Aprendizajes de la digitalización y el indizado de los dictámenes históricos de la Procuración del Tesoro de la Nación

Pablo Factorovich[1-2], Axel Iglesias [1], María Angélica Lobo[3], Andrea Pelecano[3], Federico del Río[1] y Pablo Rodrigo Salgán Ruiz[1]

[1] Procuración del Tesoro de la Nación, Posadas 1641,CABA, 1112, Argentina
{pfactorovich,aiglesias,fdelrio,prsalgan}@ptn.gob.ar

[2] Universidad Nacional de Quilmes, Roque Saenz Peña 352, Bernal, Provincia de Buenos Aires, 1876, Argentina

[3] Dirección Nacional de Registro Oficial, Secretaría Legal y Técnica de la Presidencia de la Nación, Campichuelo 553, CABA, 1405, Argentina
{mlobo,apelecano}@boletinoficial.gob.ar

Abstract. La Procuración del Tesoro de la Nación a la fecha ha emitido más de 35.000 dictámenes para todos los servicios jurídicos de la Administración Pública Nacional. Estos documentos se encuentran en 316 tomos en papel desde la década de 1942 y solamente desde el siglo XXI se produjeron, almacenaron y disponibilizaron para consulta en formatos digitales. En este trabajo se relata el proceso por el cual la Procuración ha escaneado los mencionados volúmenes en colaboración con la Secretaría Legal y Técnica de la Nación, para que sea posible buscar y jerarquizar los dictámenes en relación a una búsqueda textual arbitraria que puedan hacer personas de la población en general. Se destacan los problemas encontrados, en vistas a una sistematización que pueda hacer el Estado Nacional sobre los procesos de digitalización de la documentación con que este cuenta.

Keywords: Digitalización, transcripción, indizado

1 Introducción

1.1 Descripción del problema

La **Procuración del Tesoro de la Nación** (PTN) es el órgano superior del Cuerpo de Abogados del Estado, de los servicios jurídicos de cada Ministerio y de cada ente de la Administración Pública Nacional. Esto implica que la PTN es la última instancia de asesoramiento en el procedimiento administrativo. Una de los servicios que caracterizan a la PTN es la potestad de **emitir dictámenes**. Tomando la definición elaborada en Broquel, la revista de la PTN: "Un dictamen es el análisis exhaustivo de una situación jurídica determinada, realizado conforme las normas vigentes y los principios generales, para recomendar conductas acordes con la justicia y el interés legítimo de las partes involucradas en la consulta."[2] De este modo, la jurisprudencia administrativa

que emana de los dictámenes de la PTN es vinculante y obligatoria para los servicios jurídicos que componen el Cuerpo de Abogados del Estado. Los dictámenes otorgan un pronunciamiento definitivo no sujeto a debate o a una revisión posterior, salvo que concurran nuevas circunstancias de hecho, o que el contexto legal tenido en cuenta haya sufrido modificaciones, es ahí donde la Procuración puede revisar su opinión ya emitida [4]. Por tales motivos, el acceso a los dictámenes emitidos y su correspondiente doctrina (que abstrae la información del dictamen, además de proveer una justificación teórico jurídica al mismo) se vuelve indispensable para el trabajo jurídico de toda la Administración Pública Nacional (APN), permitiendo consultar cómo se debe proceder ante determinadas situaciones. A su vez, poder dar acceso público y sencillo a esta información se vuelve una necesidad, si queremos acercar a la población general a las tareas y definiciones que toma el organismo. La PTN a la fecha ha emitido más de **35.000 dictámenes** que se encuentran en **316 tomos**, algunos de dos o más volúmenes. Los **primeros de estos documentos datan de 1933**, siendo recopilados a partir de 1942, año de inicio de la Colección de Dictámenes de la PTN. La biblioteca de la PTN registró 7680 búsquedas en 2019 (4800 internas y 2880 externas) correspondientes a dictámenes del siglo XX y XXI y esto es una cota inferior de las búsquedas que podrá recibir en el futuro, ya que la disponibilidad digital permite que surjan más interesados. Por otra parte, el decreto 1131/16 dispuso en su art. 2º que "Los documentos y expedientes producidos en primera generación en soporte papel deberán ser digitalizados siguiendo el procedimiento que fije la SECRETARÍA DE MODERNIZACIÓN ADMINISTRATIVA", obligando a la PTN a digitalizar su colección de dictámenes.

El problema de recuperación de la información o RI (*information retrieval*, en inglés) parte de una base de documentos y una búsqueda de texto realizada por un usuario y consiste en devolver un subconjunto de la base que responden a la búsqueda. Para mayor utilidad, el subconjunto se presenta como una secuencia jerarquizada de acuerdo a la similitud de cada documento a la búsqueda realizada. Típicamente este problema se afrontó constituyendo un índice para la base de documental que es usado en cada búsqueda a realizar. Cada vez que un conjunto de documentos se incorpora a la base documental, se reconstruye el índice. Se espera que esta tarea sea menos frecuente que las búsquedas.

El **problema de RI para los dictámenes del siglo XXI de la PTN** ya ha sido abordado en las 49 JAIIO y mejorado a posteriori para permitir a los usuarios filtrar por ciertas secciones comunes a todos los dictámenes [5]. **Extender la base documental a todos los dictámenes de la PTN** resulta de utilidad para letrados de la APN y otros estados subnacionales, así como para la población en general e incluso historiadores.

Para que los documentos se encuentren en el índice usado para RI es necesario que se encuentren en formato digital. Si bien es teóricamente posible realizar la transcripción, para un volumen de decenas de miles de páginas esto insumiría demasiado tiempo y sería propenso a errores, además de no permitir contar con una imagen del documento original. Por esta razón **se ha optado por un proceso digitalización y transcripción automática** a través del uso de escáneres y aplicación posterior de OCR (*op-*

tic character recognition). Estos procesos presentan una cierta cantidad de fallos por documento, que es mayor en la medida en que el mismo es más antiguo o la calidad de papel y/o tinta es baja. **El objetivo no es contar con una transcripción perfecta de un dictamen, sino con una que permita encontrarlo al hacer una búsqueda donde este sea relevante.**

A continuación, explicaremos el proceso por el cual preparamos el material en papel para su escaneo (sección 2), cómo configuramos los escáneres y por qué (sección 3), cómo postprocesamos los escaneos y cuáles fueron los resultados obtenidos (sección 4) y las conclusiones y los pasos futuros (sección 5).

2 Preparación para escaneo

Los tomos de dictámenes son libros donde los documentos se encuentran cosidos en cuadernillos y estos pegados al lomo, por lo general con un pegamento muy resistente. Para pasar las hojas por un escáner industrial estas deben ser separadas de manera individual y no deben contener restos de pegamento ya que trabarían el delicado mecanismo de los escáneres. Realizar este trabajo sobre toda la colección ha insumido decenas de horas de trabajo. Dado que las hojas se encontraban en buen estado ha sido posible en general **quitar el pegamento mediante el uso cuidadoso de una amoladora**, cómo se puede ver en la figura 1.

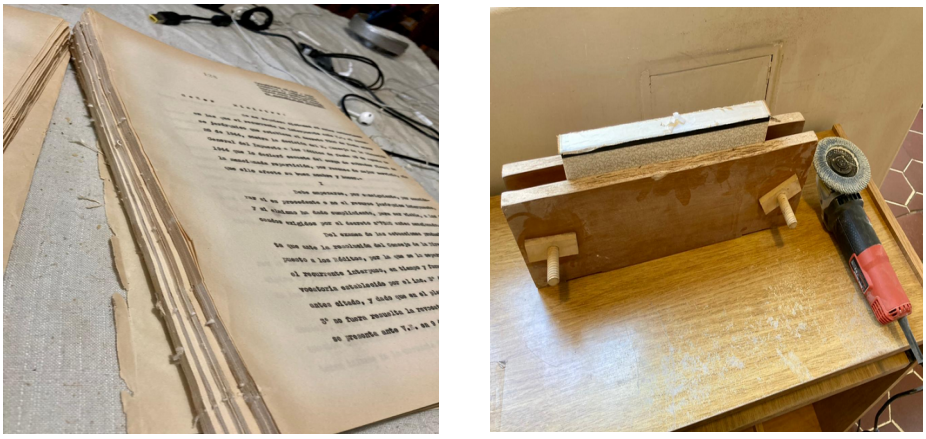

Fig. 1. Preparación de dictámenes para su escaneo

Los escáneres industriales permiten escanear una sucesión de documentos y nombrar automáticamente a los archivos, siempre que antes de comenzar las hojas correspondientes al mismo se inserte un **código de barras** con el respectivo nombre. Por lo tanto, luego de separar cada tomo fue necesario insertar una hoja con un código de barras previamente generado. Cada dictamen se identifica de manera unívoca con el par (número de tomo, número página), de modo que se debió tomar el índice de cada

tomo para, de forma manual volcarlos en una base a fin de generar los códigos de barra y luego imprimirlos. Dado que la cantidad de dictámenes implicaba imprimir más de cien resmas de códigos de barra, se decidió **reusar códigos de barra**: una base de datos identificaba qué dictamen correspondía a cada código de barras en cada envío a escanear.

Los envíos se organizaban -con todas las prevenciones del caso por tratarse de ejemplares únicos- de a varios tomos para **ser escaneados en el centro que a tal fin posee la Dirección Nacional del Registro Oficial, de la Secretaría Legal y Técnica de la Presidencia de la Nación**, a la que el Decreto 1131/16 encomienda la tarea de "registro y digitalización de archivos correspondientes a la Administración Pública Nacional". Allí las hojas fueron escaneados usando escáneres Kodak i4250, que **producen un archivo pdf** para cada dictamen (además de un XML con metadata), **incluyendo información textual conseguida a través de OCR**. Cada envío luego fue llevado nuevamente a la PTN para ser reencuadernado y los archivos para ser renombrados y postprocesados.

3 Configuración de escáneres y análisis de resultados

Cabe destacar que antes de hacer el escaneo masivos se probaron tres dictámenes de distinto tipo (buena calidad, impresión algo borrosa con buen papel, impresión algo borrosa con papel traslúcido), usando tres resoluciones de escaneo, siempre en blanco y negro: 300, 400 y 600 dpi (puntos por pulgada).

De las tres configuraciones de impresión, **la de 300 dpi dio los mejores resultados en cuanto a su OCR**.

Con el fin de evaluar la calidad de los dictámenes escaneados, decidimos tomar 15 dictámenes (3 documentos de 5 décadas distintas) que **transcribimos manualmente** y verificamos que no tuvieran errores. La calidad del OCR de cada documento la evaluamos tomando la **distancia de *Levenshtein*** entre la transcripción manual chequeada (documento de referencia) y el OCR generado por el escáner. Utilizamos esta unidad de distancia ya que el OCR podía haber producido, no solo caracteres transmutados, sino agregado o salteado caracteres. En la tabla 1 presentamos los resultados obtenidos, **normalizados por la cantidad de caracteres del documento referencia** y luego de eliminar los saltos de línea a la digitalización seguida de OCR.

Tabla 1. Distancia de Levenshtein normalizada entre el resultado del OCR Kodak y el dictamen transcripto.

Década	Dictamen	OCR Kodak
'40	001-001	26,60 %
	001-004	18,98 %
	004-008	11,27 %
50'	050-001	21,95 %
	050-059	20,23 %
	050-072	22,17 %

60’	080-013	14,57 %
	080-021	17,51 %
	080-023	14,25 %
70’	120-001	14,89 %
	120-005	12,55 %
	120-013	28,76 %
80’	075-302	13,02 %
	077-321	10,51 %
	078-058	15,05 %

Como se puede observar, la calidad del OCR fue muy disímil, siendo en algunos casos sumamente pobre.

4 Postprocesamiento y resultados obtenidos

A fin de mejorar la calidad del texto, probamos escanear el documento con los software de transcripción automática *Textract* de *Amazon* [1] y *Vision* de *Google* [3], ambos configurados para español. En la figura 2 presentamos los resultados obtenidos.

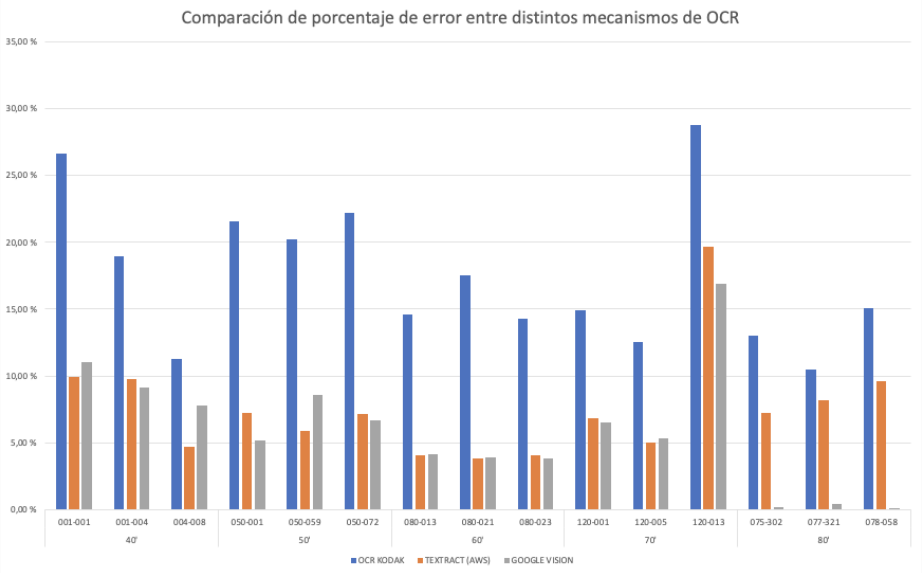

Fig. 2. Comparación de porcentaje de error entre distintos mecanismos de OCR.

Como puede apreciarse, la calidad ha tenido una mejora sustancial, quedando en casi todos los casos por **debajo del 10% de error** y siendo similar para los dos productos. En ambos casos este proceso tiene un costo económico que se puede eliminar

usando herramientas de software libre para tal fin. Resta evaluar los resultados que se obtienen al aplicar una buena herramienta de este tipo.

5 Conclusiones y trabajo a futuro

El proceso de transcripción de un gran volumen de documentos es una tarea artesanal e implica una logística no trivial, aún al automatizarla a través de escáneres y software de transcripción. **Se relata una experiencia de manejo de un gran volumen de información a digitalizar, que puede ser sistematizada.** Como parte de la misma se nombra automáticamente e identifica cada documento, reutilizando los códigos de barra y una base de datos para su identificación unívoca en los distintos envíos. **De esta menra optimizamos papel.** La calidad del OCR producido por los escáneres industriales, al menos para el español, es pobre y debe ser complementada con software específico para tal fin, que logra resultados aceptables en cuanto a la cantidad de caracteres leídos incorrectamente. **Se estableció una métrica algorítmica que permite determinar la calidad del escaneo y cuantificar la posibilidad de mejora del texto.**
Dado que es un problema habitual en dependencias del estado nacional, sería bueno que el mismo tenga protocolos para este fin y pueda asesorar en metodologías prácticas, de buena calidad y económicas para el desarrollo de estas tareas. Resta evaluar en cantidad de errores la calidad del software libre de transcripción para comparar su efectividad en relación al costo de uso. Asimismo, dado el fin último de estas transcripciones, se comparará la calidad obtenida por la transcripción de escáneres, software pago y software libre a la hora de buscar documentos. A ese fin pueden usarse la base documental de dictámenes de la PTN del siglo XXI y las búsquedas reales que realizan los usuarios sobre los mismos y luego insertar errores de manera aleatoria en los documentos, para obtener en los mismos una determinada tasa de error similar al que obtienen los distintos software de transcripción evaluados. Sobre esta base pueden repetirse las búsquedas y comparar los documentos devueltos en comparación con los que arrojaba la búsqueda sobre la base documental sin errores.

Referencias

1 https://aws.amazon.com/es/textract/
2 https://broquel.ptn.gob.ar/broquel/2020/04/12/la-procuracion-del-tesoro-y-sus-dictamenes/
3 https://cloud.google.com/vision
4 Dictámenes de la Procuración del Tesoro de la Naciión, 283:211
5 https://www.ptn.gob.ar/buscarDictamenes/page/acceder-al-buscador