

Análisis de duplicidad en la información recolectada por medio de Registros Únicos



Unidad de Científicos de Datos
2017



DNP Departamento
Nacional
de Planeación

1. Análisis de Registros Únicos

El análisis realizado sobre el conjunto de registros únicos tiene el objetivo de extraer información relevante que evidenciara de forma clara y eficaz los niveles de duplicidad en la información capturada a través de los diferentes registros.

2.2. Descripción

Se capturo información de los diferentes registros a través de búsquedas web y de la información suministrada a través de la encuesta de explotación de datos. Los siguientes datos resumen la información encontrada:

1. Entidades gubernamentales de orden nacional: 209
2. Entidades encuesta DNP: 152
3. Entidades que contestaron afirmativamente al registro único: 42
4. Entidades encontradas a través de búsqueda web: 21
5. Total entidades: 72
6. Total entidades con Registro Único (RU): 50
7. Total de RU con información de registro disponible en página web: 16
8. Total de RU incluidos en el análisis: 27 (*N*)
9. Total de campos únicos: 115 (*M*)

Características de los datos recopilados

- Confiabilidad baja, en ninguna captura se logró recopilar el total de campos del registro y en 4 solo se logró llegar a la información básica.
- La información de 32 URLs suministradas en la encuesta no corresponde a la ubicación del registro.
- Para el ingreso a un gran porcentaje del total de formularios encontrados, es necesario registrarse en el sistema correspondiente y validar a través de algún sistema de seguridad (ej. captcha) que el usuario es humano.
- No hay una diferenciación entre Sistemas de información (que se nutren de la información recopilada a través de los RU) y registros únicos.
- Hay una diferencia en el tamaño de la información capturada por los registros. Esta diferencia está en función del tipo de registro que se captura y el sector al que se encuentra dirigido: PN-Persona Natural, PJ-Persona Jurídica, Tramites, Documentos entre otros.

2.3. Análisis

Se construyó una matriz de $N \times M$, en donde N corresponde al número de registros y M al número total de campos (Unión de los campos de todos los registros). A partir de la información contenida en la matriz se consideró que era necesario diseñar un grupo de indicadores que reflejaran de forma acertada y cuantificable los supuestos del proyecto.

Índices de redundancia

Se construyeron tres Índices de redundancia para determinar el nivel de repetición de la información contenida en la matriz: índice de redundancia del campo, índice de redundancia del registro y el índice de redundancia total. Estos índices toman valores entre $[0, 1]$, en donde 0 corresponde a información no duplicada (Solo presente en una ocasión) y 1 corresponde a información duplicada en todos los escenarios en donde sea posible.

Como primer paso se diseñó una variable indicadora, z_{ij} , para determinar si el campo j está presente en el registro i . Se debe tener en cuenta que todo campo debe ser usado al menos en un registro.

$$z_{ij} = \begin{cases} 1, & \text{si campo } j \text{ esta presente en el registro } i \\ 0, & \text{cualquier otro caso} \end{cases}$$

Índice de redundancia del campo (IRC)

Se define el Índice de redundancia del campo j sobre N registros como:

$$\sigma_j = \left(\frac{1}{N-1} \sum_i^N z_{ij} \right) - \frac{1}{N-1}$$

En donde $i = 0, 1 \dots N - 1$

Este índice mide el nivel de repetición de un campo dentro del conjunto de registros.

Índice de redundancia del registro (IRR)

Se define el índice de redundancia del registro i como:

$$\gamma_i = \frac{1}{L_i} \sum_j^{L_i} \sigma_{ij}$$

En donde σ_{ij} corresponde al IRC de los campos contenidos en el registro i , y L_i corresponde al número total de campos contenidos en el registro i .

Este índice mide el nivel de repetición de un registro en función del IRC de los campos que contiene el registro.

Índice de redundancia total (IRT)

Se define el índice de redundancia total sobre un conjunto de N registros con M campos agregados como:

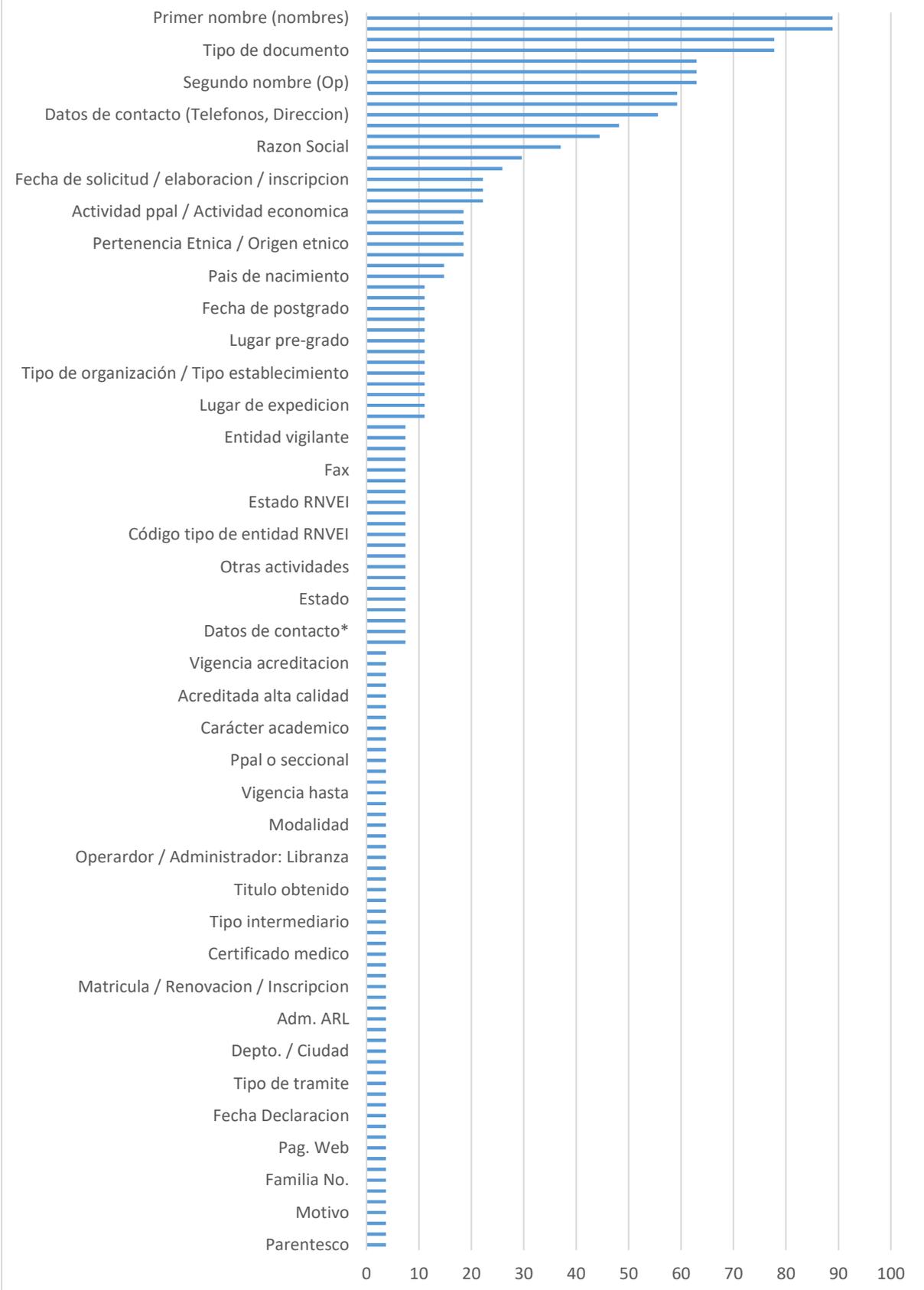
$$\varphi = \frac{1}{M} \sum_j^M \beta_j \sigma_j$$

En donde la variable β_j es utilizada para ponderar el campo j de acuerdo a un criterio específico. Por ejemplo, ponderar en función del espacio en memoria física que ocupa cada campo o ponderar en función de la relevancia del campo dentro del conjunto (es más relevante un campo con mayor IRC)

2.4. Resultados

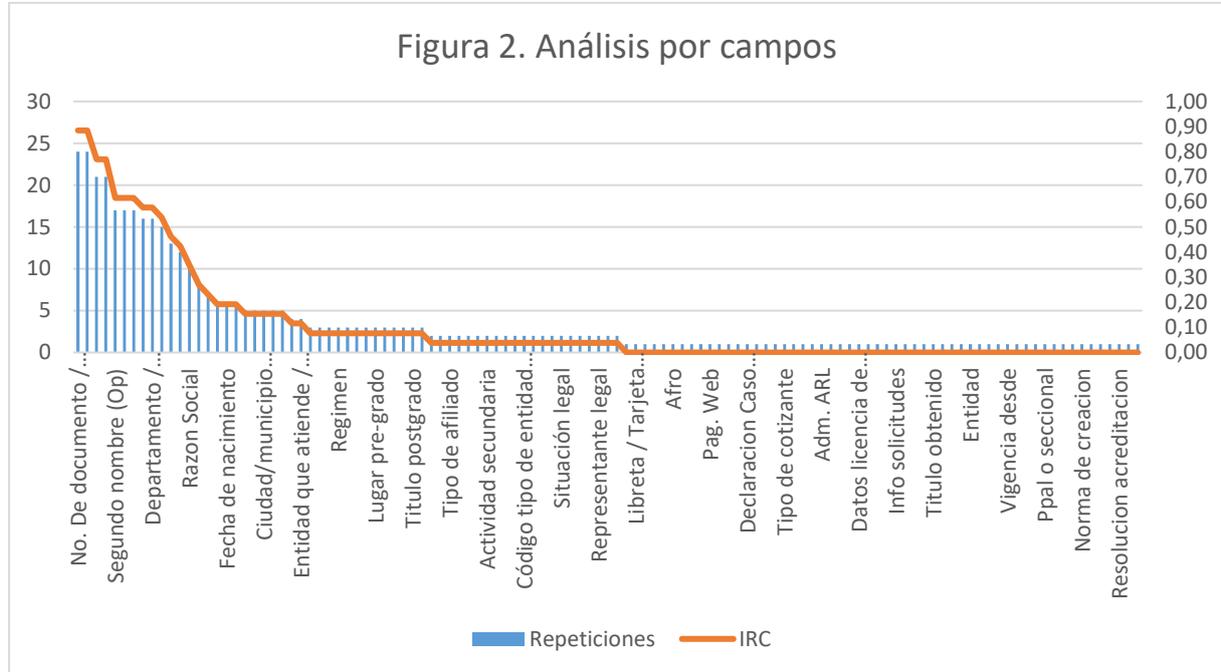
Haciendo un conteo por columnas se obtiene la Figura 1, en donde se muestra el porcentaje de aparición de cada campo en el conjunto total de registros.

Figura 1. Porcentaje de utilización de los campos



IRC - Índice redundancia del campo

Al realizar el cálculo del ICR, sobre los M campos, se obtienen los resultados de la figura 2., en donde los datos han sido organizados de acuerdo a su ICR. Adicionalmente, se incluye el número de repeticiones de cada campo.



IRR - Índice redundancia del registro

Al realizar el cálculo del IRR sobre los N registros se obtienen los resultados que se muestran en la Figura 2. Los registros están organizados de acuerdo al número de campos que contienen.

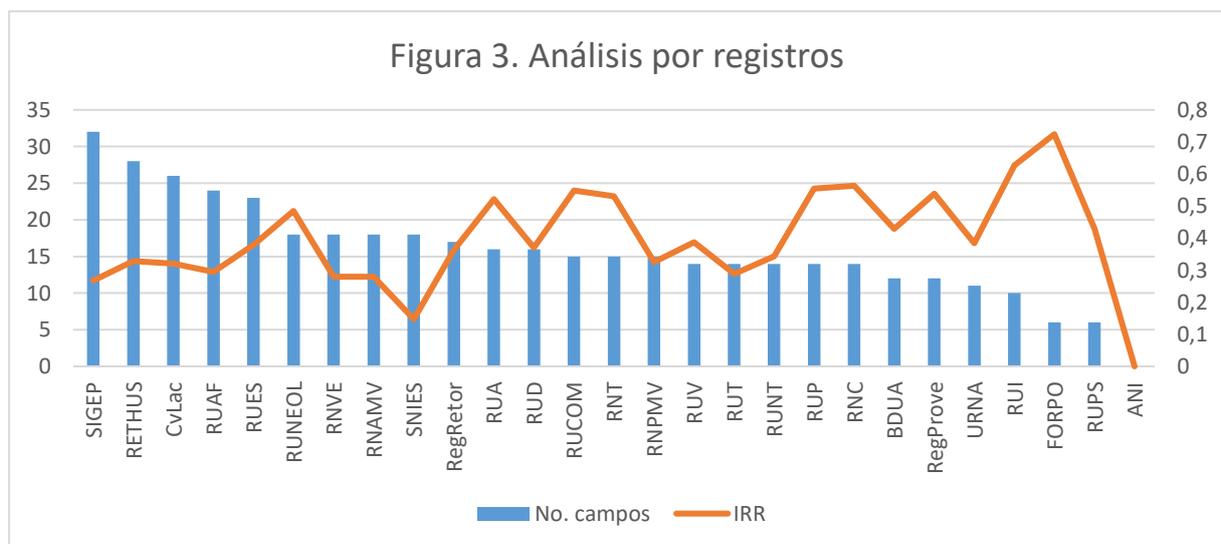


Tabla 2. IRR	
Registro	γ_i
FORPO	0,72
RUI	0,63
RNC	0,56
RUP	0,55
RUCOM	0,55
RegProve	0,54
RNT	0,53
RUA	0,52
RUNEOL	0,49
RUPS	0,43
BDAU	0,43
RUV	0,39
URNA	0,38
RUES	0,38
RUD	0,37
RegRetor	0,36
RUNT	0,34
RETHUS	0,33
RNPMV	0,33
CvLac	0,32
RUAF	0,29
RUT	0,29
RNVE	0,28
RNAMV	0,28
SIGEP	0,27
SNIES	0,15
ANI	#¡DIV/0!

IRT - Índice redundancia total

Al realizar el cálculo del IRT se obtiene,

$$\varphi = 0,104$$

Se utilizó un $\beta_j = 1/115$, que corresponde a una ponderación uniforme de los campos.

2.5. Conclusiones

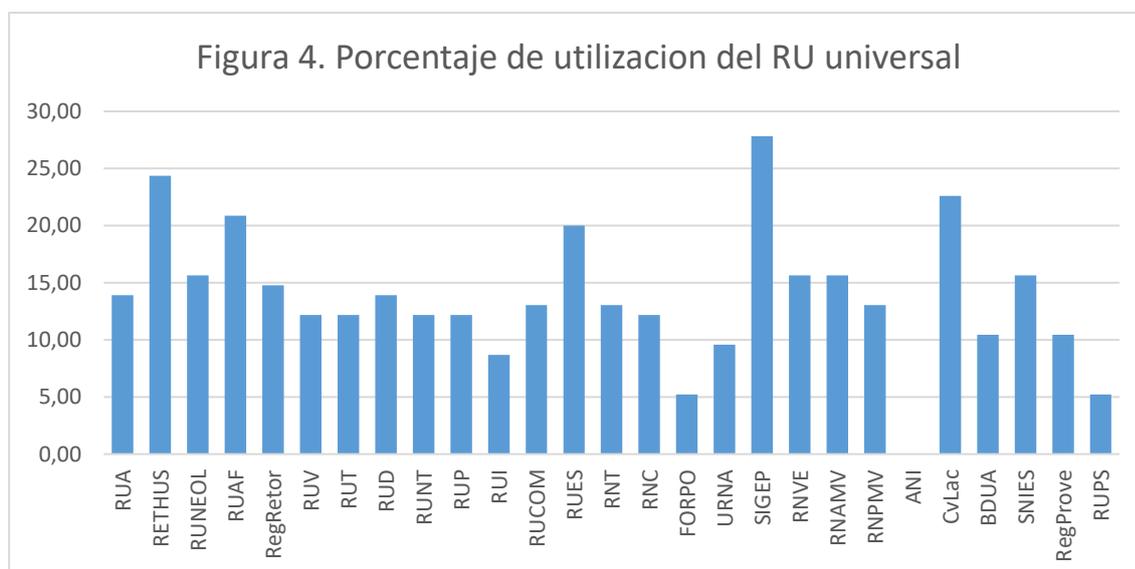
Un grupo de 10 campos tienen un IRC superior a 0,5. Lo que evidencia la existencia de una información que es altamente relevante en las instituciones para “indexar” los registros capturados.

Sin embargo, también existe un número de campos (105) con un IRC inferior a 0,5, lo que muestra el nivel de especificidad de gran parte de la información recopilada en los registros.

A nivel de registros se observa que todos los registros poseen información redundante en por lo menos un registro del conjunto. En especial 8 registros tienen IRR superior a 0,5 lo que indica que el registro tiene redundancia de la información con más de la mitad de los registros del conjunto.

Sin embargo, se encuentra que el IRT es bajo, lo que sugiere que una gran cantidad de los registros poseen campos únicos.

Asumiendo un Registro Universal - RU, aquel que contiene un total de M campos, cada uno de los registros tendría el porcentaje de utilización del RU mostrado en la figura 4.



2. Anexos

Registros únicos

Tabla 1. Resumen información de registros encuesta y búsqueda web.							
Entidad	Registro		Ori	RU ?	Acceso	Tipo reg.	Matriz
DEPARTAMENTO ADMINISTRATIVO PARA LA PROSPERIDAD SOCIAL		LLAVE MAESTRA	E	Si	No	PN	No
Parques Nacionales Naturales de Colombia	RUNAP	Registro Único Nacional de Áreas Protegidas	E/B	Si	No	Otros	No
DIRECCIÓN NACIONAL DE DERECHO DE AUTOR		Registro Nacional de Obras	E	Si	No	Obras	No
SUPERINTENDENCIA DE LA ECONOMÍA SOLIDARIA	SICES		E	No	No	Doc	No
FONDO ROTATORIO DE LA POLICÍA NACIONAL	BIOSTAR -		E	No	Registro	PN	Si
UNIDAD ADMINISTRATIVA ESPECIAL CONTADURÍA GENERAL DE LA NACIÓN	CUIN	Código Único Institucional	E	No	No	PJ	No
CONSEJO SUPERIOR DE LA JUDICATURA - DIRECCION EJECUTIVA ADMINISTRATIVA JUDICIAL	URNA	REGISTRO NACIONAL DE ABOGADOS	E	Si	No	PN	Si
DEPARTAMENTO ADMINISTRATIVO DE LA FUNCIÓN PÚBLICA	SUIT	Sistema Único de Información de Trámites	E	No	No Aplica	Trámites	No

DEPARTAMENTO ADMINISTRATIVO DE LA FUNCIÓN PÚBLICA	SIGEP	Sistema de Información y Gestión del Empleo Público	E	Si	Registro	PN	Si
MINISTERIO DEL INTERIOR Y DE JUSTICIA		Registro público de entidades religiosas no católicas	E	Si	No	PJ/PN	No
DIRECCION GENERAL DE LA POLICIA NACIONAL	RNMC	Registro Nacional de Medidas Correctivas	E	No	No	PN/Doc	No
MINISTERIO DE MINAS Y ENERGÍA	SICOM	Sistema de Información de Comercialización de Combustibles	E	No	Registro	Varios	No
UNIDAD ADMINISTRATIVA ESPECIAL DE ORGANIZACIONES SOLIDARIAS		Registro de Entidades Acreditadas	E	Si	No	PJ	No
AGENCIA NACIONAL DE MINERIA		Registro único de Comercializadores de Minerales RUCOM	E	Si	Registro	PJ/PN	No
SUPERINTENDENCIA FINANCIERA DE COLOMBIA - SIMEV	RNVE	Registro Nacional de Valores y Emisores	E	Si	No	PJ	Si
SUPERINTENDENCIA FINANCIERA DE COLOMBIA - SIMEV	RNAMV	Registro Nacional de Agentes del Mercado de Valores	E	Si	No	PJ	Si
SUPERINTENDENCIA FINANCIERA DE COLOMBIA - SIMEV	RNPMV	Registro Nacional de Profesionales del Mercado de Valores	E	Si	No	PN	Si
SUPERINTENDENCIA FINANCIERA DE COLOMBIA	SIRI	Sistema Integrado de Registro de Información	E	No	No	Trámites	No
SUPERINTENDENCIA FINANCIERA DE COLOMBIA	SICOP	Sistema de Información de Colocaciones Ofertas Públicas	E	No	No	Trámites	No

REGISTRADURIA NACIONAL DEL ESTADO CIVIL - GESTION GENERAL	ANI	Archivo Nacional de Identificación	E	Si	No	PN	Si
REGISTRADURIA NACIONAL DEL ESTADO CIVIL - GESTION GENERAL	AFIS	Sistema Automatizado de Identificación Dactilar	E	No	No	Huellas	No
SUPERINTENDENCIA DEL SUBSIDIO FAMILIAR		CERTIFICADO DE EXISTENCIA Y REPRESENTACIÓN LEGAL	E	No	No	N/A	No
UNIDAD NACIONAL PARA LA GESTIÓN DEL RIESGO DE DESASTRES	RUD	Registro Único de Damnificados	E/B	Si	No	PN	Si
INSTITUTO NACIONAL PENITENCIARIO Y CARCELARIO - INPEC	SISIPEC	SISTEMA DE INFORMACIÓN PENITENCIARIO Y CARCELARIO	E	No	No	N/A	No
ARCHIVO GENERAL DE LA NACIÓN	RUSD	Registro Único de Series Documentales	E/B	Si	No	Doc	No
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	CVLAC	Hojas de vida de investigadores	E	Si	Registro	PN	Si
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	GRUPLAC	Información de grupos de investigación de CTI	E	Si	Registro	Grupos	No
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	INSTITULAC	Información de instituciones del SNCTI	E	Si	No	PJ	No
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	SIGP	Sistema Integrado de Gestión de Proyectos	E	No			No
DEPARTAMENTO ADMINISTRATIVO	Publindex	Sistema de Indexación y	E	Si	No	Revistas	No

DE CIENCIA, TECNOLOGÍA E INNOVACIÓN		Homologación de Revistas Especializadas de CTI					
SUPERINTENDENCIA DE SOCIEDADES	RUNF	Registro Único Nacional de Factores	E/B	Si	No	PJ	No
UNIDAD ADMINISTRATIVA ESPECIAL DIRECCIÓN DE IMPUESTOS Y ADUANAS NACIONALES	RUT	Registro único tributario	E/B	Si	Registro	PJ/PN	Si
DIRECCIÓN DE ADMINISTRACIÓN DE FONDOS DE LA PROTECCIÓN SOCIAL	BDUA	Base de Datos Única de Afiliados	E	Si	No	PN	Si
MINISTERIO DE EDUCACIÓN NACIONAL	SINEB	Sistema de Información Nacional de Educación Básica y Media	E	Si	No	Varios	Si
MINISTERIO DE EDUCACIÓN NACIONAL	SNIES	Sistema Nacional de Información de la Educación Superior	E	Si	No	Varios	No
INDUSTRIA MILITAR		Registro de Proveedores	E	Si	Registro	PJ/PN	No
Unidad Administrativa Especial Migración Colombia		Registro de extranjeros	E	Si	No	PN	No
UNIDAD ADMINISTRATIVA ESPECIAL DE GESTIÓN DE RESTITUCIÓN DE TIERRAS DESPOJADAS		Sistema de Registro de Tierras Despojadas y Abandonadas Forzosamente	E	N/A	N/A	N/A	No
FIDUCIARIA COLOMBIANA DE COMERCIO EXTERIOR S.A.		REGISTRO DE PROVEEDORES	E	Si	Registro	PJ/PN	Si
MINISTERIO DE TECNOLOGÍAS DE LA INFORMACIÓN Y LAS LAS	VUTIC	Ventanilla Única de Trámites	E	No	Registro	Trámites	No

COMUNICACIONES								
INSTITUTO NACIONAL DE MEDICINA LEGAL Y CIENCIAS FORENSES	RND	Registro Nacional de Desaparecidos y Cadáveres, Registro de Lesiones no Fatales	E	Si	No	PN	No	
MINISTERIO DE COMERCIO, INDUSTRIA Y TURISMO		Registro de Usuario de la VUCE	E	Si	No	PJ/PN	No	
MINISTERIO DE COMERCIO, INDUSTRIA Y TURISMO	RUNIC	Registro Único Nacional de Importadores y Comercializadores Autorizados	B	Si	No	N/A	No	
FUERZA AÉREA COLOMBIANA	SIATH		E	N/A	N/A	N/A	No	
BANCO DE COMERCIO EXTERIOR DE COLOMBIA S.A.		Índice de Información Clasificada y Reservada	E	N/A	N/A	N/A	No	
BANCO DE COMERCIO EXTERIOR DE COLOMBIA S.A.		Índice de Información Clasificada y Reservada	E	N/A	N/A	N/A	No	
SUPERINTENDENCIA DE SERVICIOS PÚBLICOS DOMICILIARIOS	RUPS	REGISTRO UNICO DE PRESTADORES DE SERVICIOS PUBLICOS	E	Si	No	PN	Si	
INSTITUTO COLOMBIANO AGROPECUARIO		Registro Nacional de Predios Pecuarios	E	Si	No	Otro	No	
INSTITUTO DE HIDROLOGIA, METEOROLOGIA Y ESTUDIOS AMBIENTALES- IDEAM	RUA	Registro único ambiental	E/B	Si	Registro	Otro	Si	
UNIDAD ADMINISTRATIVA ESPECIAL BIBLIOTECA NACIONAL		REGISTRO ÚNICO DE BIBLIOTECAS PUBLICAS	E	Si	No	N/A	No	

Instituto Nacional de Salud	sivigila	Sistema Nacional de Vigilancia en Salud Pública	E	No	No	Otro	No
Instituto Nacional de Salud	sivicap	Sistema de Información de la Vigilancia de la Calidad del Agua para Consumo Humano	E	No	No	Otro	No
Instituto Nacional de Salud	Reddata INS	Información donantes de órganos	E	N/A	N/A	N/A	No
MINISTERIO DE CULTURA		Registro como productor permanente u ocasional de espectáculos públicos de las artes escénicas	E	Si	Registro	PJ/PN	No
MINISTERIO DE CULTURA		Autorización de Operadores de Boletería en Línea	E	No	N/A	PJ/PN	No
UNIDAD ADMINISTRATIVA ESPECIAL DE GESTIÓN PENSIONAL Y CONTRIBUCIONES PARAFISCALES DE LA PROTECCIÓN SOCIAL	RUA	Registro único de aportantes	E	Si	No	PN	No
MINISTERIO DEL TRABAJO		40.00 empleos	E	No	No		No
FASECOLDA	RUS	Registro Único de Seguros	B	Si	No	PJ	No
Ministerio de Salud y Protección Social	RETHUS	Registro Único de Talento Humano en Salud	B	Si	Registro	PN	Si
MINISTERIO DE TRABAJO	RUTEC	Registro Único de Trabajadores Extranjeros en Colombia (En construcción)	B	Si	No	PN	No
MINISTERIO DE EDUCACIÓN NACIONAL		Registro Único de Peticiones	B	Si	No	Trámites	No

CONFECAMARAS	RUNEOL	Registro Único Nacional de Entidades Operadoras de Libranza	B	Si	Registro	PJ	Si
Ministerio de Salud y Protección Social	RUAF	Registro Único de Afiliados (Tramite, afiliación, identificación)	B	Si	No	PN	Si
Cancillería	RegRetor	Registro Único de Retornados	B	Si	Registro	PN	Si
Unidad para las víctimas - RNI	RUV	Registro Único de Víctimas	B	Si	No	PN	Si
MINISTERIO DE TRANSPORTE	RUNT	Registro Único Nacional de Tránsito	B	Si	Registro	PJ/PN/Vehículos	Si
Cámaras de Comercio	RUP	Registro Único de Proponentes (Info. Ubicación)	B	Si	Registro	PJ/PN	Si
MINISTERIO DE TRABAJO	RUI	Registro Único de Intermediarios	B	Si	Registro	PJ/PN	Si
Agencia Nacional Minera	RUCOM	Registro Único de Comercializadores de Minerales	B	Si	Registro	PJ/PN	Si
CONFECAMARAS	RUES	Registro Único Empresarial y Social, Cámaras de Comercio	B	Si	No	PJ/PN	Si
CONFECAMARAS	RNT	Registro Nacional de Turismo	B	Si	No	PJ/PN	Si
INSTITUTO DE INVESTIGACIÓN ALEXANDER VON HUMBOLDT	RNC	Registro Único Nacional de Colecciones Biológicas	B	Si	No	Otro	Si

Tabla 2. Ranking de campos de acuerdo al número de repeticiones en el conjunto y al IRC respectivo

CAMPO	Rep.	σ_j	CAMPO	Rep.	σ_j
No. De documento / identificación	24	0,88	Cargo	2	0,04
Primer nombre (nombres)	24	0,88	Parentesco	1	0,00
Tipo de documento	21	0,77	Libreta / Tarjeta profesional	1	0,00
Primer apellido (apellidos)	21	0,77	Estado tarjeta prof.	1	0,00

Segundo nombre (Op)	17	0,62		Motivo	1	0,00
Segundo apellido (Op)	17	0,62		Libreta Militar	1	0,00
Ubicación / Dirección	17	0,62		Afro	1	0,00
Correo electrónico	16	0,58		Familia No.	1	0,00
Departamento / Ciudad/municipio	16	0,58		Estado salud	1	0,00
Datos de contacto (Teléfonos, Dirección)	15	0,54		Otra nacionalidad	1	0,00
NIT	13	0,46		Pag. Web	1	0,00
Notificaciones (Dir, tel, email)	12	0,42		Responsabilidades	1	0,00
Razón Social	10	0,35		Lugar Declaración	1	0,00
Nombre Comercial Sigla / Nombre establecimiento	8	0,27		Fecha Declaración	1	0,00
País	7	0,23		Declaración Caso Especial	1	0,00
Genero / Sexo	6	0,19		Se considera víctima (Tipo acto)	1	0,00
Fecha de nacimiento	6	0,19		Tipo de tramite	1	0,00
Fecha de solicitud / elaboración / inscripción	6	0,19		Tipo de afiliación	1	0,00
Tipo persona / Tipo Titular	5	0,15		Tipo de cotizante	1	0,00
Pertenencia Étnica / Origen étnico	5	0,15		Depto. / Ciudad	1	0,00
Ciudad/municipio nacimiento	5	0,15		Nivel Sisben	1	0,00
Corregimiento / Vereda / Sector / Barrio	5	0,15		Población especial	1	0,00
Actividad ppal / Actividad económica	5	0,15		Adm. ARL	1	0,00
País de nacimiento	4	0,12		Adm. Pensiones	1	0,00
Entidad que atiende / Cámara de Comercio	4	0,12		IBC	1	0,00
Nacionalidad	3	0,08		Matricula / Renovación / Inscripción	1	0,00
Lugar de expedición	3	0,08		Datos licencia de conducción (No, exp., fecha exp., estado, restricción)	1	0,00
Fecha de expedición	3	0,08		Multas e infracciones	1	0,00
Régimen	3	0,08		Certificado medico	1	0,00
Tipo de organización / Tipo establecimiento	3	0,08		Certificado de aptitud	1	0,00
Título pregrado	3	0,08		Info solicitudes	1	0,00
Universidad pregrado	3	0,08		Tipo intermediario	1	0,00
Lugar pre-grado	3	0,08		Minerales	1	0,00
Fecha de pre-grado	3	0,08		Ed. básica y media	1	0,00
Lugar postgrado	3	0,08		Título obtenido	1	0,00
Fecha de postgrado	3	0,08		Fecha grado	1	0,00

Titulo postgrado	3	0,08		Organización Jurídica	1	0,00
Universidad postgrado	3	0,08		Operador / Administrador: Libranza	1	0,00
Estado civil	2	0,04		Entidad	1	0,00
Datos de contacto (Teléfonos, Dirección, email) - Alternos	2	0,04		Estado RNPMV	1	0,00
Tipo de afiliado	2	0,04		Modalidad	1	0,00
Discapacidad	2	0,04		Especialidad	1	0,00
Estado	2	0,04		Vigencia desde	1	0,00
Numero de inscripción / Num. Matricula	2	0,04		Vigencia hasta	1	0,00
Actividad secundaria	2	0,04		Fecha inscripción	1	0,00
Otras actividades	2	0,04		Cod. institución	1	0,00
Idiomas	2	0,04		Ppal o seccional	1	0,00
Experiencia laboral	2	0,04		Naturaleza jurídica	1	0,00
Código tipo de entidad RNVEI	2	0,04		Sector	1	0,00
Código RNVEI	2	0,04		Carácter académico	1	0,00
Calidad Inscrito	2	0,04		Norma de creación	1	0,00
Estado RNVEI	2	0,04		Fecha de la norma	1	0,00
Situación legal	2	0,04		Acreditada alta calidad	1	0,00
Dirección Web	2	0,04		Fecha acreditación	1	0,00
Fax	2	0,04		Resolución acreditación	1	0,00
No. miembros junta directiva	2	0,04		Vigencia acreditación	1	0,00
Representante legal	2	0,04		Cobertura	1	0,00
Entidad vigilante	2	0,04				