El futuro es de todos

DNP
Departamento
Nacional de Planeación

El futuro es de todos

DNP
Departamento Nacional de Planeación

Clasificación de proyectos radicados en el Senado

Unidad de Científicos de Datos
Dirección de Desarrollo Digital

Agosto, 2020

Agenda

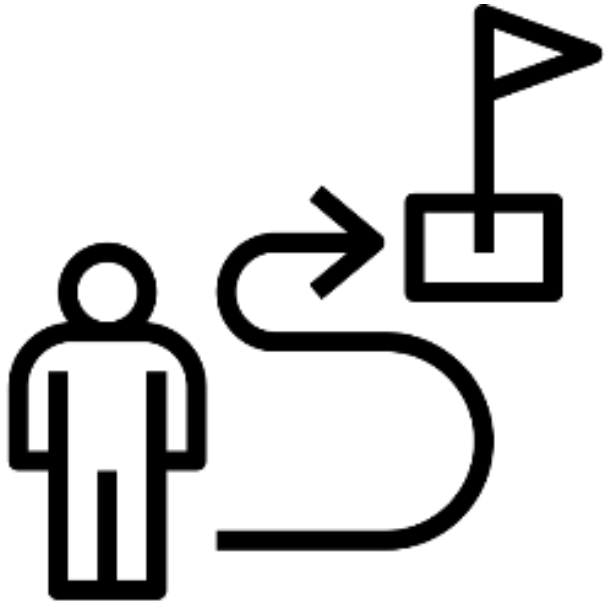
1. Introducción
2. Metodología
3. Resultados – archivos compartidos
4. Conclusiones

1. Introducción

Objetivo

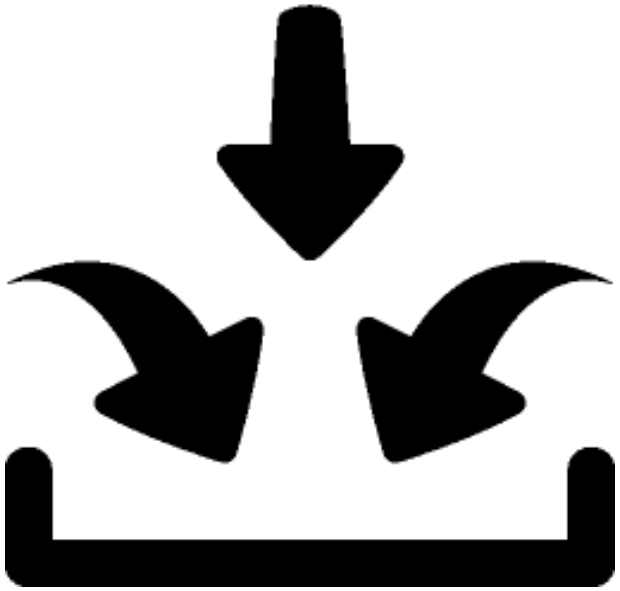
Descargar y detectar proyectos radicados en el Senado de la República que sean de interés para MinTIC

Metodología



1. Descargar documentos por *web scraping*
2. Transformar archivos a texto plano
3. Búsqueda de términos clave de interés para MinTIC
4. Generación de resultados fáciles de compartir

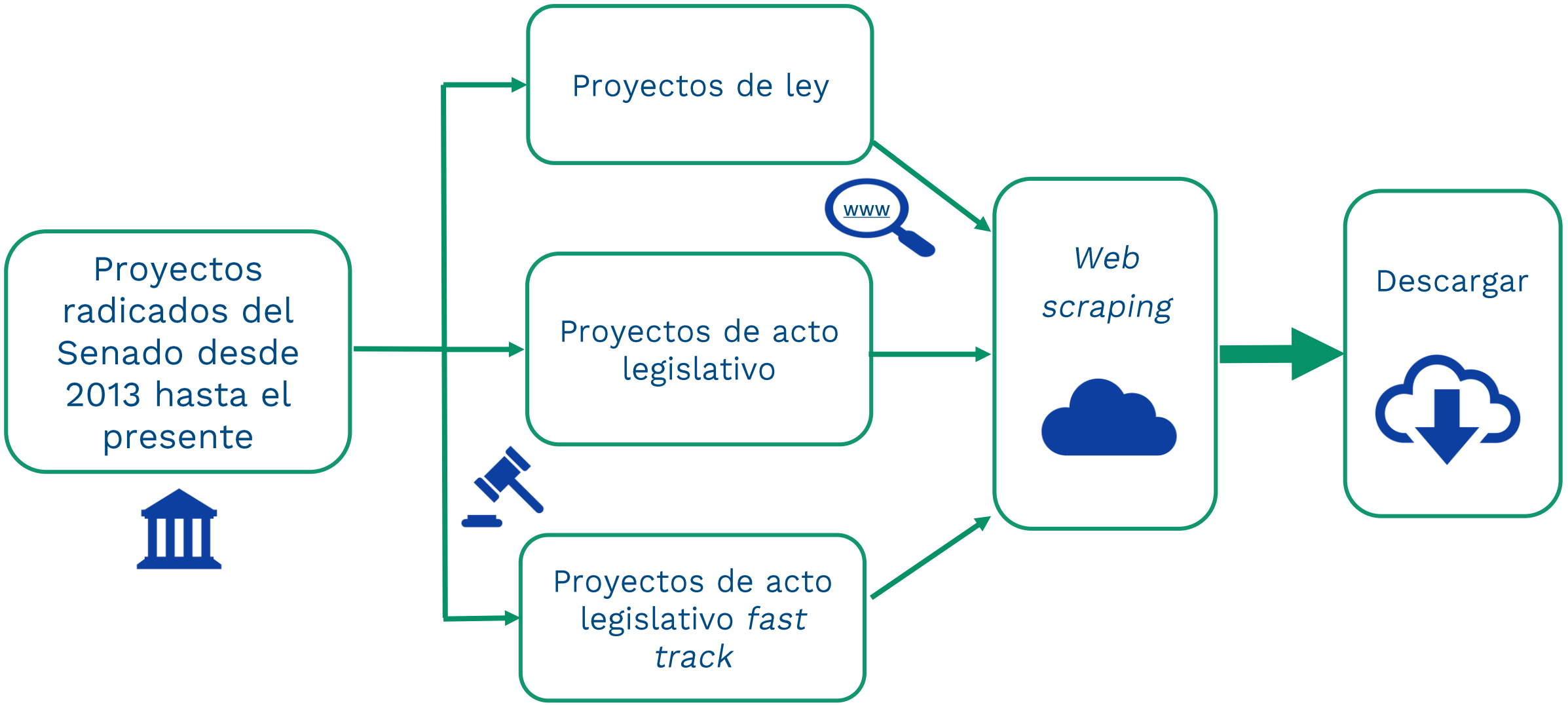
Insumos



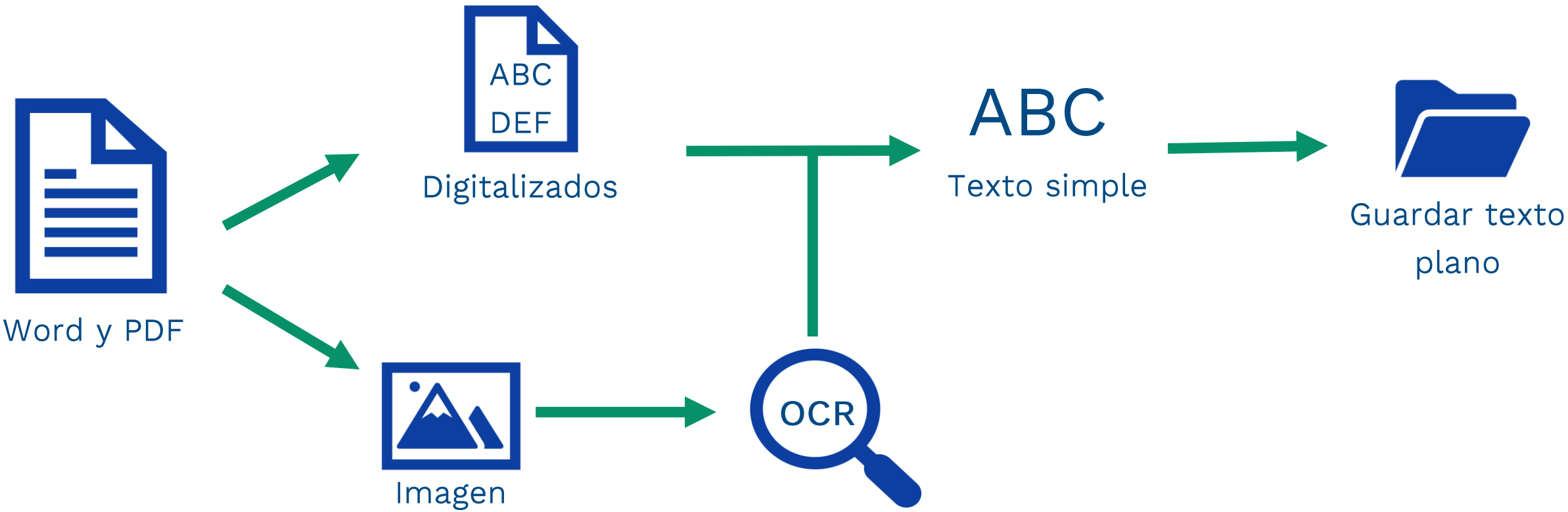
Archivos Word y PDF de la página web del Senado de la República y los metadatos de cada proyecto

2. Metodología

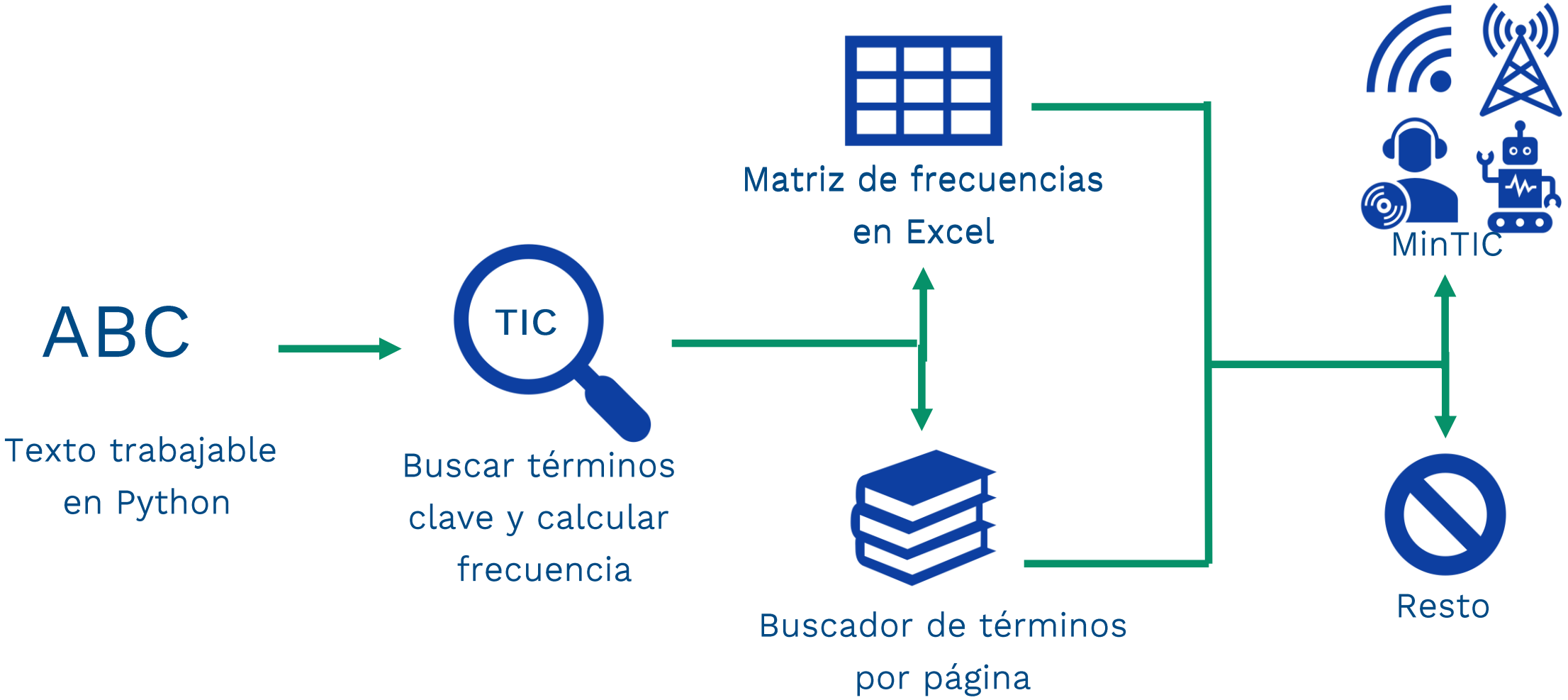
1. Descarga de proyectos



2. Transformación de archivo a texto

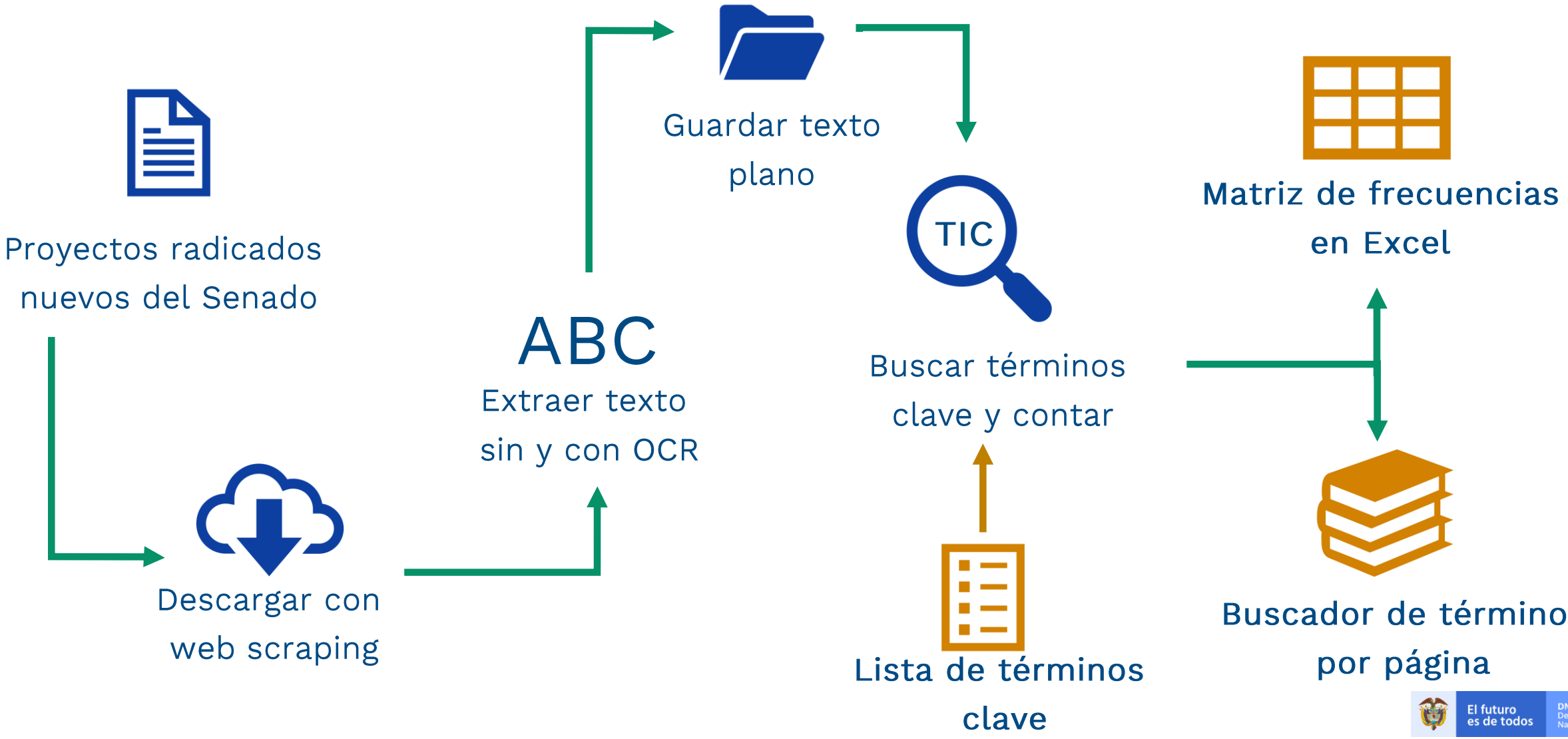


3. Búsqueda de términos clave



Resumen de proceso

Los miembros de MinTIC tendrán acceso a 3 archivos



3. Resultados – archivos compartidos

Lista de términos clave

	A	B
1	Término	Peso
2	audiovisual	5
3	brecha|digital	5
4	ciber	5
5	cobertura	1
6	comercio|electrónico	5
7	conectividad	3
8	conexión	2
9	despliege|infraestructura	2
10	digital	5
11	digitalización	5
12	dispositivo	1
13	evision	5
14	e vision	5
15	e-vision	5
16	eléctrico	1
17	electrónico	1
18	emisora	2
19	emprendimiento|digital	5
20	espectro	3
21	fondo único	2
22	gobierno|digital	5
23	imttelecomunicación	4

Hoja1

Lista de Excel con términos clave y su peso

Único insumo que se debe proporcionar desde MinTIC

Lista de términos clave

Términos que se pueden buscar

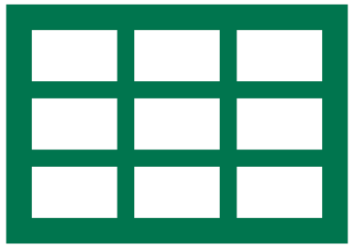
Palabras: “electrónico”
Raíces de palabras: “electr”, “ciber”
Expresiones: “gobierno digital”
Dúo de términos : “gobierno | digital”

Sensible a espacios

“ciber” no es lo mismo que
“ ciber”

Término	Peso
audiovisual	5
brecha|digital	5
ciber	5
cobertura	1
comercio|electrónico	5
conectividad	3
conexión	2
despliege|infraestructura	2
digital	5
digitalización	5
dispositivo	1
evision	5
e vision	5
e-vision	5
eléctrico	1
electrónico	1
emisora	2
emprendimiento|digital	5
espectro	3
fondo único	2
gobierno|digital	5

Resultado 1: matriz de términos clave en Excel



Matriz de frecuencias en Excel

Información de cada proyecto

Código
Año
Tipo

Frecuencia de términos clave

Total
Total ponderado
Por término

Otros

Indicador de OCR
Medidas de calidad

Resultado 1: matriz de términos clave en Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Llave	Número	Tipo	Código	Año	Número de páginas	OCR	Términos clave total	Términos clave total ponderado	audiovisual	brecha|digital	cambio|digital	ciber	cobertura	com
2	1000-PL-108-2017_2018	1000	PL	108	2017_2018	13	0	10	27	0	0	0	0	0	
3	1001-PL-109-2017_2018	1001	PL	109	2017_2018	55	0	15	19	0	0	0	0	2	
4	1002-PL-110-2017_2018	1002	PL	110	2017_2018	8	0	2	4	0	0	0	0	0	
5	1003-PL-112-2017_2018	1003	PL	112	2017_2018	16	0	3	3	0	0	0	0	0	
6	1004-PL-113-2017_2018	1004	PL	113	2017_2018	16	0	4	5	0	0	0	0	0	
7	1005-PL-114-2017_2018	1005	PL	114	2017_2018	15	0	0	0	0	0	0	0	0	
8	1006-PAL-008-2017_2018	1006	PAL	008	2017_2018	11	0	2	2	0	0	0	0	2	
9	1007-PL-117-2017_2018	1007	PL	117	2017_2018	25	0	4	4	0	0	0	0	1	
10	1008-PL-118-2017_2018	1008	PL	118	2017_2018	60	0	24	26	0	0	0	0	4	
11	1009-PL-119-2017_2018	1009	PL	119	2017_2018	29	0	12	12	0	0	0	0	0	
12	1010-PL-120-2017_2018	1010	PL	120	2017_2018	68	0	21	23	0	0	0	0	0	
13	1011-PAL-009-2017_2018	1011	PAL	009	2017_2018	22	0	0	0	0	0	0	0	0	
14	1012-PL-122-2017_2018	1012	PL	122	2017_2018	7	0	4	8	0	0	0	0	0	
15	1013-PL-123-2017_2018	1013	PL	123	2017_2018	47	0	10	19	0	0	0	0	0	
16	1014-PL-124-2017_2018	1014	PL	124	2017_2018	24	0	6	6	0	0	0	0	0	
17	1015-PL-125-2017_2018	1015	PL	125	2017_2018	16	0	1	5	0	0	0	1	0	
18	1016-PL-126-2017_2018	1016	PL	126	2017_2018	19	0	2	4	0	0	0	0	0	
19	1017-PL-127-2017_2018	1017	PL	127	2017_2018	35	0	10	16	0	0	0	0	0	
20	1018-PL-131-2017_2018	1018	PL	131	2017_2018	9	0	3	4	0	0	0	0	2	
21	1019-PL-132-2017_2018	1019	PL	132	2017_2018	6	0	1	5	0	0	0	0	0	
22	1020-PLFT-11-2017	1020	PLFT	11	2017	43	1	7	7	0	0	0	0	0	
23	1021-PL-133-2017_2018	1021	PL	133	2017_2018	6	0	32	59	0	0	1	0	0	

Resultado 2: diccionario de términos clave por página

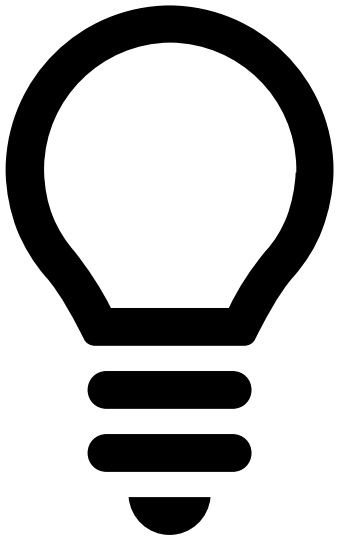
Diccionario de términos en aplicación HTML

1. Año o período legislativo
2. Código del proyecto
3. Término clave
4. Frecuencia por página


```
"1465_PL011": ⊕{1 item},
"1466_PL012": ⊖{
    "electrónico": ⊕{1 item},
    "información": ⊖{
        "1": 1,
        "3": 2,
        "4": 2,
        "5": 2,
        "6": 1,
        "8": 1,
        "11": 5,
        "12": 1,
        "14": 1
    },
    "tecnológico": ⊕{1 item}
},
"1467_PL013": ⊕{1 item},
"1468_PL014": {},
```

4. Conclusiones

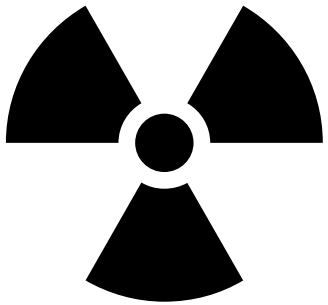
Conclusiones



- Gran ahorro de tiempo y esfuerzo en proceso de descarga y búsqueda de términos
- Aplicable para más sectores

Consideraciones y riesgos a tener en cuenta

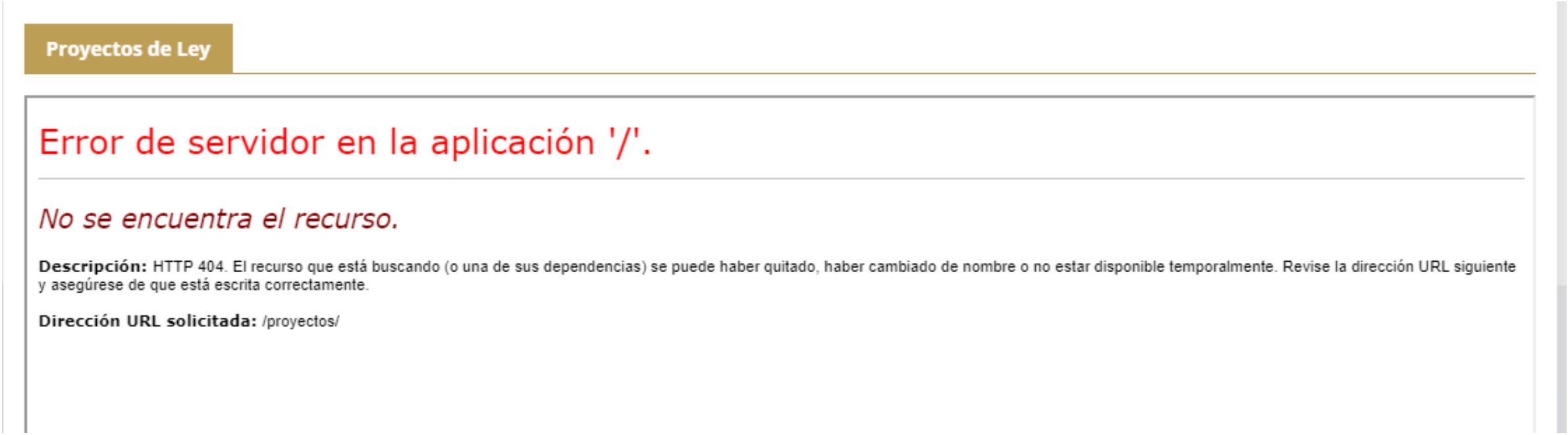
Hay errores inherentes a la minería de texto y a la práctica de *web scraping*



- Extracción de texto no siempre es perfecta
  - Textos escaneados
  - Codificación del texto
  - Archivos corrompidos
  - Enlaces dañados
- Riesgo de depender de páginas web
  - Página web inestable
  - Cambio en diseño de la página web

Aspectos y riesgos a tener en cuenta

Hay errores inherentes a la minería de texto y riesgos por depender de una página web



El futuro es de todos

DNP
Departamento
Nacional de Planeación