Revista Internacional de Ciencia y Tecnología de la Información Geográfica
International Review of Geographical Information Science and Technology

Pons, X. y Masó, J. (2021): Data type, compression and interoperability in geographic information formats / Tipos de datos, compresión e interoperabilidad en los formatos de información geográfica, *GeoFocus, Revista Internacional de Ciencia y Tecnología de la Información Geográfica (Editorial), nº 28, p. 1-4.* http://dx.doi.org/10.21138/GF.767

DATA TYPE, COMPRESSION AND INTEROPERABILITY IN GEOGRAPHIC INFORMATION FORMATS

TIPOS DE DATOS, COMPRESIÓN E INTEROPERABILIDAD EN LOS FORMATOS DE INFORMACIÓN GEOGRÁFICA

[1]XAVIER PONS, [2]JOAN MASÓ
[1,2]Grumets research group.
[1]Dep. de Geografia, Edifici B, Universitat Autònoma de Barcelona, Cataluña, España
[2]CREAF, Universitat Autònoma de Barcelona, Cataluña, España
[1]xavier.pons@uab.cat, [2]joan.maso@uab.cat

As we already commented in the last editorial (Pons, 2021), the discipline of geographic information runs the risk of reinventing the wheel in the formats it uses. In this editorial we are going to comment on the choice of data types and compression formats, as well as their implications for interoperability.

Obviously, the use of compression techniques makes it possible to use less storage space while allowing data to be transmitted in less time. It is also true that compressing is at no cost: time must be spent, sometimes not less, in compression, and when the moment comes to use the data, it must also be decompressed (although it may not be demanding, this delay exists); the tradeoff can be beneficial in many cases and for this reason different compression forms, algorithms and formats have become popular. Among the forms of compression, it is appropriate to remember the two main strategies: lossy (as we find in most files –not all– in formats such as JPEG, JPEG2000, MrSID or ECW), or lossless (such as those that we can find in files in formats like TIFF, TAR, MMZX, etc). In the case of the former, their use in the world of geographic information has been (*e.g.*, Zabala and Pons, 2011) and is (*e.g.*, Makarichev *et al.*, 2021) constantly reevaluated and reconsidered due to the interest provided by a very high compression *versus* the risk of an excessive degradation of data. On the other hand, in the case of the later there is, by definition, no data loss and the differences in the compression ratios are not so great, as long as the correct decisions are made.

Indeed, recently an institution has distributed geographic information data in TIFF files using an encoding and compression variant described in Adobe (2005). This variant, certainly not frequent in our environment, pursues, through the use of data types of fewer bits in the encoding of real numbers (16 and 24 instead of the traditional 32 and 64) and the reordering of bits with a certain criteria, to obtain a more efficient application of the deflate and LZW compression methods and, ultimately, achieve smaller files. We have carried out a small exercise consisting of converting some of these files with this encoding and compression into

Recibido: 30/12/2020

Pons, X y Masó, J. (2021): Data type, compression and interoperability in geographic information formats / Tipos de datos, compresión e interoperabilidad en los formatos de información geográfica, *GeoFocus, Revista Internacional de Ciencia y Tecnología de la Información Geográfica (Editorial), nº 28, p. 1-4.* http://dx.doi.org/10.21138/GF.767

uncompressed TIFF files and compressing them into a conventional ZIP file. The result is surprising, to say the least: Taking one of them as an example thread, the original TIFF occupies 165 Mbyte, and applying the proposed exercise a file is obtained... that is much smaller than the one distributed: 79.7 Mbyte (48.3% of the original size). If we process it as a raw file (like Envi or MiraMon IMG formats without compression), we compress it with the classic and simple Run-Length Encoding (RLE) algorithm in all the repeated cells (areas with no data) and we compress it into a ZIP (for example, as it is done in the MMZX based on the standard ISO 19165:2018 "Geographic information - Preservation of digital data and metadata" and ISO 29500-2:2021 "Open packaging conventions"), an even smaller file is obtained: 69.0 Mbyte (41.8% of the original size, clearly less than half); and this much smaller file is much more interoperable, since it can be decompressed even from the file explorer of Windows 7 and later, as it is a ZIP.

Someone might point out that the choice of a single precision data type (32 bits per value, IEEE 754) is inappropriate because of the unnecessary significant figures it allows to hold without really being part of the data encoded in each cell (which in this case were *heights of objects in meters with precision in centimeters*); and this is true: the ZIP occupies much more (111.0 Mbyte) if it contains noise in the least significant bits. But simply rounding to centimeters (forcing all decimal places beyond centimeters to zero), the file is compressed to the 69.0 Mbyte already reported.

Finally, another option that still requires less space is to distribute the data in 16 bit integers, indicating in the metadata that the values are in centimeters; in this case an even smaller ZIP file is obtained: 55.6 Mbyte (33.7% of the original size, just over 1/3!).

This exercise reflects the extent to which we are, still today, distributing data in a non optimal way, which not only needs three times more space than necessary and takes three times longer to be transmitted over the network, but this distribution is done in a variant of unconventional TIFF and, therefore, much less interoperable than other options that we have indicated. Probably in Geographical Information Science and Technology courses greater emphasis should be placed on choosing the type of data and the type of compression. Not just because size matters, but because interoperability matters too. Those ideas are in line with what we also mentioned in the previous editorial: it is everyone's responsibility to think well before distributing data since, as we have shown, and contrary to popular expression, sometimes the best option is even a friend of a good option or, in other words, we can generate data that occupies less with more standard proposals.

In this issue of GeoFocus, 28, we present interesting theoretical contributions on our discipline in historical landscape studies, methodologies on fire mapping, applications to the study of land cover dynamics in Venezuela and Chile, or the analysis of dynamic networks for access to public hospitals. We hope that they will be of interest to our readers.

Como ya comentamos en el último editorial (Pons, 2021) la disciplina de la información geográfica corre el riesgo de reinventar la rueda en los formatos que usa. En el presente editorial vamos a efectuar un comentario respecto a la elección de los tipos de datos y los formatos de compresión, así como sus implicaciones en la interoperabilidad.

Como es obvio, la utilización de técnicas de compresión permite utilizar menos espacio de almacenamiento a la vez que permite transmitir los datos en menos tiempo. Es cierto también

Pons, X y Masó, J. (2021): Data type, compression and interoperability in geographic information formats / Tipos de datos, compresión e interoperabilidad en los formatos de información geográfica, *GeoFocus, Revista Internacional de Ciencia y Tecnología de la Información Geográfica (Editorial), nº 28, p. 1-4.* http://dx.doi.org/10.21138/GF.767

que comprimir no es gratis: debe invertirse un tiempo, a veces no menor, en la compresión, y llegado el momento de la utilización de los datos, también hay que descomprimirlos (aunque pueda ser poco demandante, este retraso existe); la resultante puede ser en muchos casos beneficiosa y por ello se han popularizado diferentes formas, algoritmos y formatos de compresión. Entre las formas de compresión es adecuado recordar las dos grandes estrategias: con pérdida (como encontramos en la mayoría de ficheros –no todos– en formatos como JPEG, JPEG2000, MrSID o ECW), o sin pérdida (como los que podemos encontrar en ficheros en formatos como TIFF, TAR, MMZX, etc). En el caso de los primeros su uso en el mundo de la información geográfica ha sido (*e.g.*, Zabala y Pons, 2011) y es (*e.g.*, Makarichev *et al.*, 2021) reevaluada y reconsiderada constantemente por el interés que proporciona una compresión muy alta ante el riesgo de acabar degradando excesivamente los datos. En cambio, en el caso de los segundos no hay, por definición, pérdida de datos y las diferencias en las razones de compresión no son tan grandes, siempre que se tomen las decisiones correctas.

En efecto, recientemente una institución ha distribuido datos de información geográfica en ficheros TIFF utilizando una variante de codificación y compresión descrita en Adobe (2005). Esta variante, ciertamente no frecuente en nuestro entorno persigue, a través de la utilización de tipos de datos de menos bits en la codificación de números reales (16 y 24 en lugar de los tradicionales 32 y 64) y de la reordenación de bits con un determinado criterio, obtener una aplicación más eficiente de los métodos de compresión *deflate* y LZW y, en definitiva, conseguir ficheros de menor tamaño. Hemos efectuado un pequeño ejercicio consistente en convertir algunos de estos ficheros con dicha codificación y compresión en ficheros TIFF no comprimidos y comprimirlos en un fichero ZIP convencional. El resultado es cuando menos sorprendente: Tomando uno de ellos como hilo de ejemplo, el TIFF original ocupa 165 Mbyte, y aplicando el ejercicio indicado se obtiene un fichero… mucho menor al distribuido: 79.7 Mbyte (48.3 % del tamaño original). Si lo procesamos como un fichero en bruto (tipo *raw*, como los IMG de Envi o MiraMon sin compresión), lo comprimimos con el clásico y sencillo algoritmo codificación por longitud de recorrido (RLE por sus siglas en inglés para *Run-Length Encoding*) en todas las celdas repetidas (zonas sin datos) y lo comprimimos en un ZIP (por ejemplo tal como se hace en el MMZX basado en los estándares ISO 19165:2018 "*Geographic information - Preservation of digital data and metadata*" e ISO 29500-2:2021 "*Open packaging conventions*") se obtiene un fichero todavía menor: 69.0 Mbyte (41.8 % del tamaño original, claramente menos de la mitad); y este fichero mucho menor es mucho más interoperable, puesto que puede descomprimirse incluso desde el explorador de ficheros de Windows 7 y posteriores al ser un ZIP.

Alguien podría apuntar que la elección de un tipo de datos de precisión simple (32 bits por valor, IEEE 754) es inadecuada por las cifras significativas innecesarias que permite albergar pero que no forman realmente parte del dato codificado en cada celda (que en este caso eran *alturas de objetos en metros con precisión en centímetros*); y ello es cierto: el ZIP ocupa mucho más (111.0 Mbyte) si contiene ruido en los bits menos significativos. Pero simplemente redondeando a centímetros (forzando a cero todo decimal más allá de los centímetros), el fichero se comprime a los 69.0 Mbyte ya informados.

Finalmente, otra opción que todavía requiere menos espacio es distribuir los datos en enteros de 16 bits indicando en los metadatos que los valores están en centímetros; en este caso se obtiene un fichero ZIP todavía menor: 55.6 Mbyte (33.7 % del tamaño original, ¡poco más de 1/3!).

Este ejercicio refleja hasta qué punto estamos, todavía hoy, distribuyendo datos de forma poco óptima, que no sólo ocupan tres veces más de lo necesario y tardan tres veces más

Pons, X y Masó, J. (2021): Data type, compression and interoperability in geographic information formats / Tipos de datos, compresión e interoperabilidad en los formatos de información geográfica, *GeoFocus, Revista Internacional de Ciencia y Tecnología de la Información Geográfica (Editorial), nº 28, p. 1-4.* http://dx.doi.org/10.21138/GF.767

en ser transmitidos por la red, sino que esta distribución se hace en una variante de TIFF nada convencional y, por tanto, mucho menos interoperable que otras opciones que hemos indicado. Probablemente en los cursos de Ciencia y Tecnología de la Información Geográfica debería hacerse mayor hincapié en la elección del tipo de datos y del tipo de compresión. No sólo porque el tamaño importa, sino porque la interoperabilidad también importa. Unas ideas que van en la línea de lo que también comentábamos en el anterior editorial: es una responsabilidad de todos pensar bien antes de distribuir datos ya que, como hemos mostrado, y al contrario de lo que reza la expresión popular, a veces lo mejor incluso es amigo de lo bueno o, en otras palabras, podemos conseguir generar datos que ocupen menos con propuestas más estándar.

En este número de GeoFocus, 28, contamos con interesantes aportaciones teóricas sobre nuestra disciplina en los estudios históricos de paisaje, metodológicas sobre cartografía de incendios, aplicaciones al estudio de la dinámica de las cubiertas del suelo en Venezuela y en Chile, o al análisis de redes dinámicas para el acceso a hospitales públicos. Confiamos que resulten de interés a nuestros lectores.

References / Referencias

Adobe (2005): Adobe Photoshop® TIFF Technical Note 3, *The Internet*, http://chriscox.org/TIFFTN3d1.pdf (last accessed 04-Dec-2021).

Makarichev, V., Vasilyeva, I., Lukin, V., Vozel, B., Shelestov, A., Kussul, N. (2021) Discrete Atomic Transform-Based Lossy Compression of Three-Channel Remote Sensing Images with Quality Control. *Remote Sensing*, 14:125. https://doi.org/10.3390/rs14010125

Pons, X. (2021): The risk of reinventing the wheel in geographic information formats/ El riesgo de reinventar la rueda en los formatos de información geográfica, *GeoFocus (Editorial), Revista Internacional de Ciencia y Tecnología de la Información Geográfica*, 27, 1-3. http://dx.doi.org/10.21138/GF.738

Zabala A., Pons X. (2011) Effects of lossy compression on remote sensing image classification of forest areas. *International Journal of Applied Earth Observation and Geoinformation*, 13 (1): 43-51. DOI: 10.1016/j.jag.2010.06.005