UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

DEPARTAMENTO DE OBRAS CIVILES

METODOLOGÍA DE EXTRACCIÓN DE INFORMACIÓN DE DOCUMENTOS Y UNA HERRAMIENTA COMPUTACIONAL PARA LA GESTIÓN, VISUALIZACIÓN GEOESPACIAL Y GENERACIÓN DE REPORTES CON LOS DATOS DE LAS ÓRDENES DE TRABAJO

Memoria de Título presentada por

Vicente Alejandro Moraga Jerez

como requisito parcial para optar al título de la carrera de

Ingeniería Civil

Profesor Guía
Héctor Allende Cid

Profesor Correferente
Alelí Osorio Lird

Marzo de 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: METODOLOGÍA DE EXTRACCIÓN DE INFORMACIÓN DE DOCUMENTOS Y UNA HERRAMIENTA COMPUTACIONAL PARA LA GESTIÓN, VISUALIZACIÓN GEOESPACIAL Y GENERACIÓN DE REPORTES CON LOS DATOS DE LAS ÓRDENES DE TRABAJO.

Nombre del candidato(a): VICENTE ALEJANDRO MORAGA_JEREZ

Carrera / Grado: INGENIERIA CIVIL

Campus: SAN JOAQUÍN **Departamento:** OBRAS CIVILES

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Alelí Osorio Lird, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción):**

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 05 de julio 2026_____**Firma:** _____

Estudiante o Candidato(a): VICENTE ALEJANDRO MORAGA JEREZ

Fecha: 05 de julio 2026_____**Firma:** _____

TITULO DE LA MEMORIA:

METODOLOGÍA DE EXTRACCIÓN DE INFORMACIÓN DE DOCUMENTOS Y UNA HERRAMIENTA COMPUTACIONAL PARA LA GESTIÓN, VISUALIZACIÓN GEOESPACIAL Y GENERACIÓN DE REPORTES CON LOS DATOS DE LAS ÓRDENES DE TRABAJO

AUTOR:

Vicente Alejandro Moraga Jerez

TRABAJO DE MEMORIA, presentado como requisito parcial para optar al título de la carrera de INGENIERIA CIVIL de la Universidad Técnica Federico Santa María.

	Nombre	Firma
Prof. Guía	Héctor Allende Cid.....	.....
Miembro 1 Comisión	Alelí Osorio Lird.....	.....
Miembro 2 Comisión	.....	.....

Santiago, Chile, marzo de 2026

METODOLOGÍA DE EXTRACCIÓN DE INFORMACIÓN DE DOCUMENTOS Y UNA HERRAMIENTA COMPUTACIONAL PARA LA GESTIÓN, VISUALIZACIÓN GEOESPACIAL Y GENERACIÓN DE REPORTES CON LOS DATOS DE LAS ÓRDENES DE TRABAJO

Vicente Alejandro Moraga Jerez[1], Alelí Osorio Lird[1], Héctor Allende Cid[2]

[1] UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
[2] PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Resumen

La conservación de pavimentos urbanos constituye un desafío clave para las municipalidades, dado que el deterioro progresivo de la infraestructura vial impacta directamente en la movilidad, la seguridad y la calidad de vida de la población. Los Sistemas de Gestión de Pavimentos (SGP) han demostrado ser herramientas eficaces para optimizar recursos y planificar la conservación, sin embargo, su implementación en Chile se ha visto limitada por los altos costos de levantamiento de información y las restricciones presupuestarias de los organismos locales.

En la comuna de Santiago, la Ilustre Municipalidad de Santiago (IMS) tiene la responsabilidad exclusiva sobre la gestión de pavimentos urbanos, recurriendo a procedimientos mayormente reactivos para su conservación. En este contexto, las Órdenes de Trabajo (OT) generadas durante la ejecución de obras de conservación representan una fuente valiosa de información, actualmente subutilizada debido a que su procesamiento se realiza de manera manual.

El presente trabajo propuso una metodología de bajo costo y fácil implementación para la digitalización, extracción y procesamiento de la información contenida en las OT físicas. A través de técnicas de reconocimiento óptico de caracteres (OCR) y herramientas de análisis de datos, se buscó estructurar la información en bases de datos que permitan generar reportes, indicadores y visualizaciones que apoyen a la toma de decisiones. Finalmente, la metodología fue validada mediante la comparación con documentos oficiales, evaluando su efectividad y eficacia.

La propuesta contribuirá al apoyo de la gestión de pavimentos en la comuna de Santiago, ofreciendo un enfoque replicable y escalable para otras unidades municipales.

Palabras claves: Gestión de pavimentos urbanos; Conservación vial; Órdenes de Trabajo; Digitalización documental; Reconocimiento óptico de caracteres (OCR); Extracción de información; Análisis de datos.

1. Introducción y contextualización

La gestión de infraestructura vial está relacionada con el desarrollo económico y social de los países. Los pavimentos son un elemento básico de la infraestructura vial que requieren de recursos considerables para su construcción, conservación y mantenimiento. El deterioro en los pavimentos es esperable, por lo que es fundamental contar con Sistemas de Gestión de Pavimentos que permita asignar de manera eficiente el uso de recursos para su construcción, operación y conservación de los caminos (de Solminihac et al., 2018).

La gestión eficiente de los recursos para la conservación de las vías públicas requiere de información específica y actualizada de los caminos, la obtención de tal información se ve dificultada por el costo asociado, lo que hace que el sistema público opte por inspecciones más rudimentarias y poco eficientes (de Solminihac et al., 2018; López et al., 2020; Whiteley-Lagace et al., 2011).

En la comuna de Santiago, la Ilustre Municipalidad de Santiago se encarga de todo trabajo de ejecución, renovación, conservación, reparación y vigilancia de los pavimentos de las calzadas y aceras de las vías, siendo la única comuna en Chile capaz de administrar su infraestructura vial (Ley N° 11.150, 1953). Pese a su autonomía, el presupuesto limitado y falta de herramientas dificultan la implementación de un sistema de gestión para la conservación de los pavimentos, recurriendo a procedimientos reactivos más que preventivos para la conservación de pavimentos, además de depender del trabajo manual de los funcionarios para el análisis de datos, por lo que se hace necesario encontrar una manera alternativa con la cual se pueda gestionar el patrimonio vial que sea a bajo costo y de fácil implementación.

Los datos históricos sobre los pavimentos son de importancia para los sistemas de gestión de pavimentos urbanos y en general se tiene que, en las ciudades con redes viales densas no se cuentan con datos históricos de mantenimiento urbano debido a la falta de recursos económicos y técnicos para la recolección de datos (de Solminihac et al., 2018).

Las Órdenes de Trabajo de las obras de conservación de pavimentos de la comuna de Santiago son documentos que pueden resultar útiles para la gestión y futura aplicación de algún sistema de gestión de pavimentos urbanos, sin embargo, actualmente la información contenida en estos documentos físicos se extrae de forma manual, teniendo un alto impacto en el recurso humano, por lo que se propuso desarrollar una metodología que permita aprovechar de mejor manera este insumo en base a los recursos disponibles.

2. Marco teórico

2.1 Sistemas de Gestión de Pavimentos

Un Sistema de Gestión de Pavimentos o SGP es el conjunto de operaciones que tienen por objetivo preservar las condiciones adecuadas para la circulación, soportando las condiciones de tráfico, clima y entorno, considerando los costos y las externalidades (de Solminihac et al., 2018).

A nivel internacional se resalta el impacto positivo de los SGP en la conservación de pavimentos al reducir costos de operación y extender la vida útil de las vías (Shahin, 2005). Sin embargo, su implementación en municipios pequeños y medianos enfrenta dificultades asociadas a la disponibilidad de información y recursos técnicos (Whiteley-Lagace et al., 2011).

En Chile se han desarrollado iniciativas para adaptar estas herramientas al contexto local como el proyecto FONDEF D09I1018 titulado *"Investigación y Desarrollo de Soluciones para la Gestión de Pavimentos Urbanos en Chile"*, el cual tenía como objetivo desarrollar un sistema que apoyara a los municipios en la planificación de conservación vial. Un resultado del proyecto fue la generación de la herramienta SGPUC la cual integra modelos de deterioros probabilísticos y modelos decisionales de optimización y priorización robustos (de Solminihac et al., 2018).

Tradicionalmente la gestión de pavimentos urbanos se ha centrado en el desarrollo de modelos predictivos y en la auscultación en terreno, dejando de lado el manejo de la información histórica no estructurada por la complejidad y falta de datos existentes (FONDEF, 2015; Whiteley-Lagace et al., 2011). Por otro lado, los avances en digitalización y Reconocimiento Óptico de Caracteres (OCR) han revolucionado la gestión documental en diversos sectores (Peddarapu et al., 2025; Sciannameo et al., 2024), más no en el área vial.

En el contexto nacional, una de las pocas herramientas de apoyo disponibles para el sector público en materia de conservación de pavimentos urbanos es la Metodología Simplificada para evaluar proyectos de Mantenimiento Vial Urbano (MANVU/MANVISIMP), elaborado por el SERVIU en 1992 (de Solminihac et al., 2018). Este instrumento permite priorizar inversiones y seleccionar alternativas de conservación en base a criterios técnicos y económicos, facilitando la asignación de recursos en redes urbanas. Sin embargo, su aplicación se ha visto restringida debido a que requiere de información de inventarios y levantamientos de condición que muchas municipalidades no están en condiciones de generar, por lo que en la práctica su uso es limitado.

La existencia del MANVUSIMP y el SGPUC refleja tanto el interés por introducir metodologías de gestión a nivel urbano como la brecha existente entre la disponibilidad de herramientas y la capacidad real de los municipios para implementarlas debido a la falta de un levantamiento adecuado de la información a nivel de red.

Este antecedente refuerza la necesidad de explorar metodologías alternativas que se ajusten a las condiciones locales en términos económicos y de capacidad.

2.2 Deterioros en los pavimentos

El pavimento constituye la capa superficial de la infraestructura vial destinada a soportar las cargas del tránsito, además de proporcionar condiciones adecuadas de seguridad y confort a los usuarios. Debido a la acción combinada de factores mecánicos, climáticos y el envejecimiento de los materiales, los pavimentos presentan deterioros progresivos que afectan su desempeño con el tiempo (de Solminihac et al., 2018; Shahin, 2005).

Entre las principales causas de deterioro se encuentran:

- Factores mecánicos, como la acción del tránsito vehicular, especialmente el transporte pesado.
- Factores climáticos, como las variaciones de temperatura y la infiltración de agua, que aceleran procesos de fatiga, pérdida de adherencia y disgregación de materiales.
- Factores de envejecimiento, asociados a la oxidación y pérdida de flexibilidad de los ligantes en mezclas asfálticas o al desgaste superficial en pavimentos rígidos (Shahin, 2005).

La presencia de deterioros incrementa los costos de conservación si no se atienden oportunamente, lo que resalta la importancia de adoptar enfoques preventivos antes de que evolucionen a deterioros de mayor gravedad (Shahin, 2005).

La necesidad de reparación surge no solo para garantizar la seguridad y confort de los usuarios, también para preservar la vida útil de los pavimentos y evitar que los costos de mantención aumenten. Los trabajos de conservación se convierten en una acción fundamental para restituir las condiciones de servicio. La gestión adecuada de estas acciones resulta clave para asegurar la eficiencia en el uso de los recursos públicos destinados a la infraestructura vial (FONDEF, 2015).

2.3 Órdenes de Trabajo y Estados de Pago

Las Órdenes de Trabajo (OT) son documentos físicos mediante los cuales el Departamento de Construcción y Conservación de la Subdirección de Pavimentación de la Ilustre Municipalidad de Santiago solicita las obras de conservación a la empresa adjudicada al contrato de obras de conservación y trabajos complementarios. Estos documentos contienen información relevante organizada en formato de tablas en donde se registran datos como la fecha, la ubicación de la obra, el Inspector Técnico de Obra (ITO) municipal a cargo, las actividades a ejecutar y las cubicaciones correspondientes. Un conjunto de Órdenes de Trabajo conforma un Estado de Pago (EDP), el cual es utilizado posteriormente para permitir el traspaso de recursos desde el municipio hacia la empresa que ejecuta las obras.

En la práctica, las OT constituyen documentos oficiales que respaldan la trazabilidad de los trabajos de conservación realizados en los pavimentos de la comuna, sin embargo, la información contenida en ellas se extrae actualmente de manera manual para los análisis de gestión que realiza la Subdirección de Pavimentación, lo que representa un proceso con alto costo en recursos humanos y tiempo. Esta situación no es exclusiva de la comuna de Santiago, ya que estudios previos han señalado que las limitaciones en la sistematización de datos y la dependencia de procedimientos manuales dificultan la implementación de sistemas de gestión de pavimentos en municipios, especialmente en contextos con restricciones presupuestarias y técnicas (FONDEF, 2015; Whiteley-Lagace et al., 2011). En este escenario, la automatización de procesos aparece como una alternativa viable para la gestión de pavimentos y aprovechar de mejor manera los recursos disponibles.

La comuna de Santiago para efectos de la gestión administrativa de las obras de conservación se divide en Sector 1 (zona norte) y Sector 2 (zona sur) delimitados por la Avenida Libertador Bernardo O´Higgins. Cada sector a lo largo del año genera diferentes estados de pago (1 por mes) que contienen una cantidad variable de OT. Estos documentos son manejados íntegramente en formato físico sin un respaldo digital.

2.4 Digitalización y OCR (Reconocimiento Óptico de Caracteres)

La digitalización de documentos constituye el primer paso para transformar información física en un formato que pueda ser procesado computacionalmente. En el ámbito de la gestión pública, esta práctica no solo contribuye a la conservación de los documentos originales, sino que además permite mejorar el acceso, la consulta y el análisis de la información (Martínez et al., 2012). La preservación digital resulta especialmente relevante cuando se trata de expedientes históricos o administrativos como las Órdenes de Trabajo que representan una fuente primaria de información sobre las obras realizadas sobre los pavimentos.

El Reconocimiento Óptico de Caracteres (OCR por sus siglas en inglés) es una tecnología que posibilita la conversión de documentos digitalizados en imágenes a texto legible por una computadora. Por medio del OCR, los caracteres impresos en los documentos escaneados son detectados y convertidos en texto digital, lo que posibilita las búsquedas automáticas, la extracción de datos y su integración en bases de datos estructuradas (Sciannameo et al., 2024).

La tecnología del OCR tradicional cuenta con limitaciones como la lectura de figuras y gráficos, puede presentar problemas con textos inclinados, sellos, firmas, entre otras limitaciones técnicas que dependen de la calidad de digitalizado. Para mejorar la extracción el nivel de la extracción de información se han adoptado metodologías que utilizan modelos semánticos, de agrupamiento y de clasificación que, si bien

han mostrado un buen desempeño, requieren de un alto poder de cómputo siendo necesario el uso de soluciones basadas en la nube (Peddarapu et al., 2025).

Otras herramientas más avanzadas como los modelos de lenguaje visuales (VLM) y los modelos de lenguaje multimodales (MLLM) logran una recuperación eficiente de la información respecto al uso del OCR tradicional, sin embargo, sigue siendo necesario implementar soluciones basadas en servicios de la nube para no requerir computadoras de alto rendimiento de manera física (Peddarapu et al., 2025; Tanasa & Oprea, 2025).

En el caso de la gestión de pavimentos, disponer de información confiable y actualizada es fundamental para la correcta asignación de recursos, lo que refuerza la importancia de contar con procesos de digitalización eficientes (de Solminihac et al., 2018). En consecuencia, la digitalización y el uso de OCR no solo resguardan la integridad documental, sino que además posibilitan el desarrollo de metodologías de gestión basadas en evidencia, al transformar documentos físicos en insumos útiles para el análisis sistemático y la toma de decisiones (Sciannameo et al., 2024).

2.5 Extracción y estructuración de datos

La extracción de información corresponde al proceso mediante el cual se identifican y capturan datos relevantes desde documentos digitales para posteriormente organizarlos en un formato estructurado que permita su análisis. Este procedimiento constituye un paso fundamental en metodologías de gestión de información, al reducir la dependencia del trabajo manual y optimizar el uso de recursos humanos (Peddarapu et al., 2025).

Los enfoques tradicionales de extracción de información se basan en la definición de reglas y expresiones regulares aplicadas al texto obtenido mediante OCR. Si bien estas técnicas ofrecen resultados aceptables en documentos con formatos homogéneos, su efectividad disminuye cuando existe variabilidad en la estructura de los documentos, podrían ser requeridos modelos de aprendizaje automático y de lenguaje natural (NLP por sus siglas en inglés) que permitan extraer información de manera más robusta y flexible (Sciannameo et al., 2024).

La estructuración de datos implica convertir los resultados de la extracción en bases de datos, donde cada campo de interés queda normalizado y estandarizado. La organización posibilita la generación de reportes, indicadores y visualizaciones que sirven de apoyo a la toma de decisiones. En el caso de los pavimentos urbanos, transformar las Órdenes de Trabajo en información estructurada abre la posibilidad de contar con un registro histórico sistemático de las actividades de conservación, que puede emplearse tanto para reportes inmediatos como para el desarrollo de modelos predictivos de deterioro.

2.6 Definición del problema

Pese a su autonomía, la Ilustre Municipalidad de Santiago enfrenta dificultades para implementar un SGP debido a limitaciones presupuestarias y de recursos humanos. La gestión de la información sobre conservación de pavimentos depende de procesos manuales que consumen tiempo y restringen el análisis sistemático y automático. La ausencia de metodologías adaptadas provoca que la toma de decisiones de conservación sea reactiva y no preventiva.

En consecuencia, surge la oportunidad de investigar si es posible desarrollar una metodología de bajo costo y fácil implementación, de tal forma de que no se requiera de una inversión económica ni personal especializado y que permitiera aprovechar la información contenida en las Órdenes de Trabajo existentes utilizando los recursos disponibles del municipio.

Considerando las limitaciones de la tecnología OCR para documentos digitalizados también surge la posibilidad de explorar de qué manera se puede mejorar el manejo y almacenamiento de la información en el proceso de gestión al interior de la Subdirección de Pavimentación con enfoque en la transformación digital. Un cambio de metodología al interior de la Subdirección podría permitir que se aproveche la información contenida en las Órdenes de Trabajo en insumos estructurados y hacerlos analizables para que apoyen la gestión de pavimentos urbanos en la comuna de Santiago.

3. Objetivos

3.1 Objetivo general

Proponer y validar un sistema integral, compuesto por una metodología de extracción de información de documentos y una herramienta computacional para la gestión, visualización geoespacial y generación de reportes con los datos de las Órdenes de Trabajo de la Subdirección de Pavimentación de la IMS considerando un enfoque de bajo costo y fácil implementación.

3.2 Objetivos específicos

OE1: Evaluar la extracción de datos de Órdenes de Trabajo mediante Reconocimiento Óptico de Caracteres (OCR) desde documentos digitalizados.

OE2: Diseñar una metodología y herramienta computacional para la extracción y gestión de datos de obras de conservación que sea aplicable a la Subdirección de Pavimentación de la IMS.

OE3: Validar eficacia de la herramienta y la metodología con reportes y visualización geoespacial de datos mediante el uso de Órdenes de Trabajo reales de la comuna.

4. Metodología

La metodología de investigación propuesta se estructuró en tres etapas, cada una vinculada a un objetivo específico como se ilustra en la Figura 1.

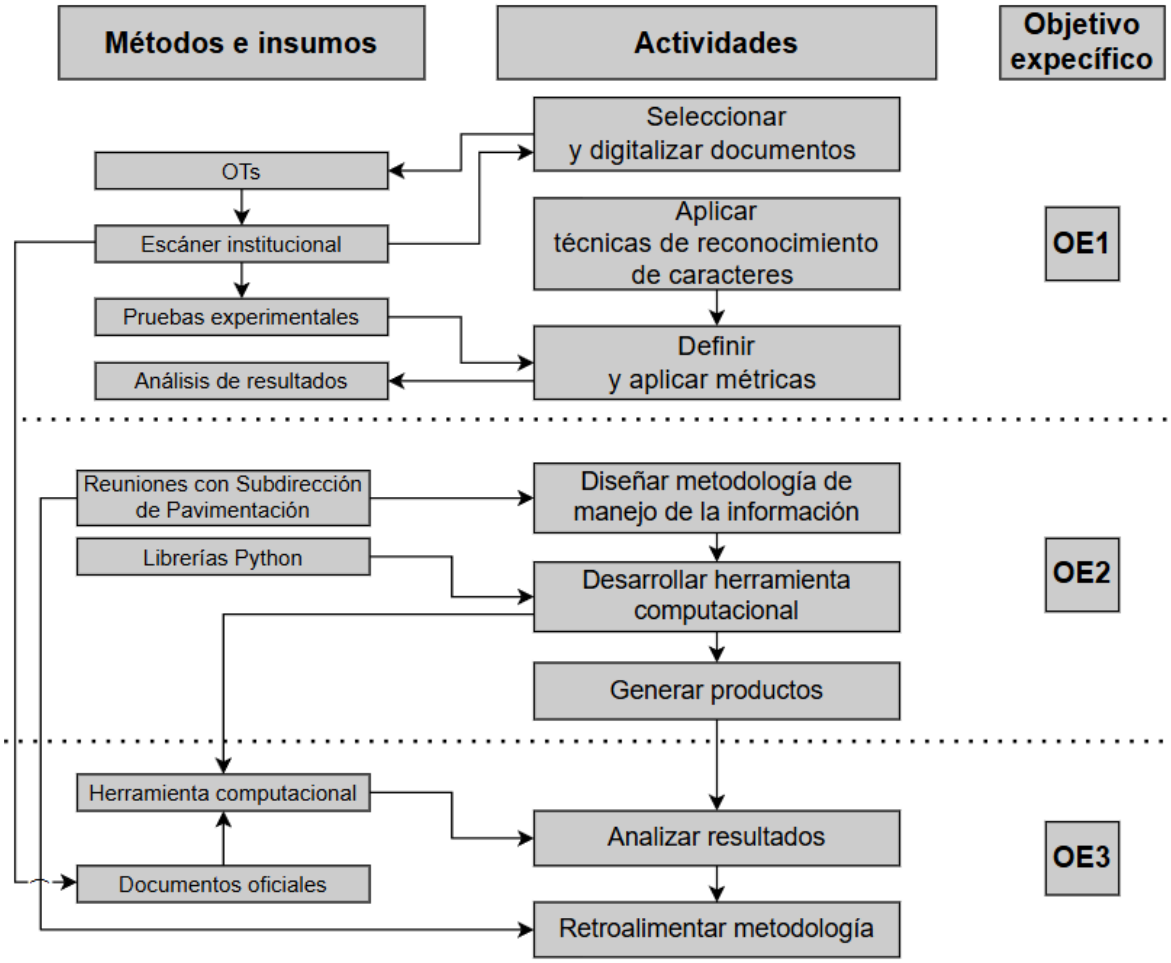


Figura 1. Esquema metodológico

4.1 Etapa 1 – Diagnóstico y Evaluación de Tecnologías de Extracción (OE1)

Actividades:

- Se seleccionaron y digitalizaron documentos disponibles y representativos de la Subdirección de Pavimentación.
- Se aplicaron técnicas de Reconocimiento Óptico de Caracteres (OCR) y de extracción de texto para evaluar la legibilidad y estructuración de la información.
- Se definieron y aplicaron métricas que reflejen el nivel de la extracción de información disponible en los documentos físicos.

Insumos: Órdenes de Trabajo digitalizadas. Equipamiento de digitalización

Métodos: Pruebas experimentales de extracción de texto. Análisis comparativo de resultados.

4.2 Etapa 2 – Diseño y Desarrollo del Sistema de Gestión de Datos (OE2)

Actividades:

- Se diseñó una metodología de manejo de la información a partir de insumos digitales y productos esperados de la Subdirección de Pavimentación.
- Se desarrolló una herramienta computacional de lectura, limpieza y extracción de campos de interés mediante expresiones regulares.
- Se desarrollaron funcionalidades de generación de productos de georreferenciación y reportes que agrupen la información.

Insumos: Requerimientos de información de la Subdirección de Pavimentación. Lenguaje de programación Python y librerías complementarias.

Métodos: Diseño de sistemas de información.

4.3 Etapa 3 – Validación Funcional y Análisis de Aplicabilidad (OE3)

Actividades:

- Se ejecutó la herramienta desarrollada con datos reales de la comuna.
- Se analizaron productos generados.
- Se realizó retroalimentación en base a las apreciaciones de la Subdirección de Pavimentación.

Insumos: Herramienta desarrollada (Etapa 2). Conjunto de datos reales acotados.

Métodos: Estudio de Caso.

Considerando el tiempo de desarrollo de la investigación, se acotó el volumen de información al Sector 1 de la comuna (zona norte delimitado por eje Alameda) para Órdenes de Trabajo generadas durante el contrato de conservación vigente que abarca entre el periodo de mayo de 2024 y agosto de 2025.

5. Herramientas para el OCR

Se desarrolla un script de Python para el análisis de los documentos digitalizados, siguiendo el lineamiento de realizar un análisis a bajo costo para emplear librerías gratuitas que sirvan para el procesamiento de imágenes y texto con la finalidad de seleccionar las más apropiadas.

Una alternativa serían herramientas más complejas como los VLM o los MLLM que han demostrado resultados favorables (Tanasa & Oprea, 2025), sin embargo, estas herramientas no siguen el lineamiento de ser de bajo costo a largo plazo ya que los servicios basados en la nube tienen un sistema de cobro, lo que podría implicar un problema en la asignación de recursos municipales. Otra alternativa sobre el uso de tales modelos es la adquisición de un equipo apto para ejecutarlos localmente, lo que requiere de una alta inversión inicial, lo que tampoco es viable para el caso en estudio de la Ilustre Municipalidad de Santiago.

5.1 pdflumber

Esta librería está diseñada para extraer información directamente de archivos PDF. Funciona analizando la estructura interna del PDF para identificar y extraer entidades de texto, así como para detectar tablas basándose en las líneas y rectángulos presentes en la página. Es de gran utilidad para trabajar con archivos PDF que contienen texto incrustado, sin embargo, pdfplumber no puede leer texto de archivos PDF que son simplemente imágenes escaneadas sin una capa de texto.

5.2 pdf2image

Esta librería no extrae texto, sino que actúa como un intermediario para convertir las páginas de un archivo PDF en objetos de imagen. Internamente esta librería utiliza la biblioteca externa Poppler que es un renderizador de PDF de código abierto, para generar una imagen ráster de cada página. Este paso es fundamental cuando se trabaja con documentos escaneados, ya que transforma el contenido visual del PDF en un formato que puede ser procesado por herramientas de Reconocimiento Óptico de Caracteres (OCR) como Tesseract. Su uso requiere que Poppler esté instalado en el sistema operativo.

Herramientas de código abierto como Tesseract OCR han permitido democratizar el acceso a esta tecnología, facilitando su aplicación en contextos de bajo costo y recursos limitados. No obstante, al utilizar OCR tradicional se requiere de un mayor cuidado en el proceso de digitalización y en la forma en la que se dispone el texto.

5.3 ocrmypdf

Esta librería se especializa en agregar una capa de texto OCR a archivos PDF que consisten principalmente en imágenes escaneadas. A diferencia de pdfplumber, ocrmypdf toma un PDF de entrada, utiliza un motor OCR (predeterminadamente Tesseract) para reconocer el texto en las imágenes de cada página, y luego genera un nuevo archivo PDF. Este nuevo PDF conserva la apariencia visual original, pero incluye una capa invisible de texto superpuesta con precisión sobre las palabras correspondientes en la imagen. El resultado principal es un PDF escaneado que ahora es detectable y cuyo texto puede ser seleccionado y copiado. Además, ocrmypdf permite realizar optimizaciones en las imágenes como corrección de inclinación y limpieza antes de aplicar el OCR, lo que puede mejorar la calidad del texto reconocido.

5.4 OpenCV

OpenCV (Open Source Computer Vision Library) es una biblioteca de código abierto, enfocada principalmente en tareas de visión por computadora, procesamiento de imágenes, video y aprendizaje automático. No está diseñada específicamente para extraer texto o manejar PDFs directamente, sino para manipular los datos visuales (píxeles) de las imágenes. OpenCV es útil para realizar pasos de preprocesamiento en las imágenes obtenidas de los PDFs después de usar pdf2image y antes de enviarlas a un motor OCR (Tesseract). Las funciones clave incluyen la conversión a escala de grises, la binarización, la eliminación de ruido y la corrección de inclinación.

6. Herramientas para la extracción y manejo de información

Para concretar los objetivos y realizar las pruebas necesarias, se optó por crear un programa en Python para la lectura, extracción y manejo de la información de los documentos digitalizados con las librerías de la Tabla 1 las cuales son de código abierto.

Tabla 1. Librerías que se utilizan

Librería	Categoría	Descripción	Limitantes	Motivo de elección
pdfplumber	Extracción de texto (PDF digital)	Permite leer texto estructurado desde documentos PDF, manteniendo saltos de línea y disposición tabular.	No reconoce texto en PDFs escaneados (imágenes).	Librería liviana, precisa y de uso sencillo para documentos digitales.
pdf2image	Conversión de PDF a imagen	Convierte las páginas de un PDF escaneado en imágenes, paso previo al uso de OCR.	Requiere instalación adicional de *Poppler*; alto consumo de memoria en PDFs grandes.	Es estable, ampliamente utilizada y compatible con *pytesseract*.
pytesseract	OCR (Reconocimiento Óptico de Caracteres)	Aplica OCR sobre imágenes para obtener texto digital legible por máquina.	Requiere instalación de *Tesseract OCR*; rendimiento dependiente de la calidad del escaneo.	Herramienta de código abierto y configurable en español.
pandas	Procesamiento de datos	Manejo y organización de los datos extraídos (OT, fechas, inspectores, coordenadas, etc.).	Puede ser lento con archivos muy grandes.	Es el estándar en ciencia de datos por su eficiencia y compatibilidad con Excel.
openpyxl	Exportación de resultados	Permite generar y editar archivos Excel desde Python.	Limitado a formatos XLSX. No permite funciones avanzadas de Excel.	Se integra nativamente con *pandas*.
geopy	Geocodificación	Transforma direcciones en coordenadas geográficas mediante el uso de servicios	Precisión dependiente del servicio	Permite automatizar la ubicación de direcciones sin software GIS.
geopandas	Análisis y visualización geográfica	Maneja estructuras espaciales (puntos, polígonos, líneas) y genera capas geográficas.	Alto uso de memoria; requiere dependencias adicionales (Shapely, Fiona).	Librería consolidada para el manejo de datos geoespaciales en entornos Python.
shapely	Geometría computacional	Define y manipula objetos geométricos (puntos, líneas, polígonos) necesarios para la creación del shapefile.	No gestiona archivos espaciales directamente.	Base de *geopandas*; permite definir geometrías con precisión.
simplekml	Exportación geográfica	Genera archivos KML y KMZ para visualizar las OT georreferenciadas.	No permite análisis espacial avanzado.	Ligera y práctica para exportar resultados de forma visual e intuitiva.

Una alternativa al uso la librería geopy es el uso de la API de Google Maps, sin embargo, esta alternativa requiere de una API Key, lo que incurre costo de contratación del servicio.

7. Metodología para la Subdirección de Pavimentación

Tomando en consideración el contexto en el cual se desarrollan las actividades de conservación de pavimentos urbanos de la comuna de Santiago, se plantea el proceso metodológico de la Figura 2, la cual requiere de insumos digitales para su implementación, dejando que la lectura, la extracción de información y la generación de productos se realice de manera automática.

Para el insumo principal contempla la identificación de archivos en formato PDF, donde se planea aprovechar las OT existentes del contrato de conservación vigente mediante un proceso de digitalización de estas y aplicando técnicas de OCR para que la información pueda ser aprovechada. Además se considera el caso en el cual el documento PDF entregado contenga información de texto digital.

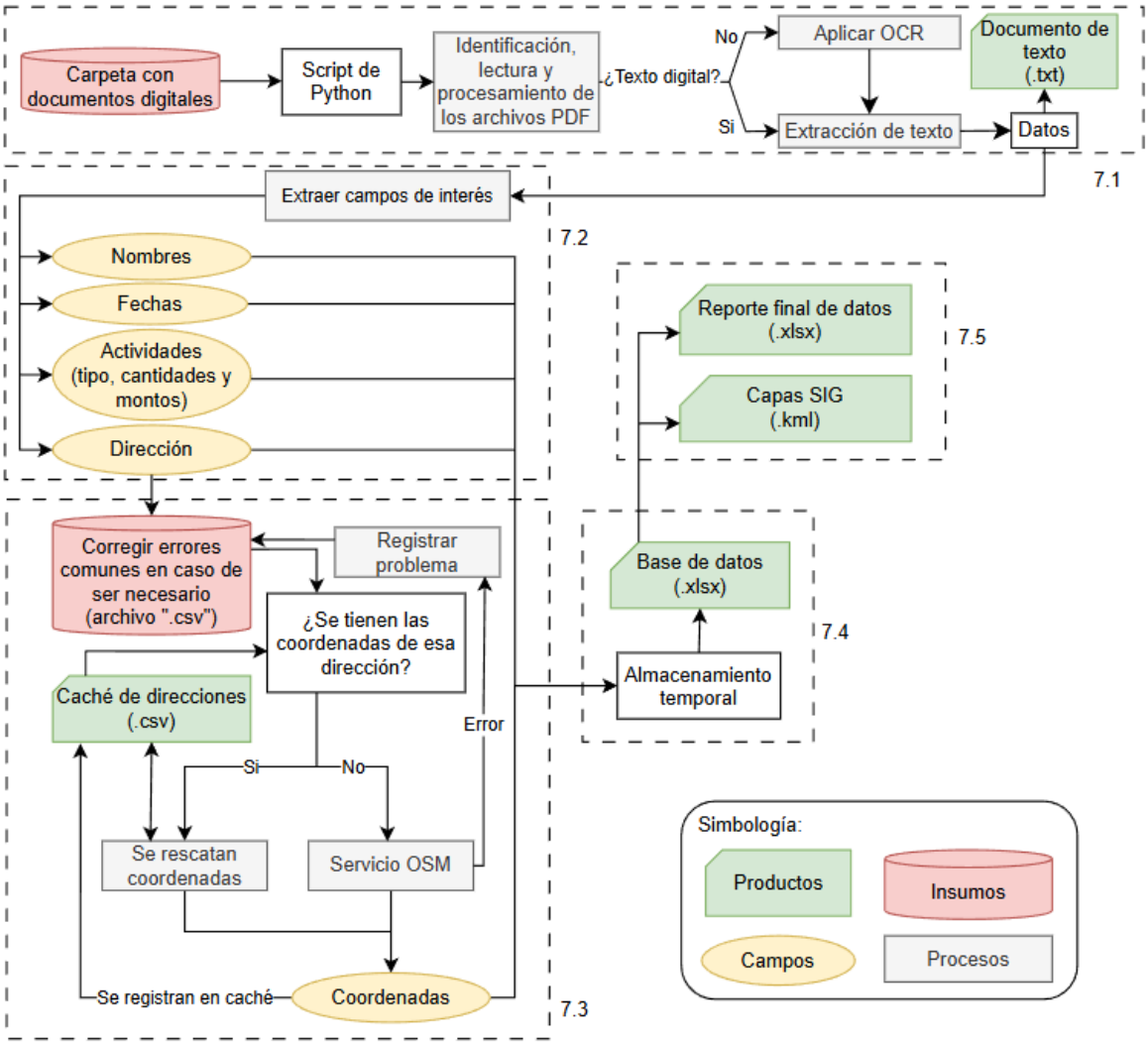


Figura 2. Metodología de procesamiento de datos

7.1 Identificación, lectura y procesamiento de documentos PDF

En el proceso de lectura se analiza la carpeta seleccionada que contenga los documentos en formato PDF para su análisis y se seleccionan para procesar únicamente los documentos PDF (.pfd), ya que es posible que en la práctica existan otro tipo de elementos como imágenes o subcarpetas.

Por cada documento PDF detectado se verifica si contiene texto digital o si es necesario realizar un OCR producto de ser una digitalización, debido a que el tratamiento y procesamiento del archivo depende de tal condición.

En el caso de que el documento contenga texto (producto de que fue generado como PDF) su información se extrae mediante la librería pdfplumber. Para el caso contrario donde no se detecta texto, se trataría de una digitalización donde es necesario aplicar técnicas de OCR. Para ambos procedimientos se genera un archivo de texto (.txt) con el mismo nombre que el archivo de origen en cual contiene todos los caracteres identificados del documento procesado.

7.2 Identificación y extracción de información

En general los campos de interés como los nombres, dirección, fecha, número de OT, entre otros, suelen estar desordenados en la información extraída en los documentos, por lo que se requieren un proceso de extracción enfocado en cada dato de interés. Para la extracción de los campos se implementan líneas de código que sean capaz de identificar expresiones regulares y almacenarlas temporalmente para su registro posterior.

Para el caso de los documentos digitalizados, los campos de interés podrían presentar variaciones producto del proceso del OCR (como confundir I.M.S por 1.M.S), por lo que estas expresiones regulares deben ser lo suficientemente flexibles para lograr identificar, corregir y extraer la información.

Por cada documento analizado se extraen los campos de interés y se almacenan en variables temporales para su registro posterior.

7.3 Georreferenciación de datos

La dirección es un campo crítico dado que debe relacionarse a un sistema de coordenadas, por lo que la dirección pasa por un proceso de validación y corrección en caso de ser necesario, ya que un error ortográfico podría afectar la georreferenciación de los datos. Lo anterior se realiza mediante un archivo .csv de corrección, el cual cambia la manera en que el campo de dirección se maneja, de tal manera que no deba modificarse el documento original.

Con el campo de direcciones detectado se extraen las coordenadas geográficas, creando campos adicionales de latitud y longitud mediante la librería geopy y el servicio de *OpenStreetMap* (OSM). Esto permite relacionar un documento con una ubicación al interior de la comuna.

Para efectos prácticos las direcciones con sus coordenadas son almacenadas en un documento ".csv" el cual ayuda a la optimización de recursos, ya que una dirección que ya fue consultada por el servicio de OSM no volverá a ser procesada, de tal forma que las coordenadas provengan del archivo y no del servidor.

Este proceso es sensible a la información de entrada por lo que es importante que el campo de dirección contenga una dirección real con su numeración respectiva, de lo contrario no se obtendrá una correcta georreferenciación de los datos. En caso de errores durante el proceso de georreferencia, la herramienta desarrollada entregará un archivo ".csv" con las direcciones problemáticas para su análisis manual y posible incorporación en el archivo de preprocesamiento de las direcciones.

7.4 Exportación a base de datos

Toda la información almacenada de forma temporal es registrada en un documento Excel el cual conforma la base de datos de los documentos analizados, donde la información contenida puede ser visualizada y ser utilizada para su libre uso.

7.5 Productos

Con la base de datos conformada, esta funciona como insumo para generar un producto de documento Excel de resumen el cual agrupa la información de las actividades realizadas junto a sus cubicaciones y frecuencia. Adicionalmente con la georreferenciación de los datos se generan dos archivos .kmz con la finalidad de que los datos puedan ser visualizados en Google Earth (plataforma gratuita), uno donde se muestran las intervenciones (direcciones) de manera puntual y otro con un buffer de 30 metros de cada punto de intervención a solicitud de la Subdirección de Pavimentación para una mejor visualización.

8. Digitalización

Debido a que las OT disponibles de la Subdirección de Pavimentación no cuentan con un respaldo digital, se hizo necesario iniciar un proceso de digitalización de estos documentos con la finalidad de contar con una muestra real de documentos a los cuales someter la metodología. Este proceso podría dar pie para la implementación de un Plan de Mejoramiento de Gestión (PMG) al interior de la Subdirección de Pavimentación.

Al momento de digitalizar documentos físicos en formato PDF generalmente son procesados como imágenes, sin interpretar el texto registrado en el documento de manera directa, de tal forma que el texto sólo puede ser leído por el usuario y no por la computadora, siendo necesario utilizar técnicas de OCR para que la información pueda ser manejada de forma automatizada.

8.1 Escáner utilizado

El proceso de digitalización hay múltiples parámetros que dependen del escáner que se utilice y la tecnología que este tenga, en las dependencias de la Subdirección de Pavimentación se cuenta con el escáner de un equipo "Canon imageRUNNER ADVANCE 500i" de la Figura 3 el cual se utiliza para la digitalización de los documentos.

Figura 3. Escáner presente en las dependencias de la Subdirección de Pavimentación

8.2 Parámetros de digitalización

Dentro de las opciones disponibles a la hora de digitalizar y mantener un estándar sobre todos los documentos, se selecciona una resolución de 300 PPP (pixeles por pulgada) y a color. La resolución empleada permite que el texto pueda ser visualizado por los algoritmos de OCR manteniendo un balance con el peso final del archivo. Que la digitalización sea a color permite que la información visual respecto a las anotaciones pueda ser visualizada por el usuario.

Al seleccionar los parámetros e introducir los documentos, el equipo de manera automática digitaliza todas las hojas entregadas, incorporándolas todas en un mismo documento PDF y enviándolo a un correo electrónico registrado previamente. Con dicha configuración durante el proceso de digitalización se observa un tiempo promedio de 4 segundos por hoja.

8.3 Gestión y almacenamiento

Al momento de digitalizar documentos en formato PDF, el nombre del archivo no es correlativo a su contenido, sino que es asignado automáticamente por el escáner, por lo que se transforma en una buena práctica establecer una nomenclatura estándar para el almacenamiento de estos documentos.

Una vez finalizado el proceso de digitalización se reasigna el nombre del archivo de la forma "SX_XXXX_EDPX_OTX.pdf", siendo los caracteres de "X" cambiados según el contenido del documento (sector-año-#EDP-#OT). De esta manera el nombre de cada archivo es único y tiene una relación directa al contenido de este, evitando tener que abrir el documento para conocer su contenido.

9. Métricas para el análisis

Frente a la variedad de métodos disponibles para realizar OCR a los documentos digitalizados, se hace necesario el uso de métricas con la finalidad comparar y determinar la mejor alternativa para el caso en estudio. Para la comparación se utilizan las métricas descritas a continuación.

9.1 Porcentaje de Detección (PD)

Durante el proceso de identificación y extracción de la información utiliza el porcentaje de detección de datos no nulos, esto es si detecta o no la información para extraer, no implica que la información sea correcta.

$$PD = \frac{campos\ detectados}{campos\ disponibles} \tag{1}$$

Considerando por ejemplo que los campos que se esperan extraer por documentos sean el número de OT, la fecha, el inspector fiscalizador, el inspector técnico y la dirección, se espera que cada documento detecte 5 campos. Si se detectan 5 el PD sería del 100%, sin ser información necesariamente correcta.

9.2 Porcentaje de Extracción Correcta (PEC)

Para los campos que lograron ser detectados y almacenados, se calcula el Porcentaje de Extracción Correcta, el cual se obtiene al relacionar los campos que fueron reconocidos correctamente respecto al total de los campos extraídos.

$$PEC = \frac{campos\ detectados\ correctamente}{campos\ detectados} \tag{2}$$

A modo de ejemplo, si el campo esperado es “Inspector Uno” y el campo extraído fue “Inspector Un0”, este último no se considera como correcto, dicho campo es clasificado como aproximado.

9.3 Porcentaje de Extracción Aproximado (PEA)

Para los campos que lograron ser detectados y almacenados, se calcula el Porcentaje de Extracción Aproximado, el cual se obtiene al relacionar los campos que fueron reconocidos correctamente y de forma aproximada respecto al total de los campos extraídos.

$$PEA = \frac{campos\ detectados\ correctamente + campos\ detectados\ aproximados}{campos\ detectados} \tag{3}$$

A modo de ejemplo, si el campo esperado es “Inspector Uno” y el campo extraído fue “Inspector Un0”, este último se considera como aproximado para el cálculo del PEA.

9.4 Tiempo por Documento (TPD)

Debido a los diferentes métodos empleados para el proceso de OCR, se considera el promedio de tiempo de ejecución en función de los documentos analizados.

$$TPD = \frac{tiempo\ total\ del\ método}{documentos\ analizados} \tag{4}$$

10. Resultados con documentos reales

Durante las reuniones iniciales con la Subdirección de Pavimentación se digitalizaron 16 Estados de Pago correspondientes al contrato de conservación vigente. El contrato de conservación cubre entre mayo de 2024 y mayo de 2026, mientras que algunos de los documentos digitalizados abarcan desde mayo de 2024 hasta agosto de 2025, resultando un total de 247 Órdenes de Trabajo (OT) para su análisis.

Al tratarse de documentos reales los cuales fueron sometidos a un proceso de digitalización, estos utilizan la rama del análisis con OCR del flujo de la Figura 2, por lo que es el caso más desfavorable del análisis dado a las limitaciones que puede presentar esta tecnología. Para el caso de los documentos con texto integrado, estos son procesados por la otra rama del flujo de procesos. siendo procesado mediante la librería pdfplumber.

Se ejecutaron 7 métodos diferentes para la aplicación del OCR de las primeras 75 OT digitalizadas, tanto con sus páginas con su orientación original y con sus páginas correctamente orientadas con la finalidad de comparar los tiempos de ejecución y diferencias en la extracción de información. Los 7 métodos principalmente se diferencian en las librerías Python utilizadas y/o variaciones en sus ajustes.

El primer método consiste en la transformación del documento PDF en una imagen mediante la librería pdf2image y posteriormente se aplica pytesseract para el reconocimiento de texto. Del segundo al cuarto método se emplea la librería opencv entre el proceso de pdf2image y tesseract del primer método, donde cada método (OpenCV1, OpenCV2, OpenCV3 y OpenCV4) realiza un procesamiento diferente a las imágenes. Para el sexto método se emplea la librería ocrmypdf, donde se realiza una corrección de la inclinación de la página de ser necesario e incorporando una capa invisible de caracteres sobre el documento, permitiendo mejores búsquedas dentro del mismo. El último método es una mezcla entre ocrmypdf y opencv, siendo esta última una herramienta de respaldo en caso de que la primera no funcione correctamente.

10.1 Reconocimiento de texto y tiempo de procesamiento

Mediante un análisis previo al momento de configurar el programa junto a las librerías, se observó que el OCR empleado realiza una lectura de izquierda a derecha y de arriba hacia abajo, simulando una lectura humana tradicional, esto genera que el proceso de reconocimiento sea altamente sensible a la orientación de las hojas de los documentos. Por lo que se compararon los documentos digitalizados con su orientación por defecto (desde el escáner) con los documentos con la orientación corregida, de tal forma que en estos últimos todas las hojas sea leídas de forma correcta.

Los resultados referentes al tiempo de procesamiento de los 7 métodos, considerando las 75 OT y los 2 casos posibles (orientación original y corregida) causó que en la práctica se debieron analizar 1050 documentos para aplicar OCR. Este análisis presentó una duración de dos horas y media en total y sus resultados se muestran en las tablas 2 y 3.

Tabla 2. Tiempo de análisis con los documentos originales

Orientación Original		Tiempo[s] vs Métodos								
EDP	Cant. OT	pytesseract	OpenCV1	OpenCV2	OpenCV3	OpenCV4	ocrmypdf	ocrmypdf + OpenCV1	Total	Total [min]
2	11	90,2	62,7	73,2	66,3	73,0	120,8	120,7	606,9	10,1
3	11	98,9	67,1	75,4	69,2	75,5	117,6	117,5	621,2	10,4
5	10	125,1	93,0	100,6	88,7	102,0	179,5	178,8	867,7	14,5
10	17	106,9	75,1	82,5	68,2	79,6	236,2	234,6	883,1	14,7
11	18	108,5	75,8	87,4	72,2	83,5	252,5	251,1	930,9	15,5
12	8	51,2	36,7	41,3	34,0	39,0	116,1	114,1	432,3	7,2
	Total [s]	580,8	410,4	460,4	398,6	452,6	1022,5	1016,8	4342,1	72,4
	Total [min]	9,7	6,8	7,7	6,6	7,5	17,0	16,9	72,4	

Tabla 3. Tiempo de análisis con la documentación orientada

Orientación Corregida		Tiempo[s] vs Métodos								
EDP	Cant. OT	pytesseract	OpenCV1	OpenCV2	OpenCV3	OpenCV4	ocrmypdf	ocrmypdf + OpenCV1	Total	Total [min]
2	11	96,4	94,9	100,8	80,9	87,6	119,4	118,3	698,2	11,6
3	11	98,6	77,9	81,0	63,9	74,7	117,4	117,7	631,2	10,5
5	10	133,6	93,7	100,6	96,6	104,6	179,5	180,7	889,3	14,8
10	17	122,1	85,6	87,2	87,1	90,3	237,7	236,8	946,8	15,8
11	18	117,5	90,4	91,2	89,7	95,6	254,4	253,2	992,0	16,5
12	8	57,4	40,9	42,5	43,3	42,3	114,1	113,9	454,4	7,6
	Total [s]	625,6	483,4	503,3	461,5	495,1	1022,4	1020,7	4611,9	76,9
	Total [min]	10,4	8,1	8,4	7,7	8,3	17,0	17,0	76,9	

En base a la Tabla 1Tabla 2 y la Tabla 3, la librería menos eficiente corresponde a ocrmypdf que requirió de aproximadamente el doble de tiempo que las otras librerías. Adicionalmente se aprecia que corregir la orientación del documento perjudica el tiempo total de análisis en un 6,2%.

10.2 Extracción de información

Al aplicar la extracción de información mediante el reconocimiento de expresiones regulares para la obtención de los campos de interés, tras aplicar las métricas a los 7 métodos para el OCR y para sus dos orientaciones (original y corregida) se obtuvieron las métricas de las tablas 4 y 5

Tabla 4. Métricas con los documentos originales

	Métricas vs Métodos						
Orientación Original	pytesseract	OpenCV1	OpenCV2	OpenCV3	OpenCV4	ocrmypdf	ocrmypdf+OpenCV1
PD	19,5%	18,1%	15,2%	14,1%	13,6%	16,3%	16,3%
PEC	56,2%	63,2%	54,4%	54,7%	49,0%	57,4%	55,7%
PEA	91,8%	95,6%	94,7%	90,6%	94,1%	98,4%	98,4%
Aciertos	41	43	31	29	25	35	34
Tiempo por documento [s]	7,7	5,5	6,1	5,3	6,0	13,6	13,6

Tabla 5. Métricas con la documentación orientada

	Métricas vs Métodos						
Orientación Corregida	pytesseract	OpenCV1	OpenCV2	OpenCV3	OpenCV4	ocrmypdf	ocrmypdf+OpenCV1
PD	46,7%	52,3%	50,4%	54,1%	53,9%	52,5%	54,7%
PEC	48,6%	42,9%	30,7%	38,4%	32,2%	35,5%	41,0%
PEA	92,0%	94,9%	92,6%	92,6%	90,1%	91,4%	87,8%
Aciertos	85	84	58	78	65	70	84
Tiempo por documento [s]	8,3	6,4	6,7	6,2	6,6	13,6	13,6

En base a la Tabla 4 y la Tabla 5, la eficacia de la librería ocrmypdf (y su variante con OpenCV1) en términos de aciertos es comparable a los métodos que utilizan pytesseract y OpenCV1 por separado, las cuales obtienen resultados iguales o superiores en un menor tiempo. Además, se observó que la orientación de los documentos aumentó los aciertos en aproximadamente un 50% por lo que el aumento del tiempo por documento se relaciona al aumento en la detección de texto.

10.3 Análisis de resultados del OCR

Según el método utilizado se obtuvieron métricas similares en tiempo, pero diferentes en lo que respecta la detección de los campos de interés.

A partir de la Tabla 2 y Tabla 3 se desprende que, en términos de tiempo el programa desarrollado para la implementación de OCR no se ve mayormente afectado por la orientación de las páginas del documento entregado como insumo. En cambio, a partir de la Tabla 4 y la Tabla 5 se tiene que el porcentaje de detección (PD) presenta una variación positiva al contar con una orientación correcta de las páginas de los documentos analizados, lo que indica que se logra un mejor reconocimiento al utilizar documentos cuyas páginas estén correctamente orientadas, lo que sugiere que es necesaria una revisión y corrección del documento previo al análisis.

En ciertos casos se tiene que el Porcentaje de Extracción Aproximada (PEA) para el caso de las páginas orientadas originalmente es mayor a las que cuentan con una orientación corregida, no obstante, en los documentos orientados correctamente detectan una mayor cantidad de texto, lo que a su vez permite que los aciertos (campos detectados correctamente) sean mayores.

Las métricas podrían variar al implementar nuevas expresiones regulares que sean más flexibles en la detección, pudiendo mejorar o empeorar. La dificultad de implementar expresiones regulares generales radica en la falta de uniformidad en los documentos reales que fueron digitalizados y utilizados como muestra, por ejemplo, en algunos existe información impresa y en otros de forma manuscrita, se presentan diferencias en los tamaños y en la calidad de los documentos al ser fotocopias o documentos originales.

De los 7 métodos empleados para el uso del OCR, los que más se destacaron en términos de la identificación de texto y en tiempo son los métodos que emplean las librerías pytesseract y OpenCV1 (variante que únicamente aplica escala de grises) respectivamente. Los documentos digitalizados si bien permiten un respaldo de la información física, las métricas indican que no es recomendable utilizar técnicas de OCR para rescatar la información de manera confiable.

10.4 Alternativas al OCR

Dados los resultados se tiene que no es recomendable utilizar digitalizaciones de documentos físicos para la extracción y manejo de la información por las limitantes de la tecnología OCR empleada, por lo que se abre una nueva línea investigativa de modificar el insumo que emplea la metodología.

Frente a la problemática del uso de los documentos originales se propone implementar un nuevo tipo de documento con formato de estructurado que contengan la información de interés, pero de manera digital, de tal manera que estos documentos sirvan para preservar la información y se permita su extracción de información de manera automatizada con la metodología desarrollada.

El nuevo formato de este documento fue diseñado en base a la información de los documentos originales que sea de interés rescatar para la Subdirección de Pavimentación. Con el diseño definido se aplicaron a todas las Órdenes de Trabajo del periodo disponible, con la finalidad de ser entregadas como insumo en el flujo de la metodología de la Figura 2 y analizar su comportamiento.

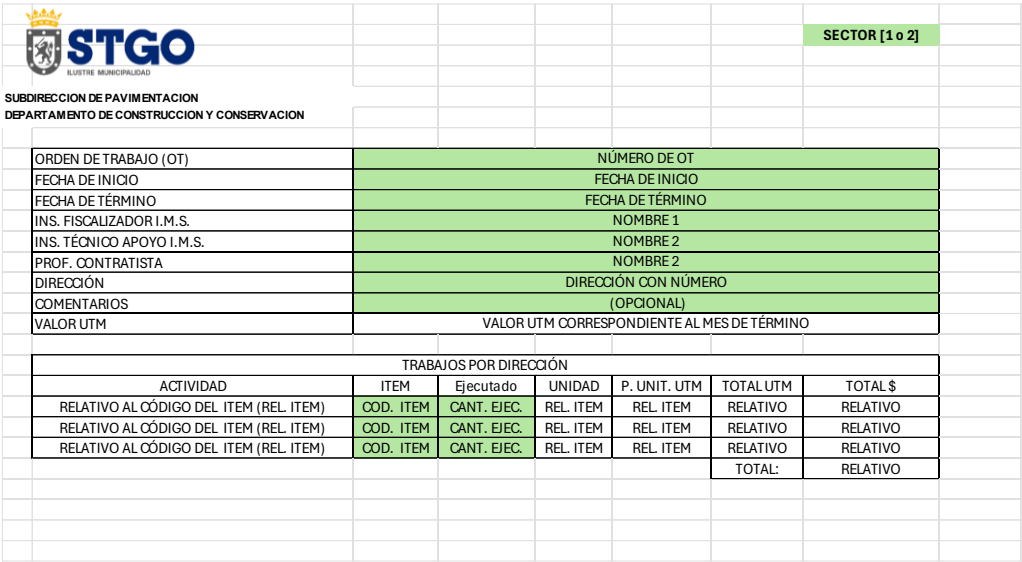
11. Nueva Ficha Digital

Dado que los resultados de la extracción directa en los documentos digitalizados mediante OCR indicaron una baja detección, se replantó la forma en la que se almacena la información de las OT y de cómo esta se entrega a la metodología que se busca implementar.

Con la finalidad de simplificar el proceso de registro y respaldo de la información contenida en las OT, se desarrolló una planilla Excel con macros la cual ayuda en los procesos de cálculo y visualización de la información, sólo siendo requerido que se indiquen los datos clave que sean de interés. Mediante las macros desarrolladas la planilla ajusta la información ingresada y se almacena en formato PDF.

11.1 Diseño

El diseño de la nueva Ficha Digital se muestra en la Figura 4 que se basa en registrar la información clave para la Subdirección de Pavimentación (celdas verdes) como el sector 1 o 2 (norte o sur de la comuna respectivamente), el número de OT, las fechas de inicio y término, los nombres de los fiscalizadores, del contratista responsable, una dirección de referencia y el código de las actividades realizadas junto a sus cantidades ejecutadas.



STGO ILUSTRE MUNICIPALIDAD

SUBDIRECCION DE PAVIMENTACION
DEPARTAMENTO DE CONSTRUCCION Y CONSERVACION

SECTOR [1 o 2]

ORDEN DE TRABAJO (OT)	NÚMERO DE OT
FECHA DE INICIO	FECHA DE INICIO
FECHA DE TÉRMINO	FECHA DE TÉRMINO
INS. FISCALIZADOR I.M.S.	NOMBRE 1
INS. TÉCNICO APOYO I.M.S.	NOMBRE 2
PROF. CONTRATISTA	NOMBRE 2
DIRECCIÓN	DIRECCIÓN CON NÚMERO
COMENTARIOS	(OPCIONAL)
VALOR UTM	VALOR UTM CORRESPONDIENTE AL MES DE TÉRMINO

TRABAJOS POR DIRECCIÓN						
ACTIVIDAD	ITEM	Ejecutado	UNIDAD	P. UNIT. UTM	TOTAL UTM	TOTAL $
RELATIVO AL CÓDIGO DEL ITEM (REL. ITEM)	COD. ITEM	CANT. EJEC.	REL. ITEM	REL. ITEM	RELATIVO	RELATIVO
RELATIVO AL CÓDIGO DEL ITEM (REL. ITEM)	COD. ITEM	CANT. EJEC.	REL. ITEM	REL. ITEM	RELATIVO	RELATIVO
RELATIVO AL CÓDIGO DEL ITEM (REL. ITEM)	COD. ITEM	CANT. EJEC.	REL. ITEM	REL. ITEM	RELATIVO	RELATIVO
					TOTAL:	RELATIVO

Figura 4. Diseño de la nueva Ficha en Excel

La ventaja de esta manera de almacenar la información se basa en que el documento PDF contenga la información esencial y esta pueda ser extraída de forma directa sin un análisis OCR mediante la librería pdfplumber, además de ofrecer una manera amigable de almacenar y preservar la información en forma digital sin ser necesaria una impresión física, lo que ayuda a que no tenga un costo físico asociado.

Una consideración importante es que cada OT física puede albergar múltiples direcciones correspondientes a diversas actividades de conservación desarrolladas en varios puntos de la comuna. Para tales casos se desarrollan múltiples fichas que, si bien están asociadas una misma OT, se diferencian en la dirección especificada y sus actividades, de tal manera que simplifica el proceso de georreferencia.

11.2 Implementación

El nuevo formato de ficha se aplicó a 247 OT pertenecientes a los estados de pago cursados y aprobados del sector 1 (zona norte de la comuna) del contrato de conservación vigente que corresponden al periodo entre mayo de 2024 y agosto de 2025, resultando en un total de 569 fichas (direcciones).

12. Resultados con Fichas Digitales

12.1 Eficiencia computacional

Al procesar las 569 Fichas por primera vez el programa desarrollado tardó 38.33 segundos en ejecutar la lectura y extracción de la información, 1049.58 segundos en la georreferenciación de datos y 0.39 segundos en la generación del reporte final agrupando toda la información procesada, resultando en un total de 1088.30 segundos (18.14 minutos).

Al ejecutar por segunda vez el programa presentó una mejora en el tiempo de ejecución debido a que las coordenadas de las direcciones fueron almacenadas en la primera vez, evitando la consulta en los servidores de OSM, resultando en un tiempo total de 36.76 segundos en la lectura, extracción, georreferenciación y generación del reporte final.

12.2 Georreferenciación

Utilizar el servicio de *OpenStreetMap* es una solución práctica para la georreferenciación de las obras de conservación de manera aproximada al utilizar direcciones de domicilios que se enfrenten a las obras de conservación ya que no es posible contar con las coordenadas exactas.

Las ubicaciones mostradas en la Figura 5 representan la ubicación puntual indicada en cada Ficha Digital analizada, por lo que suelen estar sobre propiedades o en las aceras. Cada punto de la Figura 5 contiene información referente a los trabajos realizados, de tal manera que al visualizarlo en Google Earth es posible visualizar dicha información como se muestra en la Figura 6, por lo que no solo se tiene la ubicación de las actividades, sino que de una manera rápida y sencilla se puede saber la información asociada a cada punto espacial.

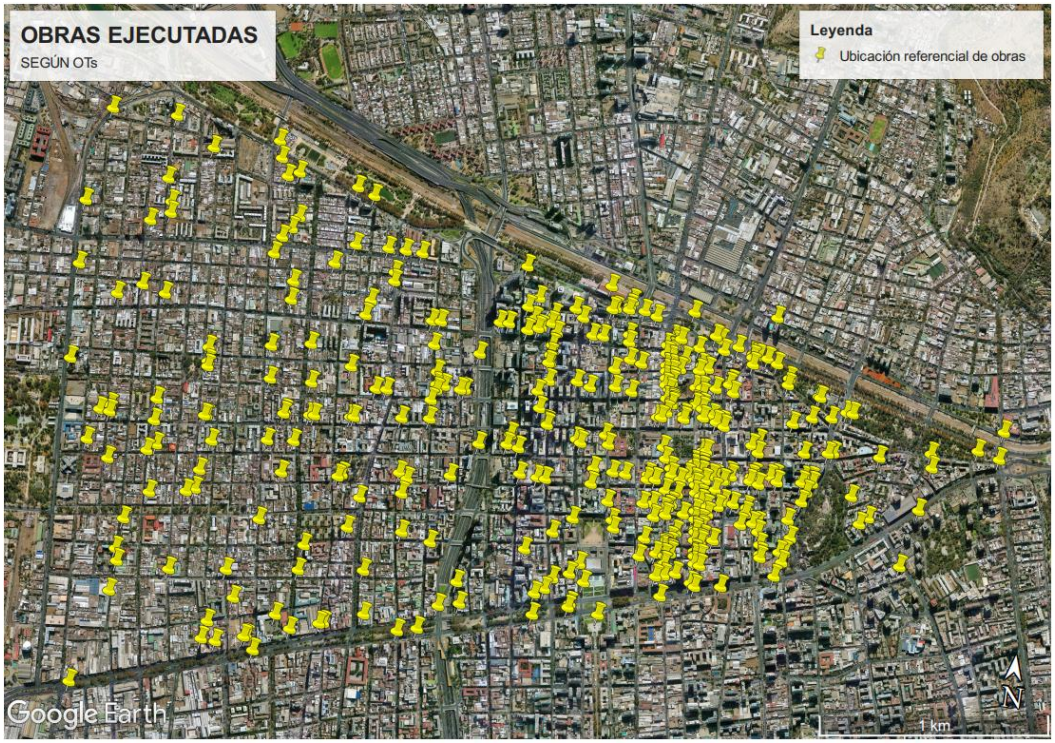


Figura 5. Ubicaciones puntuales referenciales de las obras ejecutadas

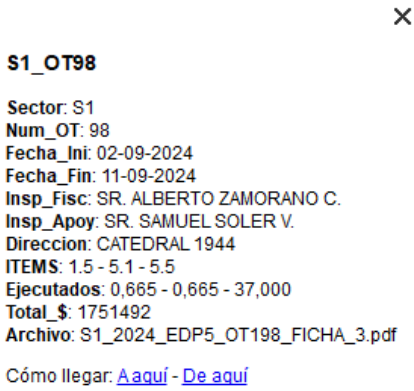


Figura 6. Ejemplo de recuadro descriptivo de actividades de cada punto georreferenciado

Debido a que los puntos corresponden a aproximaciones de la ubicación de las obras de conservación, una manera alternativa para visualizar las áreas aproximadas a las cuales se les hizo una intervención es que se genera un radio de 30 metros a partir de cada punto generado como muestra la Figura 7.

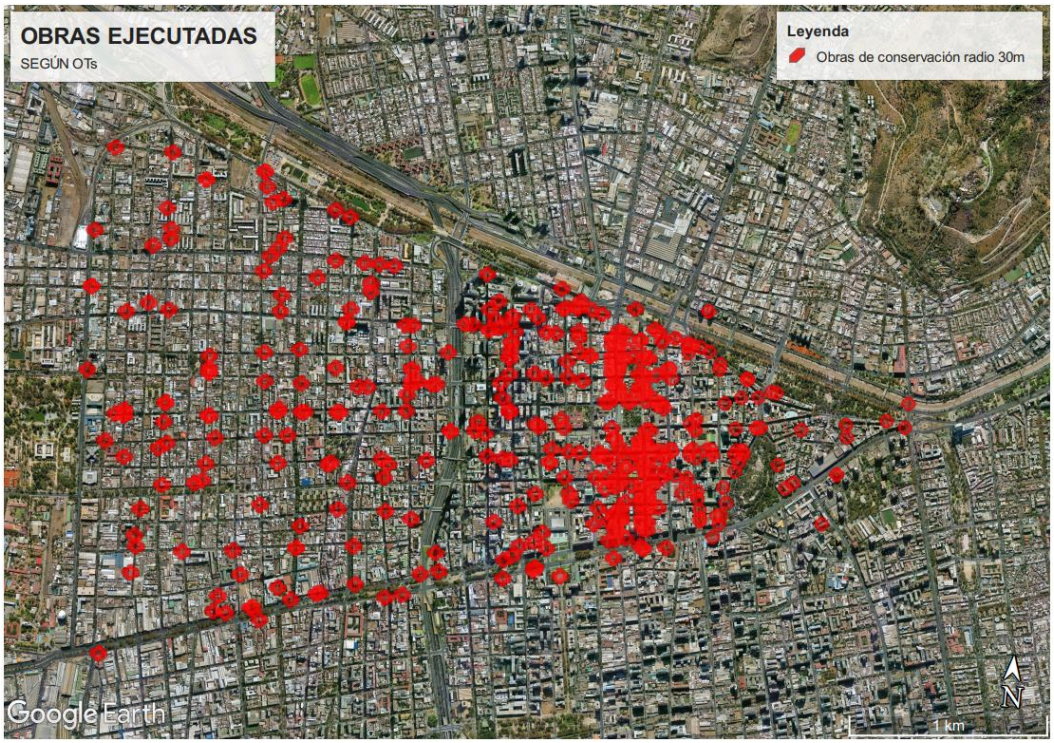


Figura 7. Superficie de zonas intervenidas de las obras ejecutadas

El radio de 30 metros es un criterio adoptado por la Subdirección de Pavimentación que permite visualizar la distribución espacial y ser considerado como radio razonable relacionado a la capacidad de inspección de los fiscalizadores que acudir a terreno para recibir las obras realizadas y visualizar los pavimentos circundantes. Un radio mayor podría ser poco realista y un radio menor poco representativo de la capacidad de visualización de los inspectores.

12.3 Reportes

Con las fichas generadas y entregadas como insumo en la metodología, el reporte global contiene el recuento de las actividades solicitadas junto a su frecuencia y total ejecutado durante el periodo mayo de 2024 y agosto de 2025. El reporte global al ser entregado en un formato de Excel, pueden ser generadas de manera práctica diversas tablas de resumen según las necesidades, como una tabla de actividades frecuentes indicadas en la Tabla 6, o que contemplan una mayor cantidad ejecutada como se indica en la Tabla 7. En caso de ser necesario el usuario podría incorporar filtros adicionales como el de partidas específicas (actividades) o por tipo de unidades.

Tabla 6. Actividades con mayor frecuencia

ITEM	DESCRIPCION	UNIDAD	Suma_Ejecutado	Conteo
1.3	Ruptura de aceras a mano, incluye retiro de escombros.	m²	1469,417	164
2.1	Excavación a mano, incluye extracción, retiro de escombros y preparación de la subrasante.	m³	110,941	143
10.30	Reinstalación de rejilla existente.	ml	274,75	126
5.1	Carpeta asfáltica de espesor variable, incluye imprimación y/o riego de adherencia y…	m³	486,111	123
7.4	Suministro y colocación baldosa especial micro vibrada de 40*40 cm de 35 a 40mm de espesor, sobre mortero de 0.04 m de espesor y base estabilizada de 0.05 m de espesor. (Cualquier tipo, cualquier color).	m²	556,605	85

Tabla 7. Actividades con mayor cantidad

ITEM	DESCRIPCION	UNIDAD	Suma_Ejecutado	Conteo
5.5	Corte de carpeta asfáltica a máquina.	ml	2278,71	68
1.3	Ruptura de aceras a mano, incluye retiro de escombros.	m²	1469,417	164
5.6	Suministro y colocación Geogrilla.	m²	1045,328	3
7.4	Suministro y colocación baldosa especial microvibrada de 40*40 cm de 35 a 40mm de espesor, sobre mortero de 0.04 m de espesor y base estabilizada de 0.05 m de espesor. (Cualquier tipo, cualquier color).	m²	556,605	85
5.1	Carpeta asfáltica de espesor variable, incluye imprimación y/o riego de adherencia y	m³	486,111	123

Adicionalmente se pueden aplicar más filtros como podría ser la unidad de trabajo o los ítems que se desean contrastar ya que pertenecen a diversos grupos. Esta clase de información es de utilidad para la Subdirección de Pavimentación ya que las actividades que son más solicitadas en frecuencia y volumen podrían tener una relación directa en el presupuesto, por lo que una estrategia técnico-económica a adoptar podría ser la selección de licitaciones de obras de conservación que ofrezcan un precio unitario más bajo en las actividades más solicitadas o importantes en términos de cantidad para la comuna.

12.4 Almacenamiento

Dado que se digitalizaron las 247 OT a modo de respaldo, estas requirieron un almacenamiento de 239 MB, mientras que las 569 Fichas Digitales creadas requirieron un almacenamiento de 88,1 MB lo que indica una reducción del 63,1% de almacenamiento respecto a las digitalizaciones. Lo anterior supone una ventaja para la escalabilidad del uso de la metodología en comparación al uso directo de las digitalizaciones que no son apropiadas para extraer información según lo visto en la sección 10.4.

12.5 Validación

Con el consolidado de las Fichas Digitales y la información contenida en los Estados de Pago se procedió a la validación de los montos oficiales y el acumulado por el programa de extracción como muestra en la Tabla 8.

Tabla 8. Validación de extracción

EDP	AÑO	MES	MONTO EDP	#OT	FICHAS	MONTO FICHAS	ERROR
1	2024	MAYO	$ 23.454.774	17	27	$ 23.454.773	0,000%
2	2024	JUNIO	$ 81.184.776	8	33	$ 81.184.644	0,000%
3	2024	JULIO	$ 192.271.706	30	52	$ 192.271.967	0,000%
4	2024	AGOSTO	$ 112.119.521	19	66	$ 112.119.457	0,000%
5	2024	SEPTIEMBRE	$ 136.224.795	10	34	$ 136.225.192	0,000%
6	2024	OCTUBRE	$ 134.824.967	21	60	$ 134.824.905	0,000%
7	2024	NOVIEMBRE	$ 118.175.483	26	71	$ 118.175.417	0,000%
8	2024	DICIEMBRE	$ 112.844.365	21	42	$ 112.844.368	0,000%
9	2025	ENERO	$ 35.474.873	15	22	$ 35.474.875	0,000%
10	2025	FEBRERO	$ 6.872.872	17	29	$ 6.872.871	0,000%
11	2025	MARZO	$ 32.023.533	20	32	$ 32.023.533	0,000%
12	2025	ABRIL	$ 8.515.572	8	14	$ 8.515.572	0,000%
1	2025	MAYO	$ 6.006.906	1	3	$ 6.007.043	0,002%
2	2025	JUNIO	$ 82.950.510	11	26	$ 82.950.511	0,000%
3	2025	JULIO	$ 56.128.686	10	25	$ 56.128.892	0,000%
4	2025	AGOSTO	$ 93.316.129	13	33	$ 93.315.920	0,000%
			$1.232.389.468	247	569	$ 1.232.389.940	0,000%

En la Tabla 8, los montos reconocidos en los Estados de Pago se registraron en la columna “MONTO EDP” entregando una cifra oficial con la cual comparar con la columna “MONTO FICHAS” la cual suma todos los montos de las fichas generadas a partir de las OT. Las diferencias en los montos totales son generadas principalmente por el criterio de aproximación a 3 decimales empleado por la Subdirección de Pavimentación en las Órdenes de Trabajo y al desglose de estas al realizar las diferentes fichas. Pese a que exista una diferencia esta no es significativa para un análisis global.

13. Conclusiones, consideraciones y recomendaciones

La presente investigación culminó con el desarrollo y validación de un sistema integral para la gestión de pavimentos, dando cumplimiento de los objetivos producto de la colaboración con la Subdirección de Pavimentación de la Ilustre Municipalidad de Santiago. Esta vinculación entre la academia y el sector público permite no solo acceder a datos reales, sino también alinear la solución tecnológica con las necesidades operativas del municipio, asegurando que el trabajo realizado posea una utilidad práctica.

13.1 Conclusiones

Respecto al OE1 se obtuvo que la extracción de información mediante el OCR tradicional sobre documentos históricos digitalizados no es viable de manera técnica ni operativamente para la Subdirección de Pavimentación. Las pruebas experimentales demostraron que la variabilidad del formato físico, sumada a las interferencias visuales (sellos, firmas o manchas), reduce drásticamente la tasa de detección confiable. Este hallazgo fue crítico para modificar el insumo de entrada que se utiliza.

En relación con el OE2 se diseñó e implementó exitosamente una metodología basada en la transformación digital del origen de datos. La creación de la "Ficha Digital" permitió eliminar la incertidumbre del reconocimiento de caracteres, garantizando una integridad de datos. La herramienta computacional desarrollada en Python demostró ser robusta, efectiva y eficiente al realizar el procesamiento automatizado de 569 fichas en menos de 1 minuto en su segunda ejecución (por optimización por caché). Lo anterior representa una alta competitividad frente a los métodos manuales o el procesamiento de OCR, manteniendo el enfoque de bajo costo.

Finalmente, para el OE3 mediante el caso de estudio real (Zona Norte, 2024-2025) se confirmó la utilidad práctica del sistema. La generación automática de mapas de cobertura, junto con los reportes de costos y frecuencias de actividades proporciona a la Subdirección de Pavimentación insumos inmediatos que pueden aportar a la gestión de pavimentos urbanos. El sistema desarrollado permite identificar visualmente focos de inversión y optimizar futuras licitaciones basándose criterios técnico-económicos considerando los ítems de mayor incidencia o las zonas de mayor demanda.

La investigación demostró que la modernización de la gestión municipal no depende necesariamente de la adquisición de software comercial costoso, sino de una ingeniería de procesos estratégica enfocada en la calidad del dato desde su origen. La transición desde el documento físico hacia el documento digital constituye una solución escalable, a bajo costo y sostenible. Este cambio de paradigma trasciende a la Subdirección de Pavimentación, ofreciendo un modelo de gestión que cualquier otro organismo municipal pueda adoptar para iniciar su propia transformación digital.

13.2 Consideraciones

En cuanto a los aspectos técnicos de implementación, es importante abordar el manejo de la ambigüedad espacial y la representación geométrica. Frente a la problemática de que diferentes comunas compartan nombres de calles y numeración, el algoritmo desarrollado incorpora una restricción de búsqueda que limita las consultas estrictamente a la superficie de la comuna de Santiago, por lo que no es necesario tomar acciones manuales para filtrar direcciones externas duplicadas.

Adicionalmente se debe considerar que una Orden de Trabajo puede contener labores que se distribuyen a lo largo de varias cuadras y que representar dicha intervención mediante una dirección única implica establecer una ubicación referencial de carácter administrativa y no representativa en términos de magnitud. Por lo que la ubicación de punto fijo en el mapa es representativa de la zona intervenida, mas no de geometría de las obras.

Durante las pruebas se observó un tiempo promedio de 2 minutos para la generación manual de cada Ficha Digital (actividades por dirección). En caso de que la Subdirección de Pavimentación decida implementar este sistema para las OT de periodos anteriores, se deberá considerar esta inversión de tiempo como un costo de transición necesario para la transformación digital de los documentos.

13.3 Recomendaciones

Debido a que entornos como parques o plazas carecen de una dirección domiciliaria convencional, se recomienda identificar alguna etiqueta específica con la que la plataforma *OpenStreetMap* asocie al lugar buscado para asegurar una georreferenciación más precisa de las actividades de conservación realizadas en dichas áreas donde un domicilio convencional podría no ser representativo.

Adicionalmente con el fin de garantizar la sostenibilidad y escalabilidad del sistema en el tiempo, se sugiere establecer la entrega de la Ficha Digital como un requisito formal dentro de futuras bases administrativas de los contratos de conservación. De esta manera se externaliza la tarea de generar las fichas al contratista, que actualmente entrega documentación complementaria similar, pero en formato físico, así evitar sobrecargar operativamente a los funcionarios municipales, permitiendo que el personal interno se enfoque en el análisis de la gestión y no en la digitación de datos.

14. Bibliografía

de Solminihac, H., Echaveguren, T., & Chamorro, A. (2018). *Gestión de infraestructura vial* (3rd ed.).

FONDEF. (2015). Sistema de Gestión de Pavimentos Urbanos. Proyecto FONDEF D09I1018. Pontificia Universidad Católica de Chile.

Ley N° 11.150. (1953). *Diario Oficial de la República de Chile*. https://bcn.cl/2jxeq

López, M., Sánchez-Robles, J., Llopis, D., García, T., & Paredes, R. (2020). SISTEMA INTEGRAL DE MANTENIMIENTO EFICIENTE DE PAVIMENTOS URBANOS (SIMEPU). *XV Jornada Nacional de ASEFMA*. https://asefma.es/wp-content/uploads/2020/11/8-Pavasal-SIMEPU-XV-Jornada-Nacional-de-Asefma-2020.pdf

Martínez, R., Portolés-Rodríguez, D., Fernández, R., Segura, M. J., Fernández, L., Cebolla, R. M., Fernández Sola, R., Maria, ´, Segura, J., & Fernández, L. (2012). *REFLEXIONES Y CONSIDERACIONES PARA UNA CORRECTA PRESERVACIÓN DE DOCUMENTOS URBANÍSTICOS Revista Catalana de Geografia Revista digital de geografia, cartografia i ciències de la Terra*.

Peddarapu, R. K., Anogna, S., Harshini, M. N. S. S., Sireesha, J., & Rathna, K. G. (2025). Document Summarizer: A Machine Learning approach to PDF summarization. *Procedia Computer Science*, *252*, 1025–1036. https://doi.org/10.1016/J.PROCS.2025.01.063

Sciannameo, V., Pagliari, D. J., Urru, S., Grimaldi, P., Ocagli, H., Ahsani-Nasab, S., Comoretto, R. I., Gregori, D., & Berchialla, P. (2024). Information extraction from medical case reports using OpenAI InstructGPT. *Computer Methods and Programs in Biomedicine*, *255*, 108326. https://doi.org/10.1016/J.CMPB.2024.108326

Shahin, MY. (2005). Pavement management for airports, roads and parking lots, 2nd ed. Springer.

Tanasa, A. M., & Oprea, S. V. (2025). Rethinking Chart Understanding Using Multimodal Large Language Models. *Computers, Materials and Continua*, *84*(2), 2905–2933. https://doi.org/10.32604/cmc.2025.065421

Whiteley-Lagace, L., Helali, K., & Dietz, J. (2011). Implementing a Pavement Management System for Low Volume Roads: Challenges and Successes. *Annual Conference of the Transportation Association of Canada*.