

Propuesta de Proceso de Ingeniería de Explotación de Información Centrado en Control y Gestión del Proyecto

Sebastian Martins^{1,2}, Patricia Pesado^{1,3}, Ramón García-Martínez²

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
<http://www.unla.edu.ar/sistemas/gisi/>
3. Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP – CIC
<http://weblidi.info.unlp.edu.ar/wp/>
martinssebastian@yahoo.com.ar, ppesado@lidi.info.unlp.edu.ar, rgarcia@unla.edu.ar

Resumen. Existen varias propuestas de procesos de ingeniería de explotación de información, pero, o se orientan a las actividades de desarrollo del proyecto estrictamente, o prevén un área de administración disociada del desarrollo. En este artículo se propone un proceso centrado en la gestión y control del proyecto que identifica las entradas/salidas de las actividades, así como sus dependencias, y que integra control, gestión y desarrollo para el concepto de ingeniería de explotación de información.

Palabras Clave. Ingeniería de Explotación de Información. Modelo de proceso. Explotación de Información. Minería de Datos.

1. Introducción

El termino minería de datos esta 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 [Maimon y Rokach, 2005]. Sin embargo, hoy existen líneas de investigación en campos tales como: minería de textos [Tan, 1999], minería de imágenes [Hsu et al., 2002], minería de patrones en flujos de información [Gaber et al., 2010], minería en la web [Kosala y Blockeel, 2000], 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 [Abran et al., 2004] 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 [García-Martínez et al., 2011] 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.

Los procesos de desarrollo de proyectos de explotación de información existentes, se centran en el descubrimiento de patrones de conocimiento en la masa de datos. Sin embargo no existe una visión de proceso que identifique fases, tareas, técnicas de representación y procedimientos de ejecución de la tarea; que permita sistematizar la concreción del proyecto.

En este trabajo se presenta el estado de la cuestión (sección 2) donde se describen abordajes a la definición de un proceso de ingeniería de explotación de información, se delimita el problema (sección 3), se formula una propuesta de solución (sección 4), se discute las instancias superadoras del modelo de proceso propuesto respecto de los abordajes previos (sección 5), y se formulan conclusiones preliminares del trabajo (sección 6).

2. Estado de la Cuestión

En esta sección se presentan tres abordajes relevantes a la definición de un proceso de ingeniería de explotación de información: CRISP-DM [Chapman, et. al., 1999], el propuesto por Vanrell [2011], y el propuesto por Marbán [Marbán et al., 2007].

El abordaje CRISP-DM [Chapman, et. al., 1999] es un modelo de proceso jerárquico, el cual se conforma por un conjunto de tareas descritas en cuatro niveles de abstracción (de lo genérico a lo específico), las cuales se ejecutan en base al modelo de ciclo de vida definido. Este modelo de proceso, utilizado para el desarrollo de proyectos de explotación de información, se centra en el desarrollo técnico del proyecto, considerando levemente algunas actividades vinculadas con la administración del mismo. Si bien CRISP-DM se define a sí mismo como una metodología, no describe cómo implementar cada una de las tareas que propone, con lo que según [Marbán et al., 2007] es un proceso.

El abordaje Vanrell [2011] propone un proceso que incorpora las actividades administrativas de un proyecto a las actividades de desarrollo propuestas en CRISP-DM, distinguiendo las mismas en 2 subprocesos. Dicho proceso parte de la estructura definida en CompetiSoft, agregando aquellas actividades de CRISP-DM asociadas a la gestión del proyecto, y recomienda la aplicación de ciertas técnicas. Para el subproceso de Desarrollo utiliza las actividades definidas en CRISP-DM, e incorpora y propone tareas y técnicas no contempladas. De forma complementaria incorpora el concepto de procesos de explotación de información [García Martínez et al., 2013] al modelo de proceso.

El abordaje descrito en [Marbán et al., 2007], propone un modelo de proceso que integra las actividades identificadas en CRISP-DM, con actividades orientadas a la gestión del proyecto, basándose en los estándares para el desarrollo de software tradicional (IEEE Std 1074 e ISO Std 12207). Define la estructura del modelo de proceso a partir del IEEE Std 1074, incorporando en el núcleo del desarrollo las actividades propuestas en CRISP-DM.

3. Descripción del Problema

A partir del desarrollo de variados proyectos de explotación de información, implementando los distintos modelos de procesos de explotación de información existentes, se identifican una serie de deficiencias y carencias para la ejecución exitosa de estos proyectos. De las insuficiencias identificadas surge como problema la necesidad de producir un modelo de proceso de desarrollo de proyectos de ingeniería de explotación de información que determine fases, tareas, técnicas de representación y procedimientos de ejecución de la tarea, que permita sistematizar el desarrollo de proyectos en el área.

4 Propuesta de Solución

Se propone como solución al problema identificado en la sección anterior, un modelo de proceso orientado en el control y la gestión, que no solo adapta y organiza a las actividades orientadas al desarrollo de un proyecto de explotación de información, sino que además integra aquellas actividades destinadas a la administración de este tipo de proyectos, identificando las vinculaciones y los elementos de dependencia entre las mismas.

El proceso propuesto se articula a través de 2 subprocesos: el subproceso de desarrollo, enfocado a las actividades técnicas del proyecto y el subproceso de gestión, orientado al control y la administración.

El subproceso de Desarrollo se encarga de todas las tareas asociadas a la educación 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 Desarrollo, está compuesto por seis fases, las cuales están conformadas por distintas tareas generales, que identifican un conjunto de actividades con un objetivo específico dentro del proyecto. En la figura 1, se ilustran las fases que componen al subproceso de desarrollo, junto con cada una de las tareas que las conforman, identificando además, las dependencias existentes entre las mismas (sus elementos de entrada y salida).

La fase de Entendimiento del Negocio, se compone de cuatro tareas: *Análisis del Negocio*, que se alimenta del Discurso del Usuario/Experto, y a partir del mismo, genera como productos de salida el documento de requisitos y la identificación de los objetivos del negocio; *Identificación Inicial de las Fuentes de Datos*, la cual produce a partir del documento de requisitos y del Discurso del Usuario/Experto, la identificación de las fuentes de datos y el diagrama de relación de fuentes de datos; *Entendimiento del Problema de Negocio*, que se alimenta de los elementos producidos en la primera tarea de la fase actual (documento de requisitos y objetivos del negocio) y produce como elemento de salida la Planilla de Criterios de éxito del

Problema de Negocio y la identificación del problema de negocio, la cual ingresa junto con el documento de requisitos, a la última actividad de la primera fase, denominada *Determinación del Problema de Explotación de información*, cuyos elementos resultantes son: el problema de explotación de información identificado y la planilla términos relacionados con el problema de explotación de información.

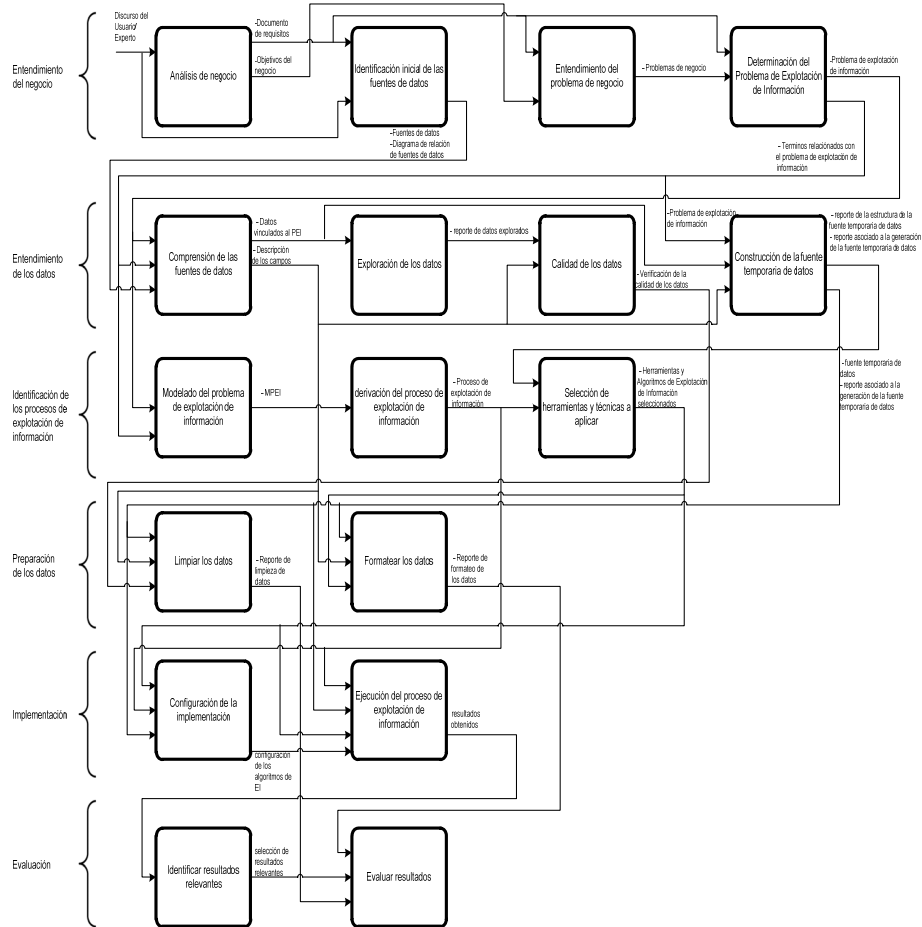


Fig. 1. Subproceso de Desarrollo

En la fase Entendimiento de los Datos, se identifican cuatro actividades generales: *Comprensión de las Fuentes de Datos*, cuyos elementos de entrada son la fuente de datos, el diagrama de relación de las fuentes de datos, el problema de explotación de información y los términos relacionados con el problema de explotación de información, y produce como elementos de salida las planillas: Datos Vinculados al Problema de Explotación de Información (PEI) y Descripción de los Campos; *Exploración de los Datos*, cuyo elemento de entrada es la planilla de Datos Vinculados al PEI, obtenida de la actividad anterior y genera el Reporte de Datos

Explorados; *Calidad de los Datos*, produce a partir de la planilla de Descripción de los Campos y el Reporte de Datos Explorados, la planilla de Verificación de la Calidad de los Datos; y *Construcción de la Fuente Temporal de Datos*, la cual se retroalimenta de los elementos: problema de explotación de información, y las planillas de Datos Vinculados al PEI y Descripción de los Campos, y genera como elementos de salida: la construcción de la fuente temporal de datos, y los reportes de la estructura de la fuente temporal de datos, y asociado a la generación de la fuente temporal de datos.

La fase Identificación del Proceso de Explotación de Información, se compone de 3 actividades generales: *Modelado del Problema de Explotación de Información*, en la cual ingresa el Problema de Explotación de Información, identificados en la primera fase, y produce como salida el Modelo del Problema de Explotación de Información (MPEI), a partir del cual la siguiente actividad denominada *Derivación del Proceso de Explotación de Información*, identifica el Proceso de Explotación de Información a utilizar. Por último, la actividad *Selección de Herramientas y Técnicas*, tiene como elementos de entrada el Proceso de Explotación de Información, obtenido en la actividad anterior, y los reportes de la estructura de la fuente temporal de datos y asociado a la generación de la fuente temporal de datos, producidos en la última actividad de la fase anterior, y produce como elementos de salida la selección de la Herramienta y de los Algoritmos de Explotación de Información a utilizar.

La fase Preparación de los Datos, está conformada por dos actividades generales: *Limpiar los Datos*, se alimenta de la fuente temporal de datos, y el reporte asociado a la generación de la fuente temporal de datos, y las planillas Descripción de los Campos, y Verificación de la Calidad de los Datos, y produce como salida el reporte de Limpieza de Datos; y la actividad *Formatear los Datos*, tiene como elementos de entrada la fuente temporal de datos, el reporte asociado a la generación de la fuente temporal de datos, y las planillas Descripción de los Campos, y Herramientas y Algoritmos de Explotación de Información seleccionadas, produciendo como resultado el Reporte de Formateo de los Datos. Las actividades pertenecientes a la fase actual, además de los elementos de salida producidos, pueden modificar la estructura de la fuente temporal de datos.

La fase implementación está conformada por dos actividades generales: *Configuración de la Implementación*, cuyos elementos de entrada son la fuente temporal de datos, y el reporte asociado a la generación de la fuente temporal de datos, producidos en la fase de Entendimiento de los Datos, el Proceso de Explotación de Información y las Herramientas y Algoritmos de Explotación de Información, identificadas en la fase anterior, y produce como salida el reporte de configuración de los Algoritmos de Explotación de Información (EI); y la actividad *Ejecución del Proceso de Explotación de Información*, se alimenta de la fuente temporal de datos, el reporte asociado a la generación de la fuente temporal de datos, las Herramientas y Algoritmos de Explotación de Información, el Proceso de explotación de información, y el reporte de configuración de los Algoritmos de EI, y produce como salida los resultados obtenidos de aplicar Explotación de Información.

La fase Evaluación, se compone de dos actividades generales: *Identificar los Resultados Relevantes*, cuyos elementos de entrada son la Planilla de Criterios de éxito del Problema de Negocio y los resultados obtenidos en la fase anterior, y produce como elemento de salida la selección de los resultados relevantes asociados

al problema de negocio; y la actividad *Evaluar los Resultados*, cuyos elementos de entrada son los reportes de Limpieza y Formateo de los Datos, y de configuración de los Algoritmos de MD, y produce como salida la planilla de Verificación de los resultados y el reporte del proyecto.

El subproceso de gestión se produce como un elemento de soporte a las actividades de desarrollo, donde la 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. En la figura 2, se ilustran las fases que lo conforman, junto con cada una de las tareas que las integran, y sus dependencias existentes. Algunas tareas de gestión, se alimentan de salidas producidas en alguna tarea perteneciente al subproceso de desarrollo.

La fase de gestión del entendimiento y planificación del negocio, está compuesta por cuatro actividades: *Evaluación de la situación*, tiene como elemento de entrada el documento de requisitos (generado por la actividad *Análisis del negocio*, perteneciente al subproceso de desarrollo), y produce como elementos de salida cinco elementos: el reporte de recursos requeridos, la identificación de los recursos disponibles, los elementos tercerizables, los riesgos asociados al proyecto y el plan de contingencia del proyecto; los primeros tres elementos de salida de la actividad anterior, junto con el documento de requisitos y el calendario de actividades e hitos entregables, ingresan a la actividad *planificación de recursos*, la cual produce como salida el Plan de incorporación de RRHH, el Plan de incorporación de recursos materiales, la Planificación de la tercerización, el Plan de capacitación de RRHH y la planificación de la necesidad de recursos; *Definición del Ciclo de Vida*, se alimenta del documento de requisitos y define el ciclo de vida seleccionado para el proyecto; *Planificación de las Actividades*, tiene como elementos de entrada el documento de requisitos y el ciclo de vida seleccionado, y genera como salida las planillas de Métricas asociadas al proyecto y de actividades asociadas al proyecto, y el calendario de actividades e hitos entregables; y *Costos, Alcances y Planificación*, cuyos elementos de entrada son el documento de requisitos, el Reporte de recursos requeridos y la identificación de los Recursos disponibles y produce como elemento de salida el costo estimado y el contrato del proyecto.

La fase Gestión de la Realización, la cual se compone de tres actividades generales: *gestión de las actividades*, cuyos elementos de entrada son el calendario de actividades e hitos entregables, las actividades asociadas al proyecto, el documento de requisitos, el Plan de incorporación de RRHH, el Plan de incorporación de recursos materiales, el Planificación de la tercerización, el ciclo de vida seleccionado, el reporte de evaluación de cierre de ciclo y el Reporte de seguimiento de tareas tercerizadas, y produce como elementos de salida el informe de inicio del ciclo, el Reporte de integración de tareas tercerizadas, las responsabilidades del personal, el Informe de cierre de ciclo, los Contratos de recursos, la Planilla de contratos de recursos y el Reporte de tercerización de tareas; *implementación del cambio* que se alimenta de los Riesgos asociados al proyecto, el Plan de contingencia del proyecto, el reporte de

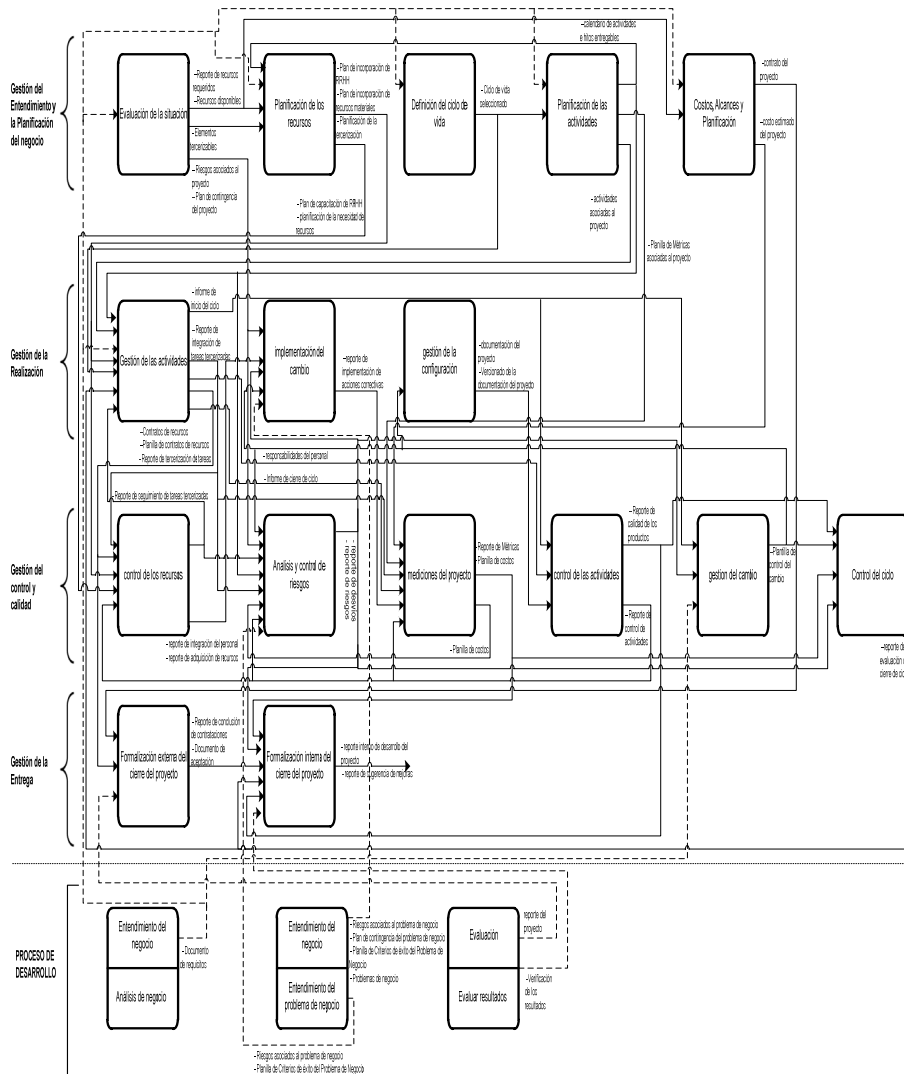


Fig. 2. Subproceso de Gestión

integración del personal, el reporte de adquisición de recursos, el reporte de desvíos, el reporte de riesgos, la Plantilla de control del cambio y el Reporte de seguimiento de tareas tercerizadas; *implementación del cambio*, tiene como elementos de entrada los Riesgos asociados al proyecto, el Plan de contingencia del proyecto, los reportes de integración del personal, de adquisición de recursos, de desvíos, de riesgos, la Plantilla de control del cambio, los Riesgos asociados al problema de negocio, el Plan de contingencia del problema de negocio, la Planilla de Criterios de éxito del Problema de Negocio, y los Problemas de negocio y produce como salida el reporte de implementación de acciones correctivas (los últimos 4 elementos producidos por la

actividad *entendimiento del problema de negocio* del subproceso de desarrollo); y *Gestión de la Configuración*, se alimenta de la Plantilla de control del cambio y genera la documentación del proyecto y el Versionado de la documentación del proyecto.

La fase Gestión del control y calidad, compuesto por seis actividades generales: *control de los recursos*, tiene como elementos de entrada el Reporte de integración de tareas tercerizadas, los Contratos de recursos, la Planilla de contratos de recursos, el Reporte de tercerización de tareas, el Plan de incorporación de RRHH, el Plan de incorporación de recursos materiales, la Planificación de la tercerización, el Plan de capacitación de RRHH, la planificación de la necesidad de recursos y el Reporte de control de actividades, y produce los reportes de seguimiento de tareas tercerizadas, de integración del personal y de adquisición de recursos; *Análisis y Control de Riesgos*, se alimentan de la Plantilla de control del cambio, Riesgos asociados al proyecto, el Plan de contingencia del proyecto, el Reporte de seguimiento de tareas tercerizadas, el calendario de actividades e hitos entregables, el Reporte de integración de tareas tercerizadas, la Planilla de costos, el Reporte de control de actividades, los Riesgos asociados al problema de negocio y la Planilla de Criterios de éxito del Problema de Negocio (los últimos 2 elementos producidos por la actividad *entendimiento del problema de negocio* perteneciente al subproceso de Desarrollo), y genera los reportes de desvíos y de riesgos; *Mediciones del proyecto*, cuyos elementos de entrada son el costo estimado del proyecto, la Planilla de Métricas asociadas al proyecto, el Reporte de integración de tareas tercerizadas, el Informe de cierre de ciclo, el reporte de implementación de acciones correctivas y el Reporte de control de actividades, y produce como elemento de salida el Reporte de Métricas y la Planilla de costos; *Control de Actividades*, en la cual ingresan el informe de inicio del ciclo, las responsabilidades del personal, la documentación del proyecto y el Versionado del mismo, y producen como elementos de salida los Reportes de calidad de los productos y de control de actividades; *Gestión del Cambio*, cuyos elementos de entrada son el informe de inicio del ciclo, los reportes de desvíos y de riesgos, y el Documento de requisitos (producido por la actividad *Análisis del negocio* perteneciente al subproceso de Desarrollo) y produce como elemento de salida la Plantilla de control del cambio; y *Control del Ciclo*, cuyos elementos de entrada son el Reporte de calidad de los productos, la Plantilla de control del cambio, el Reporte de Métricas, la Planilla de costos y los reportes de desvíos y de riesgos, y produce como elemento de salida el reporte de evaluación de cierre de ciclo.

La última fase Gestión de la entrega, se compone de 2 actividades generales: *Formalización externa del cierre del proyecto*, cuyos elementos de entrada son el contrato del proyecto, los Contratos de recursos, la Planilla de contratos de recursos, el Reporte de tercerización de tareas, y el reporte del proyecto y produce como elementos de salida el Reporte de conclusión de contrataciones y el Documento de aceptación; y *Formalización interna del cierre del proyecto*, la cual se alimenta del Reporte de Métricas, la Planilla de costos, el reporte de desvíos, el reporte de riesgos, el Reporte de conclusión de contrataciones, el Documento de aceptación, el reporte de evaluación de cierre de ciclo, el Reporte de calidad de los productos, y la Verificación de los resultados (producido por la actividad *Evaluar Resultados* perteneciente al subproceso de Desarrollo) y produce como elemento de salida los reportes interno de desarrollo del proyecto y de sugerencia de mejoras.

5. Aportaciones del Proceso Propuesto en Relación a los Abordajes Previos

A partir de la solución propuesta en la sección previa, se resumen las aportaciones realizadas con respecto a los abordajes existentes:

- Se define un modelo de proceso para proyectos de explotación de información, el cual identifica los productos de entrada y salida de cada una de sus actividades y las dependencias conceptuales entre las mismas. La necesidad del mismo surge a partir de las carencias existentes en los abordajes previos, con respecto a la identificación de las entradas y salidas de cada una de las actividades del proyecto, es decir, la definición integra de un proceso [Marbán et al., 2007], que a su vez, contemple no solo las actividades técnicas de dicha ingeniería, sino también aquellas orientadas a la gestión y control del proyecto.
- Se reestructuran y redefinen las actividades de desarrollo, a partir de la identificación de carencias e inconsistencias en el modelo de proceso de facto para el desarrollo de proyectos de explotación de información (CRISP-DM), las cuales no fueron resueltas por los abordajes posteriores. Se elimina la necesidad de iteración a causa de la dependencia existente entre la preparación de los datos y la selección de las herramientas y técnicas a utilizar, a partir del ordenamiento de las fases. Se adapta el subproceso al objetivo principal de este tipo de proyectos, la recolección de conocimiento, y no la construcción de un software [Marbán et al., 2007], por lo cual se eliminan aquellas actividades asociadas a dicho propósito, y se incorporan actividades orientadas a la identificación, preparación y ejecución de los procesos de explotación de información y las técnicas de Explotación de Información asociadas [Garcia-Martinez et al., 2013], así como a la evaluación y entrega del producto. Se redefine el concepto de modelado, a partir de la incorporación de los procesos de explotación de información y de derivación del mismo a partir del modelado del problema de negocio [Martins et al, 2014].
- Se define al subproceso de gestión y control, como elemento soporte del subproceso de desarrollo, el cual se determina específicamente bajo los objetivos de la ingeniería de explotación de información, desestimando actividades incorporadas en abordajes previos que no cumplen con dichos objetivos. Adicionalmente, se incorporan actividades destinadas al análisis y definición de la viabilidad y factibilidad del éxito del proyecto, identificación, determinación y control de la tercerización de trabajo, controles formales asociados a la posibilidad de seleccionar el modelo de ciclo de vida a utilizar, la realización y seguimiento de los contratos, y actividades destinadas a la generación de conocimiento asociado al desarrollo del proyecto y su reutilización en proyectos futuros.

6. Conclusiones

La Ingeniería de Explotación de Información, requiere un modelo de proceso que abarque todas las actividades asociadas a los proyectos, desde la especificación de requisitos del proyecto hasta la redacción del informe final sobre resultados. Los modelos de procesos existentes presentan deficiencias sobre la profundidad con la que se definen las actividades de control y gestión y como estas interactúan con las

actividades de desarrollo. En este trabajo se define un modelo de proceso que busca definir la totalidad de actividades vinculadas, focalizándose no solo en las actividades de desarrollo, sino también en las actividades de gestión y control de proyecto y sus productos asociados.

Como línea de trabajo futura, se prevé construir las técnicas asociadas al desarrollo de cada fase y tarea, y los formalismos de documentación de las entradas/salidas de las mismas.

7. Financiamiento

Las investigaciones que se reportan en este capítulo han sido financiadas por Beca BENTR13 de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, y parcialmente por el Proyecto de Investigación 33A167 de la Secretaría de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

8. Referencias

- 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.
- García-Martínez, R., Britos, P., Rodríguez, D. 2013. *Information Mining Processes Based on Intelligent Systems*. Lecture Notes on Artificial Intelligence, 7906: 402-410.
- Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 1999. CRISP-DM 1.0 Step by step data mining guide.
- 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.
- 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 ISBN 978-958-8692-32-6. Pág. 83-99.
- Hsu, W., Lee, M., Zhang, J. 2002. Image mining: Trends and developments. Journal of Intelligent Information Systems, 19(1): 7-23.
- Kosala, R., Blockeel, H. 2000. Web mining research: A survey. ACM SIGKDD Explorations Newsletter, 2(1): 1-15.
- Maimon, O. y Rokach, L. (Eds.). 2005. Data mining and knowledge discovery handbook. Springer.
- Marbán, O., Mariscal, G., Menasalvas, E., Segovia, J. 2007. An Engineering Approach to Data Mining Projects. Lecture Notes in Computer Science, 4881: 578-588. Springer.
- Martins, S., Rodríguez, D., García-Martínez, R. 2014. *Deriving Processes of Information Mining Based on Semantic Nets and Frames*. LNAI, 8482: 150-159.
- Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. 2010. Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información. XVI CACIC: 664-673.
- 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.
- Vanrell, J. A. 2011. Un Modelo de Procesos para Proyectos de Explotación de Información. Tesis de Maestría en Ingeniería en Sistemas de Información. FRBA-UTN.