**..UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ**

**Colegio de Ciencias e Ingenierías**

**WEBCRAWLER DE PROCESOS DE SUBASTA INVERSA ELECTRÓNICA DEL PORTAL DE COMPRAS PÚBLICAS DEL ECUADOR**

.

**Sebastián Alejandro Taraguay Silva**

**Ingeniería en Sistemas**

Trabajo de fin de carrera presentado como requisito

para la obtención del título de

Ingeniero en Sistemas

Quito, día de mes de año

**Universidad San Francisco de Quito USFQ**

**Colegio de Ciencias e Ingenierías**

HOJA DE CALIFICACIÓN

DE TRABAJO DE FIN DE CARRERA

WEBCRAWLER DE PROCESOS DE SUBASTA INVERSA ELECTRÓNICA DEL PORTAL DE COMPRAS PÚBLICAS DEL ECUADOR

Sebastián Alejandro Taraguay Silva

**Nombre del profesor, Título académico Xxxxxxxxxxxxxx, xxxxxxxxxxxxx**

Quito, día de mes de año

**© DERECHOS DE AUTOR**

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Sebastián Alejandro Taraguay Silva

Código: 139248

Cédula de identidad: 1723028096

Lugar y fecha: Ciudad, día de mes de año

**ACLARACIÓN PARA PUBLICACIÓN**

**Nota:** El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en <http://bit.ly/COPETheses>.

**UNPUBLISHED DOCUMENT**

**Note:** The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on <http://bit.ly/COPETheses>.

RESUMEN

Dentro de los Procesos de Contratación Pública, el mecanismo de subasta inversa electrónica ha conformado una parte significativa del total de contrataciones públicas, llegando en el año de 2019 según la SERCOP al 34.1% del total de procesos de contratación pública del país. Todos los registros detallados de este tipo de contratación se encuentran publicados en el portal web de la SERCOP, sin embargo, existen ciertos detalles que no se pueden visualizar en ciertas etapas del proceso, además, la información dentro del portal web no se encuentra presentada de una manera clara y sencilla para los usuarios finales, debido a esto se dificulta la transparencia de los procesos. La motivación principal de este proyecto es facilitar y presentar de manera clara y sencilla los detalles más relevantes de cada proceso de subasta inversa electrónica con el fin de servir de base a posibles trabajos futuros que auditen este tipo de contrataciones en busca de posibles irregularidades. A través de la implementación de un Web Crawler escrito en Python y con ayuda del módulo Sellenium, se pretende cumplir con el objetivo principal.

**Palabras clave:** contratación pública, subasta inversa electrónica, web crawler, corrupción, Python, base de datos

ABSTRACT

Within the Public Procurement Processes, the electronic reverse auction mechanism has made up a significant part of the total public procurement, reaching 34.1% of the total public procurement processes in the country in 2019 according to SERCOP. All the detailed records of this type of contracting are published on the SERCOP web portal, however, there are certain details that cannot be viewed at certain stages of the process, in addition, the information within the web portal is not presented in a clear and simple way for end users, due to this, the transparency of the processes is difficult. The main motivation of this project is to facilitate and present in a clear and simple way the most relevant details of each electronic reverse auction process in order to serve as a basis for possible future work that audits this type of contracting in search of possible irregularities. Through the implementation of a Web Crawler written in Python and with the help of the Sellenium module, it is intended to meet the main objective.

**Key words:** public procurement, electronic reverse auction, web crawler, corruption, Python, database.

TABLA DE CONTENIDO

Introducción 10

Antecedentes 10

Objetivos 11

Desarrollo del Tema 12

Conclusiones 13

Referencias bibliográficas 14

Anexo A: Título 15

Anexo B: Título 15

Anexo C: Título 15

ÍNDICE DE IMÁGENES

[Imagen 1: Página de búsqueda de procesos de contratación. 14](#_Toc65263408)

[Imagen 2: Tabla de resultado de procesos subasta inversa electrónica. 14](#_Toc65263409)

[Imagen 3: Página con los detalles de cada proceso. 15](#_Toc65263410)

[Imagen 4: Códigos únicos de cada proceso. 17](#_Toc65263411)

[Imagen 5: Información relevante para ser extraída. 17](#_Toc65263412)

[Imagen 6: Información relevante para ser extraída. 18](#_Toc65263413)

[Imagen 7: Información relevante para ser extraída. 18](#_Toc65263414)

[Imagen 8: Información relevante para ser extraída. 18](#_Toc65263415)

[Imagen 9: Información extra del proceso 19](#_Toc65263416)

[Imagen 10: Lista de proveedores invitados a participar 19](#_Toc65263417)

[Imagen 11: Información relevante del resumen de contrato. 20](#_Toc65263418)

ÍNDICE DE ILUSTRACIONES

[Ilustración 1: Flujo de proceso SIE 16](#_Toc65263444)

[Ilustración 2: Proceso cíclico de obtención de procesos. 20](#_Toc65263445)

[Ilustración 3: Flowchart del Web Crawler 21](#_Toc65263446)

[Ilustración 4: Módulos del Web Crawler 23](#_Toc65263447)

# Introducción

La subasta inversa electrónica hace referencia a los procedimientos dinámicos que son utilizados en la adquisición de bienes y servicios normalizados, en donde los proveedores pujan hacia la baja del precio ofertado. En el año 2019 este tipo de procesos de contratación pública abarco el 34.1% del total. A pesar de que la información se encuentra de manera pública en el portal web de la SERCOP, existen ciertos detalles que no se pueden visualizar en ciertas etapas del proceso por lo cual se puede dar el caso de algún tipo de corrupción. De ahí nace la necesidad de comprender la estructura y funcionamiento del portal web de consulta sobre los procesos de contratación pública para posteriormente automatizar la obtención y almacenamiento de la información relevante, mediante la implementación de un WEBCRAWLER(bot) desarrollado en base a Python3 y la librería Selenium, el cual de manera automática se encargará de consultar en el portal web los procesos de Subasta Inversa Electrónica y se la información será almacenada de manera local en una base de datos con el fin de a futuro poder analizar los registros obtenidos en busca de irregularidades en cuanto a temas de contratación pública.

# Antecedentes

En el Ecuador actualmente existen 35 tipos de procesos de contratación pública, de los cuales según el *Análisis Anual de la Contratación Pública – 2019* publicado por la SERCOP - SOCE el 34.1% del total corresponde al método de subasta inversa electrónica es decir un monto correspondiente a 1.626.8 millones de dólares. Si bien el SOCE tiene como obligación facilitar al público en general la información acerca de los procedimientos de contratación pública, salvo ciertas excepciones las cuales deben ser debidamente justificadas por parte del gobierno central. En el caso de Subasta Inversa Electrónica luego que el procedimiento pasa a los estados En *Recepción* o *Finalizados* no es posible visualizar a detalle la calificación obtenida de cada proveedor participante ni el historial de puja o negociación del proceso en específico, lo cual puede llevar a casos de corrupción en la adjudicación de procesos debido a la falta de transparencia total. Por ello nace la necesidad de automatizar la recopilación y organización de los procesos publicados en el portal web de la SERCOP facilitando a posibles investigaciones futuras.

# Objetivos

Objetivo General

* Recopilar, automatizar y organizar los registros referentes a los procesos de subasta inversa electrónica de dominio público, alojados en el portal web de la SERCOP

Objetivo Específicos

* Automatizar la recopilación de los registros alojados en el portal web de la SERCOP mediante la implementación de un WEBCRAWLER desarrollado en Python y Sellenium.
* Optimizar el almacenamiento y la organización de los registros obtenidos mediante la implementación de una base de datos correctamente estructurada para la facilitación a futuras consultas.
* Comprender los niveles de navegación, la estructura y el funcionamiento de búsqueda de procesos en el portal web de la SERCOP con el fin de implementar correctamente el WEBCRAWLER propuesto y sentar una base para extensiones a futuro del proyecto.

# Desarrollo del Tema

Subasta inversa electrónica

La contratación mediante el procedimiento de subasta inversa se rige a los lineamientos de la Ley Orgánica del Servicio Nacional de Contratación Pública (LOSNCP) en la misma se señala en la sección 2, artículo 47 que:

“*Para la adquisición de bienes y servicios normalizados que no consten en el catálogo electrónico, las Entidades Contratantes deberán realizar subastas inversas en las cuales los proveedores de bienes y servicios equivalentes pujan hacia la baja el precio ofertado, en acto público o por medios electrónicos a través del Portal de COMPRASPUBLICAS*” (SERCOP, 2018)

Se entiende como bienes y servicios normalizados “Aquellos cuyas características o especificaciones técnicas han sido estandarizadas u homologadas por la entidad contratante; y, en consecuencia, dichas características o especificaciones son homogéneas y comparables en igualdad de condiciones” (SERCOP, 2016).

Todo el proceso de subasta inversa electrónica es llevado a cabo en el portal de compras públicas de la SERCOP y se entiende como un procedimiento dinámico que es utilizado en la adquisición de bienes y servicios normalizados siempre y cuando su cuantía supere el monto equivalente al 0.0000002 del Presupuesto Inicial del Estado. La principal diferencia de este tipo de subasta inversa con el mecanismo de una subasta corriente radica en que los ofertantes pujan hacia la baja del precio ofertado inicial, siendo el ganador aquel que haya enviado la propuesta con el costo económico más bajo. Según los lineamientos de la SERCOP cada nueva puja tiene un tiempo de duración entre 15 a 60 minutos y su variación se recomienda entre el 0.5% y el 2% del precio referencial. En caso de que se presente una única oferta técnica calificada, no se continuara con la subasta y en su lugar se procederá a realizar la sesión única de negociación entre el contratante y el oferente, el objetivo de la sesión será el mejorar la oferta económica y en caso de llegar a un acuerdo, se procederá a realizar la contratación con el único oferente. (SERCOP, 2016)

El sistema oficial de contratación pública

Las páginas web de la SERCOP, en donde se encuentran los registros de los procesos de contratación pública son muy poco intuitivas y amigables con el usuario, debido a esto puede resultar complejo para un usuario no experimentado la consulta y obtención de la información de los procesos. La página principal cuenta con un recuadro a la derecha en la cual se encuentra la opción de búsqueda de procesos, al seleccionar dicha opción el usuario es redirigido a una segunda página web en la cual se puede consultar todo tipo de procesos en base a diferentes filtros como: palabras clave, entidad contratante, fechas de publicación y tipo de contratación. En este caso todos los registros serán obtenidos en base al filtro de tipo de contratación estableciendo el valor en *subasta inversa electrónica*. Es importante señalar que el sistema cuenta con un sistema de verificación CAPTCHA definido como:

“Completely Automated Public Turing test to tell Computers and Humans Apart", que literalmente significa Test Público Completamente Automatizado de Turing para Diferenciar a los Seres Humanos de las Computadoras. Por lo tanto, viene a ser un Test de Turing inverso, en el que son los humanos los que tienen que demostrar con esta tecnología que son humanos y no máquinas” (FERNÁNDEZ, 2018)



Imagen 1: Página de búsqueda de procesos de contratación.

Para concretar la búsqueda de los procesos de subasta inversa electrónica es necesario completar dicho CAPTCHA el cual consta de una imagen en la que se presentan una serie de caracteres alfanuméricos. Una vez establecido el rango de fechas dentro del periodo valido de 6 meses antes y 15 días después de la fecha de publicación del proceso y completado con éxito el filtro CAPTCHA se presenta en la parte inferior de la pantalla en formato de tabla los 20 procesos de subasta inversa electrónica más recientes de un total de 11 281.

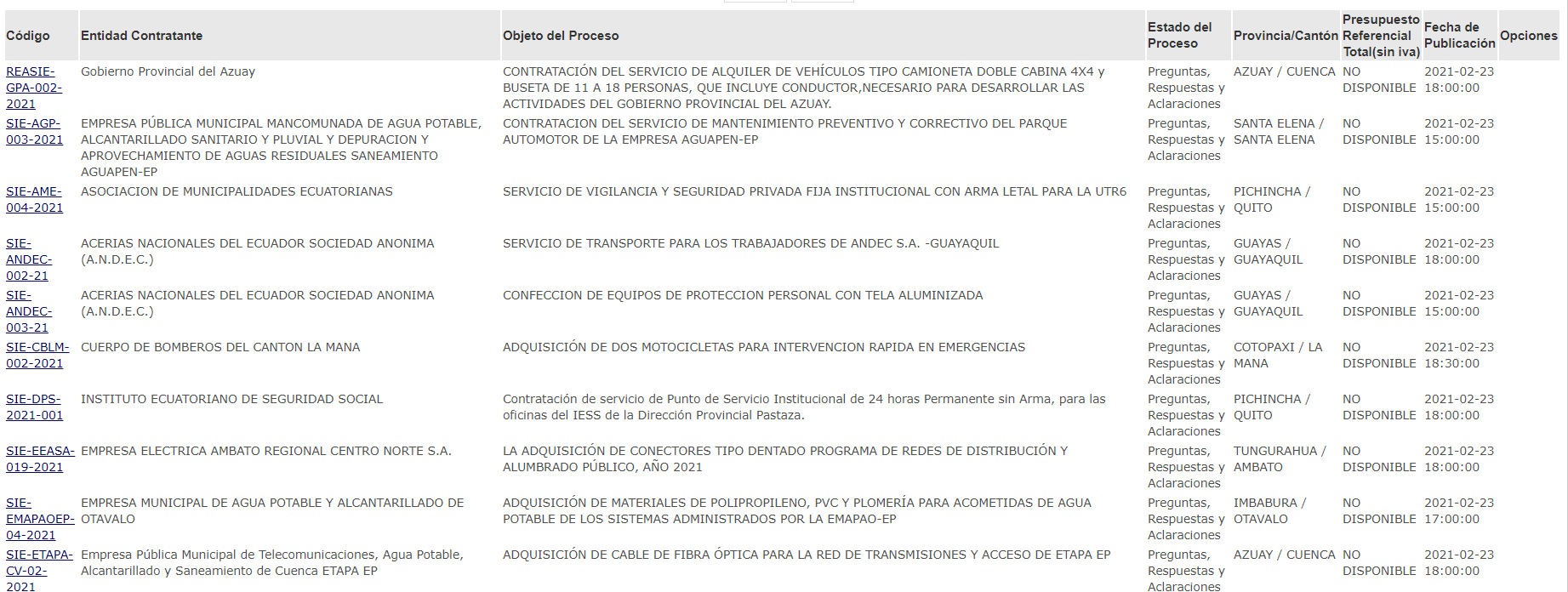


Imagen 2: Tabla de resultado de procesos subasta inversa electrónica.

Cada proceso listado en la tabla cuenta con los siguientes campos como valores de columna: código, entidad contratante, objeto del proceso, estado del proceso, provincia/cantón, presupuesto referencial total y fecha de publicación. El identificador único de cada proceso listado en la columna de *código* hace la función de un enlace el cual redirige al usuario a una página externa la cual contiene información detallada y relevante del proceso de subasta inversa electrónica tal como: autoridades a cargo, miembros de la comisión técnica, vigencia de la oferta, estado del proceso, valoración mínima de la puja, fechas relacionadas al proceso, archivos de respaldo y demás información. Cabe señalar que en la fase precontractual se añaden dos enlaces extras los cuales redirigen a otra página en la cual se encuentran preguntas y aclaraciones sobre el proceso y un segundo enlace conduce a otra página en la cual se lista en un formato de tabla los proveedores que fueron invitados a participar. Si el proceso fue finalizado se añade un enlace el cual contiene todo el resumen del proceso; todo el flujo de estados del proceso es representado en la ilustración 1.

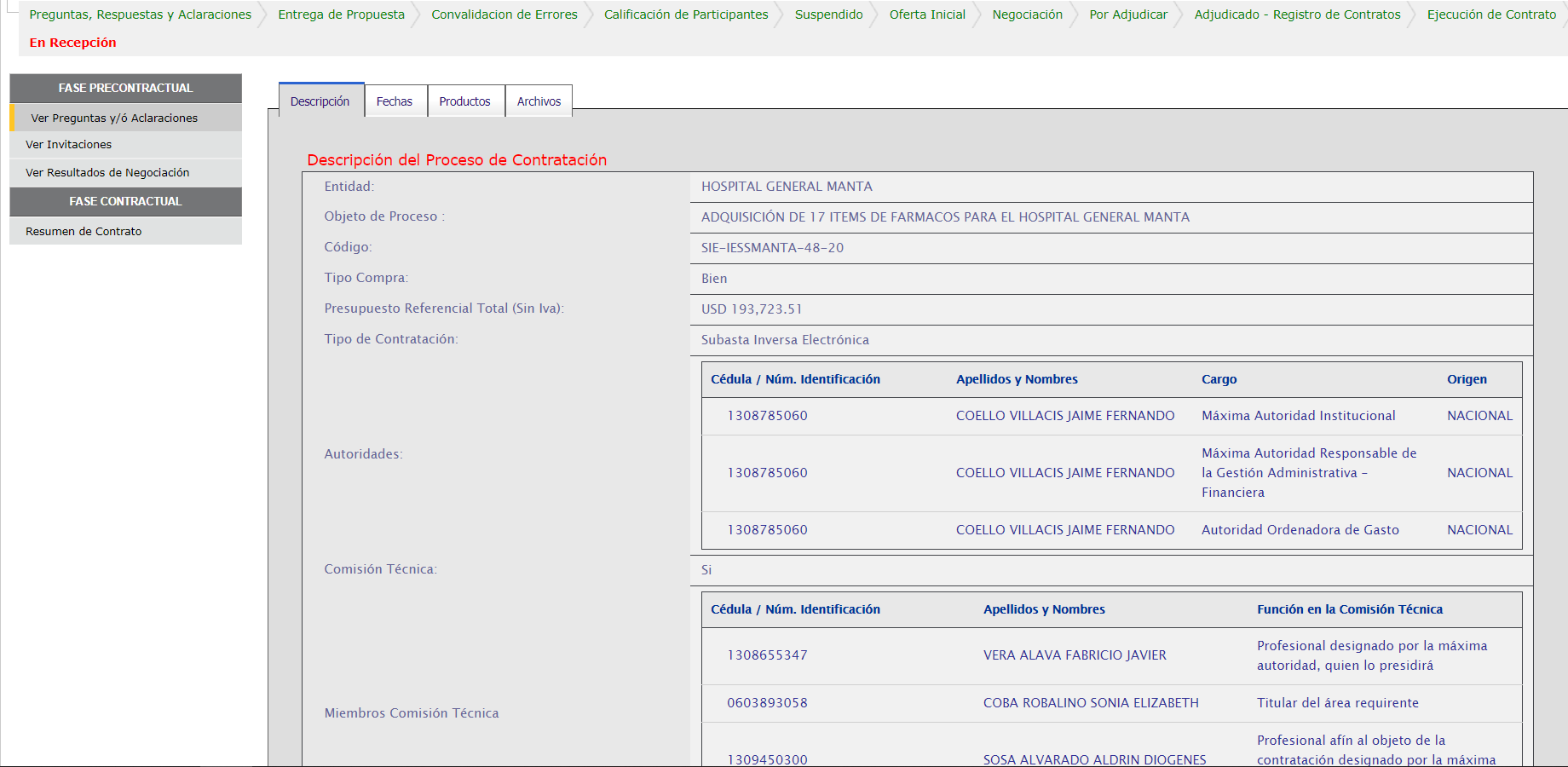


Imagen 3: Página con los detalles de cada proceso.

Ilustración 1: Flujo de proceso SIE

Estructura de navegación en el portal web de la SERCOP

El flujo de navegación a llevarse a cabo por el Web Crawler es de manera secuencial, comenzando en la página de búsqueda de procesos la cual presenta una tabla con los veinte procesos más recientes; en cada nivel de navegación se procederá a recolectar la información más relevante para ser almacenada posteriormente en la base de datos correspondiente. Con el objetivo de evitar posibles bloqueos, en caso de que sea detectado el Web Crawler en el sistema, en cada nivel de navegación se recolectará de manera independiente la información.

El inicio del Web Crawler será a partir de la página de búsqueda ( nivel Index) en la cual se presenta la tabla de procesos y su primer elemento de cada fila es un código que hace la función de enlace a otra página web ( nivel 1 ) en la cual se detalla la información del proceso de contratación. Posteriormente el Web Crawler se encargará de la recolección y almacenamiento de los datos más relevantes, mediante sus módulos segundo y tercero.

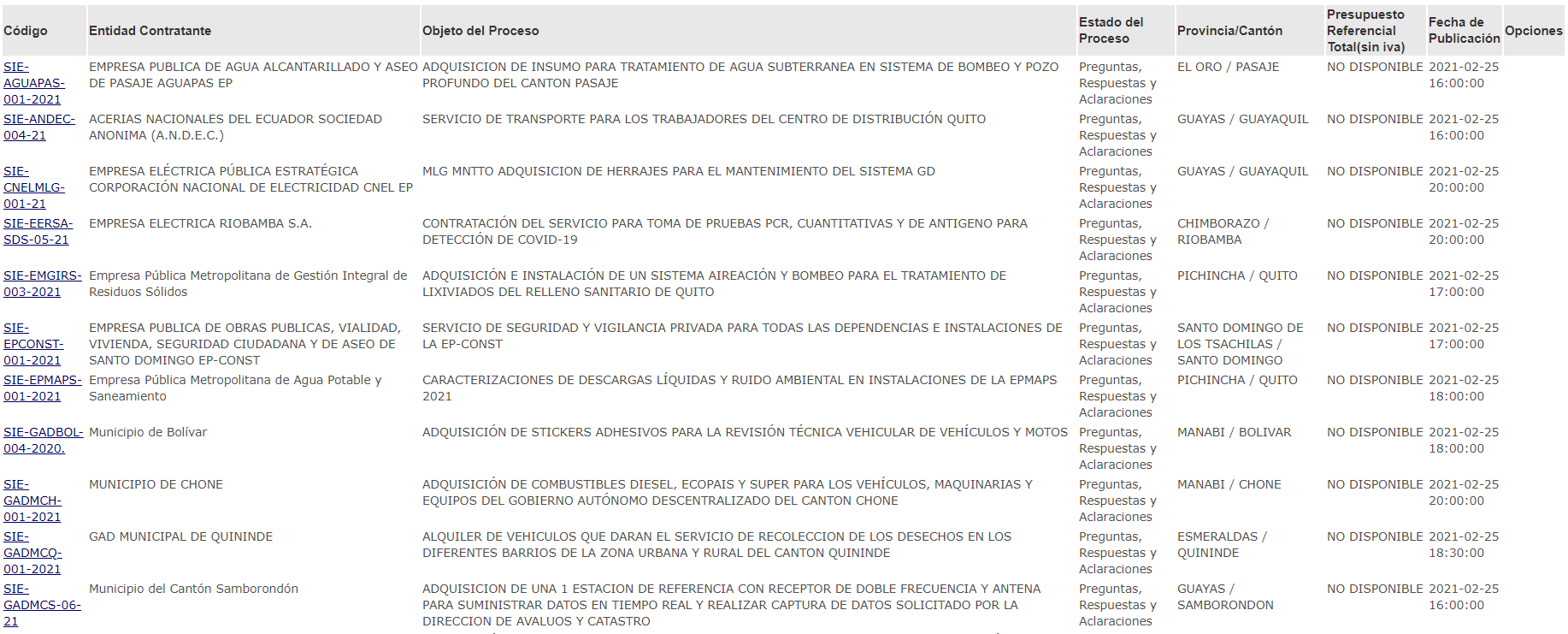


Imagen 4: Códigos únicos de cada proceso.

Según el estado en el que se encuentre el proceso existirá información extra dentro de las fases precontractuales y contractuales. Por ello la función del Web Crawler será detectar el estado y en base al mismo obtener la información disponible necesaria.

Una vez analizada la pantalla de descripción del proceso (nivel 1) se estableció como información relevante para la extracción a los campos como: entidad, objeto del proceso, presupuesto referencial total, autoridades, comisión técnica, funcionario encargado del proceso, estado del proceso, tipo de adjudicación, fecha de publicación, fecha de inicio de puja, fecha posible de adjudicación. (imagen 5 – 8).



Imagen 5: Información relevante para ser extraída.



Imagen 6: Información relevante para ser extraída.

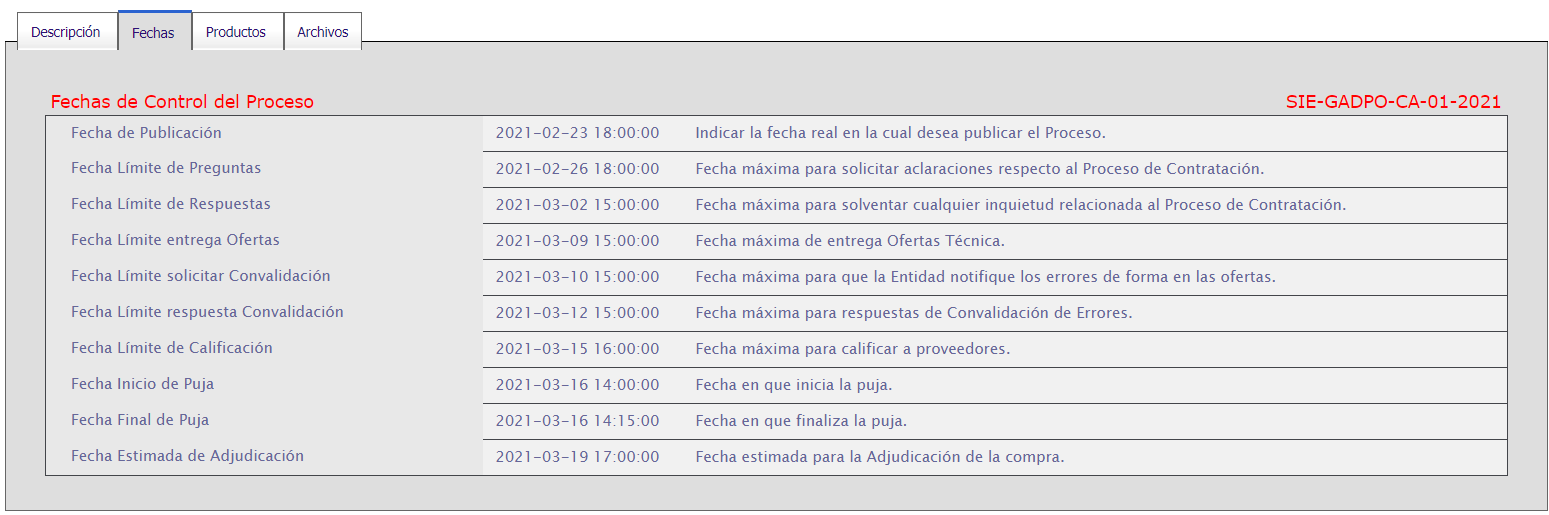


Imagen 7: Información relevante para ser extraída.

Continuando con la navegación web ( nivel 2 ) se recopilará información extra de los procesos como invitaciones a proveedores, los cuales son presentados en una tabla de 20 filas, además si se encuentra disponible se ingresará a la página de resultados de negociación y resumen del contrato, en los cuales se recopilará información puntual de gran relevancia y de igual manera será almacenada para uso a futuro.



CV

CV

Imagen 8: Información extra del proceso



Imagen 9: Lista de proveedores invitados a participar



Imagen 10: Información relevante del resumen de contrato.

A partir del nivel 1 no se requiere la solución de un sistema de CAPTCHA por ende la tarea del Web Crawler se ve facilitada y con posibilidad de navegar libremente entre los niveles de información de cada proceso. En la ilustración 2 se simplifica el proceso cíclico a seguir por el Web Crawler obtener cada proceso. En la ilustración 3 se presenta el flowchart del funcionamiento del Web Crawler

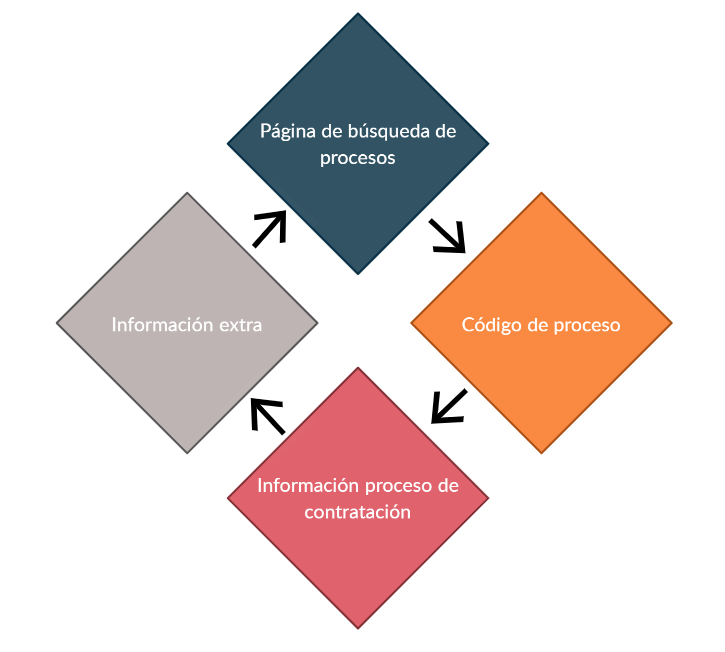


Ilustración 2: Proceso cíclico de obtención de procesos.

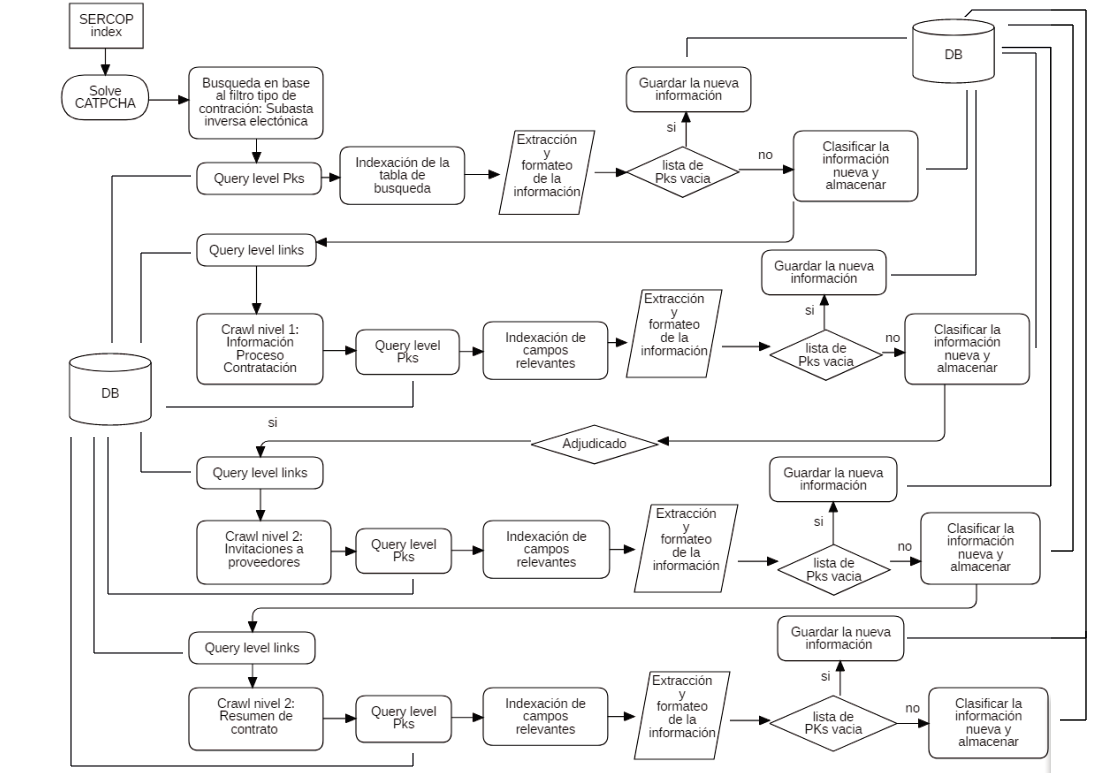


Ilustración 3: Flowchart del Web Crawler

Web Crawler

Se entiende como Web Crawler a un Bot de internet cuya funcionalidad principal es el escaneo automático de la información contenida en una web para posteriormente obtener los datos mediante técnicas de scraping y finalmente crear un índice de los datos obtenidos, generalmente mediante la implementación de bases de datos. (octoparse, 2019) En el presente proyecto se elaboró un Web Crawler específico para el portal de compras públicas, el desarrollo fue en base al lenguaje de programación Python v3.6 y el módulo de automatización web Selenium.

Para el funcionamiento del Web Crawler es necesario establecer los filtros correctos y solucionar el CAPTCHA inicial, ya que en el presente proyecto no se aborda la solución automatizada de este tipo de filtros de seguridad.

Dentro de las funcionalidades del Web Crawler implementado se encuentran: la navegación automatizada en las páginas web de la SERCOP, extracción y almacenamiento de información y documentos relevantes de cada proceso de contratación mediante subasta inversa electrónica.

Con el objetivo de optimizar el funcionamiento del Web Crawler implementado, se procedió a dividir el trabajo en 3 módulos principales cada uno encargado de una tarea específica. El primer módulo será el de navegación, el cual se encargará de realizar el crawling[[1]](#footnote-1) por los distintos niveles del portal web de la SERCOP y localizará la información de relevancia para ser extraída. El segundo módulo será el encargado de la extracción de información relevante previamente señalada por el primero modulo y mediante técnicas de scrapping[[2]](#footnote-2) se encargará de estructurar todos los datos recuperados, finalmente el tercer y último modulo será el encargado de almacenar toda la información recopilada por el segundo modulo en las bases de datos correspondientes para su futuro uso. En cuanto al desarrollo del software se tiene en cuenta el caso en que el formato de la página web de la cual se desea extraer la información cambie, es por eso que se trató de hacer lo más flexible posible al funcionamiento del wecpsie con el objetivo de poder parchar rápidamente a futuro en caso de ser necesario. En la ilustración 3 se detalla el funcionamiento de cada módulo implementado en el Web Crawler.

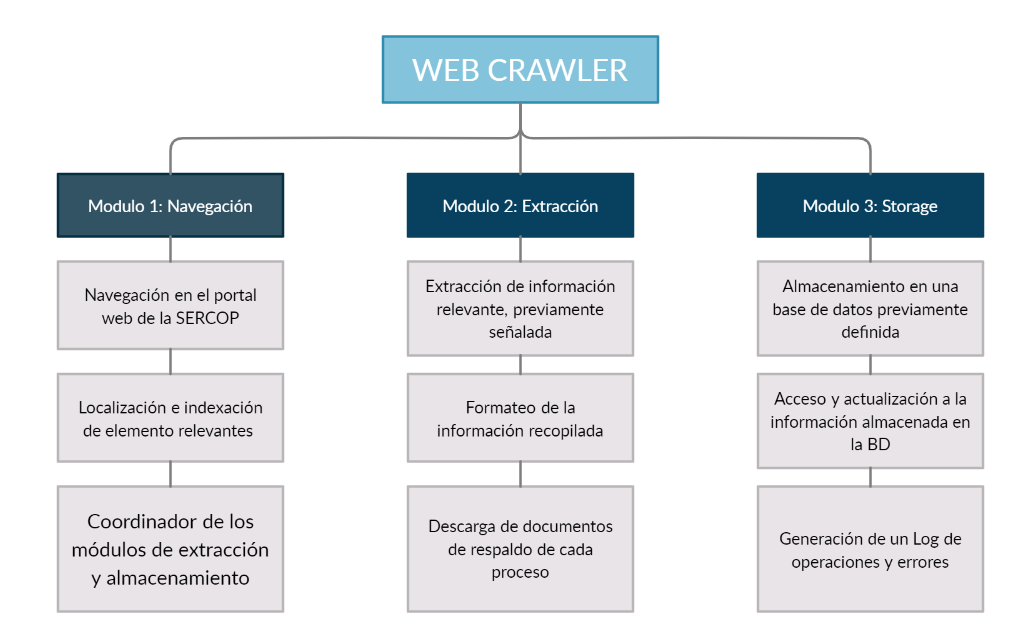


Ilustración 4: Módulos del Web Crawler

Módulo de navegación

Este módulo es el encargado de la apertura y navegación en las diferentes páginas web correspondientes a cada proceso de SIE, en cada ventana se encargará de la búsqueda de elementos claves los cuales a su vez serán dirigidos al módulo de extracción y posteriormente al módulo de almacenamiento. El presente módulo se encuentra compuesto de cuatro funciones cada una de ellas correspondiente al nivel de navegación en el que se encuentre el software.

|  |  |  |
| --- | --- | --- |
| Nivel | Método | Página |
| Index | swing\_index\_pages | Búsqueda de procesos SIE |
| 1 | crawl\_pcpsie\_page | Información Proceso Contratación |
| 2 | crawl\_pcpsie\_contrato\_page | Resumen Información Esencial del Contrato |
| 2 | crawl\_pcpsie\_invitacion\_page | Invitaciones a proveedores |

Tabla 1: Funciones del módulo de navegación

El software WECPSIE comienza con el nivel denominado Index en el cual se encuentra los parámetros de búsqueda para los diferentes tipos de procesos de contratación, en este caso el enfoque es en los procesos SIE. Una vez abierta la página web correspondiente, el software selecciona de manera automática el tipo *Subasta Inversa Electrónica* en la casilla *Tipo de contratación* y además establece el rango de fechas dentro del cual se obtendrán los procesos, cabe mencionar que el límite máximo a partir de una fecha final es de 6 meses atrás en el tiempo, es decir si se selecciona 1 de enero de 2021 como fecha final solo se lograran obtener los procesos hasta el 1 de julio de 2020. Una vez establecidos los parámetros se otorga al usuario un periodo de 12 segundos en los cuales se deberá completar correctamente el Captcha, caso contrario el programa finalizará.

Una vez completado de manera correcta el Captcha el software ejecutará el botón de búsqueda, dando como resultado el despliegue de una tabla de al menos 20 filas en la parte inferior de la pantalla, la cual contendrá la información principal de cada proceso, además, se localizará los índices de dicha tabla lo cual será útil para indicar al software cuando finalizar la primera etapa. Una vez localizada la tabla de procesos presentada, el módulo de navegación procede a enviar la misma al módulo de extracción; concretamente a la función get\_main\_page\_table\_info la cual será la encargada de extraer el valor de cada columna y devolver en una representación estructurada todos los datos de cada uno de los registros.

Una ves terminada la función del módulo de extracción de inmediato el módulo de navegación se encarga de pasar los datos devueltos al modulo de almacenamiento, previamente comprobando si en la base de datos ya existen registros de los procesos obtenidos o no. Con ello se determina el Query correcto a ejecutar; Update ó Insert.

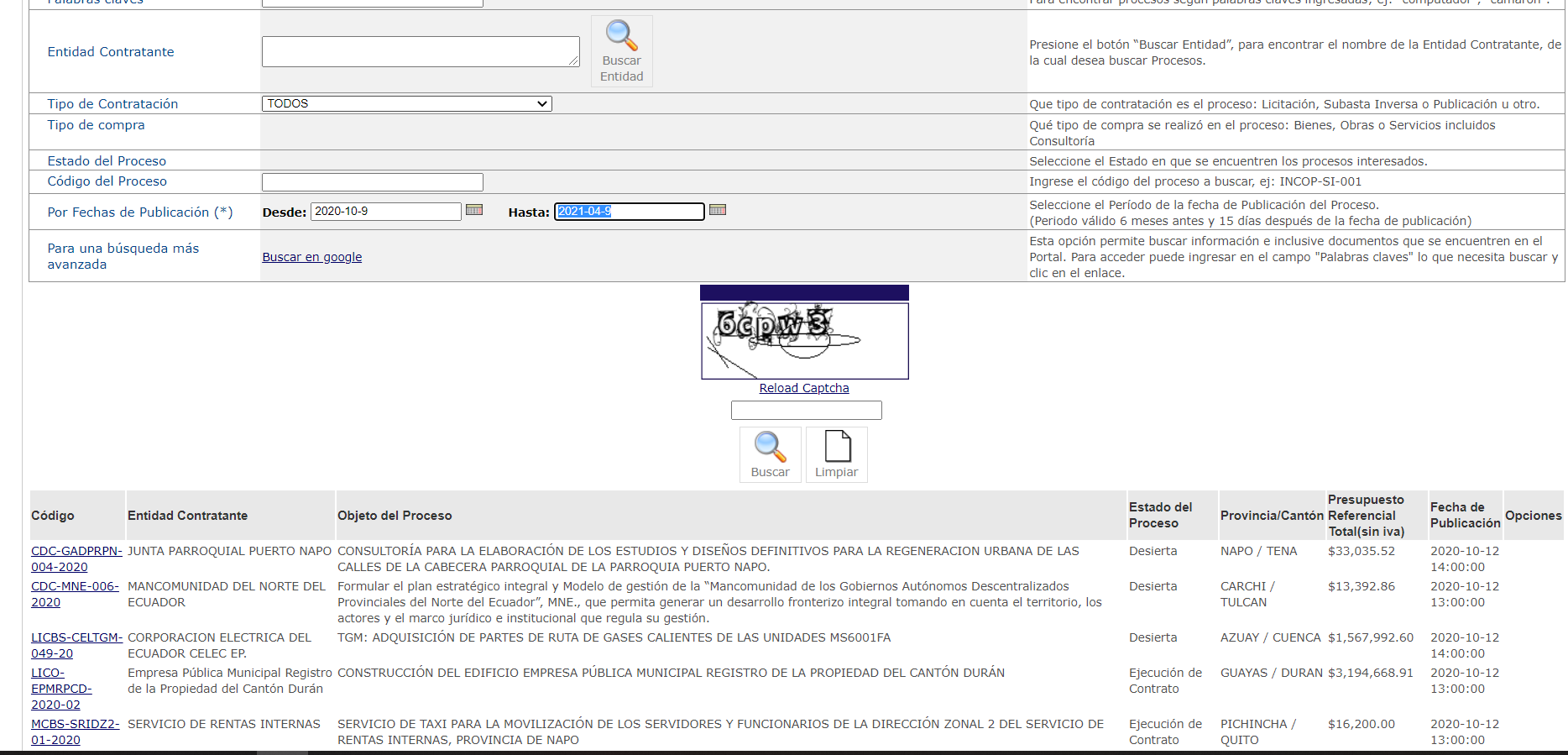


Imagen 12: Información a extraer en el nivel Index

Finalizada la extracción del nivel Index el módulo de navegación obtiene de la base de datos los enlaces correspondientes a la información de cada proceso y procede a abrir cada uno de ellos de manera iterativa. Esta ventana corresponde al nivel 1 de navegación, en la cual se presenta dentro de una tabla toda la información correspondiente al proceso. Dependiendo del estado en el que se encuentre el mismo, será factible encontrar un enlace a la información acerca del contrato de adjudicación y un enlace a los proveedores que fueron invitados a participar en el proceso. De igual manera una vez localizada la tabla correspondiente el crawler la envía al módulo de extracción más precisamente a la función *extract\_SIE\_lv1*, la misma será la encargada de extraer la información del proceso que fue considerada de mayor relevancia y al finalizar devolverá correctamente estructurada y organizada toda la información recopilada. Finalmente, el módulo de navegación se encargará de enviar al módulo de almacenamiento la información obtenida de cada proceso, previamente comprobando si en la base de datos ya existe los datos correspondientes al registro o no y en cada caso se ejecutará un Query de Update o Insert correspondientemente.

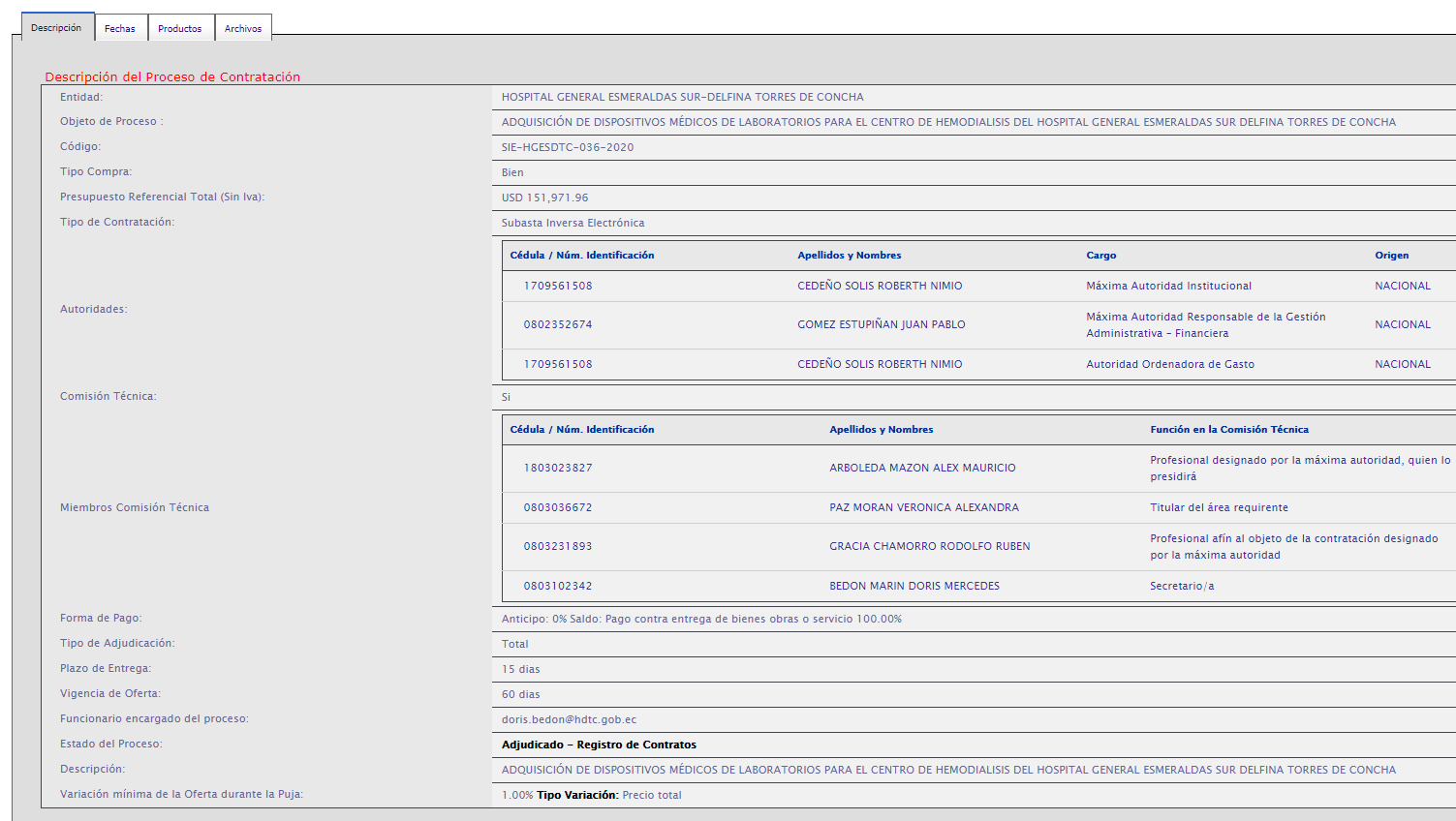


Imagen 13: Página correspondiente al nivel 1 de navegación

El nivel 2 de navegación corresponde a las paginas de información del contrato de adjudicación en caso de existir y la página de invitaciones a proveedores. El módulo de navegación será el encargado de consultar a la base de datos los enlaces correspondientes a cada página y en caso de existir los procederá a abrir en una ventana del navegador.

Dentro de la página Invitaciones a proveedores se presenta una tabla con la información de cada proveedor muy similar a la tabla del nivel Index en la cual el número de filas puede variar entre cero y miles. El crawler enviará a la función *extract\_SIE\_lv2* dentro del módulo de almacenamiento la tabla con los proveedores y al finalizar la extracción de los datos de cada fila la función devuelve correctamente estructurada la información obtenida. Posterior a ello el módulo de navegación comprueba en la base de datos si existen registros previos correspondientes al proceso y almacena la información, con la ayuda del módulo de almacenamiento, mediante un Query de Update o Insert según sea el caso.

Por otro lado, dependiendo del estado en el cual se encuentre el proceso de contratación puede existir un enlace a la página con la información acerca del contrato de adjudicación. Por lo cual el módulo de navegación consulta todos los enlaces disponibles en la base de datos y los procede a abrir. En la página de información del contrato se presentan varias tablas sin embargo en este caso se toma los datos contenidos dentro de la tabla información de adjudicación ya que se los considera muy relevantes. La función *extract\_SIE\_lv2\_inv* ubicada en el módulo de extracción es la encargada de obtener y devolver de manera estructurada todos los datos presentes dentro de la tabla.

De igual manera el módulo de navegación comprueba en la base de datos si existen registros previos correspondientes al proceso en turno o no y el módulo de almacenamiento procede a guardar la información mediante un Query de Update o Insert según sea el caso.

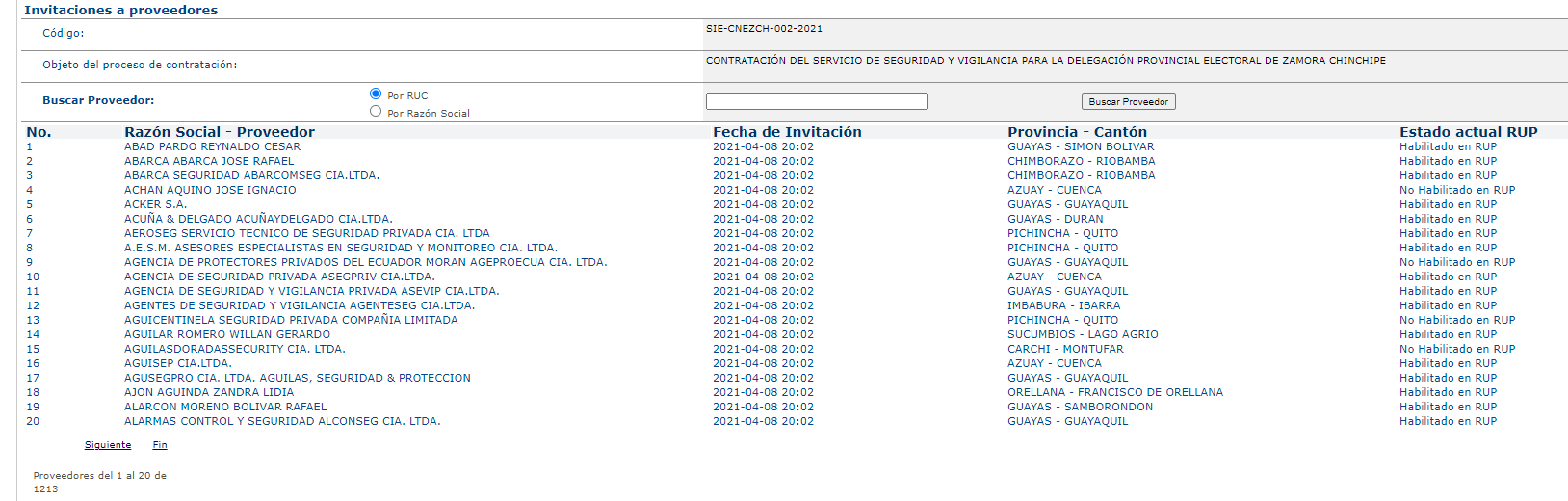


Imagen 14: Página de invitaciones a proveedores

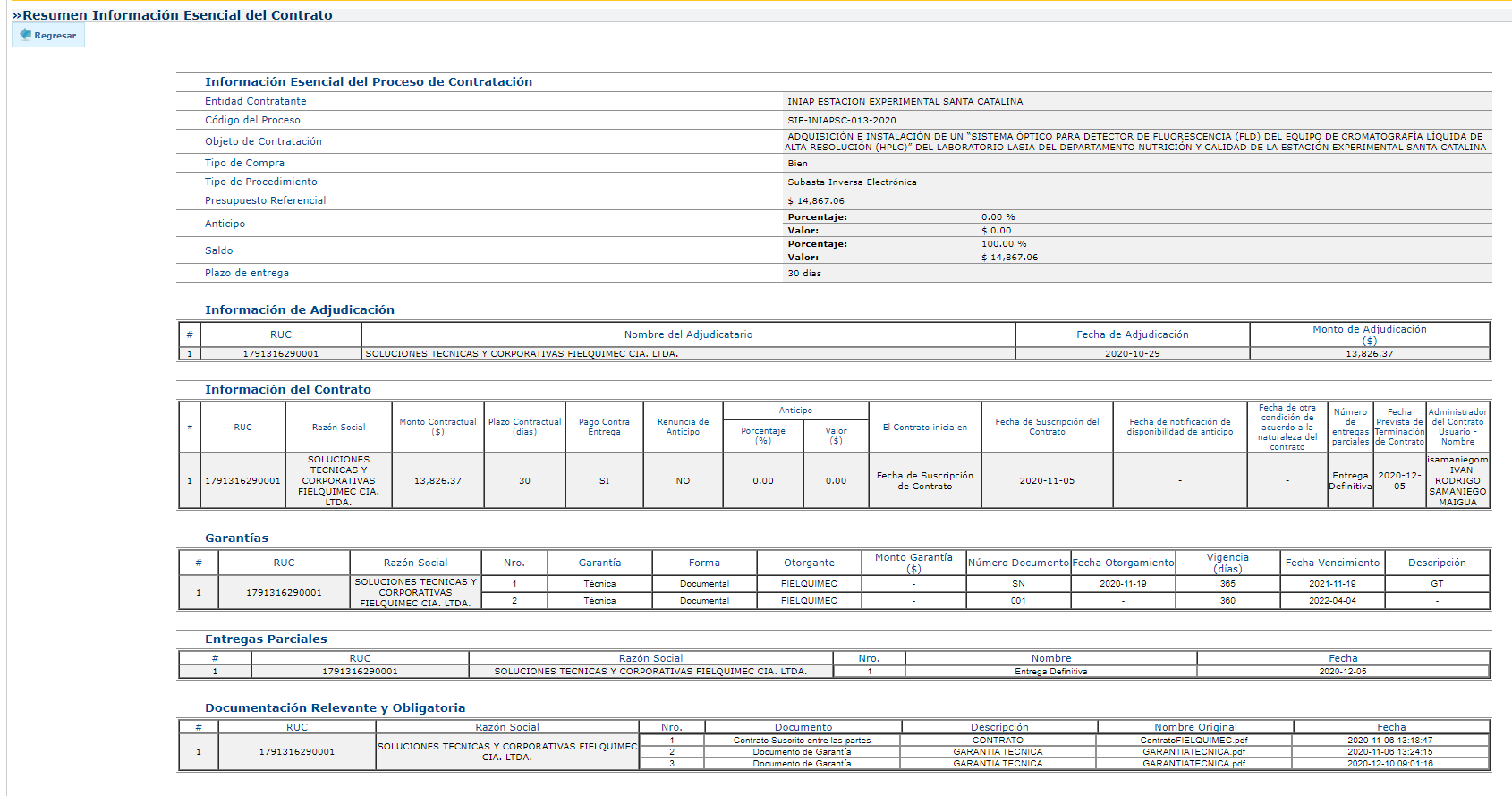


Imagen 15: Página de información esencial del contrato

Módulo de extracción

A partir de los elementos web entregados por parte del módulo de navegación; el módulo de extracción se encargará de obtener la información señalada como relevante. Dentro del módulo existen cuatro funciones las cuales serán invocadas en cada nivel en específico, al finalizar la extracción se devolverá de manera estructurada los datos recopilados mediante diccionarios para que la manipulación de los datos en el módulo de almacenamiento sea mucho más sencilla.

|  |  |  |  |
| --- | --- | --- | --- |
| Nivel | Método | Objeto Web | Información Extraída |
| Index | Get\_main\_page\_table\_info | Tabla de busqueda | Texto y enlaces |
| 1 | Extract\_SIE\_lv1 | Tabla descripción del proceso y tabla de fase contractual | Texto y enlaces |
| 2 | Extract\_SIE\_lv2 | Tabla información de contrato de adjudicación | Texto |
| 2 | Extract\_SIE\_lv2\_inv | Tabla de invitaciones a proveedores | Texto |

Tabla 2: Funciones en el módulo de extracción

Dependiendo el objeto web del cual se necesitó extraer la información se utilizaron varios métodos de localización y extracción, en caso de ser una tabla se itero por todas sus filas y columnas extrayendo toda la información necesaria. Por otro lado, la mayoría de elementos web en el portal de compras públicas no cuentan con un ID único por lo cual se dificultaba la extracción de los datos en ciertas ocasiones, es por ello que se utilizó los Xpath de los elementos directamente para proceder a la extracción del campo.

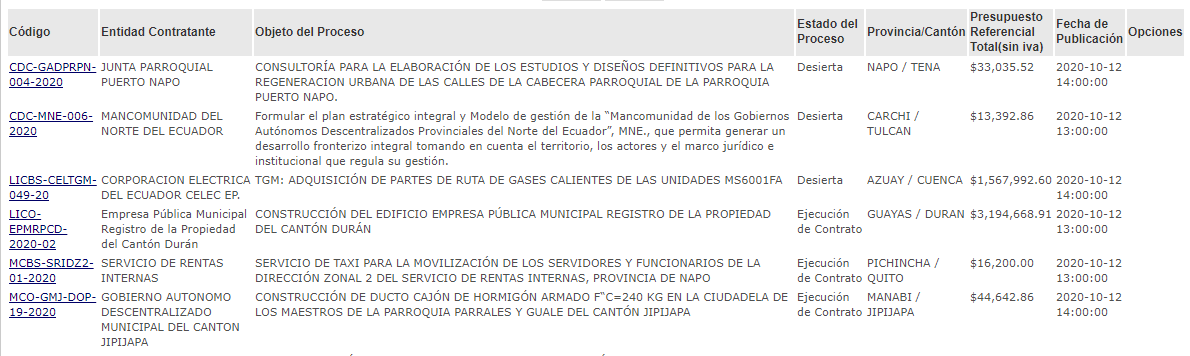


Imagen 16: Información extraída del nivel Index



Imagen 17: Información extraída del nivel, descripción del proceso



Imagen 18: Información extraída del nivel 1, fechas de control del proceso



Imagen 19: Información extraída del nivel 2, Invitaciones a proveedores.

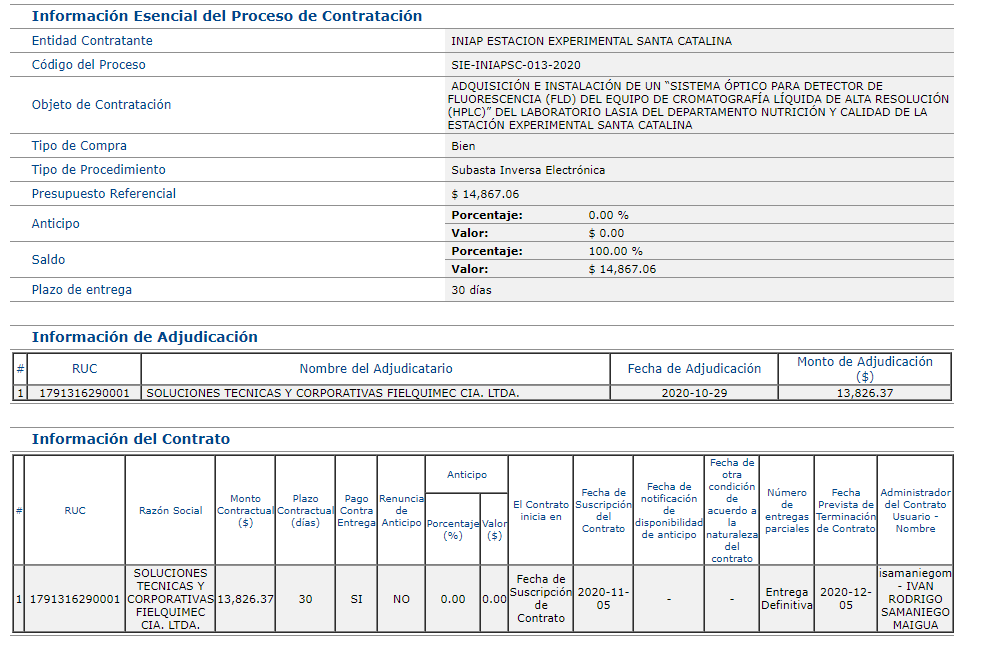


Imagen 20: Información extraída del nivel 2, información esencial del contrato

Módulo de almacenamiento

Se implemento una base de datos relacional mediante el gestor MySQL 8.0.17; la cual es operada por las distintas funciones contenidas en el módulo de almacenamiento. Una vez recibida la información estructurada por parte de la función de extracción correspondiente, el módulo de navegación es el encargado de invocar el método necesario para poder ingresar dentro de la tabla correspondiente de la base de datos la información extraída. En total son once funciones pertenecientes al módulo de almacenamiento, entre las cuales se encuentra métodos de consulta, almacenamiento, conexión y desconexión. De suma importancia es la implementación de una relación específicamente para ámbito de almacenamiento de logs generados durante la ejecución del software La imagen 21 indica el esquema de la base de datos propuesta.

|  |  |
| --- | --- |
| Método | Tipo |
| Connect\_to\_wpsie\_db | Conexión |
| Dissconect\_db | Desconexión |
| Get\_triad\_pks | Consulta |
| Get\_procesos\_page\_link | Consulta |
| Get\_contrato\_links | Consulta |
| Get\_invitacion\_links | Consulta |
| Save\_SIE\_info | Almacenamiento |
| Save\_contrato\_SIE\_info | Almacenamiento |
| Save\_invitacion\_SIE\_info | Almacenamiento |
| Add\_entry\_log | Almacenamiento |

Tabla 3: Funciones del módulo almacenamiento

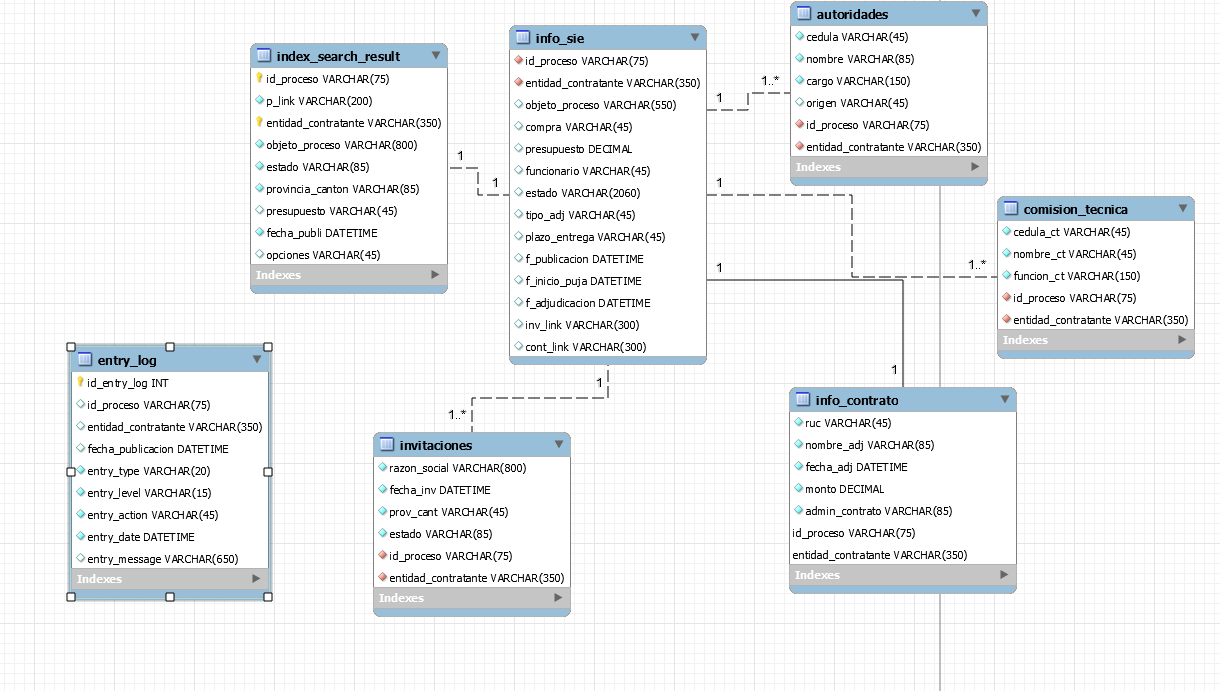


Imagen 21: Esquema de la base de datos

**mejoras propuestas**

Plantear un sistema de configuración para que el crawler extraía solo ciertos parámetros y los guarde en la BD (continúo trabajando en ello, documentación próximamente)

**mejoras implementadas**

facilidad de adaptación en caso de cambio de formato del portal de la SERCOP (continúo trabajando en ello, documentación próximamente)

# Conclusiones

Presenta los aportes de este trabajo con base en lo investigado, es importante que como autor puedas analizar el tema y su relevancia para la profesión dentro del contexto nacional e internacional (presenta similitudes, diferencias entre los diferentes enfoques del tema investigado). En el caso de presentaciones artísticas o creativas se debe describir de qué se tratan y justificar sus elementos, obligatoriamente incluir anexos con fotos, evidencias (partituras, enlaces a videos, etc.) del producto elaborado. Realiza un análisis de lo que has aprendido en este trabajo, incluye sugerencias de estudios posibles que se realicen en el futuro para comprender de mejor manera el tema, menciona alguna dificultad que hayas tenido para realizar este trabajo y sus razones.

# Referencias bibliográficas

# Referencias

FERNÁNDEZ, Y. (30 de Octubre de 2018). Captcha y reCaptcha: qué son, cuáles son sus diferencias y qué cambia con el reCaptcha v3.

Ionos. (10 de Septiembre de 2020). ¿Qué es el web scraping?

Neoattack. (s/f). ¿Qué es el Crawling?

octoparse. (30 de Diciembre de 2019). ¿Qué Es Web Scraping (Web Crawler) y Cómo Funciona?

SERCOP. (2016). Ecuador.

SERCOP. (21 de Agosto de 2018). LEY ORGANICA DEL SISTEMA NACIONAL DE. Ecuador.

# 

# Anexo A: Título

(Si hubiere anexos. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes)

# Anexo B: Título

(Si hubiere anexos. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes)

# Anexo C: Título

(Si hubiere anexos. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes)

Se recomienda iniciar cada anexo en una nueva hoja. Se puede incluir anexos adicionales (ANEXO D: TÍTULO, ANEXO E: TÍTULO, ANEXO F: TÍTULO, etc.) de conforme la necesidad de presentación de los mismos en el trabajo. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes.

1. Hace referencia a cualquier Bot de indexación enviado por los motores de búsqueda, con el fin de detectar, leer y analizar todo el contenido y el código que compone a una página web realizando los correspondientes saltos a través de los enlaces de cada web. (Neoattack, s/f) [↑](#footnote-ref-1)
2. El web scraping (del inglés scraping = arañar/raspar) se extraen y almacenan datos de páginas web para analizarlos o utilizarlos en otra parte. Por medio de este raspado web se almacenan diversos tipos de información. (Ionos, 2020) [↑](#footnote-ref-2)