ACTA DE REUNIÓN N°3 AYUDA DE MEMORIA

Ciudad: Bogotá
Lugar: Sesión virtual, plataforma Teams
Tema: Presentación proyectos DANE y aportes expertos

Hora: 8:00 a.m. a 10:00 a.m.
Fecha: 13/07/2021
Dependencia responsable: Secretaría Técnica CASEN (DIRPEN)

Participantes

Miembros de la Sala Especializada de Modernización Tecnológica

León Darío Parra Bernal
Jorge Andrés Gallego Durán
Valérie Gauthier Umaña

Departamento Administrativo Nacional de Estadística (DANE)

Asesora DIRPEN y Dirección General
Elizabeth Moreno Barbosa

Oficina de Sistemas
Marly De Moya Amaris

Oficina de Sistemas
Karin Stefany Muñoz

Oficina de sistemas
Gilberto Villalba Gamboa

Oficina de Sistemas
Laura Milena Muñoz

Asesor de la Dirección General
Juan Sebastián Ordoñez

Asesora de la Dirección general
Liliana Morales Hurtado

Asesora de la Dirección General
Karen Lizeth Chávez

Coordinadora - Planificación y Articulación Estadística
Mónica Patricia Pinzón

Profesional GIT Planificación y Articulación Estadística (DIRPEN)
Ruth Constanza Triana

Profesional GIT Planificación y Articulación Estadística (DIRPEN)
Juan José Galeano

Profesional GIT Planificación y Articulación Estadística (DIRPEN)
María José Prada

DIRPEN – Gobierno de Datos
Mateo Cardona

DIRPEN
José de Jesús Lobo

DIRPEN
Cristhyan Leonardo Naranjo

DIRPEN
Anderson Leal Vélez

Prospectiva y Análisis de Datos
Raúl Andrés Gómez

Prospectiva y Análisis de Datos, Equipo ODS
Grace Andrea Torres

Orden del día

1. Solicitud autorización para grabar la reunión

2. Verificación del quórum
   I. León Darío Parra Bernal
   II. Valérie Gauthier Umaña
   III. Jorge Andrés Gallego Duran
   IV. Mónica Patricia Pinzón
   V. Elizabeth Moreno Barbosa
   VI. Marly Esther De Moya Amaris
   VII. Juan Sebastián Ordoñez

3. Síntesis y conclusiones reunión del pasado 9 de junio (Elizabeth Moreno, Raúl Andrés Gómez, Grace Andrea Torres 20 minutos)

4. Presentación Proyecto de Educación – D4N (Karen Chávez, Liliana Morales, Juan Sebastián Ordoñez, Mateo Cardona, Gilberto Villalba – 20 minutos)

5. Presentación Comité de Administración de Datos - CAD (Juan Sebastián Ordoñez – 20 minutos)

6. Presentación expertos aportes solicitados para todos los proyectos (Expertos de la sala – 45 minutos)

7. Conclusiones y cierre (Mónica Pinzón - 15 minutos)

Desarrollo

Objetivo

Presentar los proyectos desarrollados por el DANE, a fin de recibir realimentación por parte de los expertos y fortalecer estos procesos.

1. Solicitud autorización para grabar la reunión

Se inició la grabación de la sesión con previa autorización de los participantes.

2. Verificación del quórum

Se verificó el quórum tanto de los expertos de la Sala, como del Departamento Administrativo Nacional de Estadística (DANE).

3. Síntesis y debate de conclusiones reunión del pasado 9 de junio (Elizabeth Moreno, Raúl Andrés Gómez, Grace Andrea Torres - 35 minutos)

Elizabeth Moreno inició la segunda reunión con una invitación a trabajar de forma conjunta en el ecosistema de datos del SEN como una relación simbiótica en donde todas las partes se ven beneficiadas. Recordó el rol del CASEN dentro de la simbiosis, haciendo hincapié en la importancia de la Sala, puesto que esta es transversal a las demás salas y apoya los temas de seguridad de la información e interoperabilidad, entre otras temáticas.

Mencionó los tres proyectos expuestos en la segunda reunión de la sala: Desarrollo y distribución de aplicaciones web para la recolección de datos; gestión de anomalías temáticas y operativas en el censo económico; y medición proxy para el indicador de los Objetivos de Desarrollo Sostenible (ODS) relacionado con discriminación, utilizando redes sociales. Los dos primeros presentados por Raúl Andrés Gómez y el último presentado por Grace Andrea Torres.
En este punto, Elizabeth Moreno mencionó que para el primer proyecto se identificaron posibles aportes por parte del profesor León Parra y que para los otros dos proyectos Valerie y Jorge pueden presentar aportes.

El primer proyecto es un método alternativo de recolección de información web que se puede aplicar a través de Facebook, usando como piloto el ODS que habla sobre discriminación y teniendo en cuenta las facilidades de los formularios web que permiten reducir costos al utilizarlos. Sobre este proyecto León Darío se postuló para brindar apoyo en el diseño de la aplicación web. Además, Jorge y Valérie presentaron algunos puntos de mejora y alertas que existen al respecto.

Elizabeth finalizó su intervención sobre este proyecto afirmando que en la reunión pasada se definió que se deben tener en cuenta los términos de búsqueda de información que se van a usar y que también es necesario revisar los retos de la propuesta ya que pueden existir restricciones de acceso a los datos y tener baja representatividad.

Destacó que uno de los productos de la mesa es la propuesta de los expertos sobre desarrollar Capstone con la Universidad del Rosario como nicho de proyectos, lo cual, genera una simbiosis entre estudiantes, academia y DANE. Una de las tareas de la reunión es realizar una mesa de trabajo que permita delimitar el alcance de la propuesta y formular un plan de trabajo.

Posteriormente, Raúl Gómez presentó las conclusiones sobre los dos primeros proyectos. Sobre el primer proyecto aceptó la propuesta de apoyo de León Darío, ya que el uso de metodologías ágiles efectivamente ayuda a reducir costos. Enfatizó en que Facebook es de las mejores opciones para difundir contenido por el deseo de conexión social por estudios ya hechos, siguen existiendo retos y no se debe perder de vista que esto es por ahora un proyecto piloto. Finalmente, reiteró la factibilidad de establecer alianzas con universidades (Capstone) para el desarrollo de proyectos de interés para el DANE.

Para el cierre de conclusiones, Grace Torres dio respuestas a las inquietudes planteadas por los expertos en la segunda reunión con relación al tercer proyecto que se presentó. Precisó que el proyecto de medición de discriminación usando redes sociales se puede hacer bajo análisis no supervisado; para ello, se definen los términos objetivo y se buscan en el conjunto de datos para identificar estos términos. Luego se correlacionan utilizando n-gramas y al final se hace análisis de espacio vectorial para definir los campos semánticos sobre las diversas formas de "discriminación", creando así, un modelo que "aprende" a identificar esos conceptos en determinados contextos y lo asocia con la discriminación.

Grace comentó que no es eficiente hacer un modelo supervisado por costos y por temas de metodología técnica. Por último, Grace mencionó que para la recolección de datos se están desarrollando dos enfoques diferentes, teniendo en cuenta la red social: Para datos de Twitter se están utilizando los servicios de la plataforma Azure y para datos de Facebook se desarrolló un algoritmo en Python para obtener comentarios de algunos perfiles de influencers. Asimismo, se mencionó que es de interés del proyecto contar con la participación de León Darío en la asesoría y capacitación en temas de metodologías ágiles (SCRUM) para fortalecer los procesos de análisis de los datos.

4. Presentación Proyecto de Educación – D4N (Karen Chávez, Juan Sebastián Ordoñez, Mateo Cardona, Gilberto Villalba – 50 minutos)

La presentación del proyecto D4N de educación inició con la intervención de Karen Chávez, quien hizo una breve introducción a los temas ODS que se vienen trabajando desde el DANE. Explicó que normalmente se hacen medidas continuas a los indicadores con el fin identificar cuellos de botella, para lo cual, se diseñó un barómetro; que permite a su vez, priorizar indicadores de la agenda 2030, que oscilan alrededor de diez mil.

Posteriormente, Karen habló sobre las estadísticas experimentales y el objetivo que existe frente a promover el conocimiento de estas para todo el público, haciendo una diferenciación entre estas y las estadísticas oficiales. Además, mencionó que dentro de los proyectos de innovación se incluye el proyecto D4N, por usar técnicas no convencionales para lograr los ODS involucrando a otros actores del SEN, como la ciudadanía quienes pueden aportar con la generación de datos en el proceso estadístico. En esta sección de la reunión, se mencionó también que esto hace parte del proceso de interoperabilidad de los datos, lo cual es un punto clave del CASEN.

Posteriormente, Juan Sebastián Ordoñez inició su intervención, quién habló sobre el módulo de gestión de calidad de los datos y mencionó la definición de gobierno de datos, con el cual se busca hacer un seguimiento más exhaustivo al ciclo de vida de los registros administrativos. En este caso, usando el caso del proyecto de educación, pero haciendo la claridad de que es un proceso transversal que se puede aplicar a muchos temas. También realizó la explicación del cuadro de flujo del valor público de la información y mencionó que los datos tienen valor gracias a ciertos procesos como la seguridad de la información, la divulgación de los resultados, la depuración y difusión de datos y el uso y reutilización de la información.

Luego, Juan Sebastián habló sobre el manejo de la calidad de los procesos y sobre la necesidad de estandarizarlos para lograr procesos colaborativos coordinados de registros administrativos coherentes. De acuerdo con lo anterior, la gestión de calidad es un gran módulo de la gobernanza de datos. Continuó la discusión sobre el lago de datos, como herramienta que permite un análisis más detallado de los registros administrativos con lo cual se mejora la eficiencia, la productividad y el diseño de roles en los niveles del ciclo de vida de los datos; segmentando accesos y asegurando la información para tener datos abiertos de forma parcial.

Juan Sebastián hizo referencia al marco teórico para la gestión de calidad, en la que se usa la clasificación de medición de los datos holandesa. Según esta, se deben hacer chequeos técnicos y un primer diagnóstico al conjunto de datos, revisar la elegibilidad de los archivos y evaluar no solo la calidad de los registros sino de todos los indicadores, que, para el caso del proyecto de educación, el registro es el SIMAT y el indicador es la distancia del hogar al colegio.

Teniendo en cuenta lo anterior, comentó que cada registro debería tener un proceso de evaluación de calidad y que este debe ocurrir apenas ingresan al sistema. Además, explicó que entre dependencias hay problemas de comunicación y diferencias en las evaluaciones de calidad que se hacen a los registros, por lo cual es necesario concertar nuevos lineamientos siguiendo el estándar internacional que permitan crear un sistema coordinado para mejorar el diseño de indicadores.

En ese sentido, se introdujo el proyecto de Automatización de Procesos de Calidad, mezclando a su vez los avances logrados con el proyecto de educación y mencionado también las técnicas de cruces usadas dentro de los procedimientos. De esta manera, Mateo Cardona, intervino en la reunión explicando el formulario de seguimiento a la gestión de calidad – SIMAT, el cual se divide en tres secciones que caracterizan al usuario e incluye glosarios y diferentes términos que deben ser tenidos en cuenta para la creación de los registros administrativos. Sin embargo, comentó que este formulario tiene un problema de entendimiento de las preguntas y que por eso se requiere un acompañamiento para la recolección de la información. En consecuencia, es clara la necesidad de automatizar los procesos de calidad para gestionar todos los proyectos del DANE que incluyen el proyecto D4N de educación.

Posteriormente, Gilberto Villalba intervino en la reunión explicando el proceso de automatización el cual inicia con el proceso de chequeos técnicos a través de Excel, en este caso, haciendo uso del SIMAT, con el fin de generar un diccionario estándar. Aquí, explicó que el proceso consiste en

identificar las variables que presentan problemas de tipo, longitud o categoría. Luego de identificar los problemas, se pasa a la plataforma Zeppelin para ejecutar scripts y se ingresa al lago de datos a través del perimetral para garantizar la seguridad de la información y se hace el tratamiento de los datos, con lo cual se logra reducir los tiempos de 4 horas a tan solo 5 minutos.

Para terminar su intervención, Juan Sebastián habló sobre los protocolos de integración que existen en los procesos de gestión de calidad de las llaves de integración de los datos. Mencionó que hay protocolos dentro del lago de datos para garantizar limpieza en las llaves de integración y para gestionar duplicados de información luego de la limpieza de los datos. Según esto, ~~se~~ explicó que para hacer análisis de calidad de las llaves según las bases de referencia DANE, se pueden recuperar datos por pegues probabilísticos a través de los datos de identificación y por emparejamientos multietápicos, validando los cruces para ver si son adecuados. En este caso, también hay protocolos con parte determinística y difusa, usando diferentes cruces y etapas, haciendo uso de las variables tipo y número de documento. Aquí, Juan Sebastián concluyó la exposición haciendo un ejemplo con nombres, explicando que en una base de datos puede parecer que no hay duplicados si se revisa solo el nombre, pero cuando se cruza también el número de identificación o el tipo de identificación se encuentra que los datos aparecen diligenciados dos o más veces.

Por temas de tiempo se dio paso a la intervención de los expertos sobre los temas presentados.

En primer lugar, Valérie expuso que le parecía extraño que la primera etapa del proceso de automatización utilizara Excel como plataforma de análisis de datos; sin embargo, Juan Sebastián contestó diciendo que allí se facilita el contraste con los metadatos y que esta solo se usa para la primera etapa, después se hace uso de otros softwares de programación más sofisticados.

Posteriormente, Jorge Gallego mencionó que las herramientas para trabajar los ODS son muy importantes y comentó sobre un proyecto que utiliza imágenes de Google Street View de las manzanas de los barrios en tiempo real y el uso de esos datos como fuentes alternativas. También, informó el trabajo en un proyecto del TIC Tank de la Universidad del Rosario con la alcaldía de Barranquilla bajo el cual, se desarrollaron mesas temáticas con comunidades para plantear necesidades para luego sistematizarlas y relacionarlas con los ODS y proponer soluciones y proyectos de política pública. Lo anterior podría complementarse en el barómetro de ODS.

Sobre el proyecto de automatización, Jorge preguntó por el tema de revisión de pares y sobre cuáles serían los incentivos para lograr que las entidades llenen los formularios, ya que, al parecer, estos requieren de mucho tiempo adicional de trabajo.

Por último, León Darío indicó que hace falta hacer hincapié en la metodología de la calidad de los datos y revisar otras metodologías como la evaluación por priorización de dígitos, ya que hay que tener claridad antes de hacer la integración para identificar las variables donde puede haber problemas. También hizo alusión a las preguntas orientadoras: ¿Qué mecanismos de georreferenciación podrían ser útiles para ajustar las mediciones entre censo y censo? ¿Podrían

recomendarnos modelos predictivos que nos permitan proyectar los movimientos de población utilizando información georreferencial? A lo cual, el respondió que para llegar a más personas para lograr aumentar la consecución de información se pueden utilizar los sistemas de ubicación de los celulares personales con códigos de seguimiento para mejorar las estimaciones.

Asimismo, sobre los modelos predictivos, León Darío mencionó que participó en tres procesos de automatización celular para revisar ruido en integración de datos, en la medida que estos se van limpiando y que puede aportar desde su conocimiento. También, indicó que conoce sobre redes neuronales cuando hay sesgos observacionales, en caso de que esto pueda ser útil para alguno de los proyectos. Además, recomendó hacer pilotos pequeños para encontrar errores de cada etapa y luego resolverlos y escalarlos, pues hacer todo el tratamiento de la información a la vez genera mucho ruido.

Por último, sobre la pregunta orientadora "¿Considera que la calidad de las llaves de integración es un insumo importante para la construcción de identificadores sintéticos únicos?", León Darío contestó que es clave este proceso antes de la integración.

Conclusiones y cierre (Mónica Pinzón - 15 minutos)

Mónica Pinzón cerró la reunión con la propuesta de iniciar en el próximo encuentro con las intervenciones de los expertos. De igual modo, los invitó a reflexionar y dar su posición sobre las preguntas planteadas de los temas expuestos.

Compromisos

1. **Tarea:** enviar el acta de la sesión
   Responsable: DANE
   Fecha entrega: a más tardar 16 de julio

2. **Tarea:** envío Doodle para agendar cuarta reunión
   Responsable: DANE
   Fecha entrega: 13 julio

3. **Tarea:** envío documentación proyectos a trabajar en la sala
   Responsable: DANE
   Fecha entrega: antes de la realización de la cuarta reunión

4. **Tarea:** revisión del formulario SIMAT (https://cnprivado.shinyapps.io/Calidad_SIMAT/)
   Responsable: Expertos de la sala
   Fecha de entrega: Próxima reunión

Expertos CASEN	DANE

Próxima reunión:

Responsable de convocar: Secretaría Técnica CASEN
Fecha: 3 de agosto, 9 a 11 am