

BrEarth: Tecnología de almacenes de datos para el cerebro de la tierra^{*}

David Martínez, José R.R. Viqueira^[0000-0002-1539-3746], and José A. Taboada^[0000-0003-1897-1537]

Computer Graphics and Data Engineering (COGRADE), Centro Singular de Investigación en Tecnologías Intelixentes (CiTIUS), Departamento de Electrónica e Computación, Universidade de Santiago de Compostela (USC), Santiago de Compostela, A Coruña. {david.martinez.casas, jrr.viqueira, joseangel.taboada}@usc.es
<https://citi.usc.es/investigacion/programas-cientificos/procesamiento-aproximado>

Resumen En este artículo se describen de forma breve los requisitos de un framework de desarrollo de almacenes de datos en el ámbito de la observación y modelado ambiental.

Keywords: Datos Inteligentes · Datos Geoespaciales · Datos Medioambientales · Análisis Exploratorio

1. Introducción

La principal fuente del actual diluvio de datos geoespaciales es la gran variedad de infraestructuras existentes de observación y modelado ambiental. Estas infraestructuras incluyen plataformas complejas de sensorización y modelado, normalmente implementadas por administraciones públicas. Pero en la actualidad, también se incorporan iniciativas colaborativas de crowdsensing, algunas de ellas explotando la gran cantidad de datos proporcionados por las interacciones humanas en redes sociales.

Desafortunadamente, existen todavía barreras importantes que restringen el uso de los datos generados a comunidades reducidas y muy especializadas de usuarios. Estas barreras incluyen la dificultad o imposibilidad para acceder a algunos conjuntos de datos y la dificultad de su interpretación y análisis debido a la falta de los metadatos necesarios.

Debido a todo lo anterior, se necesita una nueva tecnología de almacenes de datos que permita el desarrollo progresivo de las soluciones de analítica inteligente geoespacial necesarias en cada dominio de aplicación. En otras palabras, es necesario el desarrollo de la base tecnológica para la *Memoria de la Tierra*

^{*} Este trabajo ha sido cofinanciado por (i) el proyecto TRAFair (017-EU-IA-0167), cofinanciado por CEF EU, (ii) el proyecto RADAR-ON-RAIA (0461_RADAR_ON_RAIA_1.E), cofinanciado por FEDER a través del programa INTERREG V-A España-Portugal (POCTEP) 2014-2020 y por (iii) la Xunta de Galicia a través de las ayudas de grupo de potencial crecimiento (ED431B 2018/28).

que dé soporte a los procesos de razonamiento y aprendizaje llevados a cabo por el *Cerebro de la Tierra*, basados en todos los *Sentidos de la Tierra*, incluyendo los sentidos naturales de los seres humanos y los sentidos artificiales de todos los procesos de observación de la tierra. Esta idea puede verse gráficamente representada en la Figura 1.

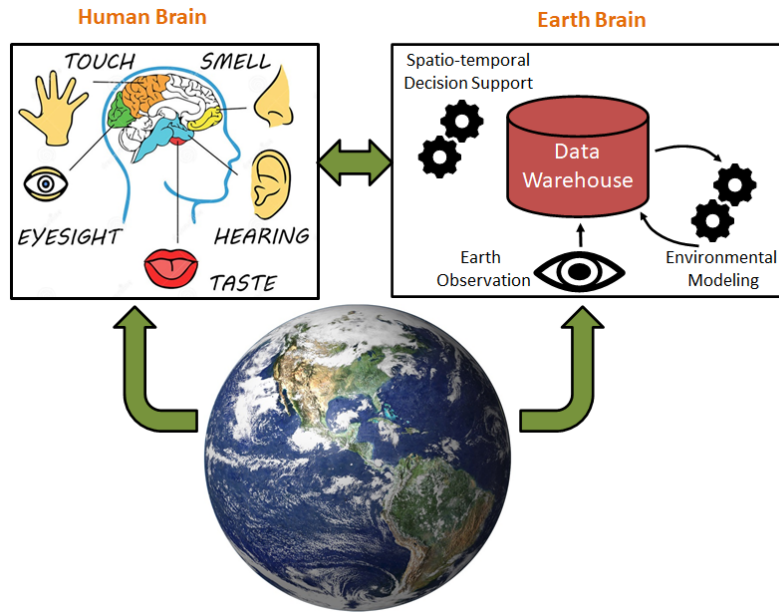


Figura 1. Un almacén de datos para el cerebro de la tierra.

2. Antecedentes

Clásicamente, los métodos principales de estimación de las principales variables geofísicas que determinan el estado de nuestro medioambiente son las plataformas de observación de la tierra (sistemas de sensorización) y las infraestructuras de modelado ambiental. Hoy en día sin embargo, el uso generalizado de dispositivos móviles por un porcentaje elevado de la población está fomentando la adopción de soluciones de sensorización centradas en las personas, que aprovechan tanto los sensores de los propios dispositivos móviles como los sentidos de sus propietarios. En este contexto, el término *Crowdsensing Móvil* se utiliza cuando una comunidad grande de usuarios se involucra en un proceso de observación, y un subconjunto importante de los entornos de este tipo está apoyado por la monitorización de las interacciones en redes sociales.

Una característica importante de los datos generados por los procesos anteriores es su naturaleza compleja de tipo espacio-temporal. Existen dos grandes

tipos de entornos de modelado de datos para representar estos tipos de datos, que o bien consideran objetos con propiedades espaciales representadas con geometrías construidas mediante vectores de coordenadas (entidades vectoriales), o bien representan la variabilidad espacio-temporal de las propiedades usando mallas regulares (coberturas raster). El modelo de referencia del Open Geospatial Consortium (OGC) considera las dos aproximaciones anteriores, dado que se adaptan mejor a distintos tipos de conjuntos de datos geoespaciales. No existe un modelo, formato y codificación único que sea eficaz y eficiente con ambos tipos de datos, y por lo contrario se utilizan soluciones distintas para cada tipo de conjunto de datos. Así, por ejemplo, los formatos GML y GeoJSON y el estándar SFS, implementado por sistemas gestores de bases de datos espaciales, son apropiados para la representación de entidades vectoriales. Por otro lado, las coberturas raster se suelen representar utilizando formatos como GeoTIFF o NetCDF.

La correcta interpretación y análisis de los datos solo se puede realizar si estos están acompañados de los metadatos que aporten la semántica necesaria, tanto relacionada con la procedencia de los datos como con su contexto. En el ámbito de los datos medioambientales se han propuesto varios modelos para la representación de estos metadatos, cuyas características principales se han incorporado ya al modelo estándar Observations and Measurements (O&M) del OGC [2].

Por último, el OGC ha definido también interfaces de servicios web estandarizadas para el descubrimiento y acceso a datos geoespaciales, que incluyen el estándar WFS para acceder a datos de entidades vectoriales y el estándar WCS para acceder a datos de coberturas raster. También se han propuesto estándares de acceso en el ámbito de los datos medioambientales, como por ejemplo el OpenNDAP o el NetCDF Subset. Ninguno de los estándares anteriores incorpora la semántica específica definida por los metadatos arriba mencionados. Por el contrario, esta semántica sí está incorporada en la definición del estándar de servicio Sensor Observation Service (SOS) del OGC [1].

3. Objetivos

Las tecnologías disponibles en la actualidad en el ámbito de los datos geoespaciales y medioambientales permiten a expertos en TIC implementar infraestructuras de almacenamiento y acceso a datos eficientes. Así, por ejemplo, no es complicado crear un almacén de datos geoespaciales para entidades vectoriales utilizando un sistema gestor de bases de datos espaciales que implemente el estándar SFS del OGC. Tampoco es complejo hoy en día combinar los datos vectoriales almacenados en la base de datos con datos de coberturas raster para proporcionar implementaciones eficientes de servicios WFS y WCS. Sin embargo, para incorporar la semántica demandada por los procesos de análisis e integración de datos, es necesario utilizar modelos y vocabularios que proporcionen los metadatos necesarios.

La implementación directa de las estructuras de datos del modelo O&M en sistemas gestores de bases de datos espaciales ya se ha intentado y existen unas pocas herramientas con esta aproximación. Pero dan lugar a soluciones con capacidades limitadas (no proporcionan soporte para observaciones remotas con forma de coberturas raster), difíciles de extender y adaptar a las necesidades de aplicaciones concretas, y con una eficiencia muy pobre en el acceso a los datos. Por otro lado, sí que existen infraestructuras de datos específicas en proyectos concretos basadas en implementaciones a medida que son eficientes, pero su coste es alto y son poco reutilizables.

El principal objetivo de este trabajo es el diseño e implementación de una nueva tecnología de almacén de datos que permita la integración de distintas fuentes de datos de observación y modelado ambiental. La solución debe ser capaz de integrar datos generados por cualquier tipo de proceso, incluyendo estimaciones complejas con forma de cobertura. Debe ser flexible en la incorporación de propiedades específicas de cada dominio de aplicación, tanto en los metadatos como en los datos. Para conseguir esto, se definirá un Lenguaje Específico de Dominio (DSL) para la definición del esquema del almacén. Dicho esquema tendrá la capacidad de almacenar todos los metadatos considerados por el estándar O&M del OGC, permitirá representar todas las propiedades de cada dominio de aplicación y habilitará el acceso eficiente. La gestión de los datos estará basada en una extensión del estándar SOS del OGC. Para demostrar la capacidad de reutilización del almacén en distintas aplicaciones, se implementará un cliente genérico de descubrimiento, búsqueda y exploración de datos, que aprovechará la semántica específica incorporada en el almacén. La solución se probará en tres casos de uso relacionados con tres infraestructuras de datos abiertos de tres proyectos distintos, desarrolladas en el grupo de investigación COGRADE, del CiTIUS.

4. Conclusiones

Se estima que este trabajo aporte una primera experiencia para analizar las ventajas e inconvenientes que tiene el movimiento de la semántica del dominio de los datos medioambientales dentro del almacén de datos, contribuyendo al debate entre mover el esquema y la semántica más hacia la aplicación o más hacia el almacén. Es decir, lo que llamamos *Big Data* con un *Data Lake* que solo tiene datos, incluso sin esquema, y los que podemos llamar *Smart Data*, con un almacén que además de tener esquema, tiene estructuras con una semántica estandarizada y conocida en el dominio de aplicación.

Referencias

1. Bröring, A., Stasch, C., Echterhoff, J.: OGC sensor observation service interface standard. version 2.0. OpenGIS implementation standard., Open Geospatial Consortium Inc. (2012)
2. Cox, S.: Observations and measurements. version 2.0. The OpenGIS abstract specification., Open Geospatial Consortium Inc. (2013)