Departamento Administrativo Nacional de Estadística

Censo de población y vivienda de Colombia
Año 2005

Análisis de la tecnología de información del Censo

Informe final

Grupo Mixto Nro. 2.1

Avantis Carlos Ardila
Ligia Galvis
(Consultor-coordinador del grupo mixto)

Septiembre 2008

advan+is
consultoría gerencial

Grupo Mixto de TI

Informe final

CONTENIDO

INTRODUCCIÓN

Objetivo

Este informe contiene las conclusiones del grupo mixto a cargo del análisis de la tecnología del Censo de 2005. Resume los informes de avance presentados durante el proceso y presenta el nuevo modelo tecnológico que se propone sea aplicado en el próximo censo.

Estructura del documento

Este documento está estructurado en dos capítulos: uno de diagnóstico, que contiene las conclusiones de las pruebas hechas por el grupo mixto, para evaluar las hipótesis formuladas; y uno nuevo, no contenido en los informes de avance previos, con el modelo tecnológico propuesto para el nuevo censo.

1. DIAGNÓSTICO

1.1 Proceso

El grupo de trabajo consolidó y evaluó la información entregada por personal del DANE, para generar un plan de trabajo dedicado al análisis de la Tecnología de Información –TI– del censo. Se logró establecer que el proceso de TI que diseñó el DANE, inicia con el alistamiento de los DMC y termina con la producción de la base de datos final.

El objetivo del grupo de trabajo fue analizar la coherencia sintáctica y semántica de los datos durante su flujo a través del proceso censal.

El plan de trabajo se dividió en cinco etapas, cuya ejecución tomaría cinco meses: (1) entendimiento del problema y plan de trabajo, (2) revisión de la documentación, (3) definición de la estrategia, (4) pruebas y análisis de resultados y (5) conclusiones y recomendaciones.

Los esfuerzos de este informe se enfocaron en los siete puntos del comité de expertos internacionales, de los cuales el primero aplica a la tecnología. “Las consecuencias de la utilización de máquinas digitales para la captura de datos y de los mecanismos de transmisión de datos utilizados”; sobre esta base, el grupo desarrolló la metodología de trabajo.

1.2 Hipótesis

El grupo planteó tres hipótesis principales durante el proyecto: hubo fugas de información en su tránsito de los DMC hasta la base de datos central, el reproceso generó duplicados en la base de datos final y el software de los DMC era inmaduro.

- *Fuga de información*: la ausencia de un sistema de monitoreo y control, el colapso de la red en medio del operativo y los múltiples pasos que las encuestas deberían dar antes de llegar a la base de datos central, generaron

la hipótesis de que se había perdido una cantidad significativa de información.

- *Duplicados*: los reprocesos a los que se recurrió, como medida de emer-gencia por el colapso de la red, que implicaron el transporte de todos los medios magnéticos disponibles, generaron información duplicada en la base de datos. Se diseñaron unos algoritmos de detección de duplicados; sin embargo, el hallazgo de cifras de población superiores en la base de datos central, con respecto a los reportados en los formatos del operativo (específicamente en el formato 15), hicieron suponer la existencia de duplicados a pesar de la aplicación de los algoritmos de detección de los mismos.

- *Software de captura de los DMC* y su sincronización con los repositorios intermedios hasta la llegada a la base de datos central, no tenía la madurez necesaria, dado el corto tiempo que se dedicó a su puesta a punto.

1.3 Estrategia para confirmar/desechar hipótesis

A continuación se presentan las estrategias desarrolladas para confirmar o desechar las hipótesis planteadas.

1.3.1 Fugas de información

- Se estudió el sistema de monitoreo y control.

- Se analizó el inventario creado durante el proceso de planeación del censo.

- Se compararon las cifras de población obtenidas en el formato 15 del operativo, con las contenidas en la base de datos final. El formato 15 contiene los totales por municipio, obtenidos de la suma de otros formatos (3, 8 y 10) que presentan los supervisores.

1.3.2 Duplicados

- Se identificaron los municipios en los que había excesos importantes de información en la base de datos central con respecto al formato 15.

- Se efectuaron consultas en la base de datos del operativo para esos municipios.

- Se observaron visualmente los registros resultantes, en busca de duplicados; se compararon los registros por persona cabeza de hogar.

1.3.3 Inmadurez del software de los DMC

- Se observó la operación de las validaciones especificadas para cada uno de los campos del software de captura usado en los DMC –unidades censales–, con el fin de establecer inconsistencias en su aplicación.

- Se ejecutaron pruebas al software de sincronización, en cada uno de los puntos donde se realizó transmisión de información, con el apoyo de la

Oficina de Sistemas; la estrategia usada fue montar todo el esquema para adelantar pruebas de sincronización en cada uno de los puntos, desde el DMC hasta el cargue de información en la Base de Datos Oracle, se realizó seguimiento a la transmisión de la información en cada una de las etapas.

1.4 Hallazgos

1.4.1 Fugas de información

- El sistema de monitoreo y control no cubrió la captura en los DMC y su traslado a los centros de acopio; pero no es útil para la confirmación de la hipótesis.

- El inventario producto de la planeación del censo, no tiene la precisión necesaria para ser utilizado como herramienta de comparación con los volúmenes de la base de datos final.

- Al comparar las cifras del formato 15 con la base de datos del Censo, se encontraron diferencias apreciables en municipios medianos y grandes. Sin embargo, el comité internacional consideró que los formatos del operativo no eran herramientas de suficiente calidad para confirmar la hipótesis. Consecuentemente, se suspendieron las pruebas adicionales que se habían planeado con los formatos 3, 8 y 10.

1.4.2 Duplicados

- El número de registros duplicados encontrados en la base de datos de personas no fue significativo.

- Se confirma que los formatos del operativo no son suficientemente confiables.

1.4.3 Inmadurez del software de los DMC

- La última versión del software de captura cumple con la totalidad de las validaciones solicitadas por el DANE.

- En entrevistas adelantadas a la empresa que tuvo a su cargo el desarrollo del software, se pudo establecer que existieron 45 versiones del sistema de captura; sin embargo, dada la inexistencia de bitácoras o software de manejo de versiones, no se pudo establecer con certeza el número de versiones del software de captura generadas durante el Censo General 2005.

- El software de los DMC y su sistema de sincronización resultó ser altamente robusto; este hecho alivió fallas del operativo.

1.5 Conclusiones

- La ausencia de un sistema de monitoreo y control, así como de un software de control de versiones y de bitácoras, impidió confirmar o rechazar categóricamente la hipótesis de fuga de información, sólo se puede afirmar que no se encontraron evidencias de fuga de información, de acuerdo con las pruebas realizadas.

- A pesar de los problemas detectados en los procesos del operativo, el reproceso mezclado con la robustez del software de sincronización, mejoraron la probabilidad de que los bits capturados llegaran a la base de datos.

- Los duplicados generados por el reproceso fueron correctamente detectados y eliminados.

- El uso de la tecnología en el Censo de 2005, hizo sus resultados más confiables que los de ediciones anteriores.

- La captura por medio de DMC significó un salto en la calidad de los datos.

- Es necesario asegurarse que en el próximo censo se capturen las oportunidades de mejora identificadas en este proceso; con este objetivo es necesario pasar de un proceso censal decenal discreto, a uno continuo. Será necesario mantener un flujo continuo de inversión en la construcción y mantenimiento de un archivo maestro de direcciones y de la cartografía.

- El proceso de planeación, diseño, desarrollo y prueba del software de captura y sincronización, así como el sistema de monitoreo y control, deben iniciarse desde ahora para el próximo censo.

- Los procesos de gestión de la información censal deben ser madurados, para garantizar un manejo confiable de los datos.

- Se debe explorar la utilización de mecanismos alternos de captura, tales como la web y la voz.

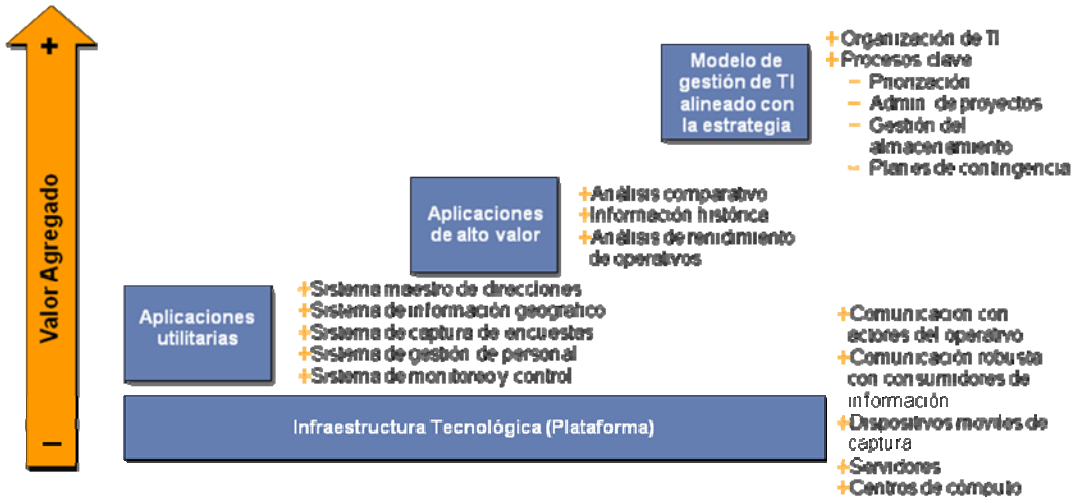
- Se deben iniciar proyectos piloto de autoempadronamiento comenzando por los estratos superiores y los jóvenes.

2. NUEVO MODELO TECNOLÓGICO

2.1 Marco de referencia

Con el fin de formular el nuevo modelo de una manera estructurada, se presenta el modelo desarrollado por Advantis. Los modelos tecnológicos se componen de tres capas: infraestructura, aplicaciones y procesos de gestión. A su vez, las aplicaciones se dividen en dos tipos: utilitarias, las cuales automatizan los procesos misionales; y las de alto valor, que apoyan el análisis y la toma de decisiones.

Diagrama 1
Modelo tecnológico



La capa de infraestructura no agrega valor, es simplemente un requerimiento para poder obtener valor en las capas superiores. No obstante, una infraestructura pobre se convierte en un obstáculo en la obtención de valor de la TI.

Las aplicaciones son la razón de ser de la tecnología; sin embargo, para obtener su valor es necesario contar con procesos de gestión adecuados. La capa de gestión es la que permite obtener el valor de la tecnología; por eso, el comité hace especial énfasis en la necesidad de madurar los procesos de gestión asociados al censo.

2.2 Directrices

A continuación se presentan un conjunto de directrices para el diseño del modelo tecnológico del censo.

- Se buscará que los elementos componentes tecnológicos del censo sean transversales a todas las investigaciones del DANE, de tal manera que se optimicen las inversiones en TI.

- Se considerará la utilización de múltiples medios de captura de información, incluyendo dispositivos móviles y papel como en los censos pasados; se adicionará la web.

- El mecanismo de encuesta será combinado con autoempadronamiento; en el próximo censo se hará un proyecto piloto de autoempadronamiento, que se enfocará en los estratos económicos altos y en la población joven.

- Se llevarán a cabo varias pruebas masivas e integradas antes del censo.

- Se continuará usando dispositivos móviles de captura, ahora mezclados con la capacidad de GPS.

- Para las rutas se considerará el reemplazo del papel por la grabación en voz de la encuesta.

- Se usarán herramientas de control de calidad de los datos, especialmente para las direcciones.

- La infraestructura se compondrá solamente de dos niveles: captura y procesamiento central. Los datos viajarán directamente del dispositivo móvil al centro de cómputo, sin atravesar centros de acopio o regionales.

- Se creará un sistema de monitoreo y control en tiempo real, que permita controlar el operativo y el flujo de la información producida, de manera precisa, incluido la ubicación de los encuestadores y supervisores. Se conservará toda la información producida por el sistema durante el operativo.

- Los dispositivos móviles mantendrán al menos 200 encuestas en su memoria local, para habilitar la solicitud de retransmisión por parte del sistema de monitoreo y control.

- Antes de iniciar el desarrollo de los componentes de software y la adquisición del hardware, se hará un proyecto de arquitectura que profundice el modelo propuesto en este documento.

- Se creará un sistema de gestión de direcciones, que se actualizará de manera continua y será transversal a las investigaciones del DANE.

2.3 Procesos de gestión de la TI

La mejora más importante propuesta por el modelo de TI para el próximo censo colombiano, consiste en la transición de un proceso censal discreto en intervalos de diez años a uno continuo, que desarrolla y evoluciona los componentes centrales del modelo:

- El sistema de monitoreo y control del operativo
- El software de captura, validación y transmisión de las encuestas
- El sistema de direcciones
- El sistema de información geográfica
- El sistema de gestión de personal

- La bodega de datos, hoy inmersa dentro de la Infraestructura Colombiana de Datos –ICD
- El sistema de difusión, hoy inmerso dentro de Colombiestad.

Además de desarrollar estos componentes, será necesario madurar los procesos centrales de gestión de TI:

- *Planeación*: este proceso, de un censo particular, se iniciará al menos con cinco años de anticipación.

- *Arquitectura*: antes de desarrollar o adquirir cualquier componente tecnológico del censo, será necesario crear su arquitectura; de esta forma, se logrará una visión holística que garantice el encaje suave de todas las piezas.

- *Pruebas*: el nuevo proceso censal será sometido al menos a dos pruebas generales: 1 y 2 años antes del operativo, de tal manera que se disponga de tiempo suficiente para hacer los ajustes requeridos. De ser necesario, se pueden programar pruebas adicionales 18 meses y 6 meses antes.

- *Gestión de riesgos y plan de contingencia*: parte central de los procesos de gestión de la TI será la gestión de los riesgos asociados a la tecnología. Esta gestión inicia con el establecimiento de un mapa de riesgos, el cual es mantenido de manera permanente. De otro lado, será necesario establecer un plan detallado de contingencia, que anticipe las circunstancias anómalas que pueden ocurrir dentro del proceso censal y determine con claridad las acciones y responsables asociados a cada una. El plan de contingencia implica además un entrenamiento a la organización antes del inicio del operativo censal.

- *Procesos de desarrollo de software*: el DANE contratará externamente una proporción importante de los desarrollos de software requeridos para el operativo censal. Sin embargo, una contratación ordenada, implica el establecimiento de un marco metodológico que debe ser transmitido y exigido a los contratistas.

- *Procesos de gerencia de proyectos*: los procesos tecnológicos asociados con el censo, implican la ejecución de múltiples proyectos, que deben estar controlados con métodos maduros a través de mejores prácticas internacionalmente aceptadas, en particular las del Project Management Institute.

- *Procesos de gestión de la infraestructura*: el DANE requiere madurar sus procesos de gestión de infraestructura, adoptando las mejores prácticas descritas por ITIL.

- *Procesos de gestión del almacenamiento*: deben asegurar la posibilidad de recuperar las bases de datos en múltiples puntos del operativo. El sistema debe garantizar la posibilidad de hacer seguimiento futuro al proceso, de tal manera que la calidad de las cifras pueda ser claramente establecida.

- *Contratación de terceros*: este proceso y el control sobre la ejecución de esos contratos, deben estar detallados en métodos relacionados.

2.4 Aplicaciones de valor

El nuevo modelo define dos aplicaciones de valor: una bodega de datos con información histórica y un conjunto de servicios de agregación de datos.

2.4.1 Sistema de difusión (Infraestructura Colombiana de Datos –ICD)

La ICD ha sido definida como la bodega de datos donde reposan las investigaciones certificadas por el DANE, y el censo es una de ellas. Lo más importante es que la ICD conservará los censos anteriores armonizados, para acceder a las investigaciones orientadas a detectar tendencias, lo que habilitará la toma de decisiones estratégicas del estado.

La bodega tendrá un conjunto de reportes predefinidos, que mostrarán la información del censo, sin interpretación.

Se tendrán *datamarts* con distintos niveles de consistencia y latencia, para obedecer a distintos objetivos. Unos contendrán información que debe ser publicada en plazos muy cortos y que, por lo tanto, no puede ser armonizada con información histórica ni su consistencia asegurada de una manera muy rigurosa. Otros, por el contrario, contendrán información menos actualizada, pero con procesos rigurosos de calidad, consistencia y armonización histórica.

2.4.2 Servicios de agregación más aplicaciones de académicos

Los servicios de agregación son componentes computacionales (*web services*), que permitirán a cualquier investigador, independiente de su localización, obtener agregados de los microdatos, de tal forma que se mantenga la reserva estadística, pero que al mismo tiempo dé conformidad para la obtención de los datos, de acuerdo con las necesidades específicas de cada investigación.

Estos servicios minimizarán la necesidad de desplazamiento de los académicos hasta el DANE, para obtener datos agregados de los microdatos. De esta manera, el investigador podrá incorporar datos del censo en su investigación desde su propia sede, lo cual mejora la flexibilidad para la comunidad y minimiza los riesgos de fuga de información.

2.5 Aplicaciones transaccionales

La automatización de los procesos asociados al censo, se hará con base en un conjunto de aplicaciones que automaticen la gestión, los marcos de área y de lista, el proceso de encuesta, el monitoreo y control del operativo, el sistema de almacenamiento, la gestión presupuestal y de recursos humanos.

2.5.1 Sistema de direcciones

Los procesos censales del DANE han recurrido históricamente al proceso de enumeración, antes del inicio del operativo, para hacer un inventario de manzanas o Áreas Geográficas –AG–; sin embargo, la experiencia de otros países muestra que el mantenimiento de un archivo detallado, facilita no sólo el censo, sino las encuestas de otras investigaciones del DANE.

Hoy el DANE maneja una base de datos de direcciones georreferenciadas, que debe ser extendida a toda la población. El DANE tiene la base de direcciones recolectadas en el Censo 2005 pero no actualizadas, lo que es urgente así como su georreferenciación para evitar su obsolescencia.

Existe software en el mercado que permite aplicar reglas de calidad a las direcciones; esta herramienta será de gran utilidad para el DANE.

El sistema de direcciones debe ser aplicable a cualquier encuesta de cualquier investigación del DANE.

2.5.2 Sistema de información geográfica (SIG)

El censo no requiere cartografía de precisión, lo cual hace menos costosa la gestión de su cartografía. De otro lado, es suficiente mantener cartografía a una escala 1:5 000.

Es clave mantener un flujo continuo de inversiones en la actualización de la cartografía, para optimizar el uso que se le dio durante el operativo de 2005. La encuesta misma se hará con celulares que tienen capacidad de GPS y, por lo tanto, será posible hacer una gran actualización de la georreferenciación durante el próximo censo.

2.5.3 Sistema de recolección de datos (SRD)

En el Censo 2005 se usaron dos medios de recolección: DMC y papel. En el próximo censo, el papel debe ser usado solamente como medio de contingencia, para lugares con dificultades de orden público. Se espera que para las rutas no sea necesario el uso de papel, pues los encuestadores de rutas podrán llevar dos o más DMC, dado su bajo precio.

La transmisión desde la encuesta, se hará inmediatamente el encuestador diga que ha concluido y el DMC tenga acceso a la red celular de tercera generación o superior. Las velocidades que se logran en las redes celulares de tercera generación actuales, son suficientes para la transmisión *near real time* de la encuesta en tiempos razonables. Usando los volúmenes de la encuesta del Censo 2005 de 10 KB por encuesta y la velocidad actual de las redes celulares de 3 GB, el tiempo que tomaría transmitir una encuesta es:

t = 10 * 1 024 byte * 8 bit/byte / 3 000 000 bit/seg = 27 milisegundos.

Si consideramos que la información de control puede ser del 40% de la información efectiva (*payload*) y que el *throughput* real de la red puede ser de una magnitud inferior a la anunciada, es decir, 300 000 bps en vez de 3 000 000, el tiempo de respuesta sería de menos de medio segundo. El PDA guardará en su memoria al menos las últimas 200 encuestas realizadas (aproximadamente, el último mes de trabajo), para poder responder a solicitudes de retransmisión por parte del sistema de monitoreo y control.

En los casos en que se use papel, el mismo encuestador deberá encargarse de diligenciar la encuesta vía web a través de un *browser*, ya sea en el PDA o en un PC estándar ubicado en un café internet.

2.5.4 Sistema de Monitoreo y Control (SMC) del operativo

El sistema de monitoreo y control tendrá dos componentes esenciales: un sistema gráfico de monitoreo y un *call center inbound/outbound* que habilite una comunicación fluida entre los distintos actores del operativo.

El sistema gráfico permite localizar cada encuestador en un mapa y obtener información sobre su estado (con colores del semáforo, se puede representar su estado con respecto al plan), encuestas totales realizadas, encuestas del día, próximas encuestas, etc. El mapa puede ser acercado o alejado para examinar zonas geográficas en menor o mayor detalle, hasta llegar a un barrio. Al tocar un punto móvil en el mapa que representa a un encuestador, toda su información se desplegará.

Se podrán obtener distintos reportes de cubrimiento a cualquier nivel de la geografía, llegando hasta la manzana.

El componente telefónico le permitirá al encuestador aclarar dudas, reportar circunstancias especiales y consultar acerca de acciones alternativas que se deben tomar en esos casos; permitirá también a la administración central, comunicarse con los coordinadores y supervisores, para indagar acerca del estado de su área.

Un sistema de esta capacidad es factible, pero complejo; su desarrollo debe iniciarse al menos con tres años de anticipación, su costo puede estar entre uno y dos millones de dólares y su aplicación debe ser genérica a cualquier encuesta del DANE.

2.5.5 Sistema de almacenamiento

El sistema de almacenamiento automático permitirá asegurar la información del operativo y hacerle seguimiento *a posteriori*. Este sistema generará automáticamente una copia de la imagen diferencial de las bases de datos cada hora, durante las últimas 24 horas; cada día, durante la última semana, y cada semana, desde el principio al final del operativo.

El sistema de restauración permitirá la renovación de la base de datos en cualquiera de los momentos en que una copia diferencial sea tomada.

2.5.6 Sistemas financieros/administrativos

El modelo de TI del censo tendrá dos componentes de este tipo:

1. **El sistema de personal**: permitirá hacer la gestión administrativa y financiera, de los encuestadores, supervisores y coordinadores.

   Manejará la información demográfica y de desempeño de cada uno de los funcionarios del operativo y controlará sus pagos.

   Este sistema debe ser horizontal a las investigaciones del DANE, es decir, que debe estar en capacidad de controlar cualquiera de los operativos asociados.

2. **El sistema de presupuesto**: capturará el presupuesto original del operativo y su ejecución, a todos los niveles de la jerarquía presupuestal. Será horizontal a todos los operativos de las distintas operaciones del DANE.

2.6 Infraestructura

La infraestructura del próximo censo tendrá solamente dos niveles: centro de cómputo y DMC. No habrá centros de acopio ni computadores intermedios en las regionales. Los PDA transmitirán directamente a un servidor en el centro de cómputo, una vez la encuesta haya sido concluida y el DMC detecte la existencia de conexión a la red.

La infraestructura de TI del censo estará compuesta por los siguientes elementos:

- Teléfonos celulares 3 GB o superiores con capacidad de GPS.

- Red celular de tercera generación, con una velocidad de 3 Mbps de bajada y al menos de 1 Mbps de subida. Puede ser que en el año del operativo ya exista la opción de redes de cuarta generación a 100 Mbps en movimiento o 1 Gbps en reposo. También es factible que existan redes WiMax a centenares de Mbps en movimiento. La adquisición de móviles de esas capacidades debe definirse con 18 meses de antelación; sin embargo, debe ser claro que el desempeño de una red de tercera generación que funcione adecuadamente será suficiente para los requerimientos del operativo censal.

- Centro de cómputo, se utilizará un centro público, al menos de categoría tres; éste no tiene que estar ubicado en Colombia.

- Servidores del centro de cómputo: la configuración detallada de los servidores que residirán en el centro de cómputo, se hará después del proceso de arquitectura, al menos un año antes del operativo. Se tendrán al menos los siguientes servidores: recepción de las transmisiones de los DMC (redundante), base de datos cruda, base de datos imputada, base de datos de la ICD.