Un modelo diferente para responder encuestas económicas

A Different Model to Respond Economic Surveys

Genaro Acevedo García* y Guillermo Aguilar Sahagún**

* Roatech Services, SA de CV, genaro.agarcia@roatech.com.mx

** Consultor independiente, gaguilar27@yahoo.com.mx

Galo Cañas/cuartoscuro.com

El presente artículo tiene como objetivo describir un modelo práctico que permite la recolección y el procesamiento de datos para la generación de información estadística útil para el Instituto Nacional de Estadística y Geografía (INEGI). El modelo aprovecha los sistemas informáticos administrativos (contabilidad y nómina) que utilizan las empresas. Los resultados presentados son producto de un proyecto de investigación aplicada, financiado a través del Fondo Sectorial INEGI-CONACYT, cuyos propósitos son impulsar la generación de conocimiento y formar recursos humanos de alto nivel a través de la atención de una problemática presentada por el INEGI.

Palabras clave: minería de datos; encuestas económicas; mejora de procesos.

Recibido: 30 de septiembre de 2016
Aceptado: 8 de marzo de 2017

This article aims to describe a practical model that allows processing and data collection to generate useful statistical information for the National Institute of Statistics and Geography (INEGI). The model incorporates administrative (accounting and payroll) computer systems that companies use. The results hereby presented are the outcome of an applied research project, financed through INEGI-CONACYT Sectorial Funding, which is aimed to promote the generation of knowledge and training of high-level human resources by focusing on a problem presented by INEGI.

Key words: Data Mining; economic surveys; process improvement.

Definición del problema

La aparición en los últimos 20 años de múltiples tecnologías de la información y comunicación —como bases de datos relacionales, internet, lenguajes de programación con mayor capacidad para el desarrollo de sistemas computacionales, sistemas distribuidos y métodos para minería de datos, entre otras— permiten pensar en explorar nuevos modelos para la recolección de las encuestas económicas con las cuales el INEGI obtiene la información; como diría esta institución en su eslogan publicitario: "Alguien tiene que contar lo que pasa aquí" (en México).

En la actualidad, las empresas hacen uso de diversos y múltiples sistemas informáticos que les permiten apoyar los procesos relacionados con nómina, contabilidad, almacén, inventarios de activo fijo, etc.; mediante su empleo, registran sus operaciones o sucesos económicos, posibilitando con ello brindar información de lo que pasa dentro de ellas.

La mayoría de los sistemas informáticos que las empresas utilizan para el registro y control de sus operaciones han transitado del uso de archivos planos a manejadores de bases de datos relacionales (*SQL Server, Mysql, Oracle, Postgres,* etcétera). Estas tecnologías permiten alojar y ordenar una serie de datos de acuerdo con un criterio común a todos ellos y así facilitar su consulta y análisis.

Internet y los sistemas informáticos distribuidos, sumados a la tecnología de bases de datos, han permitido la construcción de un modelo en una versión inicial capaz de recopilar los datos contenidos en los sistemas informáticos, procesarlos, tomando como base las preguntas que integran las encuestas económicas del INEGI, o cualquier otra pregunta, y transferir vía internet las respuestas al mismo Instituto.

Para su desarrollo, y derivado de la fase de recopilación y análisis de la información, se observaron y establecieron algunas características que nuestro modelo debería cumplir:

- Ser original tanto nacional como internacionalmente, es decir, que no existieran experiencias en otros países que permitieran abordar la problemática planteada por el INEGI.
- Aprovechar el hecho de que, como parte de los procesos de construcción de las encuestas del INEGI, se aparecían preguntas similares (semántica) con redacción diferente (sintaxis).
- Por el lado de las empresas, la heterogeneidad en la construcción y definición de sus propios catálogos contables, la diversidad de términos utilizados en la captura de las operaciones y, finalmente, el perfil del personal destinado a contestar los cuestionarios.
- Por lo que toca a los sistemas empresariales (*softwares* disponibles en el mercado), fue evidente la existencia de múltiples marcas, lenguajes de desarrollo de *software*, variados manejadores de bases de datos y múltiples sistemas operativos.

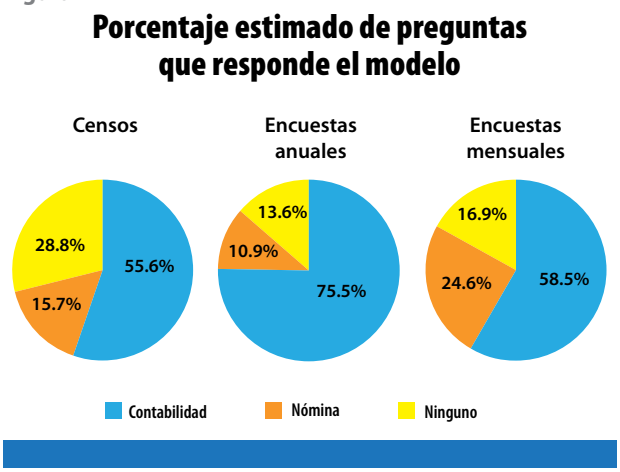
Condiciones de diseño de la solución

También, se establecieron condiciones de diseño iniciales y de frontera que el modelo debería contemplar, como: hacer uso de tecnología abierta y permitir una operación desconcentrada, con la capacidad de ser replicable y fácil de escalar.

Durante la etapa de análisis de la información, se identificó que los sistemas empresariales que apoyan los procesos de contabilidad y nómina son la fuente de información primaria para procesar y dar respuesta a un porcentaje considerable de las encuestas económicas (ver figura 1).

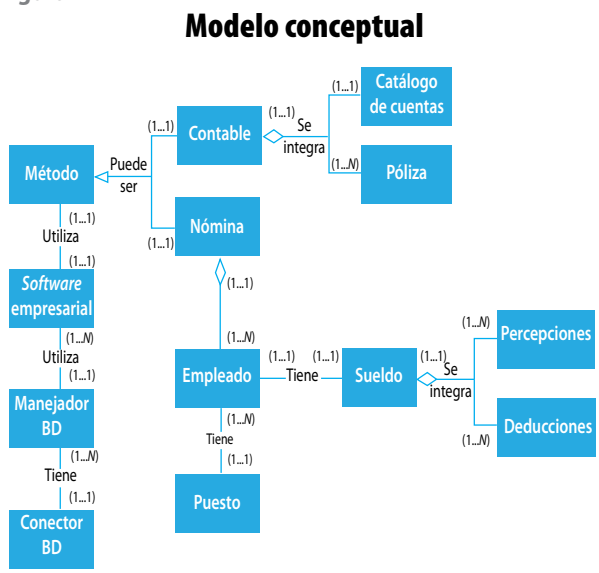
En su diseño y desarrollo, los diversos sistemas informáticos comerciales han identificado y modelado aquellos conceptos, objetos y procedimientos propios de los métodos contables y de nómina. .

Figura 1



Fuente: elaboración propia (2016).

Figura 2



Fuente: elaboración propia (2016).

Partiendo de esto, se han identificado conceptos de uso generalizado, por ejemplo: la póliza y el catálogo contables, entre otros.

Una de las características relevante de nuestro modelo es que posee la capacidad de generalizarse para hacer uso de múltiples sistemas comerciales a través de la identificación de conceptos y su integración en un modelo conceptual, lo cual ha sido posible gracias a que se ha modelado en estructuras de almacenamiento en las bases de datos (tablas y sus relaciones) y los atributos de los que se componen los conceptos involucrados (campos) donde se al ojan los datos correspondientes de los eventos que registran las empresas, de modo tal que esto constituye un proceso de abstracción y generalización (ver figura 2).

La solución

Los retos de semántica y sintaxis de las preguntas, la heterogeneidad en la construcción y definición de los catálogos contables, así como la diversidad de términos utilizados en la captura de las operaciones, se enfrentan en el modelo mediante la identificación y construcción de palabras clave y sinónimos asociados a los principales conceptos que figuran en las preguntas por responder que integran la encuesta.

Todo lo mencionado es el fundamento para mostrar que el modelo desarrollado contiene o se compone de una serie de elementos que interaccionan para hacer una extracción y procesamiento de la información contenida en los sistemas informáticos, para así dar respuestas y enviar el resultado del procesamiento a través de internet al INEGI.

A continuación, se describe cada uno de los elementos relevantes para el modelo:

- Cuestionarios: permiten la construcción de interrogantes mediante la integración de atributos propios y la asociación de n preguntas a un caso en particular.
- Preguntas: capa que permite la construcción de las interrogantes y la asociación de sus conceptos y sinónimos; de acuerdo con la semántica de éstas, se puede saber qué se desea buscar y cómo es posible dar las respuestas correspondientes.
- Métodos: cada una de las preguntas y sus sinónimos buscan responder un cuestionamiento específico, que en muchas de las ocasiones su respuesta está contenida en algún método contable, de nómina o de inventarios; la identificación y asociación de la pregunta con el método fuente es un aspecto importante para la automatización de la solución.
- Empresa: otro elemento es la construcción de instancias de empresas y sus respectivas asociaciones con los conceptos establecidos en el modelo conceptual; su implementación permite construir respuestas por industrias, empresas, situación geográfica, etcétera.
- Sistemas: la identificación de éstos o *software* empresarial, el método y los patrones de conceptos que aplica son importantes para definir e identificar los datos que aportará el sistema a la solución.
- Bases de datos: sus manejadores, las estructuras de almacenamiento y su asociación directa con los conceptos y su correspondiente mapeo son fundamentales para establecer el patrón de extracción y procesamiento de los datos; conocer cómo y dónde se almacena la información producto del día a día de las empresas convierte a las bases de datos en una mina de información que puede ser aprovechada por la solución propuesta; lo anterior se afirma partiendo de la premisa de que el Instituto desea conocer a las empresas en cifras, las cuales se almacenan diariamente.
- Algoritmos de extracción de datos: otro de los elementos fundamentales es la implementación de algoritmos que recurren al uso de técnicas de minería de datos y su interacción con las estructuras de éstos mediante conectores para, una vez obtenidos, proceder a su procesamiento; estos métodos debieron contemplar la heterogeneidad de las bases de datos que manejan los diferentes sistemas utilizados por las empresas, lo que representó un reto para obtener resultados acertados.
- Aplicación: el último elemento del modelo de solución es la construcción de una aplicación o sistema informático mediante lenguajes abiertos que permite la implementación de algunas de las capas previamente mencionadas.

Métodos de búsqueda de datos

La construcción de éstos es algo fundamental para la consulta y procesamiento de los datos fuente. Para ello, se recurre a algunos de los métodos empleados en la minería de datos con el propósito de encontrar palabras dentro de otras estructuras y que, en este caso particular, se puedan extraer mediante el uso de la explotación de las bases. Los que se aplicaron son:

- Método sintáctico.
- Método relacional declarativo.
- Método de búsqueda flexible.

La minería de textos es un proceso para descubrir nueva información a partir de un conjunto de documentos de cualquier tipo, en los que se lleva a cabo la búsqueda de patrones dentro de la redacción. Para esto, es necesaria la realización de diversas tareas, como: la categorización, la clasificación y el agrupamiento.

Los métodos de búsqueda están ligados a las estructuras dentro de los sistemas empresariales que almacenan los datos generados a través de los métodos administrativos y contables. Para definir este enlace, se requiere comprender las dimensiones en que se puede extraer la información relacionada con la contabilidad y la nómina para determinar a partir de cuáles registros se pueden construir las respuestas a las preguntas de los cuestionarios.

Por la naturaleza del proyecto y el tipo de proceso que debe realizarse para ubicar los datos y procesarlos de acuerdo con los criterios de tiempo, característica del dato y concepto, es que el modelo cuenta con los siguientes métodos (ver figura 3):

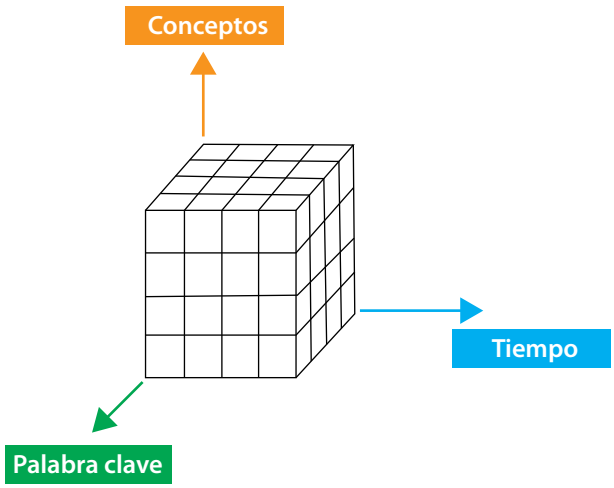
- Para la contabilidad, la búsqueda y armado de respuestas se lleva a cabo en tres dimensiones:
  - La temporalidad, esto es, en virtud de que la póliza contable cuenta con una fecha en la que se realiza la transacción es posible armar respuestas por día, mes, bimestre, trimestre, cuatrimestre, semestre y año.
  - Por profundidad, es decir, mediante la navegación del catálogo de cuentas contables se puede procesar información a distintos niveles.
  - Por un tema o dato concreto contenido en la póliza contable en el espacio destinado para referenciar la descripción de la transacción.
- Para la nómina, la búsqueda se hace de acuerdo con la información que se genera en los recibos de pago que se les entrega a los trabajadores. De este modo, el armado de respuestas se puede realizar en tres dimensiones:
  - Por tiempo, entendiendo los periodos en los que se generan los recibos de nómina.
  - Por agrupación, esto es, considerar de forma agrupada por características de los trabajadores, por ejemplo, cantidad de empleados por género.
  - Por acumulados, se puede obtener la información por rubro dentro de las percepciones y las deducciones de los trabajadores.

Algoritmos

El objetivo de construirlos es poder relacionar una pregunta, tomando los conceptos extraídos de la redacción de la pregunta, con la información que está almacenada en las estructuras de datos que utilizan los sistemas empresariales (ver figura 4).

Figura 3

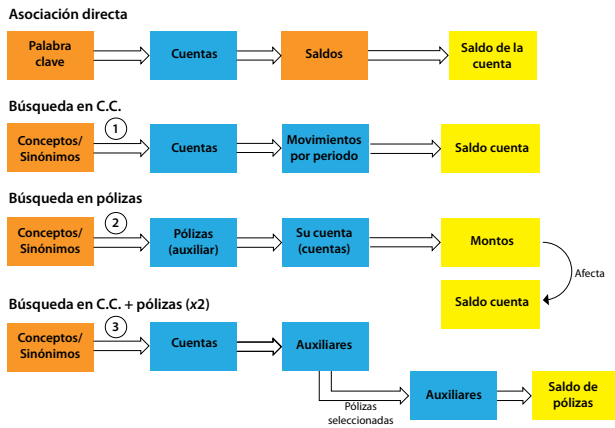
Dimensiones de búsqueda



Fuente: elaboración propia (2016).

Figura 4

Esquema de búsquedas



Fuente: elaboración propia (2016).

Para la implantación de los métodos de búsqueda en la contabilidad, se generaron diferentes algoritmos:

- El primero está centrado en correlacionar la pregunta con alguna de las cuentas acumulativas del catálogo contable; una vez efectuado esto, se puede obtener la respuesta a la pregunta a través de los saldos asociados a la cuenta.
- El segundo busca correlacionar los con-ceptos de las preguntas con las cuentas contables acotando los resultados solo a las cuentas en que se tenga un mayor número de ocurrencias de los conceptos involucrados.
- El último es para correlacionar la pregunta con la información contenida en la póliza contable. De esta acción se puede obtener la respuesta a la pregunta mediante el procesamiento de los datos contenidos en las pólizas contables (cuenta contable, descripción, monto, fecha).

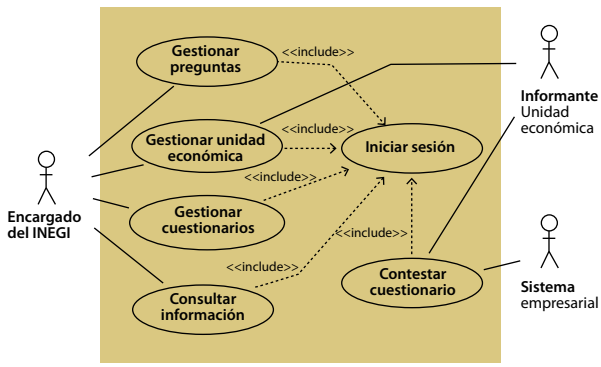
Para la implementación de los métodos de búsqueda en la nómina, se describen los siguientes algoritmos:

- El primero busca una correspondencia de los conceptos dentro de una pregunta con los de las percepciones y deducciones registradas en el sistema; así se obtienen los montos acumulados de los trabajadores a una fecha delimitada.
- El segundo está centrado en correlacionar la pregunta con atributos generales de los empleados, como: estudios, género, edad.

La aplicación se integra de una serie de módulos (casos de uso) diseñados y programados ex profeso a un sistema informático, el cual permite interactuar con los diversos actores (informante, INEGI y sistema empresarial); esta integración se muestra esquemáticamente en la figura 5.

Figura 5

Casos de uso proyecto INEGI



Fuente: elaboración propia (2016).

El sistema está construido sobre lo que se denomina arquitectura tecnológica, que se integra por protocolos de comunicación (TCP/IP), manejadores de bases de datos, *Web servers* y *Web services*, así como aplicaciones *stand alone*.

Esta arquitectura está organizada en dos secciones, la primera es el Nodo INEGI, que opera con base de datos *SQL Server o Mysql* y aplicación diseñada para funcionar en *Tomcat, Jboss, TomEE* y *Windfly* (*opensource jboss*), y se puede ejecutar en sistemas operativos *Windows*, *Linux* y *Mac*.

Por lo que respecta a la segunda sección, es el Nodo Cliente, que funciona con base de datos *SQL Lite* (base de datos embebida); los algoritmos de extracción funcionan para bases de datos *SQL Server*, *Firebird* y *Mysql,* y pueden ejecutarse en sistemas operativos *Windows*, *Linux* y *Mac* (ver figura 6).

Dentro del modelo se incorpora lo que se ha denominado *Base de conocimiento*, la cual se ubica por un lado en el Nodo Cliente, donde se genera un banco de información que concentra los datos y mantiene un histórico que puede ser reutilizable. Por el otro lado se encuentra el Nodo INEGI, donde la *Base de conocimiento* podrá recolectar la información de todas las empresas y, mediante un *Data Warehouse*, generar información sobre cómo mejorar en la homologación de los conceptos contables y administrativos la redacción de las preguntas.

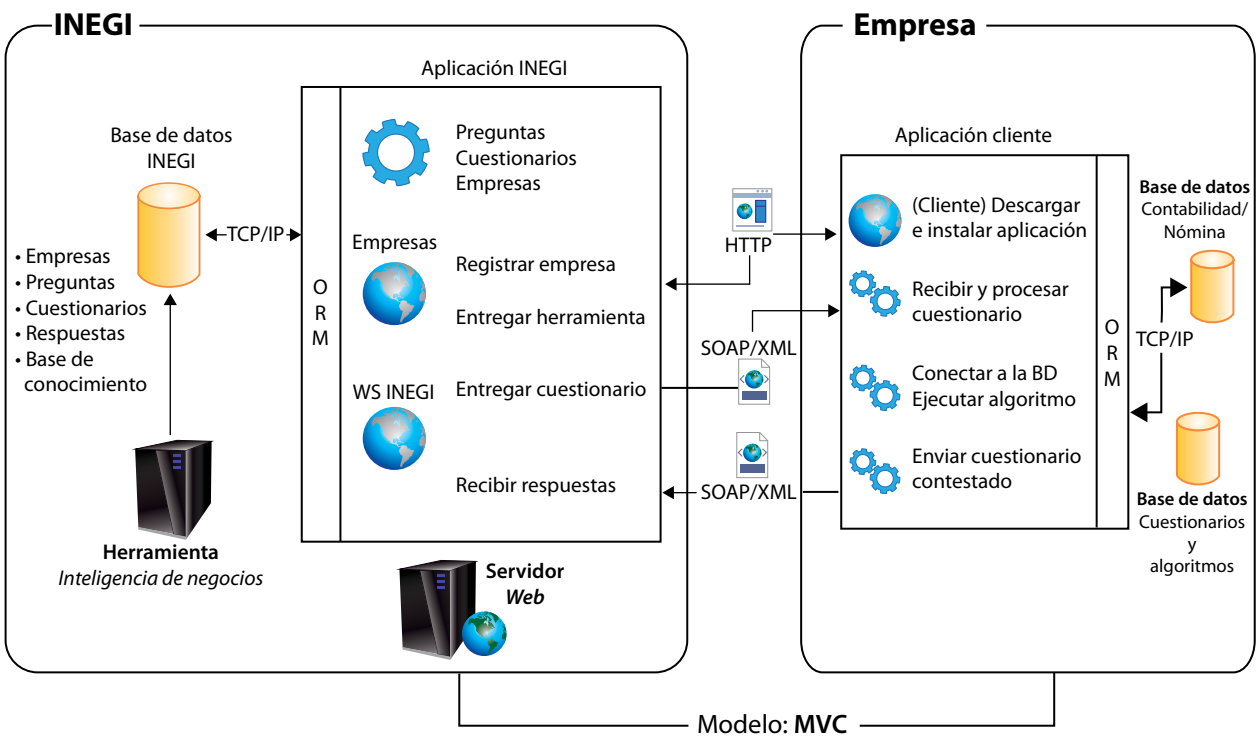
Finalmente, con el propósito de poder comunicar el grado de madurez del modelo, se recurre al apoyo de los niveles de disponibilidad de tecnología (TRL, por sus siglas en inglés), el cual se puede ubicar en el 4 y que, para su uso práctico, deberá de transitar por los siguientes niveles:

TRL 1: idea básica.
TRL 2: concepto o tecnología formulados.
TRL 3: prueba de concepto.
TRL 4: validación a nivel de componentes en laboratorio.

TRL 5: validación a nivel de componentes en un entorno relevante.
TRL 6: validación de sistema o subsistema en un entorno relevante.
TRL 7: validación de sistema en un entorno real.
TRL 8: validación y certificación completa en un entorno real.
TRL 9: pruebas con éxito en entorno real.

Figura 6

Arquitectura tecnológica



Fuente: elaboración propia (2016).

Resultados

Por último, se realizaron pruebas del modelo en su estado de prototipo con dos empresas diferentes con las contabilidades y nóminas de éstas: se determinó que la propuesta se acerca a responder las preguntas en un porcentaje cercano a las estimaciones de preguntas que podrían resolverse a través de este método, lo que estima su viabilidad de escalar.

Conclusiones

Este modelo permite, para la recolección automatizada de datos, establecer la relación desde los cuestionarios hasta la fuente de datos con el fin de atender a cada elemento e integrar las partes de la solución.

Los métodos de búsqueda logran la extracción desde las fuentes de datos de las empresas de manera exitosa; sin embargo, aún depende de que el informante realice la validación y la asignación de respuestas a cada pregunta del cuestionario que se esté aplicando.

Se plantea incorporar conceptos de contabilidad y nómina de acuerdo con la semántica y sintaxis de la redacción de las preguntas, sin modificar conceptualmente sus diferencias y particularidades.

La aplicación del modelo permitiría al INEGI disminuir costos, tiempo y esfuerzo en la recolección de datos.

Referencias

Brookshear, J., R. Escalona y J. Dorronsoro (1995). *Introducción a las ciencias de la computación*. Iberoamérica: Addison-Wesley.

IMIPE (2014). *Manual de contabilidad*. Consultado el 23 de junio de 2015 en: http://www.imipe.org.mx/pdf/transp/manualconta.pdf

INEGI (2007). *Síntesis metodológica Encuesta Anual del Comercio*. Consultado el 26 de junio de 2015 en: http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod_serv/contenidos/espanol/bvinegi/productos/metodologias/est/sm_eac2005.pdf

_____ (2014). *Síntesis metodológica de la Encuesta Mensual de la Industria Manufacturera. EMIM SCIAN 2007. Versión 2014*. Consultado el 13 de junio de 2015 en: http://www.inegi.org.mx/prod_serv/contenidos/espanol/bvinegi/productos/metodologias/EMIM/EMIM2014/SM_EMIM_2014.pdf

_____ (2015). *Encuesta Anual del Comercio. Síntesis metodológica*. Consultado el 28 de julio de 2015 en: http://www.inegi.org.mx/est/contenidos/espanol/proyectos/metadatos/encuestas/eaec_224.asp?s=est&c=10568

_____ (2014). *Censos Económicos 2014*. Resultados oportunos. Consultado el 16 de julio de 2015 en: http://www.inegi.org.mx/est/contenidos/proyectos/ce/ce2014/

Zúñiga, O. (2015). *La nómina en México*. Consultado el 8 de julio 2015 en: http://pymerang.com/administracion-de-empresas/recursos-humanos/funciones-de-recursos-humanos/evaluacion -y-retribucion/91-la-nomina-en-mexico

Rozanski, N. & E. Woods (2005). *Software Systems Architecture*. Pearson Education, Inc.

Saralegui, J., C. González & I. Arbués (2012). *Uso de fuentes administrativas para la reducción de carga y costes en las encuestas estructurales de empresas (UFAES)*. Consultado el 28 de junio de 2015 en: http://www.ine.es/ss/Satellite?L=es_ES&c=INEDocTrabajo_C&cid=1259940239219&p=1254735839320&pagename=MetodologiaYEstandares%2FINELayout

UNECE (2013). *Generic Statistical Business Process Model*. Consultado el 2 de julio 2015 en: www.istat.it/it/files/2013/12/GSBPM-5_0.pdf

Universidad de Guadalajara (2004). *Catálogo de cuentas como herramienta de aprendizaje contable*. Consultado el 11 de junio de 2015 en: http://www.cucea.udg.mx/publicaciones/pdfs/catalogo_cuentas.pdf