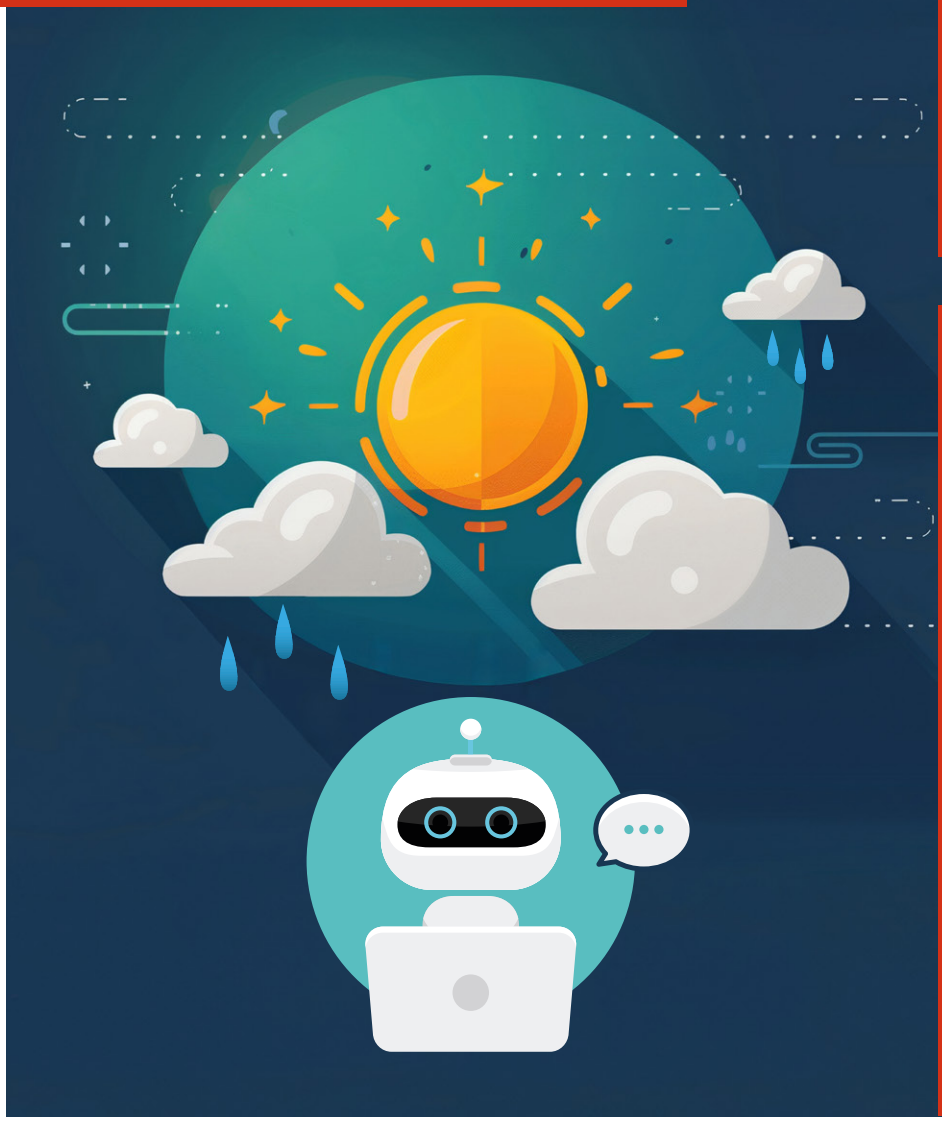
Dataclima:

Avanzando en la gestión de datos climáticos



FRANCISCA MUÑOZ BRAVO

Magister en Ciencias de la Computación por la Vrije Universiteit Brussel, Bélgica. Actualmente se desempeña como Jefa de Datos y Cómputos del Centro de Ciencia del Clima y Resiliencia (CR2) de la Universidad de Chile. Sus líneas de especialización son: gobernanza de datos climáticos, datos climáticos abiertos, acción climática basada en evidencia, visualización de datos y recuperación de datos históricos.

franmuno@uchile.cl

MARÍA CECILIA BASTARRICA

PhD Computer Science and Engineering, University of Connecticut. Profesora Asociada del Departamento de Ciencias de la Computación de la Universidad de Chile. Académica a cargo del curso Proyecto de Software. Líneas de investigación: ingeniería de software, líneas de productos de software, mejora de procesos de software, desarrollo de software dirigido por modelos.

cecilia@dcc.uchile.cl

RESUMEN. El Centro de Ciencia del Clima y Resiliencia (CR2) es un centro de investigación avanzado localizado en la Facultad de Ciencias Físicas y Matemáticas. Alumnos de Ingeniería Civil en Computación han desarrollado una herramienta basada en un ChatBot implementado en Llama 3 que le permite disponibilizar sus bases de datos y herramientas a toda la comunidad científica y encargados de tomar decisiones de políticas públicas.

Centro de Ciencia del Clima y Resiliencia CR2

El Centro de Ciencia del Clima y Resiliencia CR2 (www.cr2.cl) financiado por ANID a través de su programa de centros de excelencia en áreas prioritarias (FONDAP), inició sus actividades el año 2013 y se espera que continúe por los próximos 5-10 años. Con la Universidad de Chile como institución patrocinante, y la Universidad Austral y la Universidad de Concepción como asociadas, el CR2 se ha establecido como un referente en la investigación climática y ambiental en Chile.

La investigación del CR2 está orientada a profundizar la comprensión del sistema terrestre a través de la ciencia, contribuir a aumentar la capacidad de resiliencia en Chile y apoyar la formulación de políticas públicas y decisiones informadas por evidencia científica. El Centro aborda cinco líneas de investigación: agua y extremos, cambio de uso de suelo, ciudades resilientes, gobernanza e interfaz ciencia-política, y zona costera. Además, maneja temas integrativos transversales, entre los cuales destaca actualmente la "carbono neutralidad". Recientemente, el centro ha completado proyectos sobre floraciones algales nocivas, seguridad hídrica, e incendios.

Figura 1. Plataformas públicas de CR2.

Además de generar publicaciones científicas, el CR2 contribuye abiertamente a la comunidad con la creación y mantenimiento de bases de datos y aplicaciones para servicios climáticos abiertos. Actualmente, el centro cuenta con más de 20 plataformas y bases de datos que contienen información sobre riesgos climáticos, mitigación y adaptación desde una perspectiva tanto física como social, disponibles para investigadores, tomadores de decisiones y el público general. Estas herramientas son un aporte para facilitar el acceso a datos relevantes al clima, resultados de investigación y apoyar las estrategias de adaptación y mitigación en respuesta a los desafíos del cambio climático.

Bases de datos

El trabajo de investigación demanda la gestión de grandes volúmenes de datos para su visualización y análisis, especialmente en el campo de las ciencias de la tierra. En el caso de las ciencias sociales, los datos suelen ser muy complejos y difíciles de estandarizar. Estos datos están muchas veces dispersos y en una multiplicidad de formatos, lo que ha motivado al CR2 a compilarlos, etiquetarlos apropiadamente, estandarizarlos cuando es posible, y generar productos que los pongan a disposición del público en repositorios especializados.

Servicios climáticos

En el CR2 se han hecho importantes esfuerzos para comunicar la información e investigación científica a través de productos y aplicaciones útiles para diversos tipos de usuarios a nivel nacional. Estas plataformas no sólo proporcionan herramientas valiosas para los académicos del CR2, sino que también han sido abiertas a toda la comunidad. Constituyen uno de los legados más valiosos del centro y son un ejemplo

La investigación [climática] del CR2 está orientada a [...] apoyar la formulación de políticas públicas y decisiones informadas por evidencia científica.

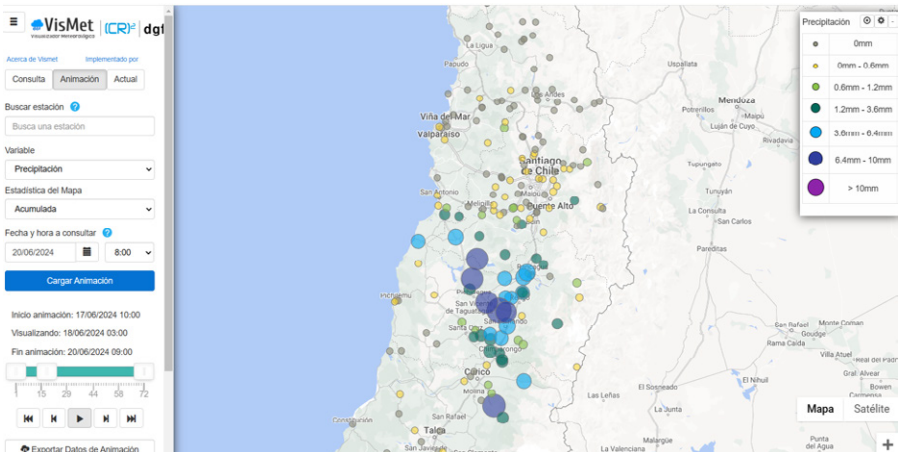


Figura 2. Plataforma VisMet.

destacado de interfaz ciencia-política y ciencia-sociedad civil. Estas plataformas son ampliamente utilizadas tanto por investigadores, servicios operativos y la comunidad en general, facilitando la comunicación entre estos distintos usuarios (ver Figura 1).

Destacamos el VisMet (vismet.cr2.cl) y el R-Explorer (rexplorer.cr2.cl), desarrollados por estudiantes de años anteriores del curso Proyecto de Software del Departamento de Ciencias de la Computación de la Universidad de Chile. VisMet (ver Figura 2) permite el acceso y visualización de datos meteorológicos en tiempo real (desde el año 2000 en adelante) provenientes de más de 800 estaciones en territorio nacional, operadas por diversas instituciones como DMC, DGA, AgroMet y CEAZA. Esta plataforma recibe más de 4000 visitas mensuales. Por su parte, R-Explorer es una herramienta más especializada que permite visualizar y descargar variables atmosféricas para momentos y áreas específicas, obtenidas del reanálisis atmosférico ERA5/ECMWF. Ambas plataformas se consideran referentes a nivel nacional en la publicación de datos climáticos relevantes con posibilidad de generar estadísticas, mapas y series de tiempo relacionados.

Dataclima

La gestión eficiente de datos climáticos es una parte fundamental para implementar las acciones dirigidas a mitigar y adaptarse al cambio climático. Los datos relevantes incluyen no sólo aquellos que describen los cambios promedios del sistema terrestre o la caracterización de eventos extremos, sino también los que apoyan acciones climáticas como la evaluación, planificación y respuesta al cambio climático. Tradicionalmente debido a la complejidad y multiplicidad de actores, se ha observado la fragmentación y a veces duplicidad de esfuerzos, junto con falta de actualización y asignación de metadatos adecuados. Esto dificulta significativamente el acceso y la usabilidad (o reutilización) de la información.

A lo largo de sus 10 años de existencia, el CR2 ha contribuido al panorama con numerosas plataformas y bases de datos, que se suman a las ya existentes. Esta abundancia, sin embargo, puede resultar abrumadora, dificultando que tanto el público general como los investigadores de diversas disciplinas, encuentren la información que les permita trabajar de manera efectiva en la investigación o en el apoyo a la toma de decisiones.

En este contexto, es crucial simplificar la conexión entre las preguntas sobre los distintos ámbitos del cambio climático, los riesgos asociados, las medidas de mitigación y adaptación relacionadas, y la información científica disponible. Además, identificar qué plataformas tienen objetivos similares o complementarios, y qué preguntas aún no pueden responderse con los datos o plataformas existentes, permite descubrir áreas clave a revisar o potenciar en investigaciones futuras.

De qué se trata el sistema

En primavera de 2023, un nuevo equipo del curso Proyecto de Software del Departamento de Ciencias de la Computación formado por Cristóbal Alcaíno, Lucas Amión, Felipe Escárate, Christopher Marín, Michael Scott, Álvaro Toro y Néstor Valenzuela, abordó el diseño y construcción de esta herramienta en conjunto con el área de Datos y Cómputos del CR2.

Dataclima es una plataforma diseñada para mejorar la accesibilidad y comprensión de datos climáticos mediante la estandarización y compilación de metadatos, y el uso de un chatbot especializado. Este sistema facilita la exploración y análisis de datos climá-

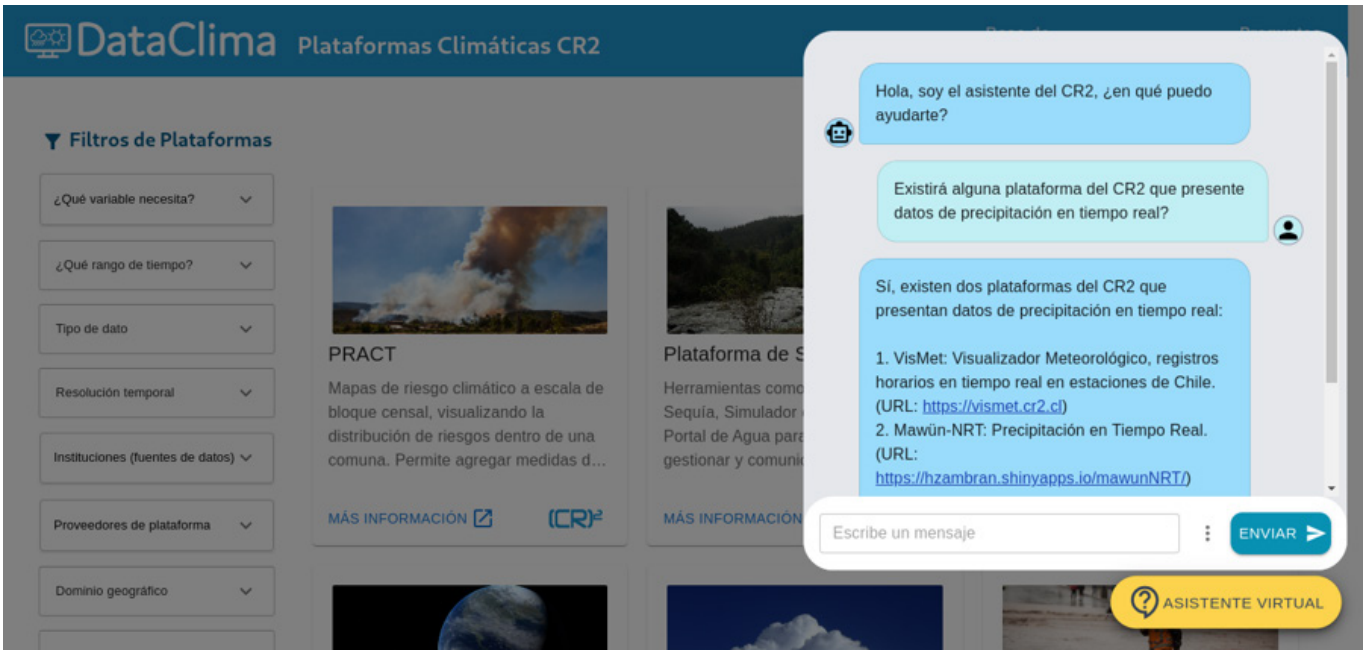


Figura 3. Chatbot de la plataforma Dataclima.

Dataclima es una plataforma diseñada para mejorar la accesibilidad y comprensión de datos climáticos.

ticos a través de la asignación y revisión de metadatos predefinidos, permitiendo a los usuarios relacionar sus áreas de interés específicas con bases de datos y plataformas relevantes. El chatbot ofrece recomendaciones personalizadas, mejorando la interacción y el aprovechamiento de los recursos disponibles (ver Figura 3).

La plataforma se enfoca en simplificar la búsqueda y el filtrado de fuentes importantes de datos climáticos, tanto nacionales como internacionales, contribuyendo a un proceso más amplio que busca optimizar el acceso a información pertinente sobre el cambio climático en Chile. Para la sección del chatbot interactivo, Dataclima.cl incorpora herramientas avanzadas de machine learning, como algoritmos que clasifican y agrupan contenido por similitud y técnicas de procesamiento de lenguaje natural, para proporcionar recomendaciones precisas y adaptadas a las necesidades de los usuarios, facilitando así una interfaz más intuitiva y accesible para una amplia diversidad de usuarios.

Cómo está construido el sistema

Para implementar el chatbot de Dataclima se experimentó con diferentes modelos de lenguaje (LLMs), incluyendo APIs externas como la de OpenAI. Finalmente, se optó por el recientemente presentado Llama 3, desarrollado por Meta, ya que es de código abierto y se puede ejecutar de manera local. Sin embargo, esto requirió la adquisición de una nueva tarjeta gráfica Nvidia T4 que se pudiera dedicar exclusivamente a la ejecución del modelo de lenguaje. Esto resulta indispensable para poder usar el modelo Llama 3 de forma fluida.

Se utiliza la técnica de Retrieval Augmented Generation (RAG) para lograr que el chatbot responda adecuadamente a las preguntas de los usuarios, utilizando como bases una serie de documentos de texto que contienen la información sobre las plataformas y bases de datos climáticas almacenadas en Dataclima. Esta técnica se basa en primero calcular los *embeddings* de cada uno de los documentos, es decir, estructuras vectoriales multidimensionales que capturan el significado semántico de los datos. Estos *embeddings* son almacenados en una base de datos vectorial diseñada para recuperar rápidamente los documentos más relacionados con cada consulta del usuario, disminuyendo así la cantidad de información relevante que debe considerar el modelo de lenguaje natural para elaborar su respuesta. En este caso se utilizó el modelo

[Dataclima] también apunta a fomentar la transparencia y trazabilidad en la toma de decisiones.

Multilingual-E5-large para el cálculo de *embeddings* y FAISS como base de datos vectorial, ya que ambos son de código abierto. Toda esta lógica es manejada mediante la librería Langchain.

Cómo luce la plataforma

Una de las características destacadas de la plataforma es la herramienta de filtro, ubicada a la izquierda de la pantalla, que permite a los usuarios mejorar la información de las plataformas seleccionadas, ya que se pueden escoger múltiples criterios a partir de la información requerida, actualizando en tiempo real las plataformas que cumplen con dichos criterios.

Además, Dataclima cuenta con un WebSocket que se conecta directamente al chatbot, el cual está preconfigurado para manejar consultas específicas sobre las plataformas y bases de datos del CR2 y sus instituciones asociadas. Esta funcionalidad enriquece la experiencia de los usuarios, entregando recomendaciones interactivas, permitiendo a los usuarios ajustar y refinar sus consultas en tiempo real para explorar diferentes aspectos en los temas de interés.

Valor e impacto

Dataclima espera facilitar el acceso a un subconjunto relevante dentro de un repositorio "curado" de información climática. También apunta a fomentar la transparencia y trazabilidad en la toma de decisiones. Al consolidar datos de diversas fuentes en un repositorio estandarizado, la plataforma espera que cada usuario pueda comprender el origen y la metodología detrás de los datos que consulta. Esto no sólo promueve la reutilización e integración de datos, sino que también facilita la identificación de oportunidades de colaboración en diversas disciplinas y en distintos niveles, desde lo local hasta lo internacional.

Curso Proyecto de Software

Dataclima es una herramienta de alta relevancia para potenciar el impacto del CR2 en la comunidad científica. Tal como ocurre con Dataclima, cada semestre, alumnos del último año de la carrera de Ingeniería Civil en Computación abordan la construcción de herramientas de software de distintas organizaciones (portalpsw.dcc.uchile.cl). Desde hace más de 20 años, los equipos de alumnos han desarrollado aplicaciones para empresas públicas o privadas, fundaciones, institutos de investigación o distintas reparticiones de la Universidad de Chile.

El Departamento de Ciencias de la Computación está involucrado de forma permanente con la industria chilena del software desarrollando soluciones de muy alta calidad, complejidad e impacto. ■

Agradecimientos.
Agradecemos a Sebastián Villalón y Lucas Amión, ingenieros de datos y plataformas de CR2, por su participación en el trabajo y escritura del artículo.