentando el conjunto de formalismos utilizados junto con las dependencias funcionales de cada actividad (sección 3), las cuales se ilustran en una prueba de concepto perteneciente al área de educación (sección 4). En la sección 5, se presentan las conclusiones y futuras líneas de trabajo.

2 Descripción del Problema

Los modelos de proceso actuales se centran en las tareas de desarrollo (aquellas vinculadas con el set de datos y los algoritmos) dejando de lado otros aspectos relevantes en el desarrollo de un proyecto, como la planificación, seguimiento y control de las actividades, la visión transversal del mismo, entre otros.

De forma complementaria, los modelos existentes brindan una visión global del proceso, pero carecen de una descripción detallada que permita comprender las vinculaciones entre las actividades y las posibles herramientas/técnicas, complejizando el desarrollo del proceso, realizándolo de manera desorganizada dificultando la trazabilidad, evaluación y reutilización de los proyectos.

Hace ya más de 10 años, se ha señalado la alta tasa de fracasos en proyectos de explotación de información (cifra superior al 60%) [10] y distintas dificultades en el desarrollo de los mismos [11-13], siendo CRISP-DM la principal metodología empleada para el desarrollo de proyectos desde dicha fecha hasta la actualidad [14]. Esto puede deberse a la carencia de artefactos propuestos por los modelos de procesos existentes, debiendo los ingenieros de explotación de información realizar de manera artesanal o poco sistematizada el desarrollo de los proyectos, contando a su vez, con un modelo guía el cual está centralizado en las actividades de desarrollo, omitiendo aspectos vinculados con la gestión del proyecto, en una disciplina cuyos proyectos han ido incrementando en magnitud y complejidad, así como en la conformación de sus grupos interdisciplinares de trabajo.

3 Subproceso de Gestión y Control en MoProPEI

El proceso propuesto está conformado por 2 subprocesos: Desarrollo, vinculado con las actividades de entendimiento y preparación de los datos y su posterior extracción de conocimiento; y Gestión, orientado al control y la administración del proyecto.

El subproceso de Desarrollo se encarga de todas las tareas asociadas a la obtención de requisitos, la comprensión del negocio y de los problemas del mismo, la identificación de recursos relevantes para el desarrollo del proyecto, particularmente de las fuentes de datos, el análisis, comprensión y preparación de los datos existentes en la organización, la identificación y selección de los procesos, técnicas y herramientas a utilizar, la implementación y su posterior evaluación y producción de los resultados obtenidos, cuyo orden se define con el objetivo de reducir la cantidad de iteraciones entre las etapas del subproceso, y favorecer la ejecución del mismo, logrando una mejor articulación entre las actividades involucradas.

El subproceso de gestión se concibe de forma transversal a las actividades de desarrollo, cuya ejecución de sus tareas no es de forma lineal, sino que se realizan en base al progreso del proyecto. El mismo abarca la administración del proyecto, comprender la situación del cliente, identificar, planificar y controlar los recursos, identificar el modelo de ciclo de vida, controlar la ejecución de las actividades, realizar las mediciones, definir la viabilidad del proyecto, y formalizar el cierre del mismo. El subproceso de Gestión, está conformado por cuatro fases, cada una de las cuales se componen de distintas tareas generales, que identifican un conjunto de actividades con un objetivo específico dentro del proyecto.

Este trabajo presenta las actividades que componen cada fase, indicando sus documentos de entrada y salida, pudiendo identificar las dependencias que ellos poseen.

La fase de **Iniciación** (figura 1), está compuesta por cuatro actividades: *definición de la comunicación*, tiene como elemento de entrada el discurso del cliente, el discurso del líder del proyecto, y las políticas de las organizaciones intervinientes definiendo a partir de ellos el protocolo del proyecto; *exploración de conceptos iniciales*, cuyos inputs son el output de la actividad previa y el discurso del cliente, produciendo como resultados parciales el reporte de recursos del cliente, el plan de adquisición de conocimiento y el conocimiento adquirido; *evaluación de la situación*, cuyos elementos de entrada están conformados por todos los elementos de salida de la actividad previa, y produce los reportes de recursos internos, posibilidad de tercerización, viabilidad, y se identifican proyectos similares que sirvan de guía para el desarrollo del proyecto y riesgos y contingencias del mismo; y *definición del ciclo de vida* que establece el patrón estructural mediante el cual el proyecto será ejecutado (ciclo de vida) a partir de las características del proyecto definidas en la actividad previa.

La fase **Planificación** (figura 2), la cual se compone de tres actividades generales: *planificación de las actividades*, cuyos elementos de entrada son los proyectos guías identificados, el ciclo de vida y el documento de requisitos, definido en el proceso de desarrollo y se acuerda el mapa y calendario de actividades; *planificación de recursos*, que a partir de los proyectos guías, el calendario de actividades y el documento de requisitos, define el plan de tercerización y el reporte de recursos requeridos; y *estimaciones y responsabilidades*, en donde a partir de los elementos de

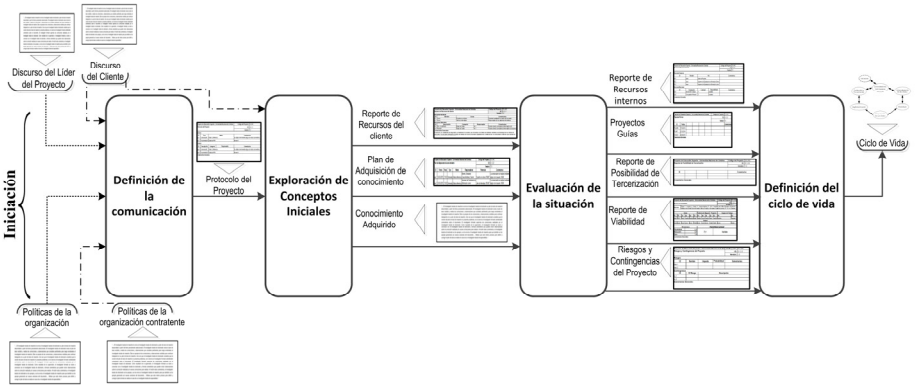

Fig. 1. Fase Iniciación

salida de la actividad previa, junto con los proyectos guías, el mapa de actividades y el documento de requisitos, se realiza la estimación de costo y se definen los alcances del proyecto y las obligaciones que las partes intervinientes acuerdan (contrato del proyecto).

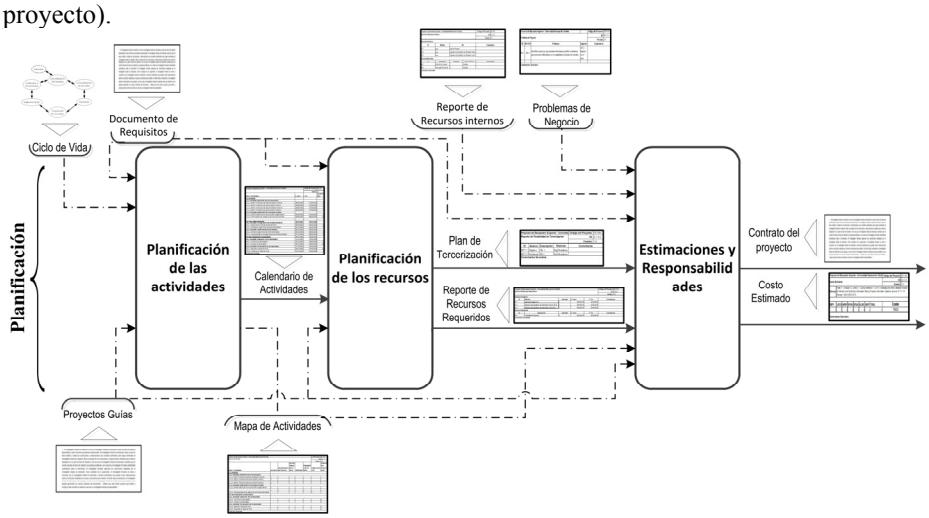

Fig. 2. Subproceso Gestión: Fase Planificación

La fase **Soporte** (figura 3), está conformada por 3 actividades: *gestión del ciclo de vida*, que a partir del calendario de actividades y el modelo de ciclo de vida escogido, se determinan los alcances del ciclo de vida y los elementos pendientes de realizar para próximas iteraciones del proyecto (en caso que hubiese) definidos en los reportes formales de inicio y fin de ciclo, los cuales son utilizados de manera global en el desarrollo del proyecto; *gestión del desarrollo*, en donde se define las responsabilidades de los recursos en el proyecto, teniendo como elementos de ingreso el calendario de actividades, el plan de tercerización (en caso que hubiese), los riesgos y contingencias del proyecto y del problema de negocio y los reportes de recursos requeridos y del progreso de las actividades, generando como productos de salida los

contratos, el reporte de asignación de responsabilidades y la comunicación del progreso del proyecto, utilizados de manera global en el desarrollo del proyecto; y *gestión de la configuración*, que a partir de los avances en el proyecto, se producen ajustes al documento del proyecto y se establece el manejo de versionado del mismo.

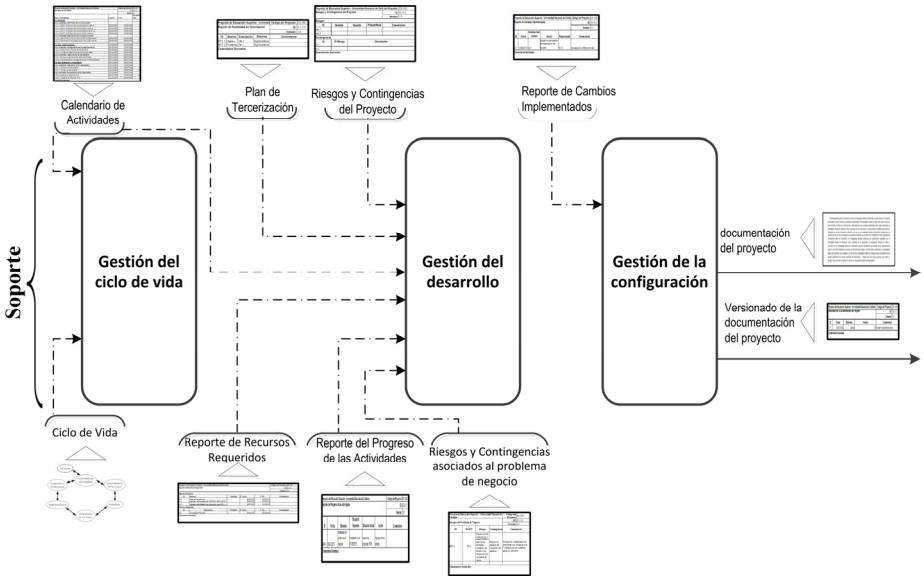

Fig. 3. Subproceso Gestión: Fase Soporte

La fase **Control y Calidad** (figura 4), integrada por 4 actividades: *control de los recursos*, cuyos elementos de entrada son los contratos de recursos y tercerización, el reporte de recursos requeridos y los problemas y objetivos de negocio, generando los controles correspondientes al incorporación de recursos y tercerización de tareas; *mediciones del proyecto*, en donde a partir del listado de métricas y los costos de las actividades del proyecto, se calculan y controlan las variables de interés para el proyecto, generando los reporte de métricas y de costos; *control de las actividades*, en donde se realizan los controles generales del progreso de las actividades, riesgos y calidad del proyecto, a partir de las salidas de la actividad previa junto con los costos y esfuerzo estimado, los riesgos y contingencias identificados, el calendario de actividades y el reporte de responsabilidades del personal, brindando un control detallado del progreso y posibles desvíos en el desarrollo del plan de proyecto; y *gestión del cambio*, donde se realiza la evaluación, implementación y control de los cambios solicitados a través del desarrollo del proyecto dejando constancia de los mismos, sus elementos de entrada son el documento de requisitos, los problemas de negocio y sus riegos asociados y las solicitudes de cambio, produciendo los reportes de cambios implementados y de control de la integración del cambio.

La fase de **entrega** (figura 5), compuesta por 2 actividades: *formalización externa del cierre del proyecto*, que se realiza una verificación y validación formal con el objetivo de constatar que las necesidades del cliente fueron satisfechas y que las obligaciones tercerizadas fueron correctamente cumplimentadas, generado a partir del discurso del

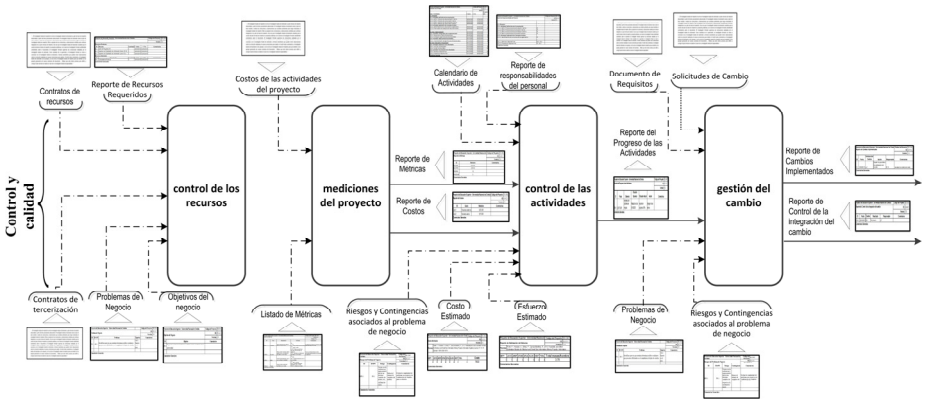

Fig. 4. Subproceso Gestión: Fase Control y Calidad

cliente, el contrato del proyecto, los contratos de recursos y tercerización y la planilla de criterios de éxito del problema de negocio, el documento de aceptación y el reporte de conclusión de contrataciones; *formalización interna del cierre del proyecto*, en donde se generan una serie de reportes para el control y mejora propia del equipo de trabajo, según la experiencia adquirida a lo largo del proyecto. Sus elementos de entrada son los generados en la actividad previa junto con los reportes de calidad del proyecto, riesgos acontecidos (en caso que hubiese), calidad del proyecto, métricas y costos, generando el reporte de sugerencia de mejoras y el documento interno de desarrollo del proyecto.

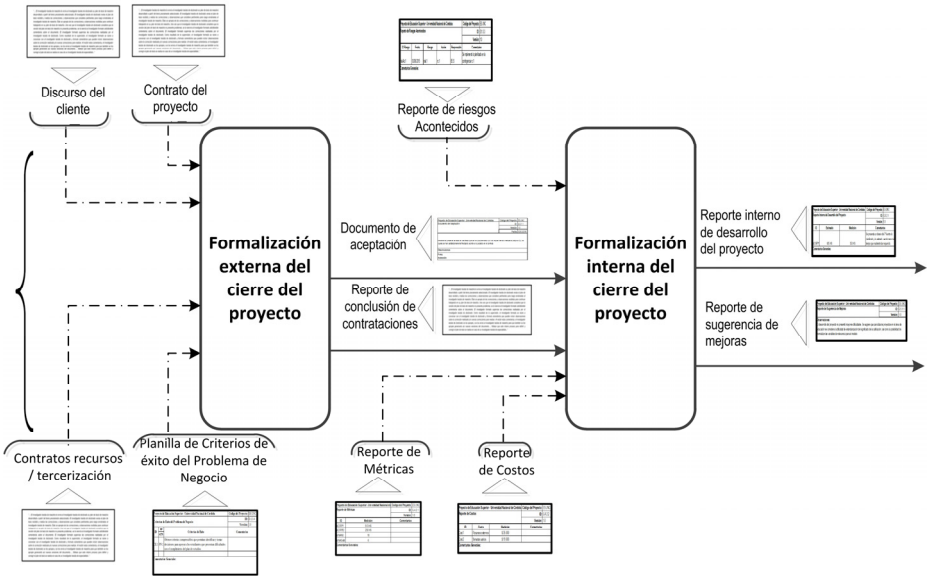

Fig. 5. Subproceso Gestión: Fase Entrega

4 Prueba de Concepto

En esta sección se presenta una prueba de concepto en la cual se introducen algunos formalismos de mayor relevancia obtenidos a partir de la aplicación de MoProPEI en el proyecto [15]. El mismo se encuentra enmarcado en el ámbito de la educación, y sus requerimientos generales obtenidos a partir del cliente pueden ser resumidos en:

> *"El propósito de esta investigación es contribuir a facilitar la apropiación del conocimiento en Educación Superior en contextos de masividad.*
> *Proveer de información para un adecuado diseño de las políticas públicas en Educación Superior, despierta el interés para contribuir con una mejor apropiación del conocimiento por parte de la sociedad. En esta dirección, una dimensión relevante es la asociada a las características del estudiante, principal actor de este complejo escenario.*
> *Se espera que esta contribución resulte novedosa, tanto en las interpretaciones sobre la información del estudiante que emerjan, como en la construcción del diseño metodológico aplicable en las muy diversas prácticas de este escenario.*
> *La Universidad Nacional de Córdoba dispone de 2 sistemas fundamentalmente en los cuales se resguarda la información del estudiante. El Sistema de Información Universitaria SIU_GUARANI, de gestión académica, el cual contiene información académica y socioeconómica de los estudiantes, y las aulas virtuales desarrolladas sobre la plataforma MOODLE las cuales contienen información de las instancias de acreditación.*
> *Se espera lograr un mejor conocimiento de las características del estudiante, protagonista principal de este escenario, permitiendo proporcionar contribuciones novedosas y valiosas que favorezcan la toma de decisiones en aspectos vinculados con la gestión de la Educación Superior en contextos de masividad."*

De los requerimientos previamente educidos, se identifican objetivos y problemas de negocio a partir de los cuales se orienta el desarrollo de la solución. Entre los elementos generados durante el desarrollo de proyecto, se ilustran de la fase de iniciación: el *Reporte de Recursos del Cliente* (tabla 1), generado en la actividad Exploración de Conceptos Iniciales, en el cual se identifican los recursos del cliente de interés para el desarrollo del proyecto; los *Riesgos y Contingencias del Proyecto* (tabla 2), producido en la actividad evaluación de la situación, identificando aquellos eventos que pueden ocurrir afectando el desarrollo del proyecto y las posibles acciones a realizar; y el *Reporte de Viabilidad* (tabla 3), generado en la actividad evaluación de la situación, determinando la posibilidad de realización con éxito del proyecto (haciendo uso del modelo definido en [16]); y de la fase de planificación: el *Listado de Métricas* (tabla 4), basadas en [17], generado en la actividad de planificación de las actividades, identificando aquellas variables de interés a medir a lo largo del proyecto.

5 Conclusiones

Los autores vienen desarrollando en los últimos años una línea de trabajo en el área de control y gestión para proyectos de explotación de información. Se han propuesto dos subprocesos: de desarrollo que abarca aquellas actividades asociadas con el entendimiento del negocio y de los datos, su preparación, las técnicas de explotación de información y el reporte de los resultados, y otro Gestión enfocado en el control y la administración del proyecto.

Tabla 1. Reporte de Recursos del cliente

Proyecto de Educación Superior - Universidad Nacional de Córdoba				**Código del Proyecto:**	ES.UNC
Reporte de Recursos del cliente				**ID:**	G.1.2.1
				Versión:	1.0
Recursos Humanos					
ID	**Nombre**	**Cargo**		**Comentarios**	
rh.1	xxxx	Experto en el área "XXXX"		Docente investigador a cargo del proyecto	
rh.1	xxxx	Jefe de Informática "XXXX"		Responsable de los sistemas informáticos	
Recursos Materiales					
ID	**Descripción**	**Categoría**	**Responsable**	**Comentarios**	
rm.1	SIU_GUARANI	Base de Datos	rh.1	Fuente donde se registra la información del desempeño del alumno en el transcurso de su carrera	
rm.2	Moodle	Base de Datos	rh.1	Fuente donde se registra la información del alumno respecto de la cursada de una materia especifica	
Comentarios Generales: Se acordó, por cuestiones de seguridad y privacidad en los datos, que el acceso a los datos se realizaría mediante una descarga de los mismos en texto plano, la cual sería entregada al equipo de trabajo con todos los campos registrados en el sistema que no contuviesen contenido que violase las normas de privacidad de los estudiantes.					

NOTA: Se resalta que por cuestiones de privacidad, el nombre del personal cliente fue omitido.

Tabla 2. Reporte de Riesgos y Contingencias del Proyecto

Proyecto de Educación Superior - Universidad Nacional de Córdoba				**Código del Proyecto:**	ES.UNC
Riesgos y Contingencias del Proyecto				**ID:**	G.1.3.3
				Versión:	1.0
Riesgos					
ID	**Nombre**	**Impacto**	**Probabilidad**	**Comentarios**	
risk.1	Demora en acceso a los datos	3	70%	Por cuestiones de políticas de aseguramiento de la privacidad de los datos personales de los estudiantes	
Contingencias					
ID	**ID Riesgo**	**Descripción**			
rc.1	risk.1	Se ajustarán los plazos del proyecto acorde a las demoras			
Comentarios Generales:					

Tabla 3. Reporte de Viabilidad

Proyecto de Educación Superior - Universidad Nacional de Córdoba													**Código del Proyecto:**	ES.UNC
Reporte de Viabilidad													**ID:**	I.1.3.7
													Versión:	1.0
Técnica	Pytel, P., Hossian, A., Britos, P., García-Martínez, R. 2015. Feasibility and Effort Estimation Models for Medium and Small Size Information Mining Projects. Information Systems Journal, 47: 01-14. Elsevier. ISSN 0306-4379.													
Datos							**Problema de Negocio**			**Proyecto**		**Equipo de Trabajo**		
P1	**P2**	**A1**	**A2**	**A3**	**E1**		**P3**	**A4**	**A5**	**E2**	**E3**	**P4**	**E4**	
todo	todo	todo	mucho	mucho	todo		mucho	poco	mucho	todo	mucho	mucho	regular	
Umbral														
poco	poco	poco	poco	poco	nada		poco	poco	poco	nada	nada	poco	nada	
Dimensiones							***Viabilidad global***							
Plausibilidad			8,24				***7,1***			***Valido***				
Adecuación			5,66											
Éxito			7,49											
Comentarios Generales:														

Tabla 4. Listado de Métricas

Proyecto de Educación Superior - Universidad Nacional de Córdoba				Código del Proyecto:	ES.UNC
Listado de Métricas				ID:	G.2.1.2
				Versión:	1.0
ID	**Tipo**	**Descripción**	**Fórmula**		**Comentarios**
M.DRPY	Proyecto	Tiempo total requerido para el desarrollo del proyecto	$\sum$ trA trA = tiempo requerido por actividad		Sumatoria de los tiempos requeridos para cada actividad del proyecto
M.DRPEI	Proyecto	Tiempo medio requerido para el desarrollo de un problema de explotación de información	($\sum$ trA.SD) / NPEI trA.SD = tiempo requerido por actividad del subproceso de desarrollo NPEI = cantidad de problemas de explotación de información		Solo se considera el tiempo de las actividades pertenecientes al subproceso de desarrollo
M.NANU	Datos	Número total de atributos que no son de utilidad en las tablas	$\sum$ aNu aNu = cantidad de atributos que no son de utilidad		
M.NASxM	Modelo	Número medio de atributos significativos por modelo	($\sum$ AtS.M) / NMOD AtS.M = cantidad de atributos significativos de un modelo NMOD = cantidad de modelos		
Comentarios Generales:					

Actualmente se está trabajando en los formalismos que encapsulan los artefactos de entrada y salida de cada una de las actividades y la implementación de técnicas, que asociadas a cada una de las actividades permiten transformar el artefacto de entrada en los correspondientes artefactos de salidas. En este trabajo se han desarrollado formalismos para el subproceso de gestión y particularmente se han detallado en cada una de las fases del proyecto cuales son los artefactos de entrada y salida. El próximo paso es desarrollar las técnicas para la trasformación de artefactos en el subproceso de gestión y seguidamente elaborar los artefactos y las técnicas para el subproceso de desarrollo de MoProPEI.
Como futuras líneas de trabajo, se prevé ampliar los casos de aplicación del modelo en cantidad y variedad de dominios.

Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por beca PROMINF-UNLa-2015-2017 del Ministerio de Educación Argentina y por el Proyecto 33A205 de la Secretaria de Ciencia y Tecnología de la Universidad Nacional de Lanús.

Referencias

1. Maimon, O. y Rokach, L. (Eds.). 2005. Data mining and knowledge discovery handbook. Springer.
2. Tan, A. 1999. Text mining: The state of the art and the challenges. In Proc. PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases. pp. 65-70.
3. Hsu, W., Lee, M., Zhang, J. 2002. Image mining: Trends and developments. Journal of Intelligent Information Systems, 19(1): 7-23.
4. Gaber, M., Zaslavsky, A. Krishnaswamy, S. 2010. *Data stream mining.* En Maimon, O. and Rokach, L. eds. Data mining and knowledge discovery handbook. Springer, Pág. 759-787.
5. Kosala, R., Blockeel, H. 2000. Web mining research: A survey. ACM SIGKDD Explorations Newsletter, 2(1): 1-15.
6. Abran, A., Moore, J. W., Bourque, P., Dupuis, R., Tripp, L. 2004. Guide to the Software Engineering Body of Knowledge (2004 version). IEEE. ISBN 0-7695-2330-7.
7. García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell. J. 2011. Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellin. ISBN 978-958-8692-32-6. Pág. 83-99.
8. Martins, S., Pesado, P., García-Martínez, R. 2014. Propuesta de Proceso de Ingeniería de Explotación de Información Centrado en Control y Gestión del Proyecto. XI Workshop de Bases de Datos y Minería de Datos. Proceedings XX Congreso Argentino de Ciencias de la Computación. Universidad Nacional de la Matanza. ISBN 978-987-3806-05-6.
9. Martins, S., Pesado, P., García-Martínez, P. 2016. Information Mining Projects Management Process. Proceedings 28th International Conference on Software Engineering & Knowledge Engineering. Pág. 504-509. ISBN 1-891706-39-X.
10. Gondar, J.E. 2005. Data Mining Methodology. Data Mining Institute. ISBN: 978-84-96272-21-7.
11. Wirth R., Hipp J. 2000. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, pp. 29–39.
12. Yang Q., Wu X., 2006. 10 Challenging Problems in Data Mining research, International Journal of Information Technology and Decision Making 5(4), pp. 597–604.
13. Lavrac N., Motoda H., Fawcett T., Holte R., Langley P., Adriaans P., 2004. Lessons Learned from Data Mining Applications and Collaborative Problem Solving. Machine Learning. 57. 13–34.
14. Kdnuggets. 2014. What main methodology are you using for your analytics, data mining, or data science projects? Poll (Oct 2014). http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-scienc e-methodology.html (Último acceso 04/07/2016).
15. Martins, S. 2016. Artefactos para el Subproceso de Gestión del Proceso MoProPEI. Reporte de Tareas RT-UNLa-DDPyT-GISI-2016-01. http://sistemas.unla.edu.ar/sistemas/gisi/papers/RT-UNLa-DDPyT-GISI-2016-01.pdf
16. Pytel, P., Hossian, A., Britos, P., García-Martínez, R. 2015. Feasibility and Effort Estimation Models for Medium and Small Size Information Mining Projects. Information Systems Journal, 47: 01-14. Elsevier. ISSN 0306-4379.
17. Basso, B., Rodríguez, D., García-Martínez, R. 2015. Comportamiento de Métricas para Proyectos de Explotación de Información en PyMEs. XII Workshop de Ingeniería de Software. Libro de Actas del XXI Congreso Argentino de Ciencias de la Computación. Pág. 485-494. ISBN 978-987-3724-37-4. Universidad Nacional del Noroeste de Buenos Aires.