FORMATO DE IDENTIFICACION DE PONENCIAS

1. **CÓDIGO DE LA COMUNICACIÓN: 108**

2. **TÍTULO COMPLETO:**

Gestión de datos e información para el análisis de la innovación a partir de encuestas: aprendizaje y retos

3. EJE TEMÁTICO: 8. HERRAMIENTAS DE APOYO A LA GESTIÓN DE LA I+D E INNOVACIÓN. 8.6. Las tecnologías de información y comunicación como soporte a la innovación.

4. AUTORES:

APELLIDO, NOMBRE: Jiménez Hernández, Claudia Nelcy
INSTITUCIÓN: Universidad Nacional de Colombia, Sede Medellín
EMAIL: cnjimenezh@unal.edu.co
PAÍS: Colombia

APELLIDO, NOMBRE: Rico Herrera, Martín Elías
INSTITUCIÓN: Universidad Nacional de Colombia, Sede Medellín
EMAIL: ericom@unalmed.edu.co
PAÍS: Colombia

APELLIDO, NOMBRE: Jiménez Ramírez, Claudia Stella
INSTITUCIÓN: Universidad Nacional de Colombia, Sede Medellín
EMAIL: csjimene@unal.edu.co
PAÍS: Colombia

5. RESUMEN

Para la política y gestión de la innovación tecnológica es fundamental contar con información de calidad que permita caracterizar y analizar las dinámicas en este campo. La necesidad de manejar eficientemente y obtener mayor provecho de los datos relacionados con la innovación en Colombia, dio origen a un proyecto de investigación que busca generar conocimiento y plantear propuestas de política en el tema a partir de encuestas e información recolectada por diferentes entidades gubernamentales. El objetivo de esta ponencia es describir el proceso de gestión de los datos e información requeridos en el descubrimiento de conocimiento sobre la

innovación. Así mismo, se presenta el aprendizaje generado en torno a este proceso, y a partir de dicha experiencia se formulan algunas propuestas para lograr mayor efectividad en el análisis de la innovación y el desarrollo tecnológico en Colombia y la validez en los resultados o inferencias.

6. TRABAJO COMPLETO

Gestión de datos e información para el análisis de la innovación a partir de encuestas: aprendizaje y retos

Introducción

Autores como Burton-Jones y Gelauff (2003) han señalado que la economía está pasando de la manufactura y la producción de bienes tangibles al manejo de información, la generación de conocimiento y la producción de bienes simbólicos o intangibles, con el empleo creciente de tecnologías de información y comunicaciones (TIC), que se despliegan como tecnologías de propósito general, junto a nuevas formas sociales basadas en el conocimiento.

En el actual contexto de la llamada *Sociedad del Conocimiento*, resulta imprescindible el análisis de información que lleve a la creación, acumulación y asimilación de nuevo conocimiento. En el proceso de gestión de datos e información, la identificación de relaciones entre diferentes situaciones, descritas por los datos recopilados, es una actividad clave para la solución de problemas en las organizaciones, permitiendo la optimización de recursos y el soporte a la toma de decisiones. Davenport y Prusack (1998) establecen que los datos se convierten en información cuando son ordenados, estandarizados bajo reglas predeterminadas y mostrados de una forma que permita la comunicación a través de representaciones del lenguaje (gráficas, tablas, etc.). Se argumenta que la información es un conjunto de datos al cual se le ha añadido valor a partir del desarrollo de tareas de contextualización, categorización, cálculo, corrección y condensación (Tiwana, 2000, citado en Baker y Badamshina, 2002). Para el procesamiento o estructuración de datos e información, las TIC son una herramienta facilitadora, debido a su capacidad para realizar las distintas operaciones de generación de valor en períodos de tiempo mucho menores y disminuyendo considerablemente la probabilidad de error en los cálculos.

La presente ponencia tiene el propósito de describir el proceso seguido en el marco de un proyecto de investigación sobre la innovación en Colombia financiado por Colciencias, respecto al manejo de los datos e información provenientes de fuentes como las Encuestas Nacionales de Innovación y Desarrollo Tecnológico, la Encuesta Anual Manufacturera y las bases de datos sobre investigadores y grupos de investigación colombianos de la plataforma Scienti. Dentro del mencionado proyecto,

se trabaja en la creación de una bodega de datos relativos a la innovación, lo cual permitirá mejorar la disponibilidad de dichos datos e información, así como aplicar diferentes tipos de técnicas para su análisis, como las conocidas bajo el nombre de *Descubrimiento de Conocimiento en Bases de Datos* (*Knowledge Discovery in Databases* o abreviadamente KDD).

El proceso que aquí se describe hace referencia a la gestión de datos e información y se ha entendido como aquel que incluye operaciones de extracción, manipulación, tratamiento, depuración y conservación de la información adquirida por una organización a través de diferentes fuentes, y que gestiona el acceso y los derechos de los usuarios sobre la misma (Information Management, 2006). Con base en la caracterización de este proceso, se evidencia el aprendizaje generado a lo largo de su desarrollo así como las falencias detectadas, incluso desde la misma captura de los datos por parte de los organismos responsables, las cuales influyen en la calidad de la información y por tanto, en la legitimidad de los análisis. A partir de allí, se formulan recomendaciones, que se constituyen en retos tanto para las entidades gubernamentales proveedoras de los datos, como para los investigadores en los temas relativos al proyecto.

1. Descripción del proceso de gestión de datos e información

En el marco del proyecto "Descubrimiento de Conocimiento sobre la Innovación en Colombia a partir de las Encuestas de Innovación y Desarrollo Tecnológico, la Encuesta Anual Manufacturera y la base de datos ScienTI", se ha desarrollado un proceso de manejo de datos e información, en el que se ha creado una bodega de datos para integrarlos y facilitar su análisis mediante técnicas estadísticas y de Inteligencia Artificial, empleando el enfoque de Inteligencia de Negocios. Este enfoque, además de utilizarse en entornos gerenciales, se considera apropiado en análisis de información de tipo institucional para el diseño y evaluación de políticas y estrategias públicas o para el trabajo investigativo en cualquier área de conocimiento (Jiménez *et al*., 2008).

En el proceso de gestión de datos e información del proyecto en mención, se identificaron cinco etapas: revisión de la estructura y semántica inicial de cada fuente de datos; diseño de la estructura de la bodega de datos y adecuación de datos; enriquecimiento inicial de datos; depuración de errores y documentación del proceso. En los siguientes acápites se describe en qué consiste cada etapa, presentando como ejemplo las características más relevantes de su aplicación a una de las fuentes de datos del proyecto: la primera Encuesta de Innovación y Desarrollo Tecnológico (EIDT I).

1.1. Revisión de la estructura y semántica inicial de cada fuente de datos

Esta primera etapa del proceso de gestión de datos e información contempló la exploración de las bases de datos y las hojas electrónicas suministradas por el Departamento Administrativo de Ciencia, Tecnología e Innovación, Colciencias, el Departamento Nacional de Estadísticas, DANE, y el Departamento Nacional de

Planeación, DNP. Se estudiaron los tipos de archivos recibidos con el objeto de describir su contenido y presentación, considerando los atributos existentes, el volumen de información y el formato de cada archivo, entre otros aspectos.

En esta primera etapa se evidenciaron algunas deficiencias en la gestión inicial de los datos, realizada por parte de los proveedores de la información, como la falta de validación de los datos o su ausencia, así como algunas inconsistencias.

Para mostrar algunos de los hallazgos, en la **Error! Not a valid bookmark self-reference.** se resumen las principales características de la base de datos correspondiente a la primera Encuesta de Innovación y Desarrollo Tecnológico (EIDT I), efectuada por el DNP y Colciencias en 1996, en donde se observa el gran volumen de datos, dada la cantidad de establecimientos industriales encuestados y de preguntas que componen la encuesta. La información fue suministrada en libros electrónicos de Excel, clasificada en nueve agrupaciones, las cuales se identificaron como capítulos de la encuesta.

Tabla 1. Principales características de la base de datos de la EIDT I

Característica	Descripción
Formato de archivo	Excel
Campos existentes	472 preguntas 885 establecimientos
Volumen de información	417.720 datos (respuestas)

Fuente: elaboración propia con base en la información de la EIDT I

Al hacer la revisión de EIDT I se encontraron problemas relativos a la integridad referencial que consistió encontrar algunas preguntas que no tenían respuesta asociada, así como la situación inversa. De igual manera, se constató la ausencia de datos para relacionar cada establecimiento encuestado con el área geográfica donde estaba ubicado, así como falta de información sobre el tamaño de algunos establecimientos.

1.2. Diseño de la estructura de la bodega de datos y adecuación de los datos

Una bodega de datos se define como una colección de datos históricos, orientados a un dominio, integrados y organizados para dar soporte objetivo a la toma de decisiones (Inmon, 2002). El diseño de la bodega de datos para el proyecto de investigación sobre la innovación en Colombia, consideró las características de las fuentes de datos disponibles, así como otras variables requeridas para un mejor manejo de los mismos.

La forma en la que se optó por almacenar los datos en la bodega de datos resultó distinta a aquella cómo estaban dispuestos en las fuentes originales, lo que implicó un importante trabajo manual y técnico para organizarlos, de tal manera que se acomodaran al diseño planteado. Durante este proceso se observó que las bases de

datos estaban incompletas o simplemente algunas variables no fueron consideradas, tal es el caso de las *temáticas* para la primera y la segunda encuestas de innovación y desarrollo tecnológico, las cuales no existían y fueron identificadas para este proyecto en el formulario de cada encuesta; de igual manera se encontraron problemas asociados a la presentación de los datos y la información, por ejemplo, la redacción de algunas de las preguntas de la EIDT I, para lo cual se efectuaron correcciones teniendo en cuenta el formulario de dicha encuesta.

1.3. Enriquecimiento inicial de datos

La bodega de datos del proyecto sobre la innovación en Colombia se diseñó para atender las expectativas de los investigadores participantes, puesto que se hizo pensando en el manejo de los datos por parte de cualquier usuario, no solo de conocedores del lenguaje y las herramientas informáticas. Además, se pretendía que esta bodega incluyera los diferentes trabajos técnicos e investigativos realizados alrededor de la temática de la innovación y que facilitara una posterior actualización.

Por lo anterior, y con el propósito de ampliar la descripción de los datos originales, se identificaron cuáles deberían ser los metadatos básicos (características o datos acerca de los datos existentes) relacionados con cada pregunta, el tipo de pregunta, las opciones de respuesta, las unidades de medida, las preguntas filtro y su descripción, así como los capítulos y subcapítulos o temáticas en las cuales están agrupadas o pueden agruparse las preguntas y respuestas.

Como ejemplo, en la Tabla 2 se presenta una de las 74 temáticas identificadas en la primera Encuesta de Innovación y Desarrollo Tecnológico, con la descripción asignada, es decir, el metadato correspondiente.

Tabla 2. Ejemplo de metadatos de la base de datos de la EIDT I

Temática EIDT I	Descripción
Inversión en proyectos de I+D entre 1993 y 1995	Identifica los recursos invertidos en actividades de I+D (humanos, económicos, tiempo) así como las fuentes de financiación empleadas. También indaga sobre la existencia de laboratorios de I+D y el impacto de las actividades de I+D en la innovación

Fuente: elaboración propia con base en la información de la EIDT I

El enriquecimiento de la bodega de datos con los metadatos permite mejorar la gestión de la información y por ende, ser más efectivos en el proceso de descubrimiento de nuevo conocimiento. Le facilita al investigador la selección de las variables o atributos que se deben considerar, la elección de las técnicas de Minería de Datos que deben ser aplicadas en un determinado análisis. Además de esto, el enriquecimiento de datos ayuda en la visualización de los resultados y, por tanto, en la comprensión del conocimiento minado. Por ello, esta fase se convirtió en parte

fundamental del proceso de gestión de datos e información, al considerar su utilidad tanto a corto como a largo plazo.

1.4. Depuración de errores

En esta etapa del proceso de gestión de datos e información se emplean diversas técnicas y procedimientos para identificar fallas o datos faltantes e implementar acciones correctivas, de manera que se incremente su calidad, como condición *sine qua non* para la validez de las inferencias obtenidas a partir de dichos datos.

Algunas de las fallas en los datos pueden deberse a la estructura de las propias encuestas. En el caso de la EIDT I, por ejemplo, se encontró que se formularon dos preguntas para determinar si las fuentes de innovación son de origen nacional o extranjero, en lugar de consultarlo usando una sola pregunta, como se muestra en la Tabla 3.

Tabla 3. Ejemplo de la detección y depuración de un error en la EIDT I

Código	Variable	Cambio o acción	Solución dada
IV416B01	Pregunta del capítulo 4, para conocer si el origen de la fuente de innovación es nacional (=1) o no (=2)	Unificación de las preguntas	El origen de la fuente de innovación es nacional (=1) o extranjero (=2)?
IV416C01	Pregunta del capítulo 4, para conocer si el origen de la fuente de innovación es extranjero (=1) o no (=2)		

Fuente: elaboración propia con base en la información de la EIDT I

A continuación se describen algunos de los procedimientos seguidos en la depuración de los datos, considerando distintos aspectos.

1.4.1 Incumplimiento de reglas de integridad

Las reglas de integridad garantizan una adecuada correspondencia y relación entre los datos, de tal manera que es parte vital para certificar la veracidad de los mismos. La integridad se maneja en gran medida gracias a las ventajas de los sistemas de almacenamiento, en asocio con unas definiciones preestablecidas.

Durante el proceso de depuración de los datos, se observó la falta de manejo de reglas de integridad, lo que refleja dificultades en la gestión inicial de la información por parte de las entidades encargadas de la recolección y almacenamiento de la misma, como ya se había mencionado previamente. Estas fallas generaron repercusiones sobre el trabajo realizado en el proyecto con respecto a la construcción de la bodega de datos, puesto que originaron demoras y en algunos casos no fue posible trabajar con la integridad referencial planteada en el diseño de esta.

El caso más particular ocurre en la primera Encuesta de Innovación y Desarrollo Tecnológico EIDT I, donde se detectó, empleando las bondades de los sistemas de almacenamiento, que existían preguntas que no tenían respuesta por parte de ninguno de los establecimiento encuestados, o que no estaban contempladas como una pregunta dentro de los archivos donde se encuentran las correspondientes respuestas; sin embargo, lo más destacado fue encontrar respuestas sin una pregunta asociada, es decir que en los archivos respectivos no existían preguntas definidas para dichas respuestas.

1.4.2 Análisis de inconsistencias

El análisis de inconsistencias busca la identificación de discrepancias en códigos, nombres o relaciones entre los datos. Para lograr esto, en el proyecto de investigación sobre la innovación se cumplieron dos pasos: el primero basado en la revisión manual de los datos, y el segundo posterior a la carga de la información en la bodega de datos, realizando consultas mediante sentencias SQL (*Structured Query Language* – Lenguaje de consulta estructurado) que constituyen tareas de Minería de Datos.

Para el caso de la EIDT I, en el primero de estos pasos, mediante la revisión manual de los datos se detectaron inconsistencias como las siguientes:

- Existencia de un código para identificar el tamaño de los establecimientos, que no estaba definido (se encontró que en realidad dicho código se asociaba a los establecimientos que no respondieron a la pregunta sobre su tamaño)
- Códigos originalmente creados para referirse a respuesta negativa, pero que fueron utilizados para expresar otras circunstancias (ausencia del dato en el establecimiento).

Luego de la carga de los datos de la EIDT I a la bodega de datos, se hicieron consultas con sentencias SQL, y a través de ello se evidenció la existencia de preguntas sin respuesta asociada y viceversa, lo cual fue señalado previamente en esta ponencia.

1.4.3 Verificación de la completitud de los datos y la información – valores atípicos

En etapas previas del proceso de gestión de datos e información se estableció que existían problemas tanto en la recolección de datos como en su transferencia al grupo de investigación, que ocasionaron la pérdida de cierta información relevante. Se comprobó que en la base de datos de la EIDT I recibida del DANE, por ejemplo, no existían datos que permitieran relacionar cada establecimiento encuestado con el área geográfica donde se ubica o conocer el tamaño de algunos de los establecimientos industriales. Esta información es muy relevante para los análisis comparativos que demanda la caracterización de la innovación en el país y que quedó incompleta, infortunadamente.

Por otro lado, una manera de detectar errores en los datos de un atributo cuantitativo consiste en determinar los valores extremos para saber si existen valores por fuera del dominio, puesto que el tipo de dato puede ser insuficiente para restringirlos. Los valores extremos pueden ser correctos aunque atípicos, y por eso, se deben analizar para determinar si corresponden a la realidad y para tomar medidas correctivas cuando no sea así. La ocurrencia de valores extremos puede tener drásticos efectos en la minería de los datos, si no son considerados (Weisberg, 2005). A manera de ejemplo, en las respuestas a la pregunta de la EIDT I sobre el porcentaje de las ventas destinado a la inversión en I+D, se detectaron dos valores extremos (92% y 87%) que pueden catalogarse claramente como errores de digitación, tal como se aprecia en la Figura 1.

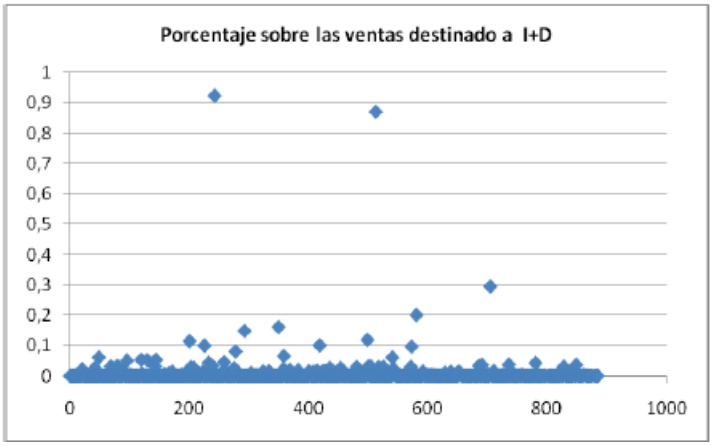

Figura 1. Diagrama de dispersión con valores atípicos

Fuente: elaboración propia con base en la información de la EIDT I

1.4.4 Medidas correctivas

En el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), se proponen algunas acciones para manejar los inconvenientes o errores en los datos. Estas acciones se realizan en la etapa de preprocesamiento de los datos conocida como limpieza de datos (*Data Clearing)* y se presentan en la **Error! Reference source not found.**.

Tabla 4. Acciones incluidas en la limpieza de datos

Denominación	Acciones incluidas
Detección de ruido	Análisis de distribución de datos con histogramas Análisis de clúster Funciones de regresión
Manejo de inconsistencia de datos	Uso de dependencias funcionales conocidas Revisión del proceso de integración de diferentes fuentes, donde frecuentemente se manejan distintos nombres para el mismo atributo
Manejo de datos ausentes	Uso de la media del atributo Uso del dato más probable basado en los datos existentes (árboles de decisión, regresión bayesiana)

Fuente: Goharian y Grossman (2003)

1.5. Documentación del proceso

Como etapa final del proceso de gestión de datos e información, se incluye aquella que permite la generación de metadatos semánticos para la bodega de datos del proyecto de investigación sobre la innovación en Colombia, que permiten asignar un significado a la información. Si bien es la última etapa descrita, no significa que solo se haya desarrollado al final del proceso, ya que la documentación se ha ido generando en etapas previas e incluso, algunos de dichos documentos fueron provistos por los organismos gubernamentales encargados de las encuestas analizadas en este proyecto.

Como ejemplos de la documentación generada, pueden mencionarse los informes y boletines del DANE sobre las encuestas de innovación y de Colciencias sobre la plataforma Scienti, las ponencias y artículos preparados por los investigadores en el marco del proyecto de investigación sobre la innovación, así como esta ponencia.

2. Aprendizaje a partir del proceso de gestión de datos e información desarrollado

En primer lugar, se evidenció que el proceso de gestión de datos e información es complejo y que para su ejecución se requiere tanto de habilidades y conocimientos técnicos en el campo informático, como de vínculos con el contexto de la información que se está gestionando, en este caso, con el tema de la innovación y la gestión tecnológica, con el fin de garantizar la correspondencia entre los datos y la realidad que se intenta abstraer a través de ellos.

Así mismo, es fundamental que los datos que se constituyen en los insumos del proceso de KDD sean de calidad, lo cual justifica la inversión requerida en recursos humanos, económicos y de tiempo para su gestión. En este orden de ideas, si se minimizan las fallas tanto en la captura de los datos como en su presentación inicial, el proceso para gestionarlos se facilitará y agilizará en gran medida.

Por otra parte, debido a los interrogantes que pueden surgir en el proceso de gestión de los datos, los cuales pueden no ser tan fácilmente solucionables, se hace indispensable el contacto permanente con los proveedores de dichos datos, que favorezca el conocimiento de aspectos clave para su comprensión y manejo por

parte de terceros, en este caso, los investigadores del proyecto sobre la innovación en Colombia.

Teniendo en cuenta que, como se mencionó anteriormente, para el proyecto ha sido necesaria la participación tanto de expertos en el campo informático como en el tema de la innovación y la gestión tecnológica, se evidenció que, si desde las primeras etapas del proceso de gestión de datos e información se hubiese contado con una interfaz de usuario para acceder a la bodega de datos, se hubiese facilitado la participación de otras personas diferentes a los técnicos, agilizando el proceso.

Por último, cabe anotar que en la gestión de datos e información, particularmente en la generación de una bodega de datos, se observa que se trata de un desarrollo continuo, es decir que dicha bodega es dinámica, lo cual implica que haya ajustes permanentes en la estructura de los datos que se van cargando.

3. Retos en la gestión de datos e información sobre innovación en Colombia

Con base en la descripción del proceso de gestión de datos e información desarrollado en el marco de un proyecto de investigación sobre la innovación en Colombia, y en la identificación del aprendizaje logrado a través de dicho proceso, es posible sugerir cambios que se relacionan con el mejoramiento o definición de nuevos sistemas de recolección de los datos, mayor coordinación entre las entidades involucradas, capacitación y asesoría para la interpretación de los datos, entre otros. Por tanto, se formulan a continuación varios retos orientados a elevar los niveles de calidad de los datos manejados y a garantizar la validez de los análisis resultantes.

El primero de estos retos se formula para las entidades responsables de la captura de los datos sobre la innovación en el país, y se relaciona con el mejoramiento de procesos de gestión de dichos datos desde su misma recolección, más aún si se considera que quienes están interesados en usarlos, deben asumir un costo económico para obtenerlos. Por tanto, se plantea la generación de interfaces de captura de los datos donde se puedan definir reglas de integridad de dominio, referencial, de valores no nulos y demás reglas particulares que requiera una encuesta o conjunto de datos particular, por parte del personal de informática. Este trabajo implica una mayor inversión en las fases tempranas de la gestión de los datos e información, pero se justifica plenamente por la calidad de las descripciones e inferencias que se deriven de los mismos.

A manera de ejemplo, en la Figura 2 se presenta un formulario de captura de datos, en donde los errores de digitación son mínimos porque no demandan que el digitador recuerde números o convenciones para codificar los valores de las variables. También se puede chequear la validez de datos como la fecha de nacimiento, así como exigir que sea un dato obligatorio para impedir que algún encuestado carezca de él.

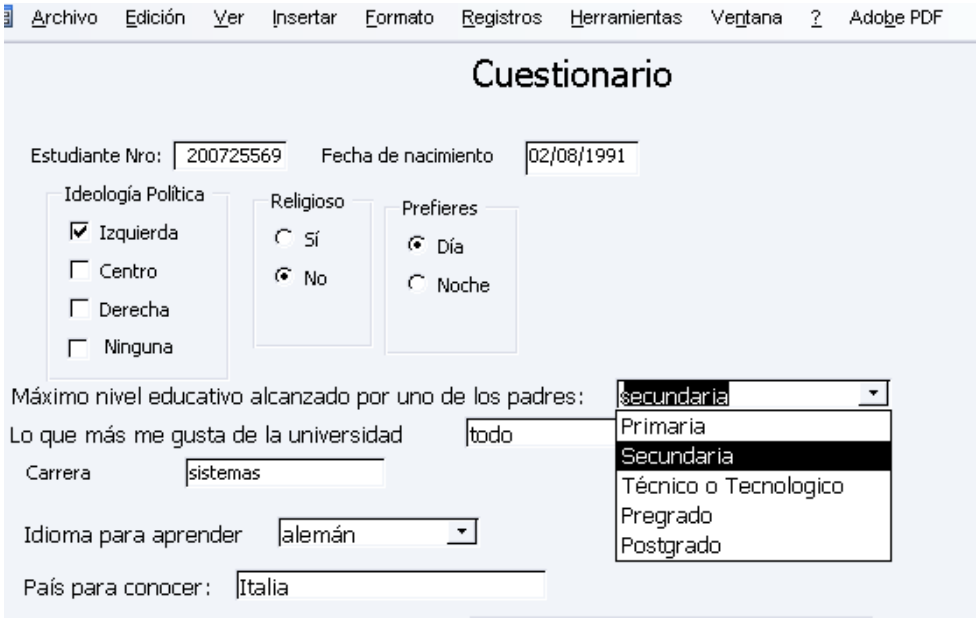

Figura 2. Ejemplo de una interfaz de captura de datos

Fuente: los autores

Por otra parte, se ha evidenciado la necesidad de mejorar la disponibilidad de los datos parte de las entidades del Estado involucradas en su captura y manejo, teniendo en cuenta que actualmente es necesario recurrir a distintas entidades y atender a sus gestiones internas, en muchas ocasiones poco ágiles, para obtener los datos de interés en la cantidad y con la calidad y presentación que estas determinen. Esto dificulta a investigadores y entidades como las participantes en el proyecto al que se hace referencia en este documento, para desarrollar de manera más eficiente sus trabajos de investigación. Por tanto, el mejoramiento que se plantea como reto, hace referencia principalmente a los sistemas de almacenamiento que se manejan actualmente, considerando además la posibilidad de centralizar la información.

De igual manera, como se anotó previamente, es fundamental contar con una vía de contacto permanente y con la asesoría de los entes proveedores de los datos. Por tanto, el reto es que estas entidades establezcan un canal de comunicación que permita que los investigadores puedan realizar consultas y recibir orientación en el manejo e interpretación de los datos, sean estos relacionados con el tema de la innovación y el desarrollo tecnológico en el país, o con otras temáticas sobre la realidad colombiana.

Finalmente, en el proceso de gestión de datos e información en el marco del actual proyecto de investigación sobre la innovación en Colombia y proyectos posteriores, se presenta el desafío de desarrollar herramientas o interfaces que faciliten la interacción de usuarios no técnicos con la bodega de datos, para participar tanto en el análisis de dichos datos, como en su misma gestión, lo cual acerca a los investigadores de diferentes campos a este tipo de repositorios y sus enormes potencialidades, a la vez que mejora y agiliza el proceso de gestión.

Agradecimientos

Los autores agradecen a Colciencias por la financiación del proyecto de investigación "Descubrimiento de Conocimiento sobre la Innovación en Colombia a partir de las Encuestas de Innovación y Desarrollo Tecnológico, la Encuesta Anual Manufacturera y la base de datos ScienTI" (código 2200-330-18872); esta ponencia es uno de los resultados indirectos de dicho proyecto. Así mismo, agradecen los aportes a este documento por parte de los demás integrantes del equipo ejecutor.

Bibliografía

BAKER, K. y BADAMSHINA, G. Knowledge Management. En: *Management Benchmark Study*. Alabama, USA: Air University, Office of Planning and Analysis. 2002.

BURTON-JONES, Alan. *Knowledge Capitalism: Business, Work, and Learning in the New Economy*. Midsomer Norton: Oxford University Press. 1999.

DAVENPORT, Thomas y PRUSACK, Laurence. *Working Knowledge: How Organizations Manage What They Know*. Boston: Harvard Business School. 1998.

GELAUFF, G.M. The Changing Position of Firms in the Knowledge-Based Economy. En: *Innovation Management in the Knowledge Economy*. London: Imperial College Press. 2003.

GOHARIAN, Nazli y GROSSMAN, David. *Data preprocessing*. Illinois Institute of Technology. 2003.

INFORMATION MANAGEMENT. *Qué es la gestión de la información?* Disponible en: http://informationmanagement.wordpress.com/category/gestion/gestion-de-la-informacion/. 2006. Acceso en: mayo 31 de 2009.

INMON, W. *Building the Data Warehouse*. New York: Wiley & Sons. 2002.

JIMÉNEZ, Claudia S., VILLA, Fernán y RICO, Martín. Metamodelo de una bodega de datos para el descubrimiento de conocimiento. En: I CONGRESO INTERNACIONAL DE GESTIÓN TECNOLÓGICA E INNOVACIÓN. Bogotá: Universidad Nacional de Colombia. 2008.

WEISBERG, Sanford. *Applied Linear Regression*. Third Edition, Wiley. 2005.