UN MODELO DE PROCESOS DE EXPLOTACIÓN DE INFORMACIÓN

Juan Ángel Vanrell, Rodolfo Bertone, Ramón García-Martínez

Escuela de Postgrado. Universidad Tecnológica Nacional (FRBA)
Facultad de Informática. Universidad Nacional de La Plata
Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús

javanrell@gmail.com, pbertone@lidi.info.unlp.edu.ar, rgarcia@unla.edu.ar

CONTEXTO

Este proyecto se desarrolla en el marco de la cooperación existente entre los proyectos de investigación "Metodología para la Especificación de Requisitos en Proyectos de Explotación de Información" de la Universidad Tecnológica Nacional (FRBA) y "Proyecto 33A081: Sistemas de Información e Inteligencia de Negocio" de la Universidad Nacional de Lanús.

RESUMEN

Los proyectos de explotación de información poseen características muy distintas a las de los proyectos de desarrollo de software tradicionales. Las clásicas etapas de análisis, diseño, desarrollo, integración y testeo no encajan con las etapas naturales de los procesos de desarrollo de este tipo de proyectos. En este contexto, se propone un marco teórico para la creación de Modelos de Procesos para Proyectos de Explotación de Información para PyMEs siguiendo los lineamientos del Modelo de Procesos para la Industria de Software (Competisoft).

Palabras clave: *Explotación de Información, Modelo, Procesos, Competisoft.*

1. INTRODUCCION

Actualmente existen en el mercado distintos modelos que ayudan a llevar a cabo proyectos con un nivel de calidad esperado en forma repetitiva como pueden ser el de la norma ISO9000:2000, el modelo CMM y su versión actual CMMI [SEI, 2006], MoProSoft [Oktaba et al., 2005] o su versión iberoamericana Competisoft [Oktaba et al., 2007]. Todos estos son modelos genéricos por lo cual pueden ser utilizados para la ejecución de cualquier tipo de proyecto.

Dentro de los distintos proyectos que son llevados a cabo por empresas dedicadas al área de tecnologías de la información se encuentra un conjunto denominado proyectos de explotación de información.

Como todo conjunto posee características propias que lo hacen diferenciarse del resto. Creemos que estas características son lo suficientemente significativas como para justificar la construcción de un modelo de procesos que se ajuste a este tipo de proyectos.

Mas y Amengual [2005] describen algunas características con las cuales se marca una diferencia entre las grandes empresas (de más de 200 desarrolladores) y las Pymes (menos de 30 desarrolladores). Estas características fueron divididas en categorías dentro de las cuales se identificaron distintos factores que diferencian a los grupos.

El primer factor es el de los recursos humanos, diferencia entre los tamaños de equipos y cantidad de los mismos, falta de roles especializados, responsabilidades no muy bien definidas y alta dependencia de los individuos ente otros. El segundo factor identificado se relaciona con los aspectos económicos, en el caso de Pymes se da mayor importancia a la obtención de beneficios a corto plazo y la inversión en investigación y desarrollo suele ser mínima. Los procesos son identificados como un factor más en los cuales las Pymes encuentran serios problemas al intentar desarrollar y reflejar los resultados de la

implantación de programas de mejora de procesos de software con el nivel de detalle y formalidad exigido por los grandes modelos. El último factor identificado tiene que ver con los proyectos, en donde las Pymes de destacan por trabajar en aquellos cuyo tamaño es reducido al igual que su duración, también se destacan en la cantidad de proyectos simultáneos que suelen ser pocos y las dificultades en alguno de los mismos suelen tener un algo grado de incidencia en la organización.

Los autores señalan algunos de los problemas que tienen las Pymes para adoptar grandes modelos de SPI como la duración promedio de los proyectos que van de 18 a 24 meses, lo cual representa mucho tiempo para una empresa de poca envergadura.

En el mismo sentido en [Oktaba et al., 2007] se remarca la complejidad de las recomendaciones para la implementación de los grandes modelos y la implementación de los modelos construidos en otros países sin adaptación, coincidiendo con los factores de costo y tiempo antes mencionados.

Relacionado con el trabajo de SPI en Pymes el artículo [Pino et al., 2006] indica que las pequeñas y medianas empresas son un engranaje muy importante en la economía mundial. En la mayoría de los países el desarrollo de software es llevado a cabo, en un gran porcentaje, por este tipo de empresas. Estas organizaciones, denominadas Pymes_DS, requieren prácticas eficientes de Ingeniería de Software adaptadas a su tamaño y tipo de negocio.

En la misma línea los autores describen que en la última década la comunidad vinculada a esta disciplina ha demostrado un gran interés en la mejora de procesos de software, buscando aumentar la calidad y productividad del software, lo cual se ve reflejado tanto en el creciente número de artículos sobre el tema como por la aparición de un gran número de iniciativas internacionales relacionadas con SPI.

Dentro de los trabajos e iniciativas a las que se hacen referencia para fortalecer SPI en Pymes_DS podemos mencionar SPIRE (*Software Process Improvement in Regions of Europe*), TOPS (*Toward Organised Software Process in SMEs*) o el programa brasilero PBQP-Software (*Productivity and Quality Software Program*) y el proyecto "mps Br" (*melhoria do processo do software brasilero*). Puede sumárseles a estos trabajos e iniciativas el modelo mexicano MoProSoft y su iniciativa iberoamericana Competisoft.

Entre las conclusiones que obtiene y citando a otros autores se encuentra que los estándares de facto (ISO y los modelos del SEI) difícilmente pueden ser aplicados a pequeñas empresas ya que "un proyecto de mejora supone una gran inversión de dinero, tiempo y recursos".

1.1. COMPETISOFT

Competisoft [Oktaba et al., 2008] es la proyección a nivel iberoamericano del modelo de procesos para el desarrollo de software MoProSoft [Oktaba et al. 2005] creado por encargo de la Secretaría de Economía Mexicana para servir de base a la norma Mexicana para la Industria de Desarrollo y Mantenimiento de Software. El modelo inicial fue modificado y adecuado a las necesidades de otros países, se le incorporó el modelo de evaluación EvalProSoft [Oktaba et al. 2004] y se definieron niveles de madurez.

Su propósito es fomentar la estandarización de las operaciones de pequeñas y medianas empresas o departamentos internos de desarrollo, a través de la incorporación de las mejores prácticas en gestión e ingeniería de software, esperando “elevar la capacidad de las organizaciones para ofrecer servicios con calidad y alcanzar niveles internacionales de competitividad”.

El modelo busca ser fácil de entender, fácil de aprender, no costoso en su adopción y ser la base para alcanzar evaluaciones exitosas con otros modelos o normas como ISO 9000:2000 o CMM.

Además de definir procesos Competisoft define un patrón que debe ser utilizado para documentar aquellos procesos que una

empresa requiere agregar a los existentes en el modelo o para documentar la adecuación de los que ya se encuentra en el mismo. Dicho patrón se encuentra constituido por tres partes: Definición general del proceso, Prácticas y Guías de ajuste. El modelo a desarrollar pretende seguir este patrón para la documentación de los procesos de explotación de información.

La estructura del modelo se encuentra dividida en tres categorías: Alta Dirección (DIR), Gerencia (GER) y Operaciones (OPE) reflejando la estructura de una organización. Estas categorías contienen los procesos de gestión de negocio (DIR), gestión de procesos, gestión de proyectos y gestión de recursos (GER) y administración de un proyecto específico, desarrollo de software y mantenimiento de software (OPE). La categoría de Alta Dirección es la "categoría de procesos que aborda las prácticas de Alta Dirección relacionadas con la gestión del negocio" y "proporciona los lineamientos a los procesos de la Categoría de Gerencia y se retroalimenta con la información generada por ellos", la categoría de gerencia es la "categoría de procesos que aborda las prácticas de gestión de procesos, proyectos y recursos en función de los lineamientos establecidos en la Categoría de Alta Dirección", además "proporciona los elementos para el funcionamiento de los procesos de la Categoría de Operación, recibe y evalúa la información generada por éstos y comunica los resultados a la Categoría de Alta Dirección" y la Categoría de Operación es la "categoría de procesos que aborda las prácticas de los proyectos de desarrollo y mantenimiento de software", además "esta categoría realiza las actividades de acuerdo a los elementos proporcionados por la Categoría de Gerencia y entrega a ésta la información y productos generados".

1.2. EXPLOTACIÓN DE INFORMACIÓN

Larose [2005] define el término explotación de información (*Data Mining*) como el proceso de descubrir nuevas correlaciones, patrones y tendencias utilizando grandes cantidades de datos almacenados en repositorios, usando tecnologías de reconocimiento de patrones así como herramientas matemáticas y de estadística.

Existen actualmente varias metodologías de para proyectos de explotación de información, entre ellas podemos nombrar CRISP-DM, SEMMA y P3TQ como las más reconocidas y algunas otras que no abarcan la totalidad de los proyectos sino que se enfocan en ciertos procesos de los mismos. Se propone utilizar las distintas metodologías existentes para identificar procesos propios de este tipo de proyectos con el fin de incluirlos en el nuevo modelo.

La metodología CRISP-DM [Chapman et al., 2000] se encuentra definida en base a un modelo jerárquico de procesos. El foco se pondrá en los procesos del nivel superior que son lo suficientemente genéricos como para cubrir todas las posibles aplicaciones de explotación de información.

Esta metodología define un ciclo de vida de los proyectos de explotación de información que define las principales fases de un proyecto de este tipo. Estas fases son: Entendimiento de Negocios, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Despliegue. Claramente estas fases difieren de las fases definidas para un proyecto de desarrollo de software clásico (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre). A continuación se presenta el concepto de cada una de las fases identificadas por CRISP-DM.

En la fase de Entendimiento del Negocio se deben entender los objetivos del proyecto y los requerimientos desde una perspectiva del negocio y luego convertir este conocimiento en una definición de un problema de explotación de información y diseñar un plan preliminar para lograr dichos objetivos.

El Entendimiento de los Datos comienza con la recolección inicial de datos y procede con las acciones para familiarizarse con ellos, identificar problemas de calidad, identificar primeras pautas en los datos o

detectar subconjuntos interesantes de las hipótesis de información oculta.

La fase de Preparación de los Datos cubre todas las actividades para construir el conjunto de datos final desde los datos iniciales, las tareas de esta fase pueden ser realizadas muchas veces y sin un orden preestablecido, incluye tanto la selección de tablas, registros y atributos como transformación y limpieza de datos para herramientas de modelado.

El Modelado incluye la selección de técnicas de modelado y la calibración de sus parámetros a los valores óptimos, suelen existir distintas técnicas para un mismo problema de explotación de información y cada una de ellas suele tener ciertos requisitos sobre los datos, muchas veces es necesario volver a la fase de preparación de los datos.

La Evaluación requiere la construcción de uno o varios modelos que aparentan tener la mayor calidad desde una perspectiva de análisis, requiere la evaluación del modelo y revisión de los pasos ejecutados para la construcción del modelo para asegurarnos de lograr los objetivos de negocio, al final de esta fase se debería poder tomar una decisión respecto de la utilización de los resultados.

Por último, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso de explotación de información repetible a través de toda la empresa.

Esta metodología define el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de datos desconocidos. Toma su nombre de las distintas etapas que conducen el proceso de explotación de información. SEMMA provee un proceso fácil de entender que permite el desarrollo y mantenimiento de proyectos de explotación de información organizado. [Britos, 2008] [Azevedo et al., 2008]

Las etapas involucradas en la metodología son: Muestreo (*Sample*) en la que se extrae la población muestral representativa sobre la cual se aplicará el análisis, Exploración (*Explore*) en donde se realiza una exploración de la información para simplificar el problema y así optimizar la eficiencia del modelo, Modificación (*Modify*) en la cual se modifican los datos de la base para que tengan el formato adecuado para la entrada del modelo, Modelado (*Model*) que permite modelar los datos permitiendo al software la búsqueda automática de una combinación de datos que predicen confiablemente las salidas deseadas y Valoración (*Assess*) que consiste en la valoración de los datos evaluando usabilidad y confiabilidad de lo encontrado en el proceso y estimando que tan bien se comporta.

La metodología P3TQ (Producto (*Product*), Lugar (*Place*), Precio (*Price*), Tiempo (*Time*) y Cantidad (*Quantity*)) según [Britos, 2008] está dividida en dos modelos, el Modelo de Negocio (MN) y el Modelo de Explotación de Información (MEI).

El primero de estos modelos “proporciona una guía de pasos para el desarrollo y la construcción de un modelo que permita identificar un problema de negocio o la oportunidad del mismo”, mientras que el segundo “proporciona una guía de pasos para la ejecución de modelos de Explotación de Información de acuerdo al modelo identificado en el (MN).

Ambos modelos poseen en su estructura los siguientes elementos: (a) una caja de actividades que indica una serie de pasos a realizar, (b) una caja de descubrimientos que provee acciones de exploración que se necesitan para poder decidir qué hacer en el próximo paso, (c) una caja de técnicas que proporciona información suplementaria sobre los pasos recomendados en las dos cajas anteriores y (d) una caja de ejemplos que dan una descripción detallada de cómo usar una técnica específica.

El modelado en (MN) depende de distintas circunstancias de negocio que promueven el planteo de 5 escenarios diferentes, Dato (el planteo comienza con una serie de datos y se debe explorar este conjunto para encontrar relaciones interesantes), Oportunidad (el planteo comienza con una

situación de negocio, problema u oportunidad, que debe ser explorada), Prospectiva (el proyecto se diseña para descubrir donde la Explotación de Información puede ofrecer un valore en el entorno de la organización), Definido (el proyecto comienza con la premisa de crear la especificación del modelo de explotación con un propósito específico) o Estratégico (el proyecto comienza con una estrategia de análisis para dar soporte a un escenario planificado por la organización).

Para el modelado en (MEI) se siguen los pasos: Preparación de los datos, Selección de herramientas y modelado inicial, Ejecución, Evaluación de resultados y Comunicación de resultados.

2. LINEAS DE INVESTIGACION y DESARROLLO

En el marco de este proyecto se investigará:

[a] Los límites, alcances y componentes del modelo Competisoft.

[b] Las distintas metodologías utilizadas para llevar a cabo proyectos de explotación de información.

[c] Los procesos que pueden ser utilizados sin modificaciones tanto en proyectos de desarrollo de software clásicos como en los de explotación de información. Además, aquellos que deben ser revisados y adecuados y, por último, los que deben redefinirse por completo.

3. RESULTADOS OBTENIDOS/ESPERADOS

El proyecto tiene como objetivo general la construcción de un modelo de procesos para proyectos de explotación de información en PyMES.

Con este modelo se espera contribuir a la mejora de la calidad de los proyectos de explotación de información buscando [Oktaba *et al.*, 2005] que:

[a] La estructura de procesos resultante esté acorde a la estructura empleada por la organización.

[b] Se custodie la integración y consistencia de los procesos y las relaciones entre ellos.

[c] Se enfatice la administración del proyecto desde un sólo proceso.

4. FORMACION DE RECURSOS HUMANOS

En el marco de este proyecto se esta desarrollando una Tesis de Maestría en Ingeniería en Sistemas de la Información en la Universidad Tecnológica Nacional y dos Tesis de Grado en Ingeniería Informática en la Universidad de Buenos Aires.

5. BIBLIOGRAFIA

Azevedo, A., Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS 2008.

Britos, P. (2008). Procesos de Explotación de Información basados en Sistemas Inteligentes. Tesis Doctoral. Facultad de Informática. UNLP.

Carnegie Mellon University, Software Engineering Institute (SEI) (2006). CMMI-DEV for Develop-ment, Vers. 1.2.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Larose, D. T. (2005). Discovering Knowledge in Data, an introduction to Data Mining. John Wiley & Sons.

Mas, A. y Amengual, E. (2005). La mejora de los procesos de software en las pequeñas y medianas empresas (PYME). Un nuevo modelo y su aplicación a un caso real. REICIS, Revista Española de Innovación, Calidad e Ingeniería del Software 1(2):7-29.

Oktaba, H., Piattini, M., Pino, F.J., Orozco, M.J. y Alquicira, C. (2008). Competisoft, Mejora de Procesos Software para Pequeñas y Medianas Empresas y Proyectos. Ra-Ma.

Oktaba, H., Garcia, F., Piattini, M., Ruiz, F., Pino y F.J., Alquicira, C. (2007). Software Process Improvement: The Competisoft Project. Computer 40(10): 21-28.

Oktaba, H., Alquicira Esquivel, C., Ramos, A. S., Martínez Martínez, A., Quintanilla Ozorio, G., Ruvalcaba López, M., López Lira Hinojo, F., Rivera López, M. E., Orozco Mendoza, M. J., Fernández Ordoñez, Y. y Flores Lemus, M. A. (2005). Modelo de Procesos para la Industria de Software. Secretaría de Economía de México..

Oktaba, H., Alquicira Esquivel, C., Ramos, A. S., Palacios Elizalde, J., Pérez Escobar, C. J. y López Lira Hinojo, F. (2004). Método de Evaluación de Procesos para la Industria de Software. Secretaría de Economía de México.

Pino, F. J., García F. y Piattini, M. (2006). Revisión sistemática de mejora de procesos software en micro, pequeñas y medianas empresas. Revista Española de Innovación, Calidad e Ingeniería de Software, Vol. 2, Nro. 1.