Una estructura de almacenamiento de series de datos en bases de datos espaciales

María Isabel Marín Morales, Fabio Alberto Vargas Agudelo, Dario Enrique Soto Durán
mmarinm2@tdea.edu.co, fvargas@tdea.edu.co, dsoto@tdea.edu.co

Institución Universitaria Tecnológico de Antioquia, Colombia
Medellín - Colombia

Resumen: *En la última década, la captura de información geo-referenciada se intensificó considerablemente gracias al rápido desarrollo y disposición de los instrumentos de adquisición de esta información. Estos datos corresponden normalmente a modelos vector, raster y a series de datos asociadas a estos dos últimos. Esta situación lleva a la necesidad de integración de la información capturada desde las diferentes fuentes con el fin de complementarla y servirla para tareas de análisis y toma de decisiones. La integración de la información se logra mediante mecanismos de interoperabilidad, por un lado en el desarrollo de los programas que la gestionan y por otro en la descripción, almacenamiento y diseminación de los datos. Esfuerzos a nivel mundial vienen logrando que esta integración sea posible, en especial el OGC, una agrupación de empresas, universidades y entidades gubernamentales, viene proponiendo especificaciones en torno a la gestión de información geográfica, consensadas y discutidas al interior del mismo. De esta manera, las especificaciones OGC proporcionan lineamientos para la gestión de los datos geográficos usando SQL en Sistemas de Gestión de Bases de Datos relacionales. Esta especificación proporciona un soporte para el almacenamiento de los modelos de datos vector y raster, aunque no así para la asociación a éstos de series de datos. Por lo anterior, en este trabajo, se presenta una propuesta de estructura de almacenamiento de series de datos en bases de datos espaciales basada en la especificación OpenGIS Simple Feature SQL. La especificación se probó en la implementación PostGIS para la gestión de datos geográficos.*

Abstract: *In the last decade capturing the geo-referenced information was considerably intensified by the rapid development and disposition of the instruments of acquiring this information. These are usually vector and raster data models and datasets associated with the latter two. This situation leads to the need for integration of information captured from many sources to supplement it and serve it for tasks of analysis and decision-making. The information integration is achieved through interoperability mechanisms, with the development of management systems and on the other hand with the description, storage and dissemination of data. Global efforts are achieving this integration possible, especially the OGC, a consortium of companies, universities and government agencies, is proposing specifications regarding the management of geographic information and discussed the consensuses inside. Thus, the OGC specifications provide guidelines for the management of geographic data using SQL Management Systems Relational Database. This specification provides support for storing models of vector and raster data but not for the association of these datasets. Therefore, this paper presents a proposed storage structure of data sets in spatial databases based on the OpenGIS Simple Feature SQL specification. The specification was tested in the implementation PostGIS for managing geographic data.*

Palabras clave: Series de datos, Sistemas de Información Geográfica, Open Geospatial Consortium.

1. Introducción

Los sistemas de información geográfica (SIG) se vienen convirtiendo en decisivos programas de gestión de información georeferenciada en dominios de aplicación como las ciencias de la tierra, la atmósfera y el océano. Esta gestión comprende procesos de análisis, predicción, estudios socioeconómicos y ecológicos, entre otros, para lo que resulta necesaria la integración con series de datos [Yang99]. La información geográfica soportada por los SIG se despliega tradicionalmente en los formatos raster y vector. Adicionalmente, se emplean series de datos para complementar esta información con el fin de realizar análisis de series asociadas a un lugar en el espacio. Las series se pueden clasificar principalmente en series de tiempo, campañas eventuales de medición y perfiles de datos en la columna de agua o de aire. Esta información se obtiene en campañas de muestreo, estaciones estáticas de medición, captura periódica de datos satelitales, entre otras.

El uso de las series de datos, especialmente las series de tiempo, es el núcleo de la toma de decisiones de grandes empresas y entidades en Colombia y en el mundo. Empresas como ISAGEN, EPM, las Corporaciones Autónomas Regionales; los Planes de Ordenamiento Territorial, los Planes de Ordenamiento y Manejo de cuencas, los proyectos de Predicción de riesgos, entre muchos otros, son ejemplo de esto. A su vez, estas empresas necesitan el análisis espacial de la información y la realizan a través de SIG, encontrando dificultades en el proceso de unir estas dos. Dificultades traducidas en tiempos adicionales en el proceso de diseño de bases de datos espaciales y el procesamiento aislado de la información que conlleva a la toma de decisiones aisladas [Anaya10].

Actualmente, el Open Geospatial Consortium [OGC12], una agrupación internacional de compañías, agencias gubernamentales y universidades que participan en un proceso de consenso para desarrollar estándares de interfaces a disposición del público, trabaja en la estandarización del almacenamiento y transporte de la información bajo los modelos de datos vector y raster. Por otro lado, existen esfuerzos enfocados en el almacenamiento de las series de datos, como es el caso

del formato NetCDF desarrollado por la comunidad Unidata [Unidata12] desligado de los sistemas de información geográfica, a pesar del carácter espacial de la información.

Algunos proyectos han tratado de incorporar las series de datos al interior de los SIG, pero este trabajo siempre requiere de un esfuerzo adicional para definir la manera para almacenar las series de datos, no así con los modelos de datos vector y raster que ya están estandarizados por el OGC, éste es el caso de HidroSIG [HidroSIG10], un SIG para la gestión de la información geográfica hidrológica, en donde se emplearon archivos XML asociados a cada estación para almacenar y desplegar las series de datos.

Por lo anterior, en este artículo, se presenta una propuesta de estructura de almacenamiento de series de datos en bases de datos geográficas, basados en las sugerencias OGC para el almacenamiento de objetos geográficos en bases de datos relacionales, recomendación implementada en la extensión PostGIS para PostgreSQL.

El resto de este artículo se estructura de la siguiente manera. En la sección 2, se muestra el marco teórico que permite contextualizar los principales conceptos SIG trabajados en este artículo y de manera resumida algunos trabajos previos en el tema. En la sección 3, se expone la estructura propuesta de almacenamiento de series de datos en bases de datos espaciales; en la sección 4, se presenta la implementación de la estructura propuesta sobre la extensión para bases de datos espaciales PostGIS y se discuten las diferentes bondades de la estructura propuesta evidenciada en la implementación realizada. Finalmente, en la sección 5, se exponen las conclusiones y el trabajo futuro que se puede derivar de este artículo.

2. Teoría del dominio y trabajos previos

Los datos geográficos representan los objetos del mundo real e.g. carreteras, ríos, divisiones políticas, la precipitación, elevación de una región y se pueden dividir en dos grupos de abstracciones: objetos discretos (e.g. una casa) y continuos (e.g. la precipitación). Por lo anterior, existen dos formas principales de almacenar los datos en un SIG: raster y vector. Cada formato presenta un mecanismo para el almacenamiento de la información descriptiva y en muchas ocasiones resulta necesaria la asociación de series de datos a éstos [Bolstad05].

El modelo de datos raster permite gestionar la información continua en el espacio, como la evapotranspiración y la precipitación centrándose en las propiedades del espacio más que en la precisión de la localización. Este modelo se representa mediante una matriz de datos en donde cada celda corresponde a una región geográfica. El tamaño de las celdas se denomina resolución espacial y determina la precisión de la información almacenada en una capa de datos raster. Así, a medida que es mayor la resolución espacial, la precisión disminuye, pues el valor de cada celda es representativo de un área mayor [Bolstad05].

El modelo de datos vector gestiona la información discreta en el espacio, como las calles, las casas y los ríos, centrándose en la precisión de localización de los elementos geográficos sobre el espacio. Representa los datos basándose en tres primitivas geométricas: el punto, la línea y el polígono. La capa de datos vector más común se denomina shapefile y cada uno de sus elementos se denominan shapes (e.g. el conjunto de casas del centro de Medellín es un shapefile y cada casa es un shape). La información asociada a cada shape se almacena en la fila de una tabla de atributos, siendo los campos de la tabla de atributos definidos por el interesado (e.g. propietario, número telefónico, estrato, etc.) [Bolstad05].

Las series de datos se dividen en dos tipos principales: las series de tiempo y los perfiles verticales de medición. Una serie temporal o cronológica es una secuencia de datos, observaciones o valores, medidos en determinados momentos del tiempo, ordenados cronológicamente y, normalmente espaciados entre sí de manera uniforme. Uno de los usos más habituales de las series de datos temporales es su análisis para predicción y pronóstico, por ejemplo de eventos climáticos o de las fluctuaciones de las acciones en el mercado energético. Los perfiles verticales de medición almacenan datos de la columna de aire o de agua, variando en la elevación o profundidad, respectivamente. Con este tipo de series de datos, se pueden asociar, por ejemplo, las variaciones de temperatura verticalmente en un punto de un embalse.

La gestión de las series de datos se realiza tradicionalmente sirviéndolas en línea a partir de formatos CDF (Common Data Format), HDF (Hierarchical Data Format) y NetCDF (Network Common Data Form). Éstos son formatos científicos para el almacenamiento, transporte y procesamiento de datos multidimensionales independientes del dominio. Son desarrollados y mantenidos por la NASA, la National Center for Supercomputing Applications (NCSA) de la University of Illinois y la University Corporation for Atmospheric Research (UCAR), respectivamente. Debido a que son formatos binarios, su gestión está dada a través de las librerías que se desarrollan para cada uno. En el caso de CDF, cuenta con implementaciones en C, Fortran, Java, Perl y C#; HDF en Java, MATLAB, IDL y Python y NetCDF en Java, Python y Matlab. Estos formatos son autodescriptivos, pues tienen la capacidad de almacenar metadatos y atributos de los datos. HDF cuenta con las versiones HDF, HDF4 y HDF5. Por otro lado, NetCDF está basado en los modelos de datos CDF y HDF [Manduchi10]. En las Figura 1, 2 y 3, se presenta un resumen de los tres modelos de datos usando esquemas preconceptuales [Zapata06].

El uso de estos formatos hace que el análisis espacial de la información pase a un segundo plano, ya que no están hechos para un acceso directo desde sistemas que gestionan la información geográfica, por ello algunos otros esfuerzos se han realizado. En la Escuela de Geociencias y Medio Ambiente de la Universidad Nacional de Colombia, sede Medellín, realizaron el almacenamiento de las series de datos al interior de las bases de datos espaciales mediante el uso de formatos XML. Sin embargo, este enfoque se queda en una implementación y no pasa a ser una propuesta general de almacenamiento de series de datos [HidroSIG10].

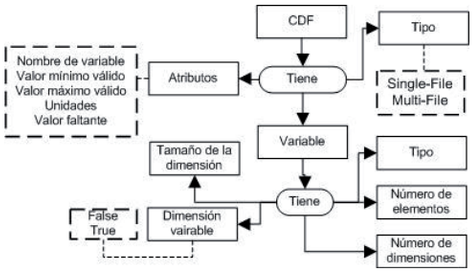


Figura 1. Formato de almacenamiento y transporte de datos Common Data Format.

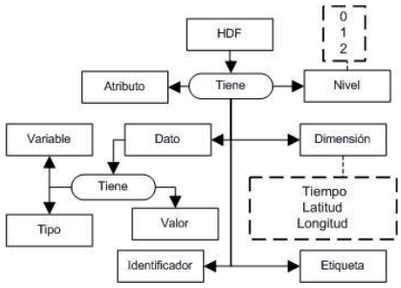


Figura 2. Formato de almacenamiento y transporte de datos Hierarchical Data Format.

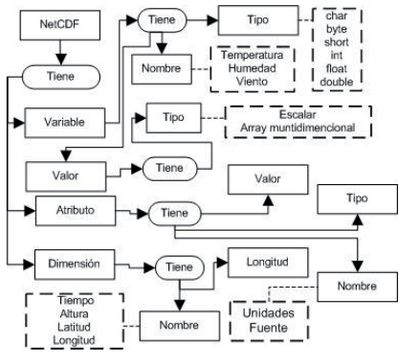


Figura 3. Formato de almacenamiento y transporte de datos Network Common Data Form.

3. Estructura propuesta

La estructura de almacenamiento de series de datos en bases de datos espaciales propuesta pretende facilitar el proceso de gestión de estas en su asociación con modelos de datos raster y vector. La estructura se basa en la especificación OGC para el almacenamiento de objetos geográficos en bases de datos relacionales. De esta manera es posible asociar a elementos de dato vector y raster, series de datos para permitir el análisis integral de la información.

La especificación en la cual se basa la estructura propuesta es la OpenGIS Simple Feature Specification for SQL. Este estándar define un esquema SQL para el almacenamiento, la recuperación, la consulta y la actualización de elementos geoespaciales. Algunas de las bondades son: a) establece una plataforma arquitectural para almacenar objetos geográficos en una base de datos relacional; b) establece un conjunto de términos con sus definiciones para ser usados en la plataforma definida; c) define un perfil para la definición de la geometría, basado en la norma ISO19107 [ISO03], y d) describe un conjunto de tipos de geometrías SQL junto con las funciones para gestionarlas. En la Figura 4, se puede observar el esquema de tablas propuesto por el OGC.

La tabla GEOMETRY_COLUMNS describe las tablas de elementos geográficos disponibles en la base de datos con los metadatos asociados; la tabla SPATIAL_REF_SYS contiene los sistemas de coordenadas y transformaciones soportadas por la base de datos; las Feature Table corresponden a todas las posibles capas de objetos geográficos, que la base de datos contendrá y finalmente GEOMETRY_TABLE almacena las coordenadas geográficas de cada objeto según el estándar definido por OGC, pudiendo ser este binario o textual. Partiendo de esta estructura para el almacenamiento de objetos geográficos se identificaron y caracterizaron los tipos de series de datos que usualmente resulta necesario asociar a éstos. Para ello se partió de las siguientes consideraciones: la caracterización se realizó a partir de la revisión de tipos de series de datos basada en casos; se asumió que cada serie está constituida por dos componentes llamadas X y Y, y la diferenciación entre un tipo de serie de datos y otro se establece a partir de las diferencias en sus componentes y atributos. En la Tabla 1 se presentan las series de datos identificadas con una breve descripción y los componentes y atributos que se le deben asociar, y en la Tabla 2, se presenta un ejemplo de cada una.

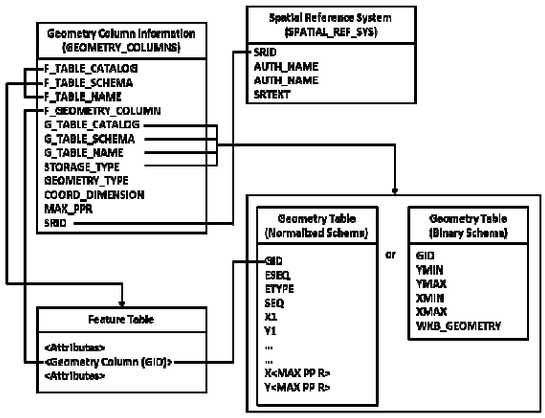


Figura 4. Esquema de tablas sugerido por el OGC.

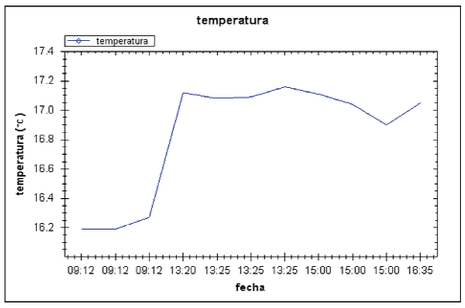
Tabla 1. Series de datos identificadas.

Tipo	Descripción	Componentes	Atributos
Serie de tiempo con resolución temporal.	Conjunto de valores de una variable a través del tiempo, con resolución temporal (e.g. cada cinco minutos exactamente).	En X: Fecha En Y: Valores de una variable dada a través del tiempo.	Tipo, Fuente, Componente geográfico, observaciones, Unidades_Y, Nodata, Resolución_ temporal .
Serie de tiempo eventual (sin resolución temporal).	Conjunto de valores de una variable a través del tiempo, sin resolución temporal.	En X: Fecha En Y: Valores de una variable dada.	Tipo, Fuente, Componente geográfico, observaciones, Unidades_Y.

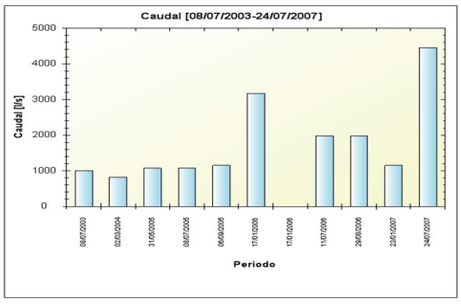
Texto vs. Valor.	Valores de varias variables para un tiempo dado. Las variables deben tener las mismas unidades de medición.	En X: Variables En Y: Valores de varias variables para un tiempo dado.	Tipo, Fuente, Componente geográfico, observaciones, Unidades_X, Unidades_Y.
Perfil vertical en la columna de agua.	Perfil vertical de valores de una variable en la columna de agua.	En X: Valores de una variable dada, para un tiempo dado. En Y: Profundidad.	Tipo, Fuente, Componente geográfico, observaciones, Unidades_X, Unidades_Y.
Perfil vertical en la columna de aire.	Perfil vertical de valores de una variable en la columna de aire.	En X: Valores de una variable dada, para un tiempo dado. En Y: Altura.	Tipo, Fuente, Componente geográfico, observaciones, Unidades_X, Unidades_Y.

Tabla 2. Ejemplos de series de datos identificadas.

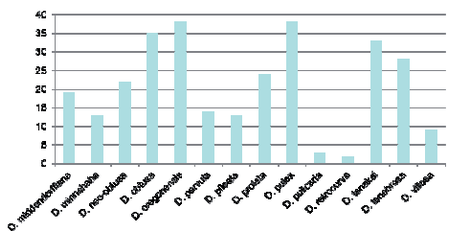
Gráfica/Atributos



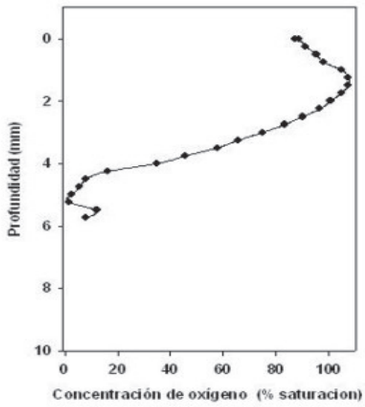
Temperatura Vs. Tiempo. Tipo: Serie de tiempo con resolución temporal; Fuente: Desconocida; Componente geográfico: lat= 06 06 42.76, long=-75 29 48.58; Observaciones: Muestreo realizado sobre el embalse Porce II el 3 de enero de 2011; Unidades_Y: Grados centígrados; Nodata: -9999; Tipo_de_estadística: Promedio; Resolución_temporal: 15 segundos.



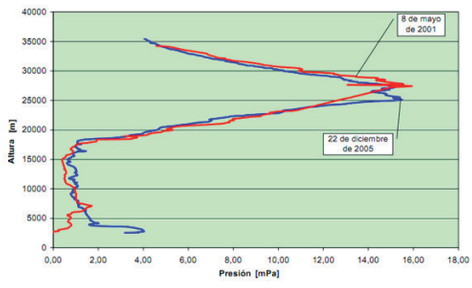
Caudal Vs. Tiempo. Tipo: Serie de tiempo sin resolución temporal; Fuente: Desconocido; Componente geográfico: lat= 06 06 42.76, long=-75 29 48.58; Observaciones: Muestreo realizado sobre el embalse Riogrande II el 23 de marzo de 2011; Unidades_Y: litros/segundos.



Daphnias Vs. Cantidad. Tipo: Texto Vs Valor; Fuente: EPM; Componente geográfico: lat= 06 06 42.76, long=-75 29 48.58; Observaciones: Muestreo realizado sobre el embalse La Fe el 15 de febrero de 2011; Unidades_Y: Unidades; Unidades_X: Daphnias.



Concentración de oxigeno Vs. Profundidad. Tipo: Perfil vertical en la columna de agua; Fuente: IDEAM; Componente geográfico: lat= 04 43 02, long=-74 03 11.2; Observaciones: Muestreo realizado sobre el embalse La Fe el 15 de febrero de 2011; Unidades: (% de saturación).



Presión atmosférica Vs. Altura. Tipo: Perfil vertical en la columna de aire; Fuente: IDEAM; Componente geográfico: lat= 04 43 02, long=-74 03 11.2; Observaciones: Muestreo realizado sobre en la estación meteorológica del aeropuerto el DORADO de Bogotá; Unidades: mPa.

Finalmente, con una base para el almacenamiento de objetos geográficos y con las series de datos identificadas, se propuso un esquema para el almacenamiento de esta información de manera integrada en una base de datos geográfica. La estructura propuesta se presenta en la Figura 5, en donde se resaltó en rojo las adiciones realizadas. En ésta se agregó una tabla (DATASET_TABLES) que permite identificar los tipos

de series de datos que puede tener asociado un objeto geográfico. La conexión entre los objetos geográficos y las series de datos está dada por medio de la tabla GEOMETRY_COLUMNS ya existente en la estructura que contiene los metadatos de cada tabla de objetos geográficos. Cada una de estas tablas de series de datos estará asociada al objeto geográfico mediante el identificador DSID y contendrá los atributos según se identificaron. Las bondades de la estructura propuesta serán discutidas en la sección 4.

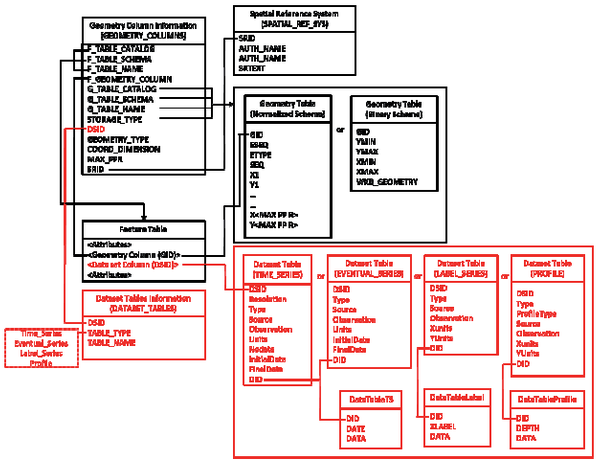


Figura 5. Estructura de almacenamiento de series de datos propuesta sobre el esquema de tablas sugerido por el OGC para el almacenamiento de objetos geográficos.

4. Experimentos, Resultados y Discusión

La estructura propuesta se implementó en la extensión para datos geográficos PostGIS del sistema de gestión de bases de datos PostgreSQL. Se agregaron los componentes propuestos a la plantilla y se creó una base de datos con series temporales asociados a puntos de estaciones de medición de precipitación.

En la Figura 6, se puede observar la curva de una serie de precipitación asociada a una estación de muestreo de lluvia. La interfaz soporta los atributos identificados en el punto anterior, como son las Observaciones, el tipo de serie, en este caso Precipitación Total – Mensual y las unidades de medida (mm) de la variable seriada. En esta experimentación se empleó el SIG MapWindow 4.6 con su extensión para la gestión y despliegue de información hidrológica HidroSIG 4.0 [HidroSIG10] vinculándolo a un proveedor de datos PostgreSQL implementado para la experimentación realizada.

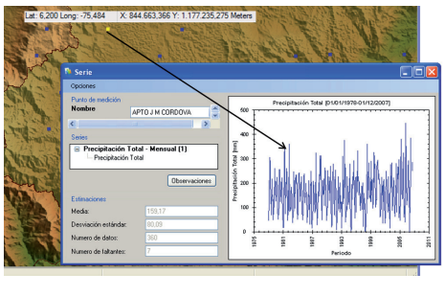


Figura 6. Despliegue de una serie de datos temporal asociada a un objeto geográfico en el SIG MapWindow 4.6.

La estructura de almacenamiento de series de datos en bases de datos espaciales propuesta y puesta en marcha, según se expone en el punto anterior, proporciona un mecanismo para el análisis integral de la información geográfica y seriada asociada a esta última. Esto facilita el análisis de esta información y la toma de decisiones con base en ello. Esta estructura incorpora elementos nuevos al esquema propuesto por el OGC sin eliminar ninguno de los ya existentes, dándole continuidad a la capacidad interoperable del mismo puesto en marcha en proyectos existentes. De esta manera. el uso de la estructura propuesta en contextos en donde las series de datos no son usadas o son usadas solo en algunos elementos geográficos no es alterado. La clasificación de los datos según los tipos de series de datos identificados permite optimizar el almacenamiento de los datos, ya que suelen corresponder a millones de registros en donde la repetición innecesaria de alguno de los atributos puede llegar a marcar la diferencia en tamaño y tiempos de consulta de los mismos. El experimento realizado evidenció que la estructura es viable, que no limita las capacidades interoperables del esquema preexistente y soporta el almacenamiento de las series de datos identificadas.

5. Conclusiones y trabajo futuro

Los datos geográficos vienen siendo utilizados con frecuencia en un amplio dominio del conocimiento. Adicionalmente, en áreas, como las ciencias de la tierra, la atmósfera y el océano, es de crucial importancia el análisis de series de datos para la toma de decisiones. Este análisis se viene realizando aislado del análisis espacial posible de realizar con objetos geográficos. La necesidad de complementar estos dos elementos en aras de un análisis integral de la información, es lo que motivó a la estructuración de un esquema de almacenamiento conjunta. De esta manera, partiendo de las recomendaciones OGC, en este artículo se presentó una estructura de almacenamiento de series de datos en bases de datos espaciales que se probó en el SIG MapWindow 4.6 usando su extensión para el análisis y despliegue de información hidrológica HidroSIG 4.0. La estructura propuesta posibilita desplegar en cualquier SIG que implementa o no los estándares OGC los objetos geográficos y las series de datos que se requieran asociar, de esta manera se logran el análisis espacial y temporal de una forma integral facilitando la toma de decisiones. La estructura se incorpora dentro de una estructura existente e interoperable a nivel mundial, como lo es PostGIS. Como trabajo futuro que se puede derivar de este trabajo se propone la implementación de herramientas en otros SIG como SAGA, GRASS y ArcGIS, adicionalmente el uso de la estructura propuesta en proyectos reales en cualquier dominio del conocimiento que incorpore el uso de objetos geográficos y series de datos.

Referencias bibliográficas

[Anaya10] Anaya, J.; Chuvieco, E., 2010. Caracterización de la Eficiencia de Quemado a partir del análisis de series de tiempo del índice de vegetación EVI, in SELPER, Guanajuato, Mexico, 2010.

[Bolstad05] Bolstad, P. (2005). GIS Fundamentals: A first text on Geographic Information Systems, Second Edition. White Bear Lake, MN: Eider Press, 543 pp.

[HidroSIG10] HidroSIG (2010). SIG para la gestión de información hidrológica.
http://www.minas.medellin.unal.edu.co/~hidrosig/.

[ISO03] ISO (2003). Geographic information – Spatial schema.
http://www.iso.org/iso/catalogue_detail.htm?csnumber=26012.

[Manduchi10] Manduchi, G., (2010). Commonalities and differences between MDSplus and HDF5 data systems. Fusion, Engineering and Design, vol. 85, Issues 3-4, pp 583-590.

[OGC12] OGC (2012). Open Geospatial Consortium. http://www.opengeospatial.org/.

[Unidata12] Unidata (2012). Network Common Data Form. http://www.unidata.ucar.edu.

[Yang99] Yang, X; Michiel C. J.; Damen, R. A.; y Zuidam. V., (1999). Satellite remote sensing and GIS for the analysis of channel migration changes in the active Yellow River Delta, China, Original Research Article International Journal of Applied Earth Observation and Geoinformation, Volume 1, Issue 2, 1999, Pages 146-157.

[Zapata06] Zapata, C; Gelbukh, A y Arango, I. (2006) "Pre-conceptual Schema: A Conceptual-Graph-Like Knowledge Representation for Requirements Elicitation". En: Lecture Notes in ComputerScience. Vol. 4293. pp. 17-27.