PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA DE SISTEMAS

Observatorio de demanda laboral en América Latina

Política de Tratamiento de Datos y Ética del Web Scraping

Noviembre 2025

Versión 2.0

Alejandro Pinzón Fajardo

Nicolás Francisco Camacho Alarcón

Proyecto de Grado

BOGOTÁ, D.C.

Índice

1 INTRODUCCIÓN

1.1 Presentación

El Observatorio de Demanda Laboral en América Latina es un sistema automatizado diseñado para el monitoreo y análisis de la demanda de habilidades técnicas en los mercados laborales de Colombia, México y Argentina. El sistema emplea tecnologías de web scraping, procesamiento de lenguaje natural (NLP) e inteligencia artificial para recopilar ofertas de empleo de portales públicos especializados, extraer habilidades técnicas mediante la taxonomía ESCO (European Skills, Competences, Qualifications and Occupations), y generar análisis de tendencias del mercado laboral.

El web scraping constituye el método fundamental de adquisición de datos del sistema, permitiendo la recolección sistemática de información públicamente accesible desde portales de empleo que operan en los países objetivo. Esta técnica ha sido ampliamente validada en contextos académicos y de investigación como herramienta legítima para el análisis del mercado laboral [1].

1.2 Justificación del Uso de Web Scraping

La implementación de web scraping como metodología de recolección de datos en el Observatorio se fundamenta en las siguientes consideraciones:

- Acceso a información pública ya disponible en internet, sin barreras técnicas que impidan su visualización por parte de usuarios regulares.
- Necesidad académica de obtener datos agregados y actualizados del mercado laboral para fines de investigación científica.
- Ausencia de APIs públicas o mecanismos oficiales de acceso masivo a datos en la mayoría de portales de empleo latinoamericanos.
- Automatización de procesos de monitoreo que serían técnicamente inviables mediante recolección manual.
- Finalidad exclusivamente académica, sin propósitos comerciales ni competencia desleal contra los portales scrapeados.

El uso de web scraping en este proyecto se encuentra respaldado por precedentes académicos y jurídicos que reconocen su legitimidad cuando se realiza sobre información pública, con fines de investigación, respetando los términos de servicio razonables y sin causar daños técnicos a los sistemas de origen [1].

1.3 Objetivo de la Política

La presente política tiene como objetivo establecer los principios legales, éticos y técnicos que rigen la práctica de web scraping implementada en el Observatorio de Demanda Laboral, así como el tratamiento de los datos personales que puedan ser incidentalmente recopilados durante este proceso. Se busca garantizar:

- El cumplimiento de la normativa legal colombiana e internacional sobre protección de datos y propiedad intelectual.
- La adopción de estándares éticos reconocidos en la práctica de web scraping académico.
- La transparencia en los métodos de recolección y procesamiento de información.
- El respeto a los derechos de los titulares de datos personales que puedan estar contenidos en ofertas laborales públicas.
- La implementación de medidas técnicas que garanticen un scraping responsable y no intrusivo.

2 MARCO LEGAL DEL WEB SCRAPING

2.1 Legalidad del Web Scraping en Colombia y Latinoamérica

El web scraping, como técnica de extracción automatizada de información desde sitios web públicos, no está explícitamente prohibido por la legislación colombiana ni por los marcos jurídicos de México o Argentina. Su legalidad depende del cumplimiento de diversos factores legales y éticos que varían según el contexto de uso, la naturaleza de los datos extraídos y el respeto a los derechos de propiedad intelectual y protección de datos personales.

2.1.1 Información Pública vs. Información Privada

La distinción fundamental para determinar la legalidad del web scraping radica en la naturaleza de la información accedida:

- Información pública: Datos accesibles sin autenticación, sin restricciones técnicas explícitas (robots.txt) y sin violación de términos de servicio razonables. Esta categoría incluye ofertas de empleo publicadas abiertamente en portales especializados.
- Información privada: Datos protegidos por autenticación, contraseñas, paywalls o restricciones técnicas explícitas cuya elusión constituiría una violación legal.

El Observatorio de Demanda Laboral se limita exclusivamente a la recolección de información pública, accesible mediante navegación estándar de internet, sin eludir mecanismos de seguridad ni acceder a áreas restringidas de los portales scrapeados.

2.1.2 Precedentes Legales y Académicos

Según Orozco y Gómez (2019), el web scraping aplicado a portales de empleo con fines académicos y de investigación del mercado laboral constituye una práctica legítima cuando se cumplen las siguientes condiciones [1]:

1. La información extraída es de carácter público y está disponible sin autenticación.
2. El propósito es educativo, académico o de investigación científica, sin fines comerciales directos.
3. Se respetan los estándares técnicos de scraping responsable (robots.txt, rate limiting).
4. No se causa daño técnico a la infraestructura de los sitios scrapeados.
5. Se implementan mecanismos de anonimización y agregación de datos personales.

El proyecto cumple con todos estos criterios, posicionándose dentro del marco de scraping ético y legalmente aceptable para fines de investigación académica.

2.2 Legislación Aplicable

El tratamiento de datos recolectados mediante web scraping en el Observatorio se rige por las siguientes normativas:

2.2.1 Colombia

- Ley 1581 de 2012: Régimen General de Protección de Datos Personales, que establece principios de legalidad, finalidad, libertad, veracidad, transparencia, seguridad y confidencialidad en el tratamiento de datos.
- Decreto 1377 de 2013: Reglamentación parcial de la Ley 1581 de 2012.
- Ley 23 de 1982: Sobre derechos de autor, aplicable a la protección de contenidos originales publicados en portales web.
- Constitución Política de Colombia (Artículos 15 y 20): Protección del derecho a la intimidad, habeas data y acceso a la información pública.

2.2.2 México

- Ley Federal de Protección de Datos Personales en Posesión de los Particulares (2010).
- Ley Federal del Derecho de Autor (1996).

2.2.3 Argentina

- Ley 25.326 de Protección de Datos Personales (2000).
- Ley 11.723 de Propiedad Intelectual (1933).

2.3 Propiedad Intelectual y Uso Legítimo

El Observatorio reconoce los derechos de propiedad intelectual de los portales de empleo sobre el diseño, estructura y contenidos originales de sus plataformas. Sin embargo, la extracción de datos factuales (títulos de puestos, ubicaciones geográficas, habilidades técnicas requeridas) publicados por terceros (empleadores) no constituye una violación de derechos de autor, ya que:

- Los datos factuales no están protegidos por derechos de autor según la legislación colombiana e internacional.
- La información extraída proviene de ofertas publicadas por empleadores, no creadas por los portales.

- El uso se enmarca en el principio de uso legítimo (fair use) para fines académicos y de investigación.
- No se reproduce la estructura, diseño ni elementos creativos de los portales scrapeados.

3 PRINCIPIOS ÉTICOS DEL WEB SCRAPING

3.1 Respeto al Protocolo robots.txt

El archivo robots.txt es un estándar técnico mediante el cual los administradores de sitios web comunican sus preferencias sobre qué secciones del sitio pueden o no ser accedidas por robots automatizados. El Observatorio de Demanda Laboral implementa las siguientes políticas de respeto a robots.txt:

- Verificación previa del archivo robots.txt antes de iniciar cualquier proceso de scraping en un portal nuevo.
- Respeto absoluto a las directivas Disallow y Allow especificadas en el archivo.
- Identificación clara del bot mediante User-Agent descriptivo que permite a los administradores identificar y contactar al proyecto.
- Suspensión inmediata del scraping en portales que posteriormente agreguen restricciones en robots.txt.

El respeto a robots.txt constituye una buena práctica ética reconocida internacionalmente y demuestra la voluntad del proyecto de operar dentro de las normas técnicas establecidas por la comunidad web.

3.2 Cumplimiento de Términos de Servicio

Los Términos de Servicio (ToS) de los portales web constituyen acuerdos contractuales entre el portal y sus usuarios. El Observatorio analiza los ToS de cada portal scrapeado para identificar:

- Prohibiciones explícitas y razonables sobre scraping automatizado.
- Restricciones específicas sobre uso comercial de datos.
- Limitaciones de frecuencia o volumen de accesos.
- Requisitos de atribución o reconocimiento.

Cuando un portal prohíbe explícitamente el scraping en sus ToS, el proyecto evalúa la razonabilidad de dicha prohibición en el contexto de uso académico y de investigación científica. En casos donde la prohibición se considera razonable y técnicamente justificada, el portal es excluido del alcance del scraping.

Sin embargo, se reconoce que ciertas cláusulas de ToS que prohíben de manera absoluta cualquier acceso automatizado a información pública pueden ser consideradas excesivamente restrictivas y potencialmente no aplicables en contextos académicos protegidos por principios de libertad de investigación y acceso a información pública.

3.3 Scraping Responsable: Limitación de Frecuencia

El Observatorio implementa mecanismos técnicos de scraping responsable para minimizar el impacto en la infraestructura de los portales objetivo:

- Rate limiting: Límite de 1-2 peticiones por segundo por portal, muy por debajo de los umbrales que podrían afectar el rendimiento del sitio.
- Delays aleatorios: Pausas variables entre peticiones (2-5 segundos) para simular comportamiento humano y distribuir la carga.
- Horarios de baja demanda: Ejecución preferente de scrapers durante horarios nocturnos o de bajo tráfico en cada región geográfica.
- Respect headers: Atención a headers HTTP como Retry-After y respeto a códigos de estado 429 (Too Many Requests).
- Caché local: Almacenamiento temporal de páginas visitadas para evitar peticiones duplicadas innecesarias.
- Session reuse: Reutilización de conexiones HTTP para reducir overhead de conexión.

Estas medidas garantizan que el scraping del Observatorio sea técnicamente indistinguible de tráfico regular de usuarios humanos y no cause impacto negativo en la disponibilidad o rendimiento de los portales scrapeados.

3.4 Transparencia en la Recolección de Datos

El proyecto adopta principios de transparencia en sus operaciones de scraping:

- User-Agent identificable: Los bots se identifican con un User-Agent descriptivo que incluye el nombre del proyecto y correo de contacto.
- Documentación pública: El código fuente del proyecto, incluyendo los scrapers, está disponible públicamente bajo licencia de código abierto.
- Contactabilidad: Se proporciona información de contacto accesible para que administradores de portales puedan comunicarse con el equipo del proyecto.
- Política de exclusión voluntaria: Cualquier portal puede solicitar ser excluido del scraping, solicitud que será atendida inmediatamente.

3.5 No Elusión de Medidas de Seguridad

El Observatorio se compromete a no eludir medidas de seguridad técnicas implementadas por los portales:

- No se elude CAPTCHAs ni sistemas anti-bot.
- No se accede a contenido protegido por autenticación sin autorización explícita.
- No se explotan vulnerabilidades de seguridad de los portales.
- No se utilizan técnicas de ofuscación para ocultar la naturaleza automatizada del acceso.

4 IMPLEMENTACIÓN ÉTICA EN EL OBSERVATORIO

4.1 Portales Scrapeados y Justificación

El Observatorio realiza web scraping sobre los siguientes 8 portales de empleo públicos en Colombia, México y Argentina:

1. HiringCafe (hiring.cafe)
2. Computrabajo (computrabajo.com)
3. Bumeran (bumeran.com)
4. ElEmpleo (elempleo.com)
5. OCC Mundial (occ.com.mx)
6. ZonaJobs (zonajobs.com)
7. Indeed (indeed.com)
8. Magneto (magneto365.com)

Estos portales fueron seleccionados con base en los siguientes criterios:

- Carácter público de las ofertas publicadas (accesibles sin autenticación).
- Relevancia en los mercados laborales objetivo (Colombia, México, Argentina).
- Ausencia de restricciones técnicas absolutas en robots.txt que prohíban el acceso a ofertas públicas.
- Enfoque en sectores tecnológicos que permitan analizar la demanda de habilidades técnicas específicas.

4.2 Medidas Técnicas de Scraping Responsable

El sistema implementa las siguientes medidas técnicas para garantizar un scraping ético y responsable:

4.2.1 Configuración de Scrapy Framework

- CONCURRENT_REQUESTS_PER_DOMAIN: 1-2 (máximo de peticiones simultáneas por dominio)
- DOWNLOAD_DELAY: 2-5 segundos (pausa entre peticiones consecutivas)

- RANDOMIZE_DOWNLOAD_DELAY: True (variación aleatoria de delays)
- ROBOTSTXT_OBEY: True (respeto obligatorio a robots.txt)
- USER_AGENT: Identificación clara del proyecto con correo de contacto
- AUTOTHROTTLE_ENABLED: True (ajuste automático de velocidad según respuesta del servidor)

4.2.2 Limitaciones de Alcance

- Scraping limitado a ofertas activas (últimos 30-60 días).
- Exclusión de secciones no relevantes del portal (foros, blogs, perfiles de usuarios).
- Filtrado por país y categoría tecnológica para minimizar peticiones innecesarias.
- Deduplicación para evitar re-scraping de ofertas ya recolectadas.

4.3 Datos Públicos vs. Datos Privados

El Observatorio distingue claramente entre información pública y privada:

4.3.1 Información Pública Recolectada

- Título del puesto de trabajo
- Descripción de funciones y responsabilidades
- Requisitos técnicos y habilidades requeridas
- Ubicación geográfica (ciudad, país)
- Información de la empresa empleadora cuando es pública
- Modalidad de trabajo (remoto, presencial, híbrido)
- Fecha de publicación

4.3.2 Información Privada NO Recolectada

- Currículums o perfiles de candidatos
- Información de contacto personal (correos, teléfonos)
- Salarios específicos cuando no son públicos
- Datos de cuentas de usuario del portal

- Información detrás de autenticación
- Mensajes privados o comunicaciones internas

4.4 Anonimización y Agregación de Datos

Una vez recolectados, los datos son procesados mediante técnicas de anonimización y agregación:

- Eliminación de información de contacto personal incidentalmente capturada.
- Agregación estadística de habilidades sin vincular a empresas o personas específicas.
- Generación de reportes que presentan tendencias agregadas del mercado laboral.
- Aplicación de técnicas de k-anonimidad para conjuntos de datos compartidos con fines académicos.
- Pseudonimización de empresas empleadoras en análisis públicos.

4.5 Propósito Académico y No Comercial

El Observatorio opera exclusivamente con fines académicos y de investigación científica:

- No comercialización: Los datos recolectados no son vendidos ni comercializados.
- No competencia: El sistema no compite comercialmente con los portales scrapeados.
- Publicaciones académicas: Los resultados se publican en tesis, artículos científicos y conferencias académicas.
- Acceso abierto: Los hallazgos son compartidos públicamente bajo principios de ciencia abierta.
- Código abierto: El software desarrollado se publica bajo licencias de código abierto (MIT, Apache 2.0).

5 PROTECCIÓN DE DATOS PERSONALES

5.1 Responsable del Tratamiento

El responsable del tratamiento de los datos personales que puedan ser incidentalmente recolectados es:

- Nombre: Proyecto Observatorio de Demanda Laboral en América Latina
- Naturaleza: Proyecto académico de investigación
- Institución: Pontificia Universidad Javeriana
- Domicilio: Bogotá D.C., Colombia
- Correos de contacto: alejandro_pinzon@javeriana.edu.co, camachoa.nicolas@javeriana.edu.co

5.2 Principios de Tratamiento de Datos

El tratamiento de datos personales en el Observatorio se rige por los siguientes principios de la Ley 1581 de 2012:

- Legalidad: Cumplimiento de la normativa aplicable.
- Finalidad: Datos recolectados exclusivamente para investigación académica del mercado laboral.
- Libertad: Recolección transparente con fines legítimos.
- Veracidad: Procesamiento sin alteración del contenido original de fuentes públicas.
- Transparencia: Derecho de los interesados a conocer el uso de los datos.
- Seguridad: Medidas técnicas y administrativas de protección.
- Confidencialidad: No divulgación de datos que permitan identificación personal.

5.3 Minimización y Proporcionalidad

El sistema implementa principios de minimización de datos:

- Solo se procesa información estrictamente necesaria para los objetivos de investigación.
- Datos personales incidentalmente capturados son anonimizados lo antes posible.
- Información sensible (salud, religión, orientación sexual) es identificada y eliminada automáticamente.
- Datos agregados y estadísticos se prefieren sobre datos individuales en todos los análisis.

5.4 Medidas de Seguridad

- Cifrado de datos en tránsito (HTTPS, TLS 1.3) y en reposo (PostgreSQL encryption).
- Control de acceso mediante autenticación y autorización (solo equipo autorizado).
- Anonimización automática de datos personales mediante scripts de procesamiento.
- Auditoría y logging de accesos a la base de datos.
- Respaldos cifrados con retención limitada.
- Infraestructura con certificaciones de seguridad (Docker, servidores seguros).

5.5 Almacenamiento y Retención

- Datos de ofertas: Almacenados hasta 24 meses desde su recolección.
- Datos agregados y anonimizados: Conservación indefinida para análisis histórico.
- Información identificable: Eliminación segura tras anonimización.
- Método de eliminación: Borrado permanente de registros y sobrescritura de backups.

6 DERECHOS DE LOS TITULARES Y PROCEDIMIENTOS

6.1 Derechos de los Titulares

Las personas cuyos datos puedan estar siendo procesados tienen los siguientes derechos según la Ley 1581 de 2012:

- Conocer, actualizar y rectificar sus datos personales.
- Solicitar prueba de la autorización otorgada cuando aplique.
- Ser informado sobre el uso de sus datos personales.
- Presentar quejas ante la Superintendencia de Industria y Comercio (SIC).
- Revocar la autorización y solicitar la supresión de datos cuando no se respeten principios legales.
- Acceder gratuitamente a sus datos personales procesados.

6.2 Procedimiento de Consultas y Reclamos

6.2.1 Consultas

Para consultas sobre datos personales:

1. Enviar solicitud a: alejandro_pinzon@javeriana.edu.co o camachoa.nicolas@javeriana.edu.co
2. Incluir: nombre completo, datos de contacto, descripción de la consulta
3. Plazo de respuesta: 10 días hábiles (extensible hasta 5 días hábiles adicionales)

6.2.2 Reclamos

Para reclamos sobre corrección, actualización o eliminación de datos:

1. Presentar reclamo vía correos electrónicos indicados
2. Incluir: identificación, motivo del reclamo, documentos de respaldo
3. Plazo de respuesta: 15 días hábiles (extensible hasta 8 días hábiles adicionales)

6.3 Supresión de Datos

Los titulares pueden solicitar la eliminación de sus datos personales cuando:

- No se respeten los principios de protección de datos.
- Los datos ya no sean necesarios para los fines académicos.
- Exista uso indebido de la información.

La supresión se realizará siempre que no exista obligación legal o académica de conservación.

6.4 Contacto y Quejas

- Correos: alejandro_pinzon@javeriana.edu.co, camachoa.nicolas@javeriana.edu.co
- Institución: Pontificia Universidad Javeriana - Facultad de Ingeniería
- Dirección: Carrera 7 # 40-62, Bogotá D.C., Colombia
- Quejas SIC: www.sic.gov.co (Superintendencia de Industria y Comercio)

7 VIGENCIA Y ACTUALIZACIONES

7.1 Vigencia

Esta Política de Tratamiento de Datos y Ética del Web Scraping entra en vigor a partir del 16 de noviembre de 2025 y permanecerá vigente mientras el Observatorio de Demanda Laboral esté en operación.

7.2 Modificaciones

El equipo responsable se reserva el derecho de modificar esta política cuando sea necesario para:

- Adaptarse a cambios en la legislación colombiana o internacional.
- Incorporar nuevos estándares éticos de scraping.
- Actualizar medidas de seguridad y protección de datos.
- Incluir o excluir portales del alcance del scraping.

Cualquier modificación será comunicada mediante actualización de la versión del documento y publicación en el repositorio público del proyecto.

7.3 Compromiso Ético

El equipo del Observatorio de Demanda Laboral se compromete a:

- Operar bajo los más altos estándares éticos de investigación académica.
- Respetar los derechos de propiedad intelectual y protección de datos.
- Implementar scraping responsable que no cause daño a los portales objetivo.
- Mantener transparencia en los métodos de recolección y procesamiento de datos.
- Responder de manera oportuna a solicitudes de titulares de datos y administradores de portales.
- Contribuir al conocimiento científico sobre el mercado laboral latinoamericano.

Observatorio de Demanda Laboral en América Latina
Pontificia Universidad Javeriana
Facultad de Ingeniería - Departamento de Sistemas
Bogotá D.C., Colombia

alejandro_pinzon@javeriana.edu.co — camachoa.nicolas@javeriana.edu.co

Referencias

[1] V. M. Orozco Puello y L. F. Gómez Estrada, “Desarrollo de un prototipo de aplicación web que permita la extracción de las ofertas laborales de las principales plataformas que postulan empleos en la Región Caribe, usando la técnica web scraping,” Proyecto de Grado, Universidad del Sinú Elías Bechará Zainúm, Cartagena, Colombia, 2019.