

**Universidad Tecnológica de la Habana**

**“José Antonio Echeverría”**

**Facultad de Ingeniería Informática**

**Módulo de búsqueda exploratoria basado en la generación automática de grafos de conocimiento a partir de documentos**

**Informe de la Práctica Profesional 2**

**Autor:** Gabriela Hernández Felipe

**Tutor**: MsC. Wenny Hojas Mazo

**La Habana,**

**mayo 2024**

Resumen

En las organizaciones contemporáneas, los intercambios internos juegan un rol crucial en la solución de problemas y la toma de decisiones. La documentación de estos intercambios mediante informes detallados no solo actúa como un registro oficial, sino que también mejora la comunicación interna y promueve la transparencia. Estos informes documentan las discusiones, decisiones y acuerdos logrados, proporcionando una base para acciones futuras. Sin embargo, debido a la complejidad y volumen de datos específicos se convierte en un proceso laborioso y propenso a errores, especialmente al buscar información relevante que permita evaluar el comportamiento de indicadores de interés definidos. La presente investigación se centra en la automatización de la búsqueda de información en estas organizaciones, utilizando la búsqueda exploratoria y la visualización de datos mediante grafos de conocimiento. El objetivo es desarrollar un módulo que permita realizar análisis a partir de documentos de forma automatizada, facilitando así la comprensión de la información requerida. Se implementa un sistema que automatiza el proceso de búsqueda y se pueda integrar a un sistema de visualización que mejore la representación de la información. Este trabajo contribuye al ámbito de la gestión del conocimiento en organizaciones. Los resultados obtenidos sí permiten obtener una compresión de datos relevantes del documento, a pesar de que para el idioma español existen ocasiones que se afecta la precisión.

Abstract

In contemporary organizations, internal exchanges play a crucial role in solving problems and decision making. The documentation of these exchanges through detailed reports not only acts as an official record, but also improves internal communication and promotes transparency. These reports document discussions, decisions and agreements achieved, providing a basis for future actions. However, due to the complexity and volume of specific data becomes a laborious and prone to errors, especially when seeking relevant information that allows assessing the behavior of defined interest indicators. The present investigation focuses on the automation of the search for information in these organizations, using the exploratory search and the visualization of data through knowledge graphs. The objective is to develop a module that allows analysis from documents automated, thus facilitating the understanding of the required information. A system is implemented that automates the search process and can be integrated into a visualization system that improves the representation of the information. This work contributes to the field of knowledge management in organizations. The results obtained do allow you to obtain a relevant data compression of the document, even though for the Spanish language there are occasions that precision affects.

**ANEXO A. PLAN INDIVIDUAL DE TAREAS DEL ESTUDIANTE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tareas** | | **PP/T** | **Fecha de entrega** | **Rol(es) que desarrolla(n) con la tarea** |
| Reunión de inicio de la práctica | | PP2 | 29/4/2024 | - |
| Revisar Introducción (con diseño de la investigación) y entregar al tutor | | PP2 | 10/5/2024 | EE |
| Elaborar Capítulo 1 | | PP2 | 17/5/2024 | EE |
| Entregar al tutor Capítulo 1 para revisión | | PP2 | 17/5/2024 | EE |
| Rectificar señalamientos del tutor sobre Introducción y Capítulo 1 | | PP2 | 31/5/2024 | EE |
| Entregar al tutor Introducción y Capítulo 1 rectificado | | PP2 | 31/5/2024 | EE |
| Diseñar un módulo de búsqueda exploratoria basado en la generación automática de grafos de conocimiento a partir de documentos. | | PP2 | 31/5/2024 | AS, DS, AR, AD |
| Implementar módulo de búsqueda exploratoria diseñado. | | PP2 | 14/6/2024 | PG |
| Diseño de las pruebas / experimentos. | | PP2 | 21/6/2024 | PB |
| Ejecución de las pruebas / experimentos | | PP2 | 28/6/2024 | PB |
| Elaborar capítulo con resultados de las pruebas y experimentos | | PP2 | 05/7/2023 | EE, PB |
| Elaborar informe de la práctica profesional | | PP2 | 10/7/2024 | EE |
| Entregar informe de PP al tutor | | PP2 | 10/7/2024 | EE |
| Rectificar señalamientos del informe de PP | | PP2 | 21/7/2024 | EE |
| Entrega del informe final de PP | | PP2 | 02/9/2024 | EE |
| Defensa de la práctica profesional | | PP2 | 4-6/9/2024 | Todos |
| Gestionar la configuración (versiones del código y entregables del proyecto) | | PP2 y T | 01/2025 | GC |
| Reunión de inicio de la tesis | | T | 09/9/2024 | - |
| Diseñar un sistema de búsqueda exploratoria basado en grafos de conocimientos generados automáticamente a partir de los documentos que sirvan de guía para la búsqueda. | | T | 16/9/2024 | AS, DS, AR, AD |
| Implementar en el sistema diseñado. | | T | 04/10/2024 | PG |
| Diseño de las pruebas / experimentos | | T | 11/10/2024 | PB |
| Ejecución de las pruebas / experimentos | | T | 25/10/2024 | PB |
| Elaborar sección con resultados de las pruebas y experimentos | | T | 01/11/2024 | EE, PB |
| Elaborar documento de tesis | | T | 08/11/2024 | EE |
| Entregar informe de tesis al tutor | | T | 08/11/2024 | EE |
| Rectificar señalamientos del tutor sobre la tesis | | T | 22/11/2024 | EE |
| Entregar documento de tesis | | T | 01/2025 | EE |
| Predefensa de tesis | | T | 01/2025 | Todos |
| Defensa de tesis | | T | 01/2025 | Todos |
| Wenny Hojas Mazo  Nombre completo y firma del primer tutor |  | | | |
| Gabriela Hernández Felipe  Nombre completo del estudiante | |  | | --- | | Firma del estudiante | | | | |
| En la columna Rol se deben poner la sigla del rol que contribuye a formar cada tarea. Las siglas de los roles son:  **AN**- Analista de negocio **AR**- Arquitecto **JP**-Jefe de proyecto  **AS**- Analista de sistema **DS**- Diseñador de software **ES**- Especialista de seguridad  **AD**- Analista de datos **DI**- Diseñador de Interfaz hombre – máquina **EE**- Escritor-expositor de trabajos técnicos  **PG**- Programador **DB**- Diseñador de base de datos **II**- Habilitador de Infraestructuras Informáticas  **PB**- Probador **GC**- Gestor de Configuración **TD**- Facilitador de la Toma de Decisiones | | | | |

Índice

[Introducción 9](#_Toc176765261)

[Capítulo 1: Marco teórico 14](#_Toc176765262)

[1.1 Documentación de intercambios internos en instituciones 14](#_Toc176765263)

[1.1.1 Tipos de informes y su función en las organizaciones 15](#_Toc176765264)

[1.1.2 Desafíos de la documentación de intercambios internos 18](#_Toc176765265)

[1.2 Evaluación de indicadores mediante búsqueda de información 19](#_Toc176765266)

[1.2.1 Métodos tradicionales de evaluación de indicadores 20](#_Toc176765267)

[1.2.2 Problemas y limitaciones de la búsqueda manual de información 21](#_Toc176765268)

[1.3 Automatización de la búsqueda de información 22](#_Toc176765269)

[1.3.1 Soluciones propuestas en la literatura 22](#_Toc176765270)

[1.3.2 Búsqueda exploratoria 24](#_Toc176765271)

[1.4 Grafos de conocimiento 26](#_Toc176765272)

[1.4.1 Conceptos y aplicaciones 27](#_Toc176765273)

[1.4.2 Herramientas y tecnologías existentes 30](#_Toc176765274)

[1.5 Conclusiones parciales 31](#_Toc176765275)

[Capítulo 2: Modulo de búsqueda exploratoria basado en generación automática de grafos de conocimiento 32](#_Toc176765276)

[2.1 Requisitos funcionales 32](#_Toc176765277)

[2.2 Requisitos no funcionales 33](#_Toc176765278)

[2.3 Diagrama de casos de uso del sistema 34](#_Toc176765279)

[2.4 Diagramas de secuencia 35](#_Toc176765280)

[2.4.1 Diagrama de secuencia de Subir documentos 35](#_Toc176765281)

[2.4.2 Diagrama de secuencia de Buscar documentos 36](#_Toc176765282)

[2.5 Implementación del módulo de búsqueda 37](#_Toc176765283)

[2.5.1 Servicio de Procesar Documentos 38](#_Toc176765284)

[2.5.2 Servicio de generación de grafo de conocimiento 41](#_Toc176765285)

[2.5.3 Servicio de Recuperación de Información 46](#_Toc176765286)

[2.6 Conclusiones parciales 47](#_Toc176765287)

[Capítulo 3: Evaluación del módulo de búsqueda 47](#_Toc176765288)

[3.1 Pruebas funcionales 48](#_Toc176765289)

[3.1.1 Extracción de texto 48](#_Toc176765290)

[3.1.2 Generación de grafo 49](#_Toc176765291)

[3.1.3 Almacenar en Solr 57](#_Toc176765292)

[3.2 Prueba con aplicación de ejemplo 59](#_Toc176765293)

[3.3 Conclusiones parciales 61](#_Toc176765294)

[Conclusiones 62](#_Toc176765295)

[Recomendaciones 62](#_Toc176765296)

[Referencias 63](#_Toc176765297)

# Introducción

En las organizaciones contemporáneas, los intercambios internos juegan un rol crucial en la solución de problemas y la toma de decisiones. Estos intercambios, a menudo caracterizados por discusiones cerradas y sensibles, no siempre son accesibles al público. El contenido de estas interacciones varía desde la planificación estratégica hasta la resolución de conflictos, y su gestión eficiente es vital para el mejor desempeño de la entidad.

La documentación de estos intercambios mediante informes detallados no solo actúa como un registro oficial, sino que también mejora la comunicación interna y promueve la transparencia. Estos informes documentan las discusiones, decisiones y acuerdos logrados, proporcionando una base para acciones futuras. Sin embargo, debido a la complejidad y volumen de datos específicos se convierte en un proceso laborioso y propenso a errores, especialmente al buscar información relevante que permita evaluar el comportamiento de indicadores de interés definidos. Esto dificulta la identificación de tendencias y la toma de decisiones basadas en datos precisos.

La generación de informes institucionales es una práctica esencial que recopila información valiosa de diversos intercambios. Estos informes no solo sirven como un registro de actividades y decisiones, sino que también actúan como una herramienta de referencia para futuras acciones y estrategias. La precisión y exhaustividad de estos informes son fundamentales para garantizar una comprensión clara de los eventos y decisiones que tienen lugar dentro de la organización. Sin embargo, el manejo eficiente de estos documentos representa un desafío significativo debido a la cantidad de información que contienen y la necesidad de extraer datos específicos de manera rápida y precisa.

La identificación de información clave en estos informes para la evaluación de indicadores de interés es una tarea compleja. Requiere una comprensión profunda del contenido textual y la capacidad de reconocer patrones y datos relevantes. Los problemas asociados con la búsqueda manual en los informes incluyen la posibilidad de pasar por alto información crucial, la interpretación incorrecta de datos y la inversión considerable de tiempo y recursos.

Para evaluar los indicadores de interés de manera efectiva, es crucial confeccionar resúmenes que describan la información encontrada en los informes. Estos resúmenes deben ser precisos y reflejar una comprensión clara de los datos y su relevancia para los indicadores definidos. Sin embargo, la elaboración de estos resúmenes puede ser igualmente desafiante debido a la necesidad de sintetizar grandes volúmenes de información de manera coherente y comprensible. La complejidad del contenido y la posibilidad de errores humanos en la interpretación y síntesis de datos representan obstáculos significativos en este proceso. Estos desafíos subrayan la necesidad de soluciones más eficientes y precisas para la extracción de información.

En respuesta a estos desafíos, se han propuesto diversas soluciones para automatizar la búsqueda en documentos y el seguimiento de la información [18], [19], [1], [20]. Estas soluciones incluyen el uso de tecnologías avanzadas como la búsqueda exploratoria, que permite a los usuarios descubrir información relevante a través de un proceso iterativo y dinámico [1]. La búsqueda exploratoria se diferencia de las técnicas de búsqueda tradicionales al permitir una mayor flexibilidad y adaptabilidad en la búsqueda de información [2], [3], lo que la hace particularmente útil en contextos complejos y dinámicos como los de las instituciones contemporáneas.

Wolfram|Alpha es un motor de cálculo y conocimiento que se utiliza para responder preguntas, realizar análisis y generar reportes automáticamente, se centra en un contexto de resolver problemas matemáticos y no es gratuita [18]. Semantic Scholar extrae significado e identifica conexiones dentro de los artículos, se basa en un motor de búsqueda de artículos académicos, mostrando un poco más de detalles de la información contenida dentro de los documentos, que a pesar de eso no te hace comprender totalmente el contenido de los artículos [19]. Una tesis de doctorado propone un sistema que analiza documentos devueltos por una consulta de Bing y construye un grafo de conocimiento, esta propuesta utiliza tecnologías online para el análisis de los datos y no presenta soporte para documentos redactados manualmente por instituciones [4]. Una tesis de grado [20] propone un sistema de apoyo a la toma de decisiones, que extrae noticias de diferentes medios, haciendo uso de la búsqueda exploratoria, analiza las noticias y muestra un gráfico de barras según los temas más abordados y una nube de términos que refleja los temas más comunes en los artículos devueltos por una consulta.

A pesar de esto la interpretación de la información es mucho más compleja, ya que es importante conocer en cuales de las noticias o artículos es que se encuentran estos términos relevantes y cuantas veces este es utilizado, o conocer cuál es la relación que existen entre ellos, poder relacionar varias noticias o artículos a partir de tener términos en común, los detalles acerca de información específica. Para lograr estas representaciones con un nivel de detalle y de comprensión más alto, se utilizan los grafos de conocimiento, que representan la información de forma estructurada, en forma de red, que conecta los nodos, que representan los términos relevantes y las aristas, las conexiones entre ellos, brindando conocimiento sobre un área determinada o dominio [5].

Sin embargo, estas soluciones no están libres de limitaciones. En el contexto de instituciones que manejan información con divulgación restringida, la implementación de soluciones automatizadas debe considerar la seguridad y privacidad de los datos. Además, la efectividad de estas soluciones depende en gran medida de la calidad de los algoritmos utilizados y la capacidad de las herramientas para interpretar correctamente el contenido textual de los documentos. Las limitaciones tecnológicas y de infraestructura también pueden representar barreras significativas para la adopción de estas soluciones.

A partir de lo enunciado anteriormente se identifica como problema de investigación:

¿Cómo facilitar a las instituciones contemporáneas la comprensión de la información obtenida mediante el uso de sistemas desarrollados para minimizar el trabajo, teniendo en cuenta el contexto de privacidad de información y tipo de documentos que se generan a partir de los intercambios?

Como objeto de estudio de la investigación: la utilización de la búsqueda exploratoria. El campo de acción: extracción de información como tarea de la búsqueda exploratoria y visualización de esta mediante grafos de conocimiento.

Para darle solución a lo planteado anteriormente se define como objetivo general:

Desarrollar un módulo de sistema que sea capaz de, mediante búsqueda exploratoria basada en la generación automática de grafos de conocimiento, realizar análisis a partir de documentos, que sirvan de guía para la búsqueda y la comprensión de la información requerida.

Para dar cumplimiento a este objetivo general se han establecido los siguientes objetivos específicos con sus correspondientes tareas de investigación:

* Caracterizar la búsqueda de información en organizaciones contemporáneas
  + Describir los procesos de documentación de intercambios y búsqueda de información en las instituciones.
  + Caracterizar la búsqueda exploratoria como forma de automatización de estos procesos.
  + Describir conceptos fundamentales de grafos de conocimiento y sus aplicaciones
  + Caracterizar herramientas existentes para el manejo de este tipo de visualización.
* Desarrollar el sistema de búsqueda exploratoria basado en grafos de conocimientos generado automáticamente.
  + Describir solución propuesta del módulo de búsqueda exploratoria.
  + Implementar el módulo de búsqueda
  + Integrar módulo con sistema de visualización basado en grafos de conocimiento.
* Evaluar el sistema de búsqueda y visualización
  + Diseñar casos de prueba funcionales para cada uno de los servicios
  + Realizar pruebas funcionales a cada uno de los servicios, comprobando errores o datos incorrectos.
  + Mostrar ejemplo de comunicación con aplicación web de ejemplo

El artefacto de salida de esta investigación es un diseño de software. El valor práctico de este trabajo es proporcionar a las organizaciones una herramienta que le facilite el trabajo a la hora de analizar los informes generados en los intercambios internos sin necesidad de tener que leer cada uno de estos documentos para obtener la información que necesitan y comprender el contenido de estos.

El trabajo está estructurado en: introducción, 3 capítulos, conclusiones, recomendaciones y referencias bibliográficas. En el capítulo 1 se tratará todo lo referente a los conceptos básicos más importantes y necesarios para lograr comprender la solución propuesta. Luego en el capítulo 2 se explicará la implementación de esta solución y el uso que se le da a los recursos disponibles en las tecnologías utilizadas. En el capítulo 3 se estará demostrando con pruebas funcionales y una aplicación web de ejemplo que la solución satisface el problema descrito.

# Capítulo 1: Marco teórico

En este capítulo se explica la importancia que tiene para las organizaciones la documentación de intercambios que tengan lugar dentro de esta. Se describe los tipos de documentos que normalmente se generan en las instituciones y los desafíos que la generación de estos conlleva. Además, se caracteriza el proceso de búsqueda de información dentro de los documentos de la organización, en búsqueda de evaluar los indicadores. Este proceso consta de las formas tradicionales de búsqueda de información, que también presenta algunas dificultades, las cuales también son mostradas aquí. En el capítulo también se ejemplifican algunas soluciones que se han propuesto para resolver las limitaciones dadas, y se hace énfasis en la búsqueda exploratoria y los grafos de conocimiento.

## Documentación de intercambios internos en instituciones

La comunicación organizacional constituye un eslabón fundamental en la cadena estratégica de cada empresa. Se compone de modelos y procesos más complejos, más allá del intercambio de información entre jefes y subordinados. Se le atribuye la condición de ser mediadora de procesos inherentes a cualquier organización [6]. Dentro de esta comunicación organizacional toma gran importancia la comunicación interna. Esta se define como los procesos comunicativos que se realizan internamente en la organización, en la que toman parte como actores los diferentes integrantes de su público interno. Mediante esos procesos se coordinan las acciones fundamentales que le dan sentido a la organización, se ejecutan tareas, se realizan cambios, y se construyen valores de la cultura organizacional [7].

La documentación de intercambios internos es una de las maneras de lograr esta comunicación. En las instituciones esta documentación juega un papel fundamental en la gestión del conocimiento, la eficiencia operativa, gestión de proyectos y la toma de decisiones [8], [9]. Este conocimiento se registra en diversos tipos de medios de comunicación, como documentos electrónicos y documentos en papel. Los informes generados durante estos intercambios, como las reuniones, proporcionan un registro detallado de las actividades, discusiones y decisiones que tienen lugar dentro de la organización. Esta información valiosa necesita ser capturada, organizada y analizada para obtener su valor. Los informes pueden variar, dependiendo de la naturaleza de la institución y los objetivos específicos de los intercambios internos [9].

### Tipos de informes y su función en las organizaciones

La mayoría de los eventos y procesos de una empresa se acompañan o formalizan mediante algún tipo de documentación y el detallado enfoque de clasificación de los documentos refleja esta situación. Algunos de los tipos más comunes de documentos internos son memorandos, actas, circulares, mensajes electrónicos, memorandos y actas de reuniones, cronogramas de trabajo, informes de progreso, informes de cifras y proyecciones de ventas, informes de actividad, de síntesis, de proyecto, de investigación, de peritaje, comercial y técnico, y documentos administrativos:[9], [10]

**Memorando:** Es un documento interno utilizado para comunicar información importante o instrucciones dentro de una organización. Se utiliza para transmitir mensajes formales o semi-formales a empleados o departamentos específicos.

**Acta:** Es un documento que registra oficialmente las deliberaciones, decisiones y acuerdos tomados en una reunión o asamblea. Se utiliza para documentar los temas discutidos, las decisiones tomadas y las responsabilidades asignadas durante el encuentro.

**Circular:** Es un documento que se distribuye a un grupo de personas dentro de una organización para proporcionar información, instrucciones o anuncios importantes. Se utiliza para comunicar de manera eficiente y uniforme información a un gran número de destinatarios.

**Memorandos de reuniones:** Los memorandos de reuniones son documentos utilizados para informar a los participantes sobre los detalles de una reunión. Su función es proporcionar información relevante, como la fecha, hora, lugar y agenda de la reunión, así como cualquier otro detalle importante que los participantes necesiten conocer. También pueden incluir la lista de asistentes, los temas a discutir y cualquier preparación necesaria antes de la reunión.

**Cronogramas de trabajo:** Los cronogramas de trabajo son documentos que contienen una planificación detallada de las tareas y actividades que deben realizarse en un proyecto o proceso específico. Su función es proporcionar una estructura de tiempo clara y organizada para que los miembros del equipo o los empleados sepan qué actividades deben realizarse, cuándo deben realizarse y quién es responsable de cada tarea.

**Documentos administrativos:** Son aquellos utilizados en la gestión y organización interna de una empresa. Su función es proporcionar información y mantener registros relacionados con actividades administrativas, como contratos, políticas internas, informes de gastos, solicitudes de permisos, registros de personal, entre otros. Estos documentos ayudan a asegurar que las operaciones cotidianas de la empresa se realicen de manera eficiente y en conformidad con las políticas establecidas. En ellos se dan jerarquías acuerdos, ordenes, obligaciones, defensas y solicitudes de los subordinados a los superiores. Son reglamentados, con formas fijas, jurídicas y convencionales, tienen una superestructura o esquema preestablecido. Siempre producen efectos frente a terceros o en la propia organización administrativa.

**Actas de reuniones:** Documentan la información clave discutida en las reuniones registrando las decisiones tomadas, puntos clave, tareas asignadas y acciones a seguir acordadas en las reuniones formales o informales. Su función principal es la comunicación, el seguimiento y la gestión de proyectos. Estos tipos de informes son de vital importancia para mantener un registro histórico de las decisiones y proporcionar información relevante para futuras referencias, además promueven la transparencia dentro de la organización.

**Informe de actividad:** Documento que resume las actividades realizadas en un período de tiempo determinado. Puede incluir detalles sobre tareas completadas, objetivos alcanzados y métricas de rendimiento.

**Informe de síntesis:** Documento que resume la información clave de un tema o proyecto de manera concisa. Suele incluir los aspectos más relevantes y las conclusiones principales.

**Informe comercial:** Documento que presenta información sobre el rendimiento financiero, las estrategias de marketing y las ventas de una empresa. Puede incluir análisis de mercado, competencia y recomendaciones para mejorar el rendimiento comercial.

**Informe técnico:** Documento que proporciona información técnica detallada sobre un producto, proceso o proyecto. Puede incluir especificaciones, diagramas, análisis de riesgos y recomendaciones técnicas.

**Informe de peritaje:** Documento elaborado por un experto en un campo específico, que proporciona una opinión o evaluación técnica sobre un tema o situación particular.

**Informes de Progreso:** Son documentos que se utilizan para informar sobre el avance y los resultados alcanzados en un proyecto o actividad específica. Muestran el avance de estos proyectos, tareas o iniciativas, incluyendo métricas de rendimiento, desafíos y próximos pasos. Su función es proporcionar una evaluación detallada de los logros, los desafíos y las próximas etapas del proyecto. Estos informes pueden incluir datos cuantitativos, análisis cualitativos y recomendaciones para mejorar el progreso en el futuro. Estos informes son esenciales para evaluar el rendimiento y la eficacia de los proyectos en curso. Presentan el avance y resultados de proyectos específicos, incluyendo objetivos, metodología, análisis de riesgos y conclusiones. Sirven para la evaluación, la rendición de cuentas y la toma de decisiones sobre la continuación de proyectos.

**Informes de Investigación:** Presentan los resultados de estudios realizados para evaluar la viabilidad de nuevos productos, procesos o estrategias, incluyendo metodología, análisis de datos y conclusiones. Son esenciales para documentar los resultados de investigaciones realizadas dentro de la organización. Estos informes proporcionan una visión detallada de los métodos utilizados, los datos recopilados y los hallazgos obtenidos en las investigaciones realizadas.

**Informes de gestión:** Ofrecen un panorama general del desempeño de la empresa, incluyendo análisis financieros, indicadores clave de rendimiento (KPIs), y planes estratégicos. Su objetivo principal es la gestión de la información, la planificación y la toma de decisiones a nivel ejecutivo. Informes de cifras y proyecciones de ventas

**Informes de cifras y proyecciones de ventas:** son documentos utilizados para analizar y presentar los datos de ventas de una empresa. Su función es proporcionar información sobre las ventas pasadas, presentes y futuras, incluyendo datos como ingresos, unidades vendidas, tendencias del mercado y proyecciones de crecimiento. Estos informes son útiles para tomar decisiones estratégicas relacionadas con las ventas y el marketing.

**Correos electrónicos:** Representan un medio de comunicación interna fundamental para la gestión de tareas, la resolución de problemas y la coordinación de actividades. Su contenido puede ser utilizado para rastrear la toma de decisiones, identificar responsabilidades y analizar las dinámicas de trabajo dentro de la organización.

Cada tipo de informe mencionado anteriormente cumple una función específica y se utiliza para diferentes propósitos dentro de la organización. Esta descripción dada sobre cada uno de ellos es general, cada organización es encargada de utilizar estos documentos correctamente.

### Desafíos de la documentación de intercambios internos

A pesar de la importancia de la documentación de intercambios internos, existen desafíos significativos que dificultan este proceso. Los encargados de la documentación deben leer y analizar cada informe detenidamente para extraer la información necesaria. Esto consume mucho tiempo y puede llevar a la omisión de información importante. Algunos de los desafíos comunes incluyen la falta de estandarización en la estructura y formato de los informes, la dificultad para extraer información relevante de los informes existentes, y la falta de herramientas y tecnologías adecuadas para facilitar el proceso de documentación.

Un problema común relacionado con la documentación no se trata de la falta de comprensión de la importancia de la documentación, sino de cometer activamente a las personas a documentar. Este problema se debe a la carencia de personas dispuestas a documentar, pero no significa que las personas carecen de la capacidad de documentar. Otro problema es reunir todos los documentos necesarios cuando hay algunas personas que no están dispuestas a compartir sus documentos. Esta voluntad de compartir el conocimiento es también uno de los desafíos más críticos en la gestión de la documentación. Otro desafío identificado está relacionado con la comisión de personas para actualizar los documentos. Una forma de abordar estos problemas para mejorar las prácticas de documentación, es aplicar la documentación como parte regular de los flujos de trabajo del proyecto. Debido a que algunos conocimientos personales y habilidades propias no siempre son posibles de documentar, se crea problemas a la hora de transferir la información y el conocimiento que esta contiene. [11].

La falta de estandarización en la estructura y formato de los informes dificulta la búsqueda y recuperación de información relevante. Cada informe puede seguir una estructura diferente, lo que dificulta la comparación y el análisis de datos. Además, la información importante a menudo se encuentra dispersa en diferentes secciones del informe, lo que dificulta su identificación y extracción. Estos desafíos afectan la eficiencia, la toma de decisiones y la gestión del conocimiento en las empresas.

## Evaluación de indicadores mediante búsqueda de información

Con aumento de la información que se recoge en las reuniones e intercambios que tienen lugar en las instituciones, el análisis y procesamiento de esta se vuelve cada día una labor más compleja [12]. Las actas de reuniones y documentos internos a menudo contienen información crucial, pero su correcta comprensión puede resultar difícil debido a la forma en que se encuentran redactados ya sea por su ambigüedad o densidad. La capacidad para extraer los puntos clave y decisiones importantes correctamente es esencial para mantener la eficiencia de la institución.

Los indicadores son algunos tipos de medidas (estadísticos o parámetros) que sintetizan o resumen los datos y sus cambios estadísticos o medidas de una cierta condición. Proporcionan información y describen el estado del fenómeno objeto de estudio. La búsqueda de información dentro de los documentos internos es crucial para evaluar el desempeño de la organización y tomar decisiones informadas. La información contenida en estos documentos puede revelar tendencias, patrones, riesgos y oportunidades que, de otra forma, podrían pasar desapercibidos. [13]. Estos indicadores varían según las necesidades de información que presente la organización y los indicadores que para cada una sea importantes de analizar, por ejemplo, indicadores financieros o de eficiencia.

### Métodos tradicionales de evaluación de indicadores

Los métodos tradicionales de evaluación de indicadores se basan principalmente en la revisión manual de los informes internos y la extracción de datos relevantes. Estos métodos implican la lectura minuciosa de los informes y la identificación de los indicadores clave a partir de la información presentada [13]. Algunos de estos métodos se presentan a continuación [9], [14]

**Búsqueda Manual:** Requiere una revisión exhaustiva de cada documento para identificar los datos relevantes. revisar cada documento individualmente para encontrar la información relevante. Consiste en revisar manualmente los documentos relevantes en busca de la información necesaria para calcular los indicadores.

**Técnicas de análisis de texto:** Aplicar técnicas como el conteo de palabras clave o la búsqueda de frases específicas para identificar información relevante. Estas técnicas pueden ser limitadas y no siempre capturan las relaciones complejas entre conceptos.

**Extracción manual de información:** La extracción manual de datos implica el uso de herramientas de hoja de cálculo o bases de datos para tabular la información recopilada. La información debe ser extraída manualmente de los documentos y organizada en tablas o bases de datos.

**Análisis manual de datos:** Los análisis se realizan mediante el uso de fórmulas estadísticas o análisis descriptivo, a menudo con la ayuda de herramientas de software. Los análisis de los datos se realizan mediante cálculos y análisis estadísticos manuales, lo que puede ser un proceso lento y propenso a errores.

**Consultas ad-hoc:** Se realizan consultas específicas a las bases de datos o archivos para obtener la información requerida.

**Tablas de indicadores:** Se utilizan tablas predefinidas que contienen los indicadores clave y los datos correspondientes.

### Problemas y limitaciones de la búsqueda manual de información

Sin embargo, este enfoque manual presenta limitaciones significativas. En primer lugar, el proceso de revisión manual es lento y laborioso, especialmente cuando se manejan grandes volúmenes de informes. Esto puede llevar a retrasos en la identificación y análisis de la información, lo que dificulta la toma de decisiones oportunas. Además, esta búsqueda manual puede ser propensa a errores ya que se puede pasar por alto cierta información relevante o interpretar de manera subjetiva los datos extraídos. Esto puede afectar la precisión y confiabilidad de los resultados de evaluación [11].

Debido a la utilidad de la información extraída, muchas veces es necesario garantizar que los expertos en el tema realicen estas tareas, para garantizar la confiabilidad de la información que será utiliza posteriormente por el personal de la organización. A pesar de esto, este personal no está libre de cometer errores humanos, sobre todo cuando hay demasiado volumen de información. Al ser un proceso tan largo en muchos de los casos se convierte en una tarea difícil que requiere mucho tiempo y esfuerzo, lo cual reduce la eficiencia de la organización en la toma de decisiones. Además, se limita la capacidad de respuesta de la organización a las necesidades de información [11] .El problema de la interpretación de la información es uno de los factores que se producen en estas tareas por lo que se precisa que sean expertos quienes realicen estos análisis. Esta situación puede llevar a discrepancias en los resultados de la evaluación lo que puede dificultar la comparación y el análisis de los indicadores a lo largo del tiempo. Una persona que no tenga suficiente conocimiento acerca del tema, puede cometer este tipo de errores, esto provoca que la información no sea confiable y no satisfaga las necesidades de la institución. Esta forma de búsqueda de información no siempre es capaz de capturar toda la información relevante afectando la exhaustividad y la precisión presente en la respuesta. La búsqueda manual puede dificultar la detección de patrones y tendencias en los datos. Esto puede limitar la capacidad de la organización para identificar oportunidades de mejora y tomar decisiones estratégicas. La búsqueda manual se vuelve impracticable para grandes volúmenes de información, lo que limita la capacidad de análisis y la toma de decisiones informadas. Cabe destacar que el proceso de búsqueda de documentos en algunas instituciones es una tarea un poco compleja, debido a la falta de sistemas que ayuden a la obtención de estos documentos [15].

Estos problemas hacen que la búsqueda manual sea inadecuada para las necesidades de análisis de información en las empresas modernas. Con esto se evidencia la necesidad del uso de las tecnologías existentes hoy en día para minimizar este tipo de trabajo sobre todo en entornos donde la información es abundante y la toma de decisiones requiere rapidez y precisión.

## Automatización de la búsqueda de información

Para superar los desafíos planteados anteriormente, es necesario desarrollar métodos y herramientas que faciliten el proceso de búsqueda y análisis. Para esto se ha adoptado la automatización de la búsqueda y extracción de información.

Con el avance de las tecnologías de procesamiento de lenguaje natural y la inteligencia artificial, se han desarrollado soluciones que buscan automatizar la búsqueda de información en los informes internos de las organizaciones. Estas soluciones buscan agilizar el proceso de evaluación de indicadores y mejorar la precisión de los resultados.

### Soluciones propuestas en la literatura

Buscar lo que weeny mando x tele

Existen diversas soluciones propuestas en la literatura para la automatización de la búsqueda y análisis de información. Algunos de estos son los Sistemas de búsqueda basados en palabras clave, Sistemas de recuperación de información, Sistemas para Análisis de sentimientos, etc.

En la revisión realizada se encontraron soluciones que presentan una técnica de búsqueda exploratoria, utilizada para sacar información de textos mediante procesamiento de lenguaje natural y aprendizaje automático. En [1] se propone un método para aprender las representaciones de preferencia de los usuarios utilizando el aprendizaje contrastivo de las acciones exploratorias. Aprovechando los datos recopilados automáticamente de un proceso de diseño señal, para aprender mejor las preferencias de los usuarios, facilitando la adaptación de robots que interactúan con humanos.

El trabajo [16] se centra en idear facetas visuales interactivas, visualizar facetas de información para respaldar la comprensión del usuario y el control del espacio de información. Propone una técnica de deslizamiento de filtro de las facetas para ver los elementos de intersección entre dos facetas. Este trabajo fue de gran utilidad para usuarios que trabajan con correos electrónicos por las sugerencias que ayudaban a la búsqueda.

Otra variante utilizada es la visualización de los resultados mediante grafos de conocimiento. Se encontraron soluciones como [17] que propone un modelo de gestión del conocimiento iterativo e incremental que oriente la resolución de problemas y un enfoque interactivo de visualización del conocimiento en la entrega de conocimiento. Además, desarrolló una estrategia de aprendizaje de conocimientos, que aborde el aprendizaje de conocimientos de dominio, el aprendizaje de conocimientos de tareas y el entorno de aprendizaje y búsqueda para respaldar la búsqueda de información del usuario. [18] presenta un sistema de búsqueda exploratoria configurable y adaptable implementa tres características fundamentales adaptabilidad, visibilidad de los detalles de procedencia, y muestra los resultados en una vista de árbol con estructura jerárquica. [4] contribuye a diseñar futuros sistemas de búsqueda al proporcionar información sobre la eficacia de las representaciones alternativas de los resultados de búsqueda para apoyar las actividades de búsqueda exploratoria, que culmina en una nueva representación híbrida llamada grafos de conocimiento jerárquico (HKG).

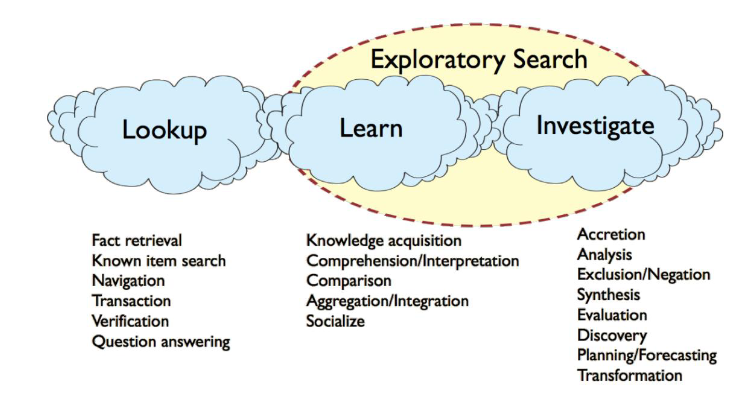
Algunas de las técnicas utilizadas en estas soluciones incluyen el análisis de texto, la extracción de entidades, el etiquetado de partes del discurso y la detección de relaciones entre conceptos. Estas técnicas permiten a las herramientas automatizadas identificar y extraer información relevante de los informes de manera más rápida y precisa que la búsqueda manual.

### Búsqueda exploratoria

La búsqueda de información es una actividad que realizamos a diario. Los navegadores modernos tienen funcionalidades de búsqueda que permiten a los usuarios ingresar consultas de búsqueda y recuperar una lista clasificada de enlaces relevantes, obteniendo así todo tipo de información. Sin embargo, a diferencia de las búsquedas simples, los usuarios suelen realizar búsquedas con necesidades de información ambiguas u objetivos inciertos [2]. Un ejemplo de esto ocurre cuando se busca literatura relacionada, donde los investigadores comienzan su búsqueda haciendo consultas de forma muy general, sin tener muy claro qué buscar [16]. En tales situaciones, los usuarios modifican dinámicamente sus direcciones de búsqueda a medida que adquieren mayor conocimiento del espacio de búsqueda [3]. A este tipo de búsqueda se le denomina búsqueda exploratoria, para hacer referencia a su naturaleza investigativa y de aprendizaje [19].

La búsqueda exploratoria se centra en la adquisición de nuevos conocimientos lo que se considera un desafío para los usuarios. Las tareas de búsqueda exploratoria se centran en objetivos de búsqueda de investigación y aprendizaje que generalmente son abiertos, dinámicos y están motivados por problemas mal definidos. Generalmente los usuarios que realizan búsqueda exploratoria tienen falta de conocimiento sobre el dominio y no saben cómo lograrán satisfacer sus objetivos de búsqueda de información [3].

También se puede definir una jerarquía de necesidades de información que va desde hechos básicos que guían acciones a corto plazo hasta redes de conceptos relacionados que nos ayudan a comprender fenómenos o ejecutar acciones complejas. Para estas respectivas capas de necesidades de información, se pueden definir tipos de actividades de búsqueda de información, cada una con estrategias y tácticas asociadas que podrían respaldarse con herramientas computacionales. Algunas de estas actividades pueden ser: buscar, investigar y aprender, donde la búsqueda exploratoria abarca las dos últimas de estas actividades, estas serían los procesos claves de la búsqueda exploratoria [19]. En la Figura 1 se muestran estas actividades y como están contenidas en la búsqueda de información y las acciones contenidas dentro de cada una de estas actividades.



1. Figura 1: Actividades de búsqueda de información[19]

Según Marchionini [19] para caracterizar la búsqueda exploratoria es importante describir dos aspectos principales: el contexto del problema y el proceso mediante el que se realiza la búsqueda. El contexto del problema se refiere a la motivación de los buscadores de resolver un problema o un deseo de aprender, este contexto generalmente se describe como mal estructurado, dado a que las búsquedas exploratorias pueden implicar situaciones complejas. El proceso de búsqueda está dado a que el objetivo de la búsqueda exploratoria suele ser crear productos o tomar decisiones, por lo que gran parte del proceso de búsqueda se dedica a comparar y examinar resultados, así como descubrir los conceptos claves del domino [4].

En los últimos años la búsqueda exploratoria ha sido utilizada en diversos escenarios. Esta aproximación no solo agiliza el proceso de búsqueda de información, sino que también facilita la detección de patrones ocultos y la generación de conocimientos más profundos a partir de los informes internos. Esta técnica se basa en la capacidad de los sistemas de procesamiento del lenguaje natural para identificar patrones y relaciones ocultas en los datos. La búsqueda exploratoria permite a los usuarios descubrir información relevante que de otra forma podría pasar desapercibida.

## Grafos de conocimiento

La visualización de resultados es un factor importante en la búsqueda de información. En la búsqueda exploratoria la visualización de estos resultados va más allá de una lista con todos los documentos que puedan estar relacionados con el tema en cuestión. Los sistemas de búsqueda exploratoria deben tener ciertas características, ya que esta búsqueda es más compleja. Algunas de estas características son: soporte de consultas y refinamiento de consultas, proporcionar agrupaciones y organización del resultado de búsqueda, ofrecer visualización para respaldar conocimiento y toma de decisiones, y apoyar el aprendizaje y comprensión [3]. El desarrollo de este trabajo se enfoca en la visualización y el apoyo al aprendizaje.

Existen diferentes formas de visualizar los resultados obtenidos mediante búsqueda exploratoria. Algunas de estas son nubes de palabras, de forma jerárquica mediante árboles, búsqueda por facetas, grafico de barras, líneas de tiempo, mapas geográficos, mapas de entidades y relaciones, entre otras. En los mapas de entidades y relaciones se encuentran los mapas conceptuales y grafos de conocimiento. Los grafos de conocimiento juegan un papel fundamental en la representación y organización de la información contenida en los informes internos. Un grafo de conocimiento es una estructura de datos que representa las relaciones entre los conceptos y las entidades presentes en los informes.

### Conceptos y aplicaciones

Un grafo de conocimiento es una representación gráfica compuesta por nodos y conexiones, donde los nodos representan conceptos o entidades y las conexiones representan las relaciones entre ellos. Estas relaciones pueden ser de diferentes tipos, como "es un", "tiene", "está relacionado con", entre otros. Los grafos de conocimiento se utilizan para organizar y estructurar la información contenida en los informes internos de las organizaciones. Estos grafos permiten visualizar las relaciones entre los diferentes conceptos y entidades, lo que facilita la identificación de patrones y tendencias.

Los grafos de conocimiento son fundamentales para muchas de las grandes empresas de hoy en día. En [20] se realiza un estudio y caracterización de como 5 grandes empresas utilizan estos grafos para realizar sus funcionalidades, como por ejemplo Google y Facebook. Los grafos de conocimiento proporcionan los datos estructurados y el conocimiento que impulsa a muchos productos, de forma general, un grafo de conocimiento describe objetos de interés y conexiones entre ellos. Por ejemplo, un grafo de conocimiento puede tener nodos para una película, los actores en esta película, el director, etc. Cada nodo puede tener propiedades, como el nombre y la edad de un actor. Puede haber nodos para múltiples películas que involucren a un actor en particular. Luego, el usuario puede atravesar el grafo de conocimientos para recopilar información sobre todas las películas en las que apareció el actor [20].

En algunos casos, los enlaces entre los nodos del grafo podrían tener sus propias propiedades: ejemplo, un enlace que conecta un actor y una película podría tener el nombre del papel específico que jugó el actor. Los grafos de conocimiento, generalmente proporcionan una representación formal compacta que los desarrolladores pueden usar para inferir nuevos hechos y desarrollar el conocimiento, por ejemplo, utilizando el grafo que conecta las películas y los actores para averiguar qué actores aparecen con frecuencia en las películas [20]. Un ejemplo de este tipo de grafo se muestra en la Figura 2.

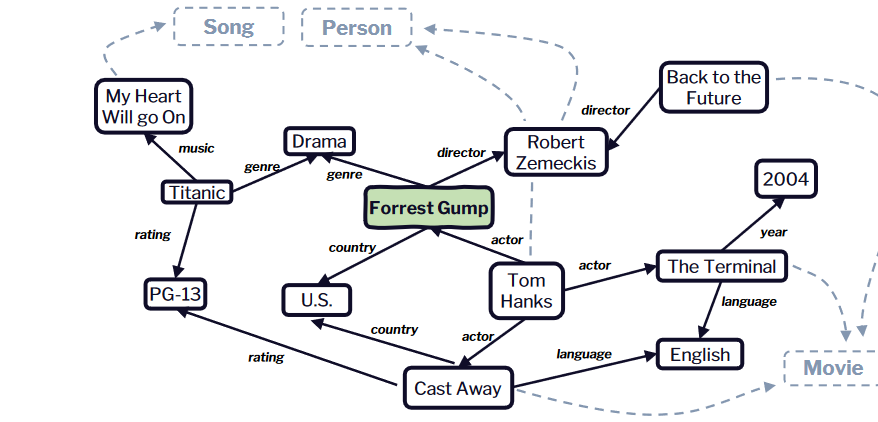


Figura 2: Ejemplo de grafo de conocimiento [21]

La variabilidad de los grafos de conocimiento es tanto una característica distintiva como una que dificulta su implementación eficiente. Debido a su amplio uso, en algunos casos pueden llegar a ser enormes y complicados, y como resultado, difícil de comprender. Por lo tanto, esto puede dificultar la comprensión para los usuarios novatos, lo que requiere la necesidad de mecanismos exploratorios para apoyar la exploración sobre este tipo de grafo [22]. La exploración del gráfico de conocimiento es el proceso de análisis automático asistido y progresivo de los grafos de conocimiento, y esta exploración cuenta principalmente de tres objetivos. El primero consiste en la comprensión de la estructura y naturaleza del gafo, segundo; identificación de que porción del grafo dado y sus datos satisfacen la necesidad de información actual, tercero; extracción de ideas que puedan ayudar a formular nuevas preguntas de investigación e hipótesis. Estos objetivos se pueden lograr mediante tres tareas principales: perfiles y resumen, análisis exploratorios y búsqueda exploratoria [21]. La relación que existe entre estos objetivos y las tareas, no es solamente de uno a uno, ya que para lograr los objetivos es necesario la combinación de varias tareas, como se muestra en las Figura 3, Figura 4 y Figura 5.

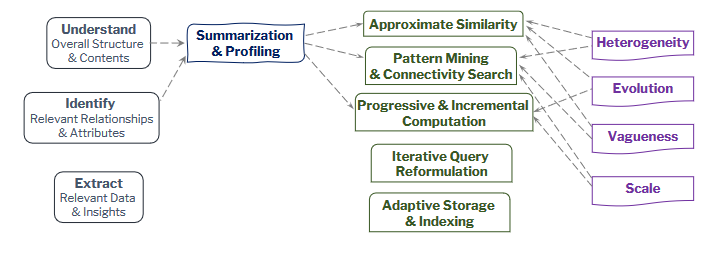


Figura 3: Relación objetivo-tarea de perfil y resumen [21]

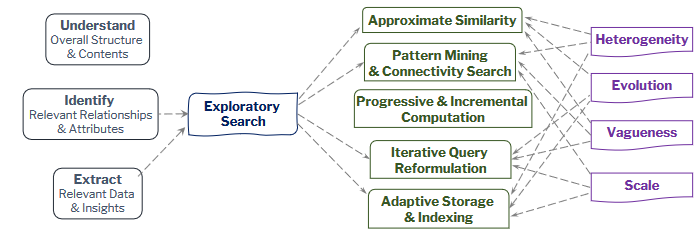


Figura 4: Relación objetivos-tarea búsqueda exploratoria [21]

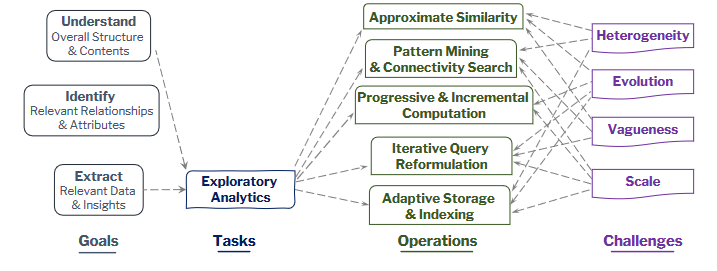


Figura 5: Relación objetivos-tarea análisis exploratorio [21]

Dada esta relación existente se puede ver que entre la búsqueda exploratoria y los grafos de conocimiento existe una relación muy estrecha, ya que una forma de visualizar los resultados en un proceso de búsqueda exploratoria es utilizar un grafo de conocimiento, y, por otro lado, una de las tareas que permite la exploración de los grafos de conocimiento y lograr sus objetivos es la búsqueda exploratoria. Este trabajo se centra en la utilización de la búsqueda exploratoria, y una forma de ver esta relación es que mientras que el grafo muestra una vista alejada, la búsqueda exploratoria se centra más en los datos, y un ejemplo de cómo lograr esto sería hacer zoom a un subconjunto de elementos de este grafo.

### Herramientas y tecnologías existentes

Existen tecnologías como la Web Semántica y los estándares de representación del conocimiento, como RDF y OWL [23], que permiten la interoperabilidad y el intercambio de información entre diferentes sistemas y aplicaciones.Debido a la existencia de tantas herramientas en la actualidad, [24] se decide solo hacer una descripcion de las que estan diseñadas para la visualizacion mediante la plataforma web.

|  |  |
| --- | --- |
| Nombre | Descripción |
| Cytoscape.js | Biblioteca JavaScript para visualizar y analizar grafos de conocimiento. |
| Graphviz.js | Biblioteca JavaScript para visualizar grafos generados con Graphviz. |
| Sigma.js | Biblioteca JavaScript para visualizar grandes grafos en tiempo real. |
| mxGraph | Biblioteca de diagramas de JavaScript que permite crear rápidamente aplicaciones de grafos y tablas interactivas |
| El Grapho | Biblioteca JavaScript de código abierto para visualización de gráficos que explota los sombreadores GLSL para generar rápidamente representaciones gráficas de grafos grandes. Los grafos renderizados se pueden ampliar y desplazar. Admite múltiples algoritmos de diseño de gráficos. Las representaciones gráficas en El Grapho son interactivas. |
| Ogma | Biblioteca JavaScript de visualización de grafos comerciales de Linkurious. Ogma funciona con WebGL, pero también admite renderizadores HTML5, Canvas y SVG. El modo geográfico permite al programador mostrar el grafo encima de un mapa web de diferentes proveedores de mapas. |
| VivaGraphJs | Biblioteca de JavaScript de código abierto que admite renderizadores WebGL, Canvas y SVG para visualización de gráficos. Está construido sobre la biblioteca de algoritmos gráficos ngraph |
| Vis.js | Framework JavaScript para crear visualizaciones de datos, incluyendo grafos y redes. |
| D3.js | Biblioteca JavaScript para crear visualizaciones de datos interactivas, incluyendo grafos. |

## Conclusiones parciales

Con el desarrollo de este capitulo se le da cumplimiento al objetivo de Caracterizar la búsqueda de información en organizaciones contemporáneas:

* Se ha discutido la importancia de la documentación de intercambios internos en las organizaciones.
* Se han descrito los desafíos asociados a la búsqueda de información relevante en los informes generados con métodos tradicionales.
* Se ha explorado la automatización de este proceso mediante el uso de soluciones basadas en búsqueda exploratoria
* Se ha explorado la visualización de resultados basado en grafos de conocimiento.
* Se ha caracterizado la aplicación de grafos de conocimiento y las herramientas y tecnologías existentes en este campo.

# Capítulo 2: Modulo de búsqueda exploratoria basado en generación automática de grafos de conocimiento

En el siguiente capítulo se propone el módulo de búsqueda exploratoria basado en la generación de grafos. Primeramente, se describen los requisitos funcionales y no funcionales que se tienen en cuenta para el desarrollo del módulo. Luego se modelan las funcionalidades que presenta el módulo propuesto, mediante diagrama de caso de uso del sistema y diagramas de flujo. Además, se muestra una vista de arquitectura por la que está conformada el módulo con una descripción de cada una de sus capas. El módulo está conformado por una arquitectura basada en microservicios que permite que este pueda funcionar de forma independiente.

## Requisitos funcionales

El módulo debe cumplir con los siguientes requisitos funcionales:

Tabla 2.1: Requisitos funcionales del módulo

|  |  |
| --- | --- |
| No | Descripción |
| Extracción de información | |
| 1 | El sistema debe Permitir al usuario subir archivos de documento en formato Word y PDF. |
| 2 | El sistema debe ser capaz de Extraer el texto contenido dentro del documento. |
| 3 | El sistema debe ser capaz de Extraer la fecha de creación del documento. |
| 4 | El sistema debe ser capaz de Extraer entidades nombradas del texto. |
| 5 | El sistema debe ser capaz de Extraer relaciones existentes entre esas entidades. |
| 6 | El sistema debe proporcionar una estructura con la información necesaria para la construcción del grafo. |
| análisis de información | |
| 7 | El sistema debe Mostrar un grafo de conocimiento donde se evidencie las relaciones entre entidades. |
| 8 | El sistema debe Permitir la navegación a través del grafo para comprender la información. |
| recuperación de información | |
| 9 | El sistema debe Permitir al usuario obtener un resultado de búsqueda de documentos subidos anteriormente |
| 10 | El sistema debe almacenar el análisis obtenido de los documentos para mayor eficiencia. |

## Requisitos no funcionales

El sistema debe tener los siguientes requisitos no funcionales:

Tabla 2.2: Requisitos no funcionales

|  |  |  |
| --- | --- | --- |
| no | Requisito | Descripción |
| 1 | Interfaz externa | **Simple de usar:** la vista del grafo debe ser comprensible de forma tal que quede clara la relación entre entidades.  **Plataforma:** debe ser web, debido a las ventajas que esta trae consigo. |
| 2 | Usabilidad | **Uso correcto de componentes:** los componentes utilizados deben cumplir sus funciones correctamente.(subir archivos o búsqueda) |
| 3 | Fiabilidad | **Retroalimentación:** debe informar al usuario en todo momento si ocurre algún problema o si necesita esperar para obtener una respuesta |
| 4 | Rendimiento | **Eficiencia:** almacenar información de los documentos en Solr, lo que permite que la recuperación de información sea más rápida ya que el análisis estará almacenado solo sería mostrarlo. |

## Diagrama de casos de uso del sistema

En la Figura 8 se muestra en diagrama de casos de uso del sistema, donde se muestra la interacción del actor Analista con el módulo diseñado. El Analista puede ejecutar la funcionalidad de subir los documentos a los que desea realizar un análisis textual. Esta interacción ocurre mediante una interfaz de usuario que brinda la posibilidad de subir los archivos que necesite. Luego estos documentos son procesados y la información recopilada de ellos se almacena en una base de datos para su análisis futuro. En otro caso para consultar esta información el Analista tiene la posibilidad de ejecutar una consulta en forma de texto plano. Con esta obtendrá los documentos relacionados a esta consulta y se mostrará el grafo de conocimiento correspondiente a la información contenida en estos documentos.

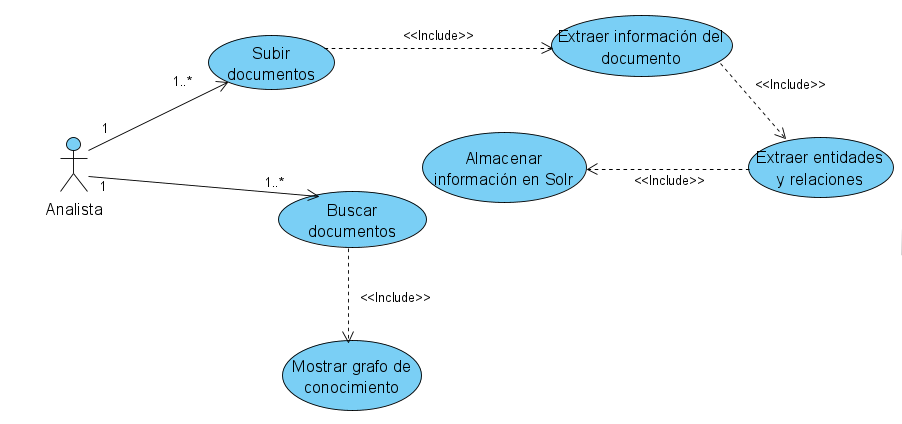


Figura 6:Diagrama de casos de uso del Sistema

## Diagramas de secuencia

En los siguientes diagramas se muestran cómo se cumplen los casos de uso de Subir documentos y Buscar documentos.

### Diagrama de secuencia de Subir documentos

En la Figura 9 se muestra la forma en la que se suben los documentos en el sistema. Primero el Analista selecciona, mediante un componente visual, los documentos que desea subir que se encuentran en el directorio local. Estos documentos son enviados a la capa de servicios mediante una solicitud web. Luego de ser procesados y se ha extraído la información necesaria de ellos, la capa de servicio devuelve una respuesta. Esta respuesta es mostrada al usuario mediante el sistema, además pueden ser de dos tipos. La primera es Subida correcta, que implica que el proceso se completó con éxito y se almacenaron los datos. La segunda seria si existe algún problema en la subida de documentos, donde se notifica cual ha sido el error ocurrido.

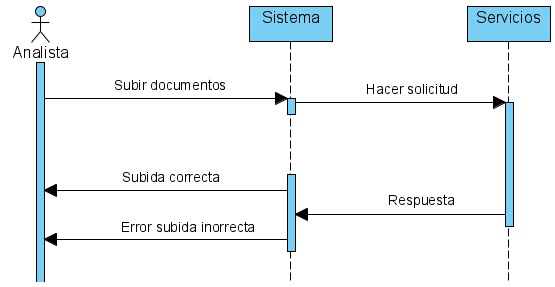


Figura 7:Diagrama de secuencia de caso de uso subir documentos

### Diagrama de secuencia de Buscar documentos

En la Figura 10 se muestra la forma en la que el Analista puede buscar los documentos que han sido subidos anteriormente. Primeramente, introduce una consulta en forma de texto plano con la información que desea obtener. Esta consulta es enviada a la capa de servicios donde es procesada en busca de coincidencias. Luego de procesar esta información, en caso de haber encontrado documentos para procesar, la capa de servicio devuelve la información en forma de grafo de conocimiento. En caso contrario se devuelve una respuesta de que no se encontraron coincidencias o no existen documentos subidos anteriormente, según sea el caso ocurrido.

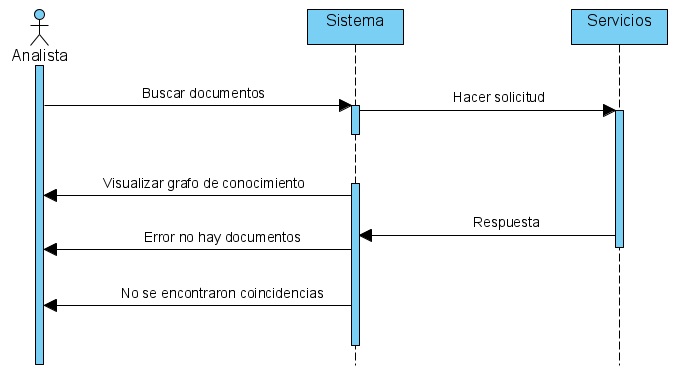


Figura 8: Diagrama de secuencia de caso de uso Buscar documentos

## Implementación del módulo de búsqueda

El módulo está orientado a la plataforma web y cuenta con una estructura basada en microservicios. Está diseñado para que cada uno de los servicios que se ofrecen sean accesibles de forma independiente. Pero para este trabajo este módulo funcionará como un todo con el objetivo de brindar una información a un sistema. Este módulo será integrado a un sistema desarrollado anteriormente siendo capaz de interactuar con una interfaz de usuario ya diseñada previamente. Para la realización de este módulo se utilizó Spring Boot, por lo que los servicios están desarrollados con esta tecnología. En la Figura 11 se define la arquitectura que presenta el módulo, observándose como es el funcionamiento interno de este, y como se realizan las comunicaciones entre las diversas capas de la arquitectura.

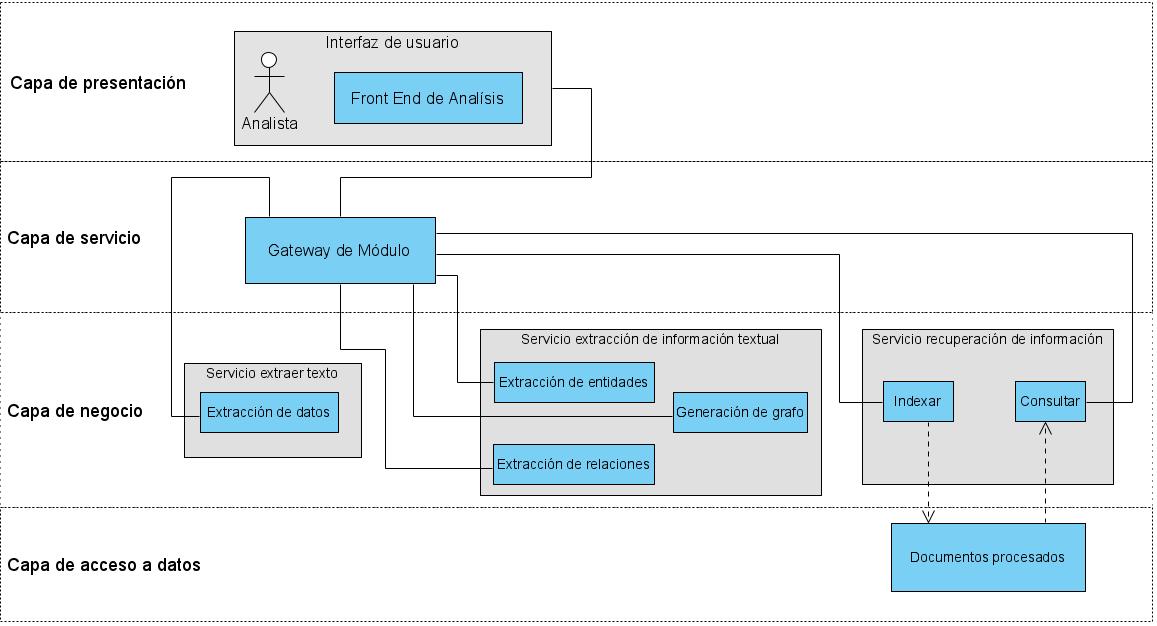


Figura 9: Vista de arquitectura del módulo

Como se observa en la figura para la comunicación con la interfaz de usuario se definió un servicio web, denominado Gateway de Módulo. Este se encarga de hacer las diferentes peticiones a cada uno de los servicios y manejar las respuestas que estos devuelven, para dar una respuesta a la capa superior. El objetivo de esto es que desde la interfaz de usuario se realicen la menor cantidad de peticiones posibles, que en este caso solo se realizan dos. En los epígrafes siguientes se describen en más profundidad cada uno de los servicios.

### Servicio de Extraer Texto

Este servicio es el encargado de procesar los documentos subidos por el Analista. Para su implementación se utiliza la biblioteca Apache Tika para procesar los documentos y extraer la información necesaria. En la Figura 12 se muestra el diagrama de clases de este servicio y los detalles del *endpoint* para poder consumir este servicio.

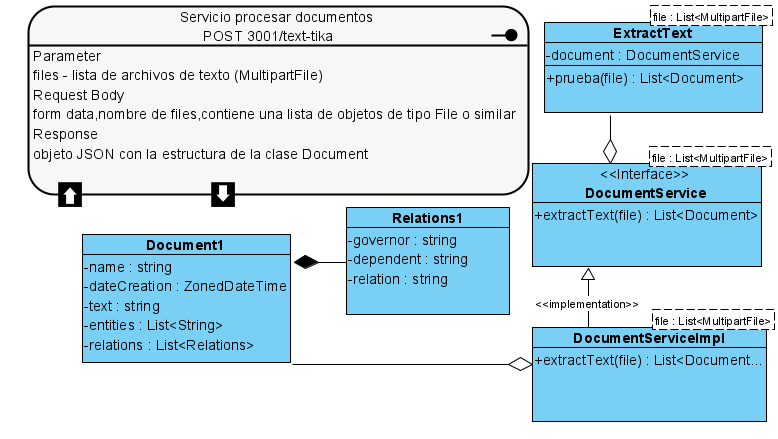


Figura 10:Diagrama de clases del servicio Procesar Documentos

En la Figura 12 se observan las clases creadas para realizar estas funcionalidades y las relaciones que existen entre estas. La clase **ExtractText** es la clase controladora la cual presenta el método **prueba** que recibe un parámetro correspondiente a la lista de archivos de tipo texto. Este parámetro está marcado con la anotación *@RequestParam(“files”)* por lo que es necesario hacer la solicitud en la forma de *formdata* con exactamente el mismo nombre. Luego esta clase llama al método de la interfaz **DocumentService** que es la encargada de buscar la clase con la anotación *@Service* que implementa su método. Esta clase es **DocumentServiceImpl**, quien se encarga de analizar cada uno de los archivos y obtener información como su nombre, fecha de creación y el texto que este contiene. Luego crea un nuevo objeto de tipo **Document**, y asigna la información obtenida a cada campo correspondiente, dando paso a la construcción del documento que será almacenado posteriormente.

En la Figura 12 también esta una descripción del *endpoint*. En la figura se muestra el tipo de método (POST), y la URI con el puerto para el que está configurado y la ruta. Como se mencionó este *endpoint* requiere un parámetro específico para su correcto funcionamiento. La información debe llegar a través de un *multipart/formdata* que tiene que ser especificado en el *header* de la solicitud. Este *formdata* debe contener un arreglo de objetos de tipo File o similares, que tiene que tener por nombre “files”, sino es así la solicitud no será enviada y dará error. La forma de obtener estas características y hacer una correcta solicitud varía en función del lenguaje.

A continuación, se muestra un ejemplo de cómo sería el acceso a este servicio mediante ***Postman***.

En la Figura 11 se muestra en la parte superior la ruta donde se ejecuta el servicio. En el atributo ***Body*** se selecciona la opción de ***form-data***, y en la pestaña ***key*** se selecciona el tipo, en este caso ***File***. Luego en la pestaña ***value*** se seleccionan los **documentos** que se quieran procesar, y se envía la solicitud.

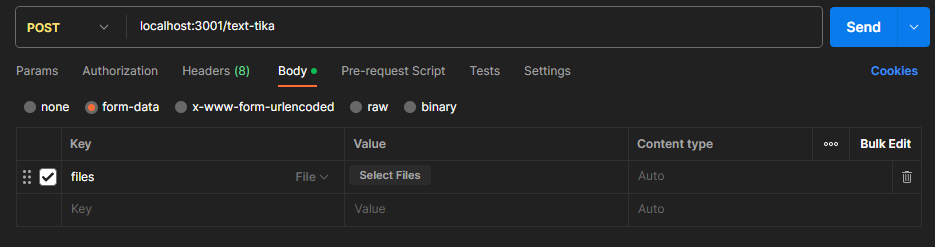


Figura 11: Llamada al servicio Extracción de texto utilizando Postman

En Figura 12 se muestra el resultado que se obtiene luego de procesar correctamente la solicitud. Esta es la estructura de la clase ***Document***, en formato JSON, que es la que se utiliza para los análisis posteriores. Los campos que están en *null* se van completando a medida que los documentos pasan a través de los diferentes servicios.

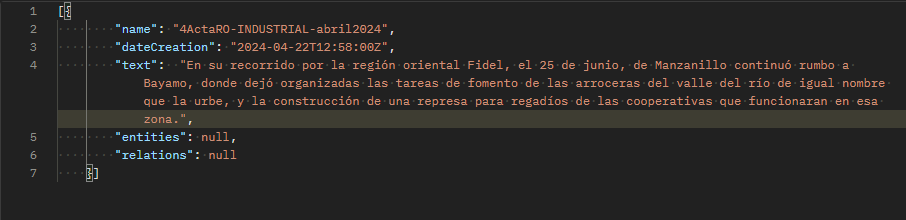


Figura 12: Resultado de llamada al servicio Extracción de texto

### Servicio de generación de grafo de conocimiento

Este servicio es el encargado de analizar el texto que se obtiene a partir de procesar los documentos, devueltos por el servicio anterior. Para su implementación se utiliza la biblioteca de Stanford CoreNLP, que es una biblioteca de procesamiento de lenguaje natural. Con ella, se reconocen las entidades nombradas presentes en el texto, y las relaciones existentes entre ellas. Para la generación de grafo se utiliza la biblioteca Jackson que permite obtener el grafo en formato JSON y dar la respuesta que se muestra en la interfaz de usuario. Este servicio está dividido en tres *endpoints.* En la Figura 13, 14, 15 se muestran las características de estos *endpoints*. Como se puede observar, en las figuras antes mencionadas, los tres diagramas tienen una clase en común, **NerController**, ya que es la que contiene los *endpoint*, y es la encargada de establecer características para el acceso a estos.

En la Figura 13 se muestra el diagrama del *endopint* que se encarga de reconocer las entidades nombradas de cada texto, obtenido de la lista de documentos creados a partir de la clase **Document**. A este se accede mediante el método *entities* de la clase controladora. En la imagen se muestra las características de este, como el puerto y ruta de acceso, método de acceso, el cuerpo de la solicitud y la respuesta que este devuelve si todo funciona bien. Este *endpoint* recibe como parámetro una lista de objetos de la clase **Document** y requiere que este contenga el atributo *text* para su correcto funcionamiento. Esta lista de documentos viaja hasta la clase **NerServiceImpl** que se encarga de reconocer las entidades existentes en el texto, de cada uno de los documentos, y asigna estos resultados al atributo correspondiente en el documento.

En el caso de este servicio recibiría en el *body* una estructura como la de Figura 12, y al texto ingresado se le haría el análisis de las entidades nombradas. En la imagen a continuación un ejemplo de entidades reconocidas en este texto.

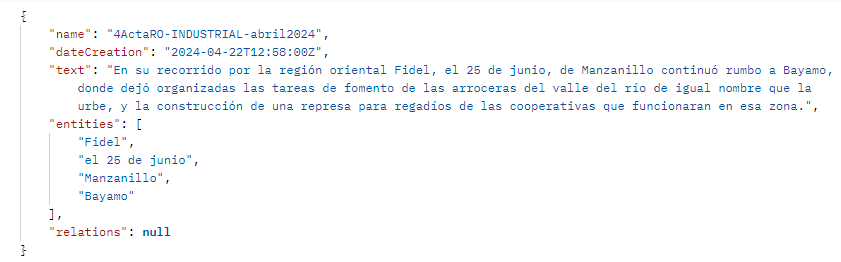


Figura 13:Ejemplo análisis NER

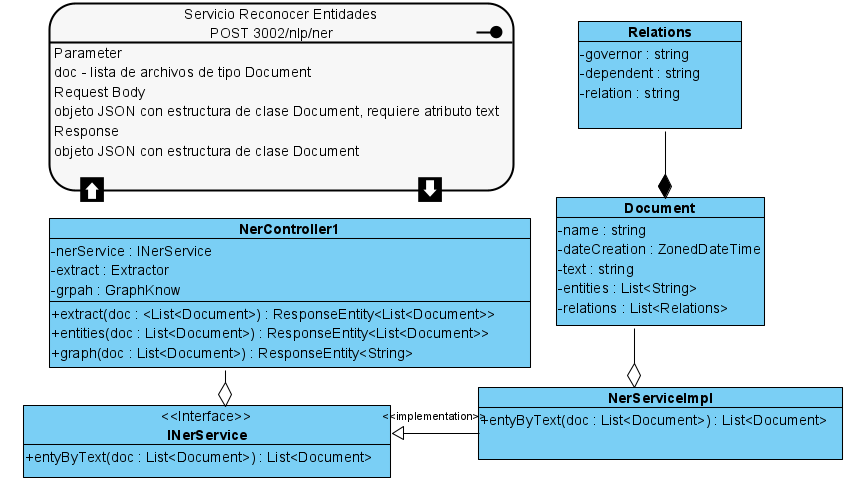


Figura 14:Diagrama de clases del endpoint de Reconocer Entidades

En la Figura 14 se muestran las características del *endpoint* que se encarga de extraer las relaciones que existen entre las entidades, anteriormente extraídas. Su funcionamiento es similar al anterior. En la figura, igualmente se muestran la ruta de acceso, el método, y el resto de características necesarias para acceder a este. La lista de documentos viaja a la clase **ExtractorImpl** donde se analiza el texto, sintácticamente. Se obtienen las relaciones entre cada palabra de cada una de las oraciones del texto. Luego se verifica que las palabras que se obtienen de este análisis concuerden con las entidades ya obtenidas, y de esta manera se busca la relación que existe entre ellas. Si se realiza el proceso correctamente esta información es devuelta, con la estructura de la clase **Document**. Terminado este proceso ya está listo el documento para ser almacenado en una base de datos. Para que funcione correctamente este endpoint necesita, el texto que se desea procesar, y la lista de las entidades obtenidas.

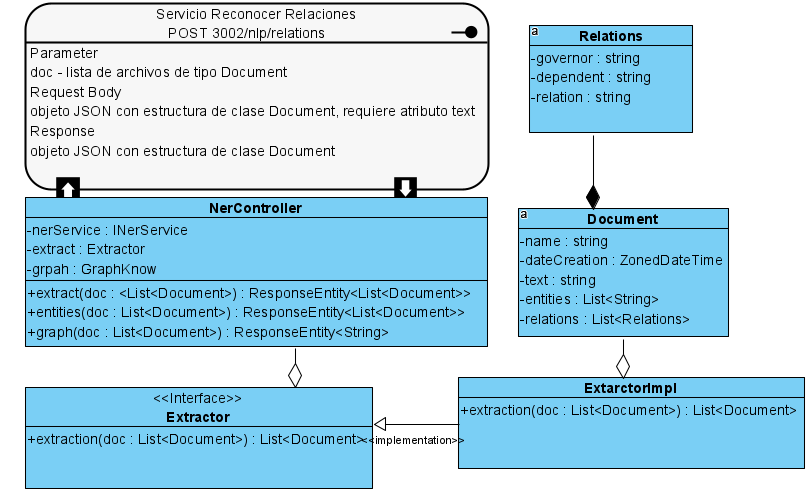


Figura 15: Diagrama de clases del endpoint de Extraer Relaciones

A continuación, un ejemplo de que se obtendría, al procesar correctamente esta solicitud.

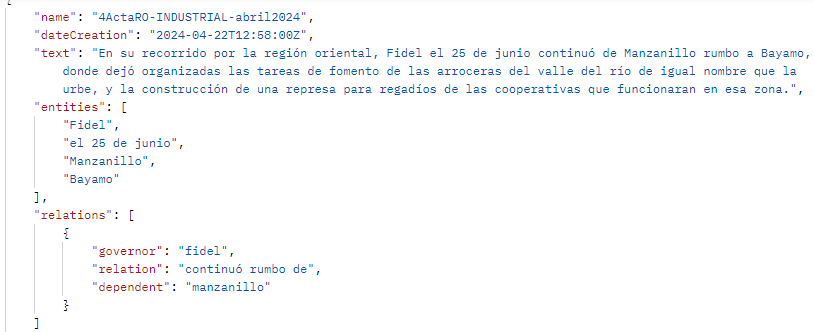


Figura 16: Ejemplo de extracción de relaciones

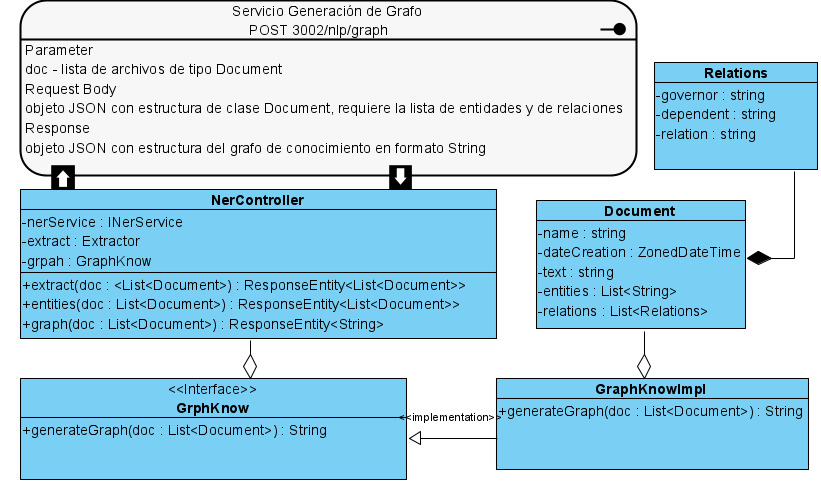


Figura 17:Diagrama de clases del endpoint de Generación de Grafo

Como se pudo observar, en la Figura 16 se muestran las características del *endpoint* que se encarga de crear la estructura que será utilizada posteriormente para la visualización del grafo de conocimiento. En este trabajo esta funcionalidad se ejecuta luego de una consulta a la base de datos, por lo que el parámetro de la lista de documentos sería el resultado de esa consulta. Este parámetro tiene que tener la estructura de la clase **Document**, y requiere que este contenga la lista de entidades y de relaciones ya obtenidas. Como se observa en la figura este *endpoint* no retorna una lista de documentos, como es el caso de los anteriores. Si todo el proceso concluye satisfactoriamente, se devuelve un objeto JSON con una estructura, que consta con la información necesaria para construir un grafo. Esta estructura es simple, donde se anota el sujeto, objeto y relación, con *source*, *target* y *label*, respectivamente. Igualmente, la información de los documentos viaja a la clase **GraphKnowImpl**, donde a partir de la lista de entidades y las relaciones, se crea una estructura para que pueda ser mostrada en una interfaz.

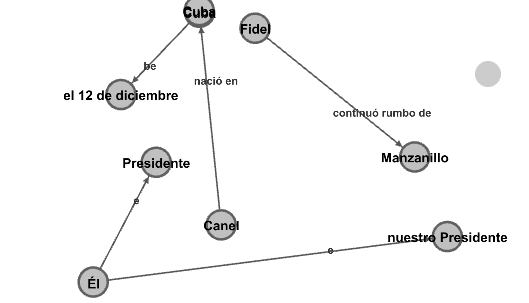


Figura 18: Ejemplo utilizando Gephi

En la Figura 18 se muestra un ejemplo de visualización de la información obtenida del procesamiento a dos textos. Esta visualización se obtiene mediante el uso de Gephi, como ejemplo de lo que se espera que se visualice. Como se puede observar se puede sacar cierta información de este análisis, como Fidel continuo el viaje desde Manzanillo, o que Diaz Canel nació en Cuba.

### Servicio de Recuperación de Información

En la Figura 16 se muestra la relación de las clases dentro del servicio de recuperación de información. Este servicio está implementado con la biblioteca *solrj*, que proporciona una API para utilizar *Solr* en *Spring Boot*. Este servicio consta de dos *endpoints,* ambos contenidos en la clase **SolrController**. El primer método de esta clase es *consultSolr()* que es el encargado de mandar la consulta a la base de datos para obtener los documentos, que luego serían enviados al servicio adecuado para obtener el grafo. En la parte superior derecha de la imagen se encuentran las características de acceso para este *endpoint*. En la parte superior izquierda, las características del segundo método de la clase controladora, *indexSolr()*. Este es el encargado de recibir la lista de los documentos ya procesados, con toda la información necesaria, para almacenarlos en la base de datos. Ambos métodos llevan la información hasta la clase **SolrImpl** que ejecuta la función correspondiente a la solicitud que se le haga al servicio.

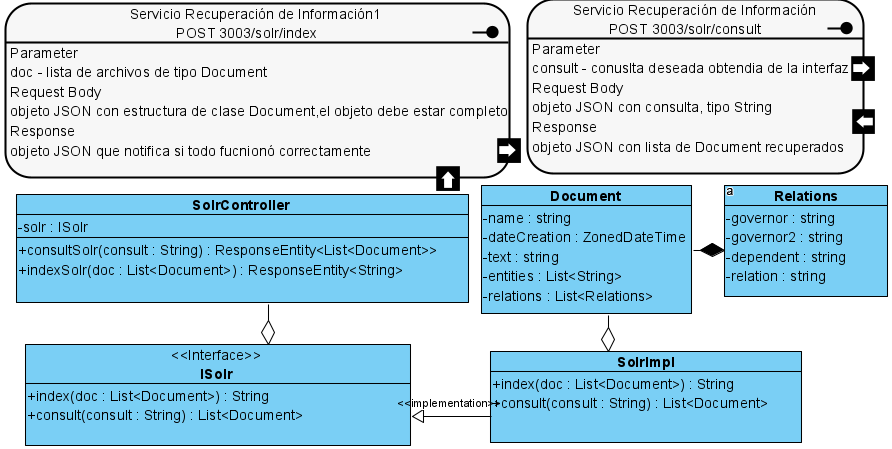


Figura 19: Diagrama de clases del servicio de Recuperación de Información

En Solr los datos se guardan en *cores*. En la clase **SolrImpl** está definido que se acceda al *core* llamado **modulo**. Por tanto, para acceder al servidor solr se debe crear un *core* con este nombre.

El objetivo por el cual se separan estas funciones es para no tener la necesidad de que siempre que el usuario quiera visualizar la información, se tenga que hacer el análisis de extracción de datos de los documentos. De esta forma mejora la eficiencia del proceso, ya que visualizar la información no demora tanto como el proceso de análisis y además el análisis se realiza una sola vez.

## Conclusiones parciales

Con la realización de este capítulo se da cumplimiento al objetivo de Desarrollar el sistema de búsqueda exploratoria:

* Queda definido cuales son los requisitos por los cuales se rige el desarrollo de este módulo de sistema.
* El diagrama de caso de uso del sistema permite entender, con una visión general del funcionamiento de este sistema.
* Los diagramas de secuencia creados describen la relación que tiene este sistema con el usuario y como este recibe la información que solicita.
* La vista de la arquitectura da la posibilidad de comprender las conexiones existentes entre cada una de las funciones presentes en el módulo propuesto, así como los servicios y la manera en la que se trabaja para obtener la información brindada que será mostrada.
* Con los diagramas de clases presentados se observa la información necesaria para poder acceder a cada uno de estos servicios de forma individual. Además, muestran cómo están estructurado estos servicios.

# Capítulo 3: Evaluación del módulo de búsqueda

En este capítulo se estarán realizando pruebas al módulo desarrollado mediante solicitudes. Se prueba el funcionamiento de cada servicio del módulo por independiente, comprobando que cumplen sus funciones. Se desarrolla una aplicación de prueba que permite comprobar el funcionamiento del módulo como un todo. En esta aplicación se recibe como entrada los documentos y se muestra el grafo que esta información genera.

## Pruebas funcionales

En este epígrafe se realizan las pruebas funcionales del módulo. Donde se evalúa el correcto funcionamiento de los mismos, y se detectan escenarios de posibles mejoras futuras.

### Extracción de texto

A continuación, se muestra cómo funciona este servicio para diferentes tipos de archivos. Este servicio solo realiza el procesamiento a documentos de tipo Word y PDF. Para el ejemplo mostrado en la Figura 20 se seleccionaron 2 archivos, una imagen y un archivo de texto con extensión txt. Como se puede observar el servicio no analiza estos archivos y devuelve una lista vacía.

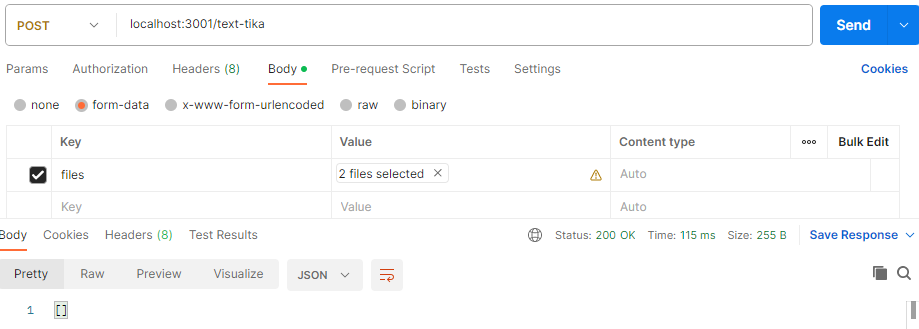


Figura 20:Prueba archivos incorrectos

En la Figura 21 se muestra la selección de 3 archivos, una imagen, un documento Word y un PDF. El servicio no procesa el archivo imagen y el resultado es el que se muestra en la figura antes mencionada. En la imagen se ven fragmentos de los dos textos, correspondientes a los documentos seleccionados, debido a la longitud de estos textos es más necesario mostrar como procesa los documentos y no tanto todo el contenido del texto extraído.

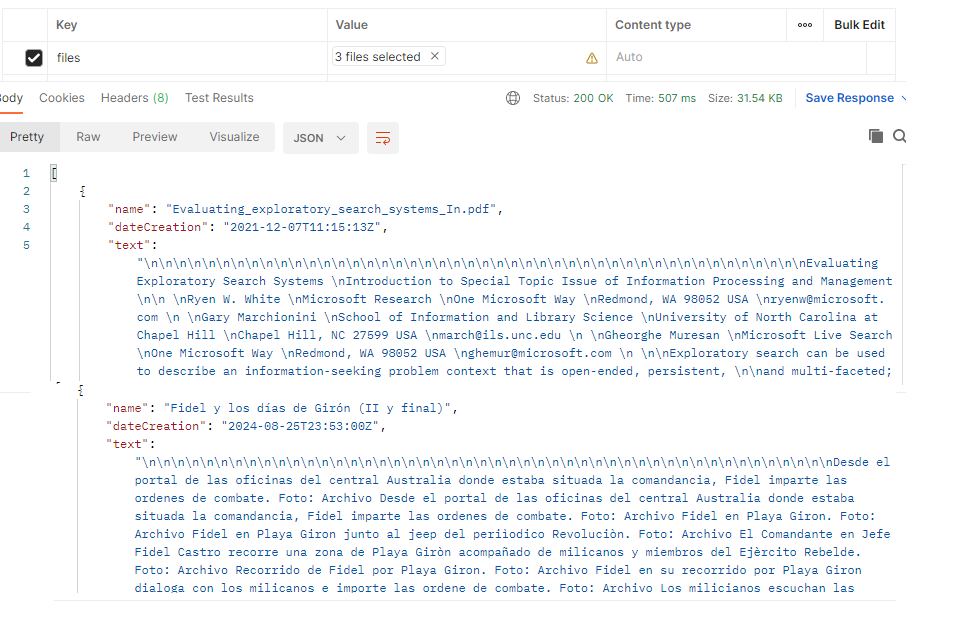


Figura 21: Resultado de procesamiento 3 documentos diferentes

Como se ha podido observar en estos ejemplos este servicio funciona correctamente, o sea, cumple con su objetivo, el cual es extraer el texto correspondiente a los archivos en el formato debido.

### Generación de grafo

Las entidades nombradas son frases o palabras en el texto que hacen referencia a personas, lugares, organizaciones, expresiones temporales, números, etc. El servicio de generación del grafo de conocimiento se encarga de extraer entidades y relaciones presentes en el texto, y devolverlo en formato JSON con una estructura simple que permita luego visualizar el grafo por alguna herramienta utilizada.

En la Figura 22 se muestra un texto de ejemplo para probar el funcionamiento de este servicio. En esta figura se señalan las entidades nombradas que se espera que reconozca el servicio.

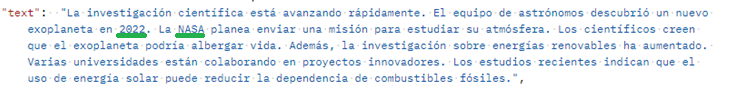


Figura 22:Texto ejemplo 1

En la imagen a continuación se muestra el resultado devuelto por la funcionalidad de extraer las relaciones del texto presentado en la Figura 22. Como se puede observar reconoce las entidades señaladas en el texto como esperadas, por lo tanto, hace una correcta extracción.



En la siguiente imagen se muestra el resultado que devuelve el análisis de relaciones asociado al texto de la Figura 22. El objetivo de esto es ayudar al usuario a comprender mejor la información que contiene el texto. En este caso la primera relación no ofrece mucha información, sin embargo, en el texto se habla de que la NASA planea enviar una misión. Con esta primera relación se puede inferir que en el texto se trata de investigaciones provenientes de la NASA (Figura 23). El resto de las relaciones extraídas si permiten comprender el contenido del texto abordado.

Para este texto se puede esperar resultados como:

* Equipo de astrónomos, descubrió, un nuevo exoplaneta
* La NASA, planea, enviar una misión
* Científicos, creen, que el exoplaneta podría albergar vida
* Varias universidades, están colaborando, en proyectos innovadores
* Estudios recientes, indican, que el uso de energía solar puede reducir la dependencia de combustibles fósiles

Estas frases ayudan a comprender el contenido del texto procesado, ya que brinda información como que los científicos creen que puede haber vida en el exoplaneta. En la imagen la palabra ***governor*** representa el sujeto, y la palabra ***dependent*** el objeto de la oración. Para entender mejor el significado de estas relaciones se muestra en la Figura 24 un ejemplo del grafo que se puede generar a partir de estos resultados.

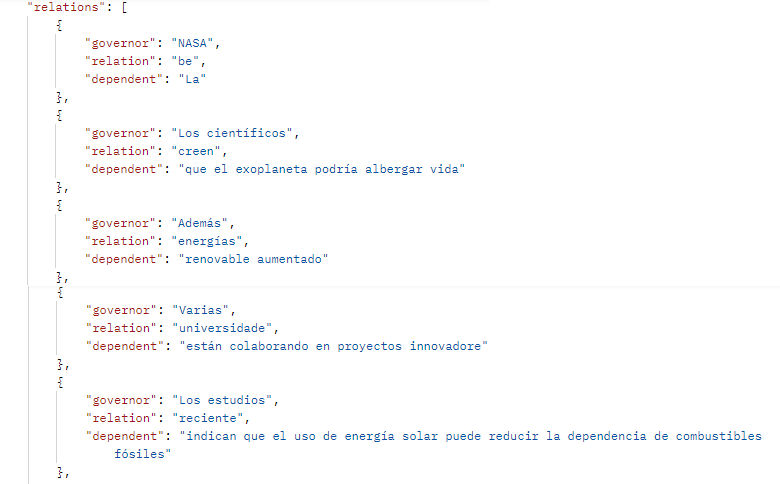


Figura 23: Relaciones texto ejemplo 1

En la imagen también se pudo observar que, con respecto a las frases esperadas, no se reconocen algunas como: “Equipo de astrónomos, descubrió, un nuevo exoplaneta” y “La NASA, planea, enviar una misión”. Esto se debe a que para el lenguaje español los modelos entrenados presentes en las bibliotecas no son tan específicos, debido a que el español es un lenguaje bastante complicado sintácticamente. Sin embargo, se reconocen otras frases como “Además, energías, renovables aumentado” y otras variedades de la frase referente a “Los estudios recientes…”. En la siguiente imagen se muestra estas características en más detalle. (Figura 24). A pesar de que las frases extraídas no son exactamente igual que las esperadas, se comprende el mensaje que estas transmiten.

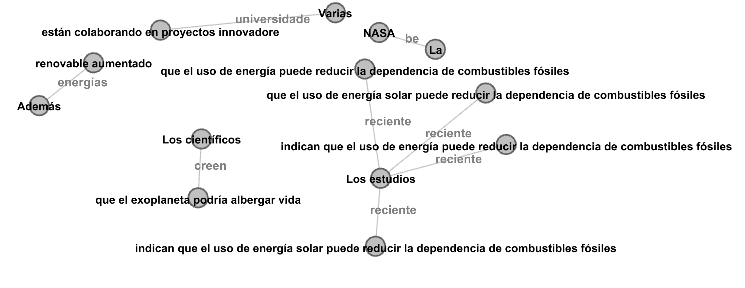


Figura 24: Grafo de texto Ejemplo 1

A continuación, se muestra con otro ejemplo el funcionamiento del servicio. El texto utilizado para este ejemplo se muestra en la siguiente figura. Las frases subrayadas son las entidades nombradas que se espera reconocer.

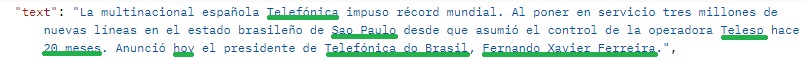


Figura 25: Texto ejemplo 2

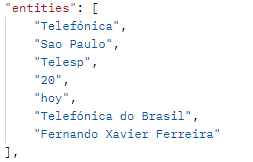
Como se observa en la imagen a continuación, Figura 26, el resultado obtenido difiere del esperado solo en un aspecto. Se espera reconocer 20 meses como una expresión temporal, sin embargo, el resultado lo reconoce como un número. 

Figura 26: Entidades extraídas texto ejemplo 2

En cuanto a las relaciones presentes en el texto (Figura 25) se puede esperar obtener algunas frases que brinden cierta información como:

* "Telefónica impuso un récord mundial."
* "El presidente de Telefónica do Brasil es Fernando Xavier Ferreira."
* "Tres millones de nuevas líneas han sido puestas en servicio."
* "Esto ocurrió en el estado brasileño de Sao Paulo."
* "Telefónica asumió el control de Telesp hace 20 meses."

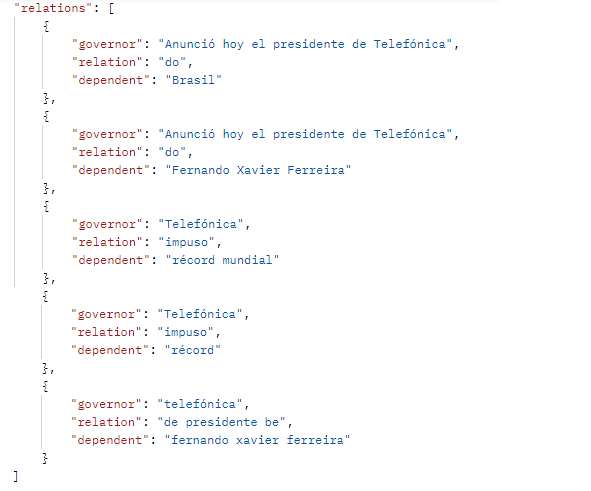


Figura 27: Relaciones texto ejemplo 2

Con el resultado obtenido se logra obtener la información referente a los 2 primeros puntos planteados en la información esperada, y en las dos primeras relaciones extraídas se entiende que “el presidente de la Telefónica do Brasil, Fernando Xavier Ferreira anunció esta noticia”. Sin embargo, no se obtiene información sobre el hecho de por qué la Telefónica obtuvo ese récord mundial, que se muestra en frases como: “poner en servicio tres millones de nuevas líneas”, “desde que asumió la operadora Telesp”.

En la imagen siguiente se muestra el grafo que puede generarse a partir de la información obtenida, para lograr comprender mejor cuales han sido las relaciones extraídas. Se representa la información descrita anteriormente. Por las conexiones entre nodos se infiere que Fernando Xavier Ferreira, es el presidente de la Telefónica do Brasil, y es quien anuncia la noticia de que la Telefónica española impuso el récord mundial. (Figura 25)

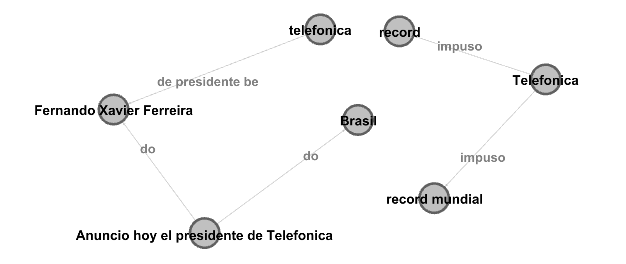


Figura 28:Grafo ejemplo texto 2

Para el siguiente ejemplo se hace una solicitud con ambos documentos y se obtienen los mismos resultados, pero ahora se muestra cómo se vería el grafo que se genera., al obtener estos dos resultados en conjunto.

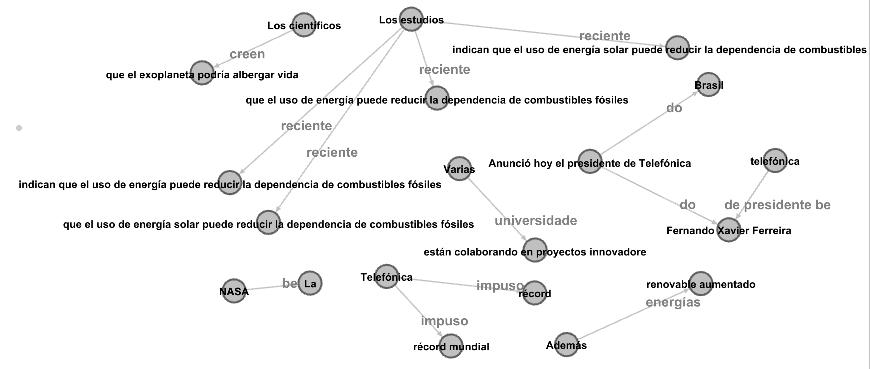


Figura 29: Grafo conjunto

En la imagen se observa una mayor cantidad de información. En este caso solo son dos textos pequeños, pero en caso de ser un mayor volumen de información, incluso una vista como esta, que facilita la extracción de información se vuelve complicada de visualizar en una pantalla. Por esta razón es importante garantizar el uso de una herramienta para visualizar el contenido, que permita al usuario la navegación a través del grafo y modificar la posición de los nodos, para poder obtener la información.

**Generación del grafo**

Para esta funcionalidad del servicio se necesita pasar por el *body* de la solicitud una estructura similar a las vistas anteriormente. Para ello se requiere que estos documentos enviados contengan la lista de entidades y la de relaciones, de no ser así, o de estas listas estar vacías, se obtiene un error. En la Figura 30 se muestra un ejemplo de lo que podría ser una solicitud incorrecta y en la Figura 31 se muestra el resultado que se obtiene al enviar esta solicitud.

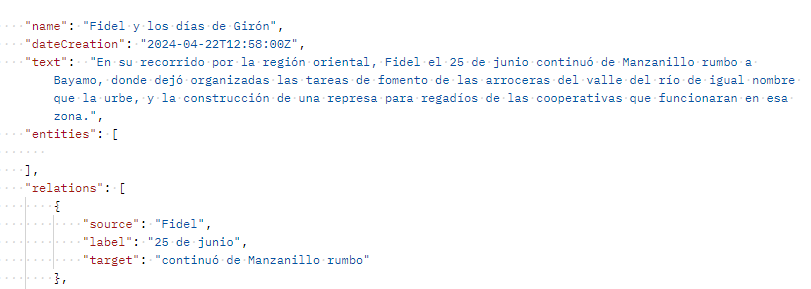


Figura 30: Ejemplo de solicitud incorrecta

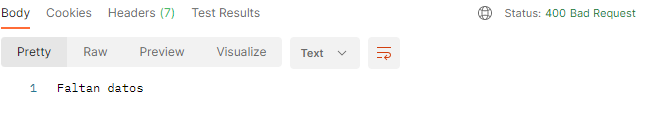


Figura 31: Resultado si no envía relaciones o entidades

Al enviar una solicitud correcta y ser procesada correctamente se devuelve un objeto que contiene una lista con todas las entidades encontradas en todos los documentos y una lista con todas las relaciones obtenidas. En la Figura 32 se muestra el resultado que se obtiene al procesar una solicitud correctamente.

Como se puede observar la propiedad *entities*, contiene todas las relaciones extraídas de los documentos, y la propiedad *relations* contiene todas las relaciones. Los atributos *source*, *target*, y *label* son comúnmente reconocidos por ese nombre, en las herramientas utilizadas para la visualización del grafo.



Figura 32: Resultado de solicitud correcta

### Almacenar en Solr

El servicio de almacenar en Solr se encarga de almacenar la información recopilada de los diversos documentos, luego del análisis completo esta información se almacena. Con una estructura como la presente en la Figura 33, el servicio luego de procesar correctamente ofrece el mensaje que se observa en la Figura 34.

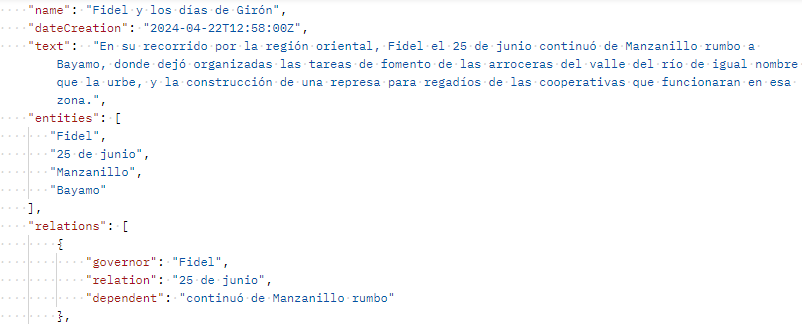


Figura 33:Estructura para enviar al servicio de Solr

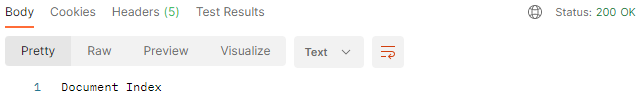


Figura 34:Resultado correcto de almacenar documentos

Los datos almacenados luego de que concluya el proceso exitosamente, se pueden observar en la Figura 35 como son estos datos almacenados en Solr. Al realizar una consulta a este servidor, los documentos serán devueltos con esta estructura. La funcionalidad de este servicio es ejecutada antes de la generación del grafo final, ya que, dada la respuesta de la consulta realizada al servidor, es que se obtiene la información real que será visualizada. Para la realización de esta prueba solo se tiene en cuenta el funcionamiento individual de esta función y no se realiza la consulta.

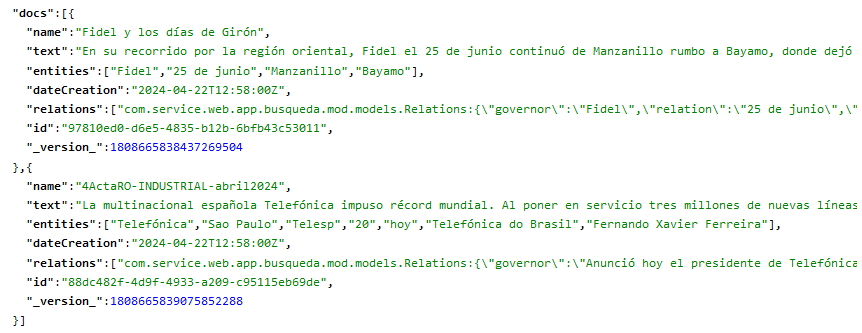
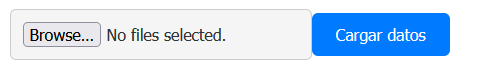


Figura 35:Documentos almacenados en Solr

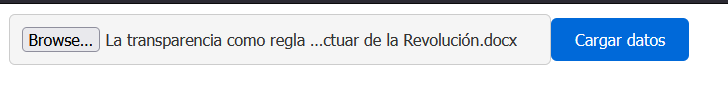
## Prueba con aplicación de ejemplo

Para probar como seria la comunicación con una interfaz de usuario se diseñó una aplicación de ejemplo. Esta aplicación está implementada en **react** y se utiliza la biblioteca **react-d3-graph** para la visualización del grafo de conocimiento. En las imágenes siguientes se muestran los aspectos la interfaz de la aplicación, es sencilla solo cuenta con los aspectos esenciales para que se ejecuten correctamente las funciones necesarias.

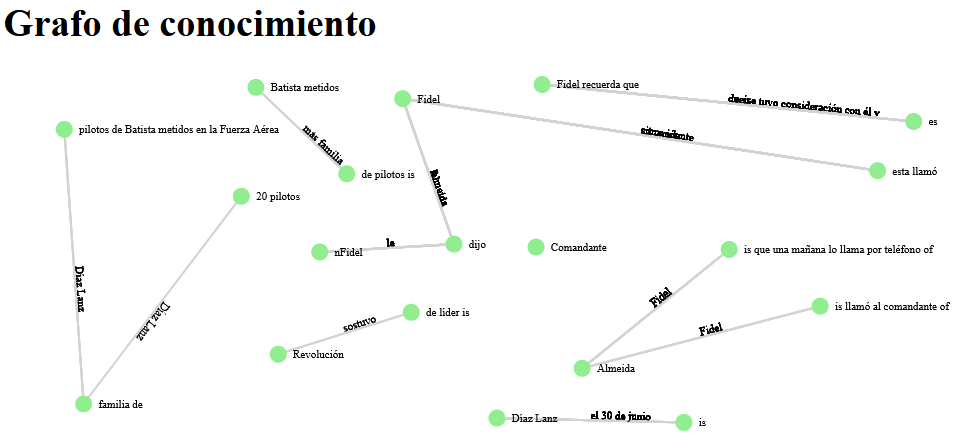
En la siguiente imagen se muestra el componente que permite seleccionar los archivos que se analizarán. Como es de plataforma web se utiliza la etiqueta ***input*** de tipo ***file.*** Se seleccionan los documentos usando el botón ***Browse*** y luego se **Cargan los datos**.



En la imagen a continuación se muestra el momento después de que se seleccionan los documentos que se desean analizar. Al presionar el botón de carga de datos, se ejecuta la llamada al servicio diseñado para realizar las solicitudes a los demás servicios y obtener una la respuesta con la información del grafo.



Luego de realiza la solicitud, y procesar el documento se obtiene la respuesta en formato JSON con los atributos necesarios para visualizar el grafo. Para esta biblioteca específicamente a cada una de las entidades existentes, se cambiaron a la forma “id:entidad” ya que para los nodos del grafo era necesario que tuviera este formato. En la siguiente imagen se muestra un grafo generado de la información de un documento de ejemplo seleccionado.



Esto es un ejemplo muy básico de cómo podría ser la comunicación del servicio con una aplicación web. Las especificaciones varían en dependencia de las herramientas y las tecnologías que se deseen utilizar, y el nivel de complejidad que se desee obtener en la interfaz de usuario. Al ser solo un a aplicación de prueba no se tiene en cuenta todos los aspectos importantes de las interfaces, como por ejemplo que la apariencia sea atractiva para el usuario o tal vez intuitiva.

## Conclusiones parciales

Con el desarrollo de este capítulo se cumple el objetivo de Evaluar el sistema de búsqueda y visualización:

* Se observa el funcionamiento de los servicios y ejemplos de las entradas que se deben enviar.
* Se muestra algunos errores que pueden ocurrir por falta de información necesaria.
* Se desarrolla una aplicación web sencilla, que muestra una forma en que puede establecerse la comunicación con los servicios, en este caso con el que se encarga de llamar a los otros servicios.
* Se muestran cuales son los componentes principales que están presentes en la interfaz, y son utilizados en la solución.

# Conclusiones

Con la realización de este trabajo se da cumplimiento al objetivo general.

* Se logra la implementación de un módulo de búsqueda exploratoria basado en la generación automática de grafo de conocimiento
* Se obtuvo una herramienta útil para organizaciones que necesiten analizar información y obtener comprensión de los datos relevantes.
* Se ha demostrado que con el uso de la búsqueda exploratoria se obtiene una reducción del tiempo y esfuerzo requerido para analizar documentos.
* Se presenta una estructura que permite la integración con herramientas de visualización de grafos de conocimiento para facilitar la representación de relaciones entre los datos.
* Se realizan pruebas para confirmar que el sistema cumple con los requisitos establecidos, que demuestra que la solución propuesta ayuda a satisfacer las necesidades de los usuarios en cuanto a extracción de información.
* Se contribuye al campo del manejo de información en organizaciones contemporáneas, proporcionando una base que puede ser utilizada por otras instituciones en la gestión del conocimiento.

# Recomendaciones

A pesar de que el trabajo cumple con los objetivos principales se plantean recomendaciones para mejoras futuras:

* Realizar pruebas de rendimiento, y ver el comportamiento de los servicios con mayor volumen de información.
* Entrenar los modelos para el idioma español, para obtener mayor precisión en el reconocimiento de relaciones entre entidades.
* Reducir el tiempo de respuesta en el análisis de relaciones.
* Realizar pruebas con documentos de organizaciones específicas.

# Referencias

[1] N. S. Dennler, S. Nikolaidis, y M. Mataric, «Using Exploratory Search to Learn Representations for Human Preferences», en *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 392-396.

[2] K. Athukorala, D. Głowacka, G. Jacucci, A. Oulasvirta, y J. Vreeken, «Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks», *Journal of the Association for Information Science and Technology*, vol. 67, n.o 11, pp. 2635-2651, 2016.

[3] R. W. White y R. A. Roth, *Exploratory search: Beyond the query-response paradigm*, n.o 3. Morgan & Claypool Publishers, 2009.

[4] B. Sarrafzadeh, «Supporting Exploratory Search Tasks through Alternative Representations of Information», 2020.

[5] R. Nararatwong, N. Kertkeidkachorn, y R. Ichise, «Knowledge graph visualization: Challenges, framework, and implementation», en *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, IEEE, 2020, pp. 174-178.

[6] A. L. Hernández, «Estrategia de Comunicación Interna de la Universidad de Matanzas», Tesis de Maestría, Universidad de La Habana, Facultad de Comunicación, 2018.

[7] I. Trelles, «Bases teóricos-metodológicas para la propuesta de modelo de gestión de comunicación en organizaciones.», Tesis de Doctorado, Universidad de La Habana, Departamento de Comunicación Social, 2002.

[8] L. D. Murphy, «Digital document in the workplace:an empirical investigation of document reuse and information technology infrastructure», Degree Doctor, Indiana University, Faculty Philosophy, 2000.

[9] P. Năstase, D. Stoica, F. Mihai, y A. Stanciu, «From document management to knowledge management», *Annales Universitatis Apulensis Series Oeconomica*, vol. 11, n.o 1, pp. 325-334, 2009.

[10] H. P. Grajales, *Documentos empresariales*. COOP. EDITORIAL MAGISTERIO, 2003.

[11] T. Tala, «Improving documentation quality management–a case study», 2016.

[12] H. Azarbonyad, R. Sim, y R. W. White, «Domain adaptation for commitment detection in email», en *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 672-680.

[13] F. R. Cobo Cuña, «METODOLOGÍA PARA EL ANÁLISIS DE LOS INDICADORES DE EFICIENCIA ECONÓMICA EN LA PRODUCCIÓN DE LECHE», Doctor Ciencias Veterinarias, INSTITUTO DE CIENCIA ANIMAL GRUPO DE BIOMATEMÁTICA, 2013.

[14] H. Morgan, «Conducting a qualitative document analysis», *The Qualitative Report*, vol. 27, n.o 1, pp. 64-77, 2022.

[15] R. H. Sprague Jr, «Electronic document management: Challenges and opportunities for information systems managers», *MIS quarterly*, pp. 29-49, 1995.

[16] C. He, L. Micallef, B. Serim, T. Vuong, T. Ruotsalo, y G. Jacucci, «Interactive visual facets to support fluid exploratory search», en *Proceedings of the 14th International Symposium on Visual Information Communication and Interaction*, 2021, pp. 1-10.

[17] Z. Huang y L. Yuan, «Enhancing learning and exploratory search with concept semantics in online healthcare knowledge management systems: An interactive knowledge visualization approach», *Expert Systems with Applications*, vol. 237, p. 121558, 2024.

[18] K. Haller, F. J. Ekaputra, M. Sabou, y F. Piroi, «Enabling Exploratory Search on Manufacturing Knowledge Graphs.», en *VOILA@ ISWC*, 2022, pp. 16-28.

[19] G. Marchionini, «Exploratory search: from finding to understanding», *Communications of the ACM*, vol. 49, n.o 4, pp. 41-46, 2006.

[20] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, y J. Taylor, «Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it’s done», *Queue*, vol. 17, n.o 2, pp. 48-75, 2019.

[21] M. Lissandrini, D. Mottin, K. Hose, y T. B. Pedersen, «Knowledge graph exploration systems: are we lost?», en *Annual Conference on Innovative Data Systems Research*, 2022.

[22] M. Al-Tawil, V. Dimitrova, D. Thakker, y B. Abu-Salih, «Emerging Exploration Strategies of Knowledge Graphs», *IEEE Access*, 2023.

[23] «W3C». [En línea]. Disponible en: https://www.w3.org

[24] A. Garcia-Robledo y M. Zangiabady, «Dash Sylvereye: A Python Library for Dashboard-Driven Visualization of Large Street Networks», *IEEE Access*, 2023.