CONAMA10
CONGRESO NACIONAL
DEL MEDIO AMBIENTE

COMUNICACIÓN TÉCNICA

El formato HDF: Un modelo de datos para el almacenamiento y gestión de información espacial de carácter ambiental.

Autor: Marcos Palomo Arroyo

Institución: Universidad Politécnica de Madrid

e-mail: marcos.palomo@upm.es

Otros Autores: Santiago Ormeño Villajos (UPM); Joaquín Alberto Rincón Ramírez (becario CONACYT, Colegio de postgraduados, México)

RESUMEN

El formato HDF es una estructura robusta de almacenamiento y distribución de datos científicos de naturaleza múltiple, y se utiliza por entidades que producen y gestionan información de carácter ambiental, así como por agencias distribuidoras de datos procedentes de la observación territorial, como es el caso de la NASA.
En la presente comunicación se describen las características de la citada estructura y se analiza su funcionalidad mediante uno de los módulos de la aplicación SOVMAP, desarrollada por los autores.
Se aplica el modelo de datos, y los procesos correspondientes, a un documento de ocupación de suelo, obtenido a partir de imágenes espaciales MODIS, con el fin de obtener y distribuir indicadores ambientales normalizados, particularmente los relativos a superficies de cultivo y otras cubiertas naturales.
La comunicación concluye, demostrando la robustez, ductilidad y versatilidad, de la estructura analizada, en modelización ambiental de datos espaciales.

Palabras Clave: Indicadores ambientales, teledetección, estructuras de datos

1. Introducción

En el ámbito científico, se genera gran cantidad de información en formato digital, que es necesario distribuir. Dada la diferente naturaleza de ésta, se utilizan distintos formatos especialmente adaptados a cada uno de los tipos de datos que se vayan a almacenar. El principal problema, radica en la estandarización y la interoperabilidad, ya que, pese a que existen numerosos estándares de almacenamiento de información, no todos son soportados por las diferentes aplicaciones informáticas, lo que puede presentar problemas y falta de eficiencia en la difusión y reutilización de dicha información.

El formato HDF (*Hierarchical Data Format*), representa una alternativa eficaz, al ser adoptado como estructura de almacenamiento de datos, cuyas características más representativas son las siguientes:

- Permite obtener información acerca de los datos de un archivo desde dentro del mismo, sin necesidad de recurrir a fuentes externas.
- Permite almacenar datos de distinta naturaleza en un mismo archivo y relacionarlos entre ellos.
- Estandariza los formatos y las descripciones de los tipos de datos más comúnmente empleados.
- Se trata de un formato abierto, con sus especificaciones publicadas, lo que permite su implementación en diversas aplicaciones informáticas, facilitando la portabilidad, así como permitiendo al usuario desarrollar sus propias aplicaciones.
- Es flexible y puede ser adaptado para almacenar cualquier tipo de dato.

2. Historia del formato HDF

Con el fin de dar respuesta a las necesidades de almacenamiento y difusión de datos científicos de diversa naturaleza, en un formato flexible e independiente de la plataforma, comenzó a desarrollarse el formato HDF en los laboratorios de la NCSA (*National Center for Supercomputing Applications*) a partir del año 1988, siendo soportado en la actualidad por el HDF Group, dependiente de la Universidad de Illinois.

Actualmente, este formato se utiliza por múltiples organismos, tanto públicos como privados, para la difusión de sus datos y resultados, entre las que se encuentran la NASA, la Agencia Europea del Espacio (ESA), que los aplican en datos procedentes de los sensores MODIS, MERIS o ETM+, entre otros.

3. Descripción del formato HDF

La estructura de los archivos HDF permite el almacenamiento de diversos tipos de datos, tales como (ver figura 1):

- *Scientific Data Sets* (SD), utilizados para almacenar estructuras n-dimensionales de datos enteros (8, 16 y 32 bits, con o sin signo) o reales (32 o 64 bits) en su formato estándar, o por medio de las Interfaces de Programación de Aplicaciones (APIs), crear datos en otros rangos de valores (1 a 32 bits), junto con sus metadatos (dimensionalidad, atributos, etc.)
- *Raster Images* (RI), que permiten almacenar imágenes de 8 bits (0-255 niveles de gris) o 24 bits (RGB) o, en su caso, utilizar las correspondientes librerías de programación para guardar imágenes en otros formatos (16 a 32 bits enteros, 32 a 64 bits en coma flotante). Además, se permite almacenar información sobre las dimensiones de la imagen, así como la paleta de color asociada a ella. También es posible utilizar ciertos algoritmos de compresión (RLE, JPEG, GZIP y adaptativa Huffman) para reducir el tamaño de los archivos resultantes.

- *Text Annotations* (TA), para almacenar cualquier tipo de información textual (etiquetas, descripciones o información de archivo).
- *VDatas* (VD), para almacenar datos vectoriales sin topología.
- *Vgroups* (VG), que nos permiten asociar datos relacionados dentro de un archivo, al estilo de carpetas.

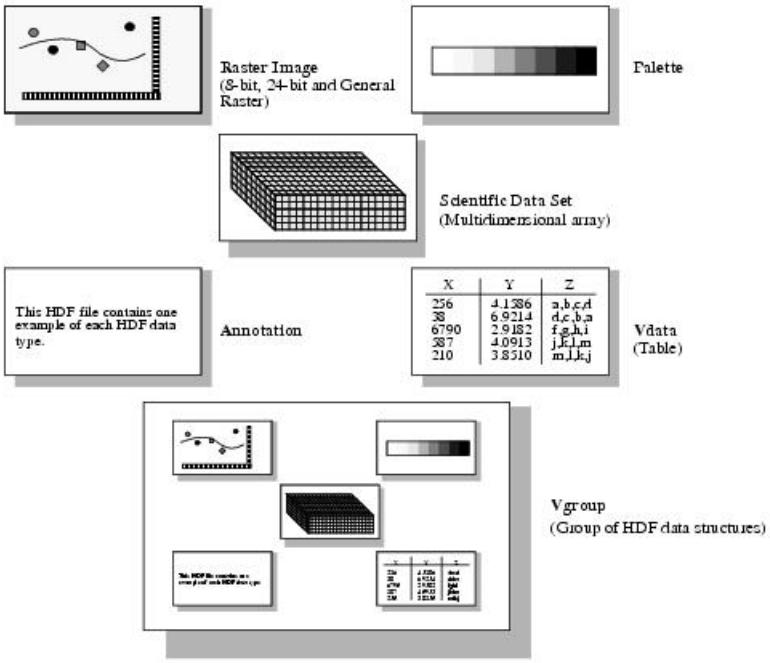

Figura 1: Estructuras de datos de un archivo HDF. (Fuente: 'The HDF Group')

4. Organización interna de los datos

Los ficheros HDF están estructurados en las siguientes partes: encabezado, bloque de descripción de datos formado por los descriptores de los mismos y los datos propiamente dichos.
A nivel binario, estos elementos se organizan dentro del archivo siguiendo este esquema:

- Encabezado. Ocupa los cuatro primeros bytes e identifica a un archivo como HDF. Su valor es siempre el mismo ^N^C^S^A
- Bloque de descripción de los datos (6 bytes). Está formado por los descriptores de los datos contenidos en el archivo. Se almacena en disco como dos valores, el primero, indica el número de descriptores de datos y el segundo define el offset del primer descriptor de datos.
- Descriptores de datos (12 bytes). Describen el tipo y la localización de los datos contenidos en el archivo. Están formados por un *tag*, un *id*, el *data offset* y el *data length*. La etiqueta (*tag*) define el tipo de datos almacenado de acuerdo a un código establecido por la NCSA, el identificador (*id*) es un número secuencial, que determina de forma unívoca a cada elemento, el *data offset* fija el desplazamiento (offset) del primer elemento del conjunto de datos con respecto al comienzo del archivo, y el *data length* indica el tamaño de ese conjunto de datos.
- Datos. Almacenados a continuación del último descriptor de datos, se almacenan en el formato correspondiente, según el tipo de datos de que se trate.

5. Acceso a los datos de un archivo HDF

Existen diversas formas de acceder a los datos contenidos en un archivo HDF. Una de ellas es el acceso a nivel binario, conociendo la estructura del archivo, o bien a través de aplicaciones disponibles que soporten este formato, de forma que el proceso se realice de manera transparente para el usuario. En este caso, el acceso se realiza a través de la aplicación SOVMAP, desarrollada por los autores, que incluye un módulo de importación/exportación con soporte para el formato HDF, permitiendo la extracción y empaquetamiento de datos y productos generados en este formato.

6. La aplicación SOVMAP

En este trabajo se ha elegido la aplicación SOVMAP, desarrollada por S. Ormeño y M. Palomo, para el proceso de los datos y cálculo de variables ambientales, así como la realización de análisis posteriores.
Se trata de un conjunto de herramientas integradas en un entorno de trabajo intuitivo y modular que permite realizar múltiples operaciones con datos ráster principalmente, si bien soporta estructuras vectoriales básicas, orientadas al ámbito de la teledetección, la fotogrametría y modelos superficiales.
Ofrece la posibilidad de trabajar con imágenes monobanda y multibanda almacenadas en diversos formatos, permitiendo la importación y exportación de datos.
Por otro lado, cuenta con una calculadora de bandas programable mediante el empleo de un lenguaje de modelado que permite realizar operaciones aritméticas y lógicas, posibilitando la obtención de diferentes variables derivadas.

7. Cálculo de variables ambientales a partir de MODIS.

Es objetivo de esta comunicación presentar el uso que se hace del formato HDF en la obtención de parámetros ambientales derivados de los datos suministrados por el sensor MODIS. Este sensor nos proporciona una cobertura diaria, en 36 bandas espectrales, con una resolución espacial que oscila entre los 250m y los 1000m, dependiendo de la banda espectral.
Particularmente, se ha realizado una clasificación, con el fin de determinar las distintas ocupaciones de suelo en una zona del interior de la provincia de Palencia. Previamente, es necesario proceder a la descarga de los datos desde el servidor de la agencia distribuidora de los mismos. Se han utilizado los productos estándar MOD09GQ y MOD09GA, reflectividad superficial, en 6 bandas, cuyo rango espectral, así como su resolución espacial se indican en la tabla 1. En ella también se indica la correspondencia entre estas bandas y las del sensor ETM+, el cual constituye una referencia en el ámbito de la obtención de este tipo de documentos.
Dada la limitada resolución espacial de este sensor, previamente a su utilización, se ha procedido a realizar un proceso de interpolación, desarrollado por los autores, con lo que el conjunto de datos multiespectrales utilizado alcanza un resolución espacial de 30m (figura 2) permitiendo aumentar el grado de detalle de la imagen.

Tabla 1: Bandas MODIS y equivalencia ETM+

Banda	Intervalo	Resolución	Equivalencia ETM+
1	620-670 nm	250 m	3
2	841-876 nm	250 m	4
3	459-479 nm	500 m	1
4	545-565 nm	500 m	2
6	1628-1652 nm	500 m	5
7	2105-2155 nm	500 m	7

Figura21: Zona de trabajo (Combinación RGB: 432)

A partir de estos datos multiespectrales, se procede a realizar una clasificación, no supervisada, de la escena, con el fin de obtener un documento de clases de ocupación del suelo presentes en la misma.

La aplicación del método con SOVMAP, requiere de varias fases:

- Entrenamiento del clasificador, en el cuál se seleccionan las áreas de entrenamiento mediante un muestreo automático y equiespaciado por toda la imagen.
- Agrupamiento o *clustering*, en la que mediante la utilización de ciertas medidas de distancia entre un punto y un grupo, las diferentes áreas de entrenamiento se agrupan entre sí formando clases.
- Filtrado de las clases iniciales, con el fin de obtener áreas homogéneas que sirvan de base para el clasificador. De este modo, se asegura la coherencia entre los perfiles espectrales obtenidos en el proceso de *clustering*. El resultado de este proceso se muestra en la figura 3**.**

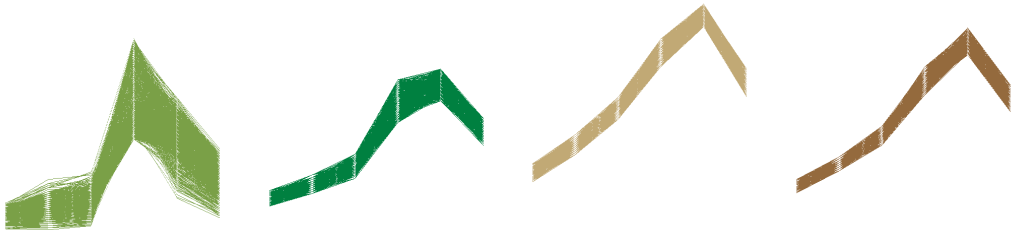

Figura 3: Firmas espectrales filtradas (Regadío, Veg. Nat. SD1. SD2)

- Asignación, propiamente dicha, partiendo de la información multiespectral de las 6 bandas y las áreas de entrenamiento, se utiliza la regla bayesiana de máxima verosimilitud. Dadas las características de la zona de trabajo, el número de clases de ocupación de suelo se ha reducido a 4, cultivos de regadío (R), zonas de vegetación natural (VN), así como dos tipos diferentes de suelos desnudos (S.D. 1 y S.D. 2) según su respuesta espectral.
- Análisis del error de la clasificación, mediante una adaptación del coeficiente de aceptación *Kappa* para estos fines, el cual nos indica el grado de exactitud del resultado de la clasificación, que es tanto mayor cuanto más próximo a 1 se encuentre. Los resultados obtenidos en este trabajo, para cada una de las clases, se indican en la **tabla 2**.

Tabla 2: Coeficiente Kappa.

Clase	Kappa
Regadío	1,0000
Vegetación Natural	1,0000
Suelo Desnudo 1	0,9979
Suelo Desnudo 2	0,9976
Global	0,9989

El proceso de clasificación nos proporciona un documento cualitativo de clases de ocupación del suelo, que servirá de base para el análisis territorial (Figura 3).

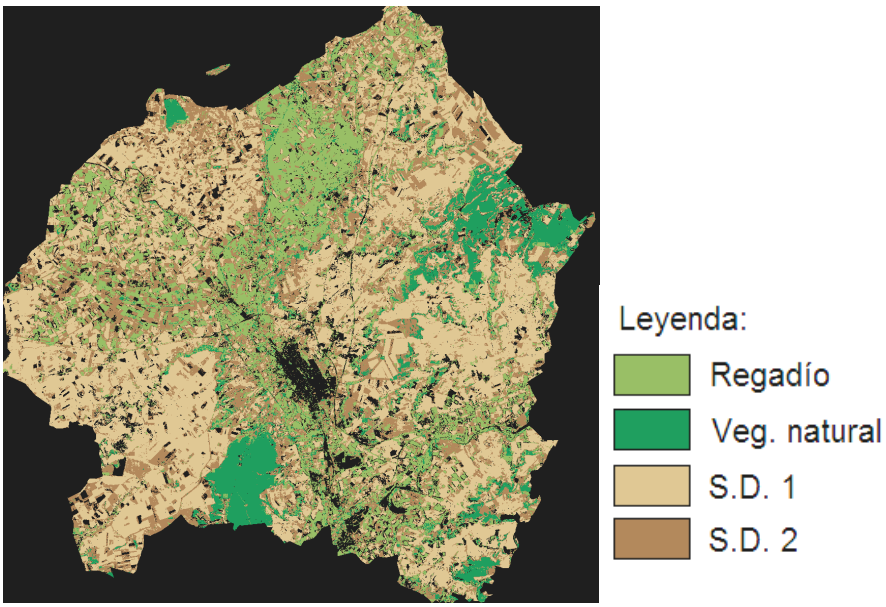

Figura 3: Documento de ocupación del suelo

Esta fase del análisis, tiene como fin, la obtención de una serie de parámetros o indicadores ambientales referidos a una o varias entidades territoriales, que pueden ser obtenidos de forma sistemática, permitiendo a los planificadores tomar decisiones orientadas a una mejor utilización de los recursos disponibles, así como a definir políticas de protección ambiental, basadas en el análisis de los resultados.
Para realizar dicho análisis, se dispone de la capa de entidades territoriales de la zona de estudio, en este caso términos municipales, en formato ráster (Figura 4).

Figura 4: Términos municipales.

Utilizando los documentos de ocupación de suelo y de términos municipales, se ha realizado una tabulación cruzada, con el fin de obtener un documento de porcentajes de ocupación de cada clase para cada una de dichas entidades (tabla 3).

Tabla 3: Porcentajes de ocupación, por término municipal

Nombre	%Sin clasif.	%Regadío	%Veg. Nat	%SD1	%SD2
Amusco	6,16	13,66	8,03	41,96	30,2
Ribas de Campos	7,04	51,88	2,87	10,65	27,55
Becerril de Campos	8,71	17,99	2,86	39,67	30,77
Monzón de Campos	7,02	30,98	15,54	25,04	21,43
Husillos	6,14	61,61	5,46	6,33	20,43
Villaumbrales	10,15	21,06	1,63	35,86	31,3
Villamediana	7,76	11	14,92	48,05	18,27
Fuentes de Valdepero	6,91	9,41	7,22	50,77	25,69
Grijota	10,38	39,81	2,86	17,24	29,71
Valdeolmillos	4,15	8,88	11,47	61,93	13,58
Palencia	14,57	16,15	19,43	26,38	23,47
Mazariegos	9,02	15,95	0,52	48,14	26,38
Villamartin de Campos	7,33	9,17	0,14	63,53	19,83
Villalobón	11,29	9,56	5,87	51,68	21,6
Autilla del Pino	10,58	3,94	1,34	68,16	15,97
Magaz de Pisuerga	14,66	22,75	3,55	35,78	23,26
Villamuriel de Cerrato	17,33	26,48	5,92	37,31	22,96
Reinoso de Cerrato	5,31	15,93	7,07	47,1	24,6
Soto de Cerrato	11,91	23,71	5,73	36,65	21,99
Venta de Baños	29,86	34,49	2,35	13,63	19,67
Santa Cecilia del Alcor	10,07	3,31	2,16	54,74	29,72
Hontoria de Cerrato	10,53	14,68	14,22	44,42	16,16

Los resultados de la tabla 3 pueden ser visualizados en forma gráfica, como se muestra en la figura 5, proporcionando un documento con la distribución espacial de cada una de las clases.

Durante el desarrollo del proceso anteriormente descrito, se han generado una serie de archivos de diferentes formatos, que incluyen documentos gráficos en formato ráster, documentos vectoriales, así como tablas de valores. En este caso, se ha considerado relevante la difusión de las imágenes correspondientes a la ocupación de suelo, los términos municipales, ambos en formato ráster, las áreas de entrenamiento, en formato vectorial, así como los perfiles espectrales de dichas áreas de entrenamiento y el resultado de la tabulación cruzada, en formato de tabla de valores. Todos estos archivos pueden encapsularse en un único fichero HDF, de forma que la difusión de los mismos se haga de manera eficiente y segura (figura 6).

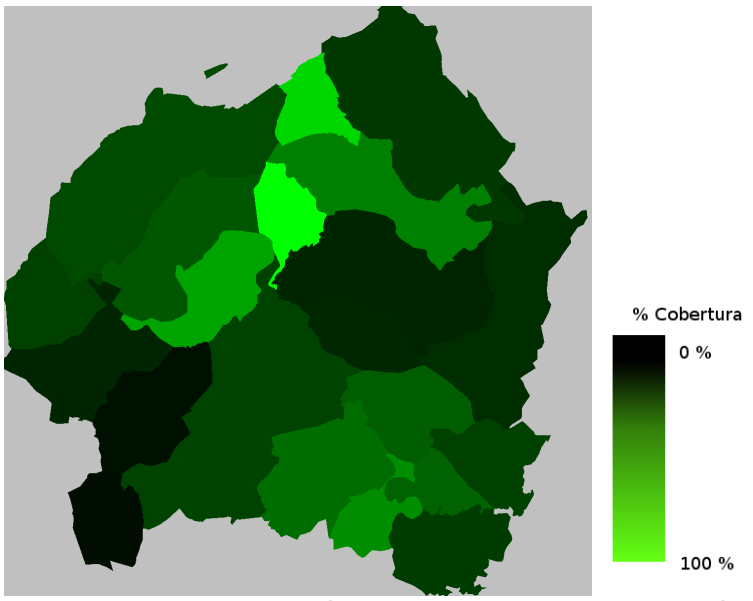

Figura 5: Distribución espacial de la clase 'Regadío'.

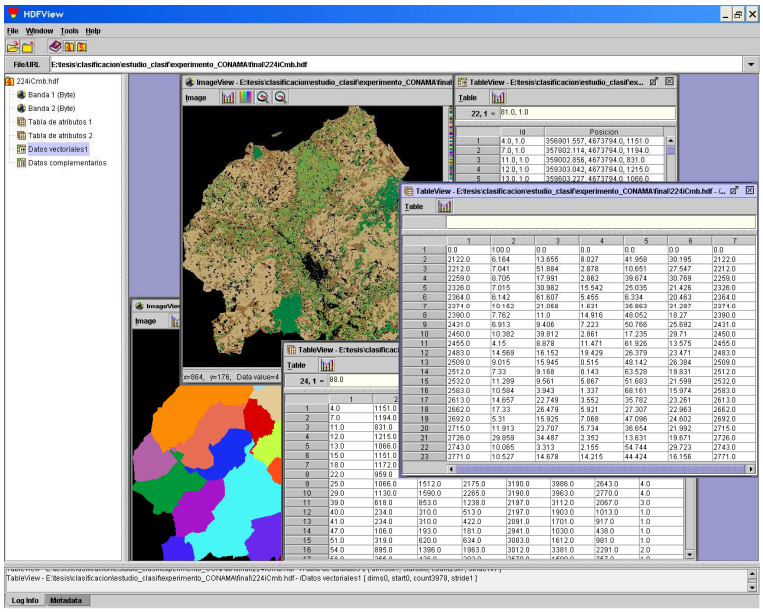

Figura 6: Archivo HDF generado.

También se incluye una tabla de correspondencia entre los identificadores de los polígonos y su nombre, que nos permitirá identificar cada uno de los términos municipales en la imagen.

8. Conclusiones

En los trabajos referidos en la presente comunicación se ha demostrado la validez del modelo de datos de los archivos HDF como estructura robusta de almacenamiento y distribución de imágenes y otros datos de interés ambiental, por su versatilidad y

posibilidades de adaptación, en este caso, referidos a un documento cualitativo de ocupación del suelo así como diversos documentos temáticos, tablas y datos vectoriales. El formato HDF se convierte en una estructura fundamental para los fines referidos, al tratarse de un formato nativo adoptado por diversas agencias, así como organismos distribuidores y productores de información, que es utilizado por un número cada vez mayor de aplicaciones software.

9. Referencias:

Ormeño, S., Arozarena, A., Martínez, M., Palomo, M., Villa, G., Peces, J. J., Pérez, L. (2008): "Los satélites de media y baja resolución espacial como fuente de datos para la obtención de indicadores ambientales". IX Congreso Nacional de Medio Ambiente. Madrid.

Ormeño, S. (2006): "Teledetección Fundamental". ETSIGC. Departamento de Ingeniería Cartográfica y Topografía. Universidad Politécnica de Madrid.

Ormeño, S. (2006): "Manual de referencia de SOV (proceds)". ETSIGC. Departamento de Ingeniería Cartográfica y Topografía. Universidad Politécnica de Madrid.

Palomo, M (2006): "El formato HDF para el almacenamiento de información relativa a imágenes de satélite. Importación y exportación en SOV de ficheros HDF" Trabajo del Curso de doctorado "Cartografía ambiental". Universidad Politécnica de Madrid.

The HDF Group (2010): "HDF4 User's Guide. HDF4 Release 2.5" [en línea] <ftp://ftp.hdfgroup.org/HDF/Documentation/HDF4.2.5/HDF425_UserGd.pdf>

The HDF Group (2010): "HDF4 Reference Manual. HDF4 Release 2.5" [en línea] <ftp://ftp.hdfgroup.org/HDF/Documentation/HDF4.2.5/HDF425_RefMan.pdf>

The HDF Group (2008): "HDF Specification and Developer's Guide. HDF4 Release 2,3" [en línea] <ftp://ftp.hdfgroup.org/HDF/Documentation/HDF42r3_SpecDG.pdf>