Central European Geology, Vol. 54/4, pp. 315–326 (2011)
DOI: 10.1556/CEuGeol.54.2011.4.1

Lessons learnt from the adoption of open source GIS for quality control of the Hungarian LPIS

János Dolleschall, Ádám Podolcsák
Compet-Terra, Szeged

Compet-Terra is carrying out QA (Quality Assurance) services for the Hungarian Agriculture and Regional Development Agency. Quality control of the LPIS is one of the important QA duties. The LPIS data set is updated yearly according to the cycles of agricultural subsidies. Compet-Terra elaborated an Open Source Software based checking method that could be adopted by the client without the financial expense of software.

QuantumGIS (as the primary tool), gvSIG, and LibreOffice were used for the quality control of the LPIS. Four primary quality types were checked: (1) the land parcel attributes, (2) the areas of the polygons, (3) the completeness of the content and (4) the topological quality. For these investigations topological GIS functions and database management functions were used. The most important functions were topological difference calculation, polygon area calculation and geometry validity check. Complex procedures were also carried out with proprietary tools to compare the results and the run-time performances.

The result of the adoption of OSS GIS tools for LPIS checking was positive. All the planned procedures could be implemented using OSS GIS. OSS tools proved to be robust, reliable, user-friendly and performed well.

Key words: Quality Assurance, LPIS, Open Source, OS GIS, QGIS, LibreOffice

Introduction

Compet-Terra contributes to the Danish BlomInfo-provided Quality Assurance (QA) services for the Hungarian Agriculture and Regional Development Agency

Addresses: J. Dolleschall: H-6753 Szeged, Budai Nagy Antal u. 69, Hungary,
e-mail: djanos@competterra.com
Á. Podolcsák: H-6721 Szeged, Kálvin tér 2, Hungary,
e-mail: adam.podolcsak@competterra.com
Received: August 28, 2010; accepted: December 16, 2011

(ARDA) regarding the IT development of Integrated Administrative and Control System (IACS) and the included Land Parcel Identification System (LPIS). This is an external QA service which is additional to the QA activities of the IT developers and ARDA. The main objective is to assure all stakeholders that the IT system is maintained according to the related quality standards and keep the project transparent for all interested bodies.

IACS is a very complex and comprehensive IT application to administer and control the agricultural subsidies that covers claim management, administrative controls, on-the-spot checks, calculation of payments (including reductions and sanctions), reporting, monitoring, business process management, document management, archiving and Management Information System. IACS requirements are defined by EU and Hungarian regulations.

IACS embraces the Land Parcel Identification System. LPIS defines and identifies the agricultural parcels that are entitled to claim agricultural subsidies. The Land Parcel Identification System (LPIS) is identified as a basic component of the IACS. The Hungarian LPIS is an orthophoto-based physical block system. The data of the LPIS have been renewed every 3–5 years. Currently the Hungarian state institute FÖMI provides an LPIS system based on data procured from private and state bodies.

The quality of the LPIS includes several aspects that can cover the following procedures: data gathering, interpretation, compilation, harmonization (Grandgirard and Zielinski 2008). The present IACS QA project is responsible for checking whether the LPIS system complies with the ARDA-imposed business requirements. These requirements are set up independently from the standard LPIS requirements and address the riskiest features.

Open Source tools

Nowadays the usage of Open Source tools in GIS is growing. In recent years GIS specialists have begun to use Open Source software for their work in all fields of GIS, from geomorphologic mapping (Mantovani et al. 2010) to health geography (Vanmeulebrouk et al. 2008). The most commonly used OS GIS software are GRASS, Quantum GIS, gvSIG, SAGA or ILWIS.

For quality control purposes, three Open Source tools were selected, based on the following factors: user friendliness, good performance and wide functionality. On this basis two GIS and one office software were chosen.

The primary tool for the investigation was Quantum GIS. The other OS GIS software (gvSIG) was used to check the calculations (when QGIS results were different from the original values). These two software packages were called "basic GIS software" and "popular Open Source software for GIS education" by Tsou and Smith (Tsou and Smith 2011). In many universities this software is used for basic GIS education. A good example is the University of Szeged in Hungary, where geography and geology students begin their GIS education with Quantum GIS (Szatmári 2010).

"Quantum GIS (QGIS) is a user friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo). It runs on Linux, Unix, Mac OSX, and Windows" (http://www.qgis.org, 2011.05.10.). In QGIS one can find all the main GIS functionalities for a widespread GIS analysis or investigation. A large variety of different vectors, rasters, database file types and WebGIS services (like WMS) can be used for spatial analysis. Several vector and raster tools can be found in this software, such as geoprocessing tools, geometry tools or a raster calculator. Digitizing and creating layouts are other strengths of QGIS.

A good alternative for QGIS is gvSIG. This software is also a "Geographic Information System (GIS), that is, a desktop application designed for capturing, storing, handling, analyzing and deploying any kind of referenced geographic information in order to solve complex management and planning problems. gvSIG is known for having a user-friendly interface, being able to access the most common formats, both vector and raster. It features a wide range of tools for working with geographic-type information (query tools, layout creation, geoprocessing, networks, etc.)" (http://www.gvsig.org/web/projects/gvsig-desktop/description2, 2011.05.10.). With gvSIG most of the quality control procedures can be repeated, so if the result with QGIS is questionable, it can be reproduced with gvSIG.

The third tool for quality control was LibreOffice Calc. This application was used for simple data table management tasks (if GIS functions were not needed). "LibreOffice is the free power-packed Open Source personal productivity suite for Windows, Macintosh and Linux that gives you six feature-rich applications for all your document production and data processing needs: Writer, Calc, Impress, Draw, Math and Base." (http://www.libreoffice.org, 2011.05.10.). In this case only the spreadsheet function was used.

The data model of the Hungarian LPIS

The data of the Hungarian LPIS are created and managed by the Institute of Geodesy, Cartography and Remote Sensing. The Hungarian LPIS has three different data types: the data of the blocks (land parcels), the data of the non-eligible areas and thematic data. The areas of blocks are continuous, containing the entire area of the country (93,036 km^2). Eligible areas are obtained by calculating the difference between the blocks and the non-eligible areas. The thematic areas are the ones affected by forestation, including those under the responsibility/ownership/control of the Ministry of Defense, the NATURA 2000 areas, the reed farms and the slopes. Most of these are non-eligible, so they must be contained in the layer of the non-eligible areas (Csornai et al. 2005).

Behind the geometry of the blocks there are many different attribute data. The most important of these are the areas of the blocks (the entire area and the eligible area), the areal tolerance of the blocks, the eligible areas and the non-eligible areas, the ID, the county and the city of the blocks and different indicators concerning the land quality (less favored, environmentally sensitive, nitrate-sensitive, high nature-valued, wind erosion-affected, flood and Natura2000 areas – Csornai et al. 2009).

The data is stored in shapefiles. "A shapefile stores nontopological geometry and attribute information for the spatial features in a data set. The geometry for a feature is stored as a shape comprising a set of vector coordinates" (ESRI 1998). A shapefile consists of three separate files, e.g. blocks.shp, blocks.shx and blocks.dbf. The main file (.shp) stores the geometric information, the database file (.dbf) stores the attribute information and the index file (.shx) stores the offset and content length for the I'th record in the main file (ESRI 1998). In shapefiles point, line and polygon objects can be stored. In the Hungarian LPIS all objects (including the roads and rivers) are polygons.

The shapefile of the blocks consisted of 373222 land parcels; the size of the file (.shp + .shx + .dbf) is 518 Mb. The shapefile of the non-eligible areas consisted of 813340 polygons; the file size is 598 Mb. The thematic layers vary from 2 to 12 Mb, except for the layer of the slopes, which is 1.2 Gb.

European LPIS conceptual model

There are many European proposals for the quality standards of the LPIS. Among these proposals the most important is the LPIS conceptual model: application scheme, feature catalog, and discussion paper. The prepared standard uses the ISO 19100 family, of which the most important standards are ISO 19113 (Geographic information – Quality principles), ISO 19114 (Quality evaluation procedures), ISO 19115 (Metadata), ISO 19138 (Data quality measures), and ISO 19101 (Reference model). Table 1 summarizes the LPIS quality model of the prepared standard.

The use of this conceptual model in the Hungarian LPIS could provide many benefits, but it requires the upgrade of the QA of the LPIS data producer and some improvement of the data model. At this time only narrower investigations could be performed. However, it is planned to gradually adopt the above-mentioned standards from the next year onward.

The investigations

Among the data of Hungarian LPIS, four primary quality types were checked: (1) the consistency of the land parcel attributes, (2) the areas of the polygons, (3) the completeness and the correctness of the geometric content and (4) the topological quality. For these investigations topological GIS functions and

Table 1
LPIS quality model (Sagris and Devos 2008)

Data Quality Element ISO 19113	Data Quality Sub-elements ISO 19113	Example of an anomaly (identified non-conformity)	Cause of such non-conformity
Completeness	commission omission	excessive data (planning) missing data	poor mapping erratic processing
Logical consistency	conceptual consistency codelist consistency format consistency topological consistency	data structure error	poor data processing
Positional accuracy	absolute or external accuracy relative or internal accuracy gridded data position accuracy	accuracy error	poor mapping
Thematic accuracy	classification correctness non-quantitative attribute correctness quantitative attribute accuracy	classification error	change of concept/ poor mapping / processing
Temporal accuracy	accuracy of a time measurement temporal consistency temporal validity	outdated value	terrain change

database management functions were used. The most important functions were topological difference calculation, polygon area estimation and geometry validity checking. Complex and performance-required procedures were also completed with proprietary tools to compare the results and run-time performances. For the performance test a Dell N5010 notebook was used. Hardware specification: Intel Core i3 M380 2.53 GHz, 4 Gb DDR3 SDRAM, 500 Gb SATA HDD. Software specification: Ubuntu 11.04 and QGIS 1.7, Windows 7 and ArcGIS 10.

(1) During the examination of the land parcel attributes the block ID, the county and city of the blocks and the land quality attributes were checked. The block IDs must contain integer numbers and strings and each must be unique. The county, city and land quality indicators must be integer numbers for all records. These requirements could easily be checked with basic data table management functions, such as ordering columns, searching for empty records and sequential comparison of the records. For this examination QGIS and LibreOffice were used as primary tools and ArcGIS to compare the performance of the Open Source and proprietary GIS software.

The performance comparison of QGIS and ArcGIS showed very small differences. While QGIS sorts a column faster, ArcGIS is quicker in finding a value. During these two functions all run-times were under 30 seconds (for a shape file with 373 222 records).

(2) The second part of the quality control was the polygon area estimation. For all the layers (blocks and thematic layers) the polygon areas are provided by the file creators. The goal was to examine if these calculations were correct. For the land parcels there are two area attributes. One is for the entire area of the block and one for the eligible area of the block. The whole area estimation is a simple geometric operation (Detrekői and Szabó 2002). For the eligible area estimation a topological difference had to be calculated between the blocks layer and the non-eligible areas layer. Then, on the resulting eligible areas layer, the areas of the polygons had to be estimated. This was a two-part operation. The first part was an arithmetic operation with geometric data (Detrekői and Szabó 2002) followed by a geometric operation. The areas were also estimated for the thematic layers (e.g. areas affected with forestation, slopes, etc.). For these examinations QGIS was used. The areas were also calculated with gvSIG, to compare the area calculation method. The performance of QGIS was compared to that of ArcGIS. In ArcGIS the "erase" function was used.

During the topological difference calculation, ArcGIS was faster than QGIS. With ArcGIS the run-time was around 22 minutes, with QGIS it was around 55 minutes. But the area estimations were faster with QGIS: it was 30 seconds at the case of the whole area of land parcels. With ArcGIS this area estimation was around one and a half minutes. The estimated areas were the same in QGIS, gvSIG and ArcGIS for 2-decimal precision.

(3) The next step of quality control was the examination of the block tolerances, followed by the examination of the completeness of non-eligible areas. Every land parcel, eligible area and non-eligible area has tolerance rate attributes. The tolerance rate must be in accordance with the eligible area and its delineation: if the eligible area is nil, 0 area can be tolerated. If the eligible area is not nil, the tolerance rate cannot be 0. These requirements are easy to test with basic data table management, such as selecting the "0" and "not 0" eligible areas and checking their tolerance values with conditional expressions in LibreOffice:

if eligible-area = 0 & tolerance = 0 then test-result must be "ok";

if eligible-area > 0 & tolerance > 0 then test-result must be "ok".

During the completeness test the non-eligible areas were examined. This layer must contain all the thematic layers whose content is a non-eligible area (e.g. areas affected with an forestation layer). If a thematic layer is contained in the non-eligible areas layer, the topological difference must be an empty layer. If the result is not an empty layer, there can be two different situations: (1) there are entire polygons in the thematic layer which do not show up in the non-eligible layer (Fig. 1) or (2) not all parts of the polygon (not the whole polygon, but a part of it) is contained in the non-eligible layer (Fig. 2). The second case is usually a topological error. Therefore the topological differences were calculated between the thematic layers and the non-eligible layer (thematic layer – not-eligible area layer).

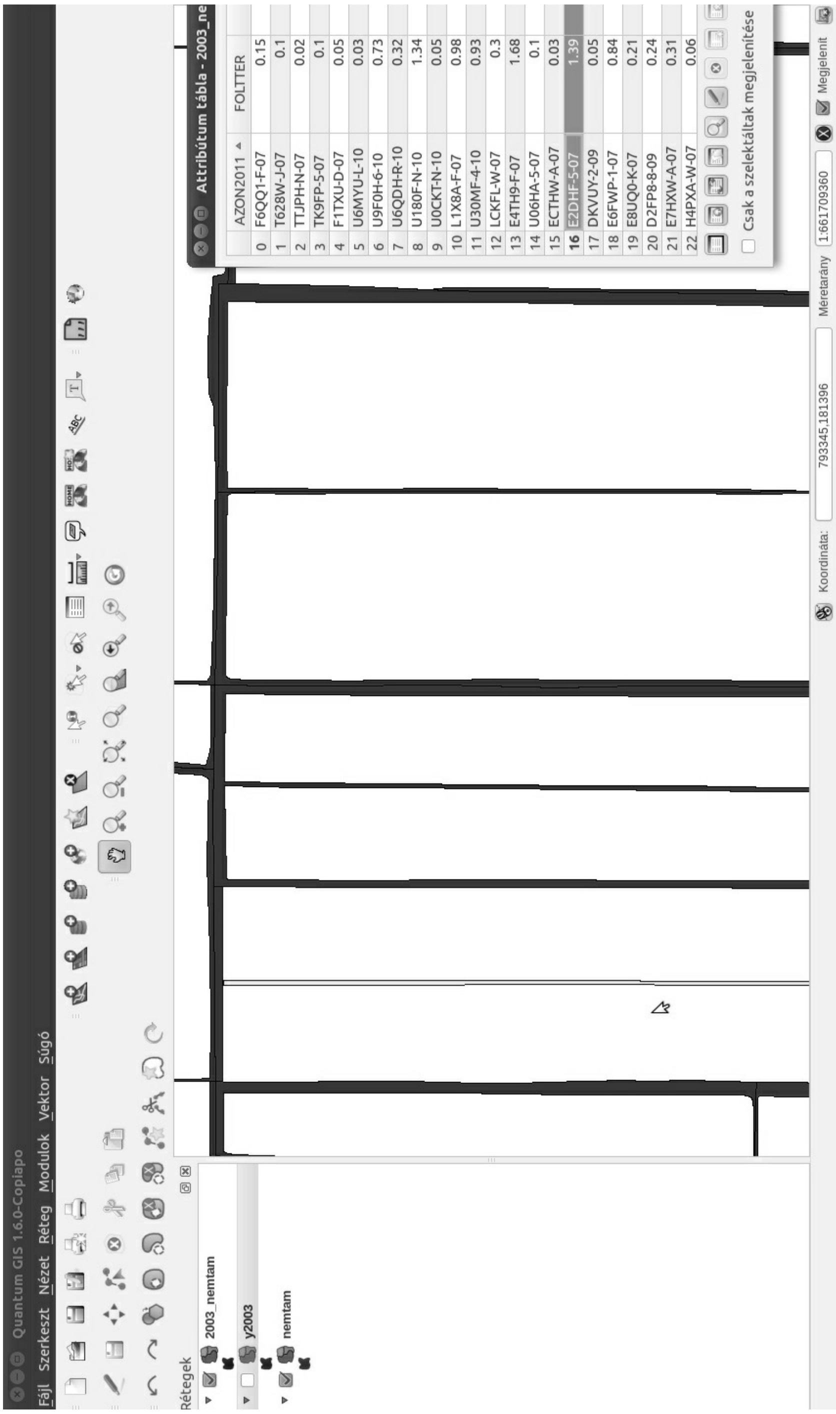

Fig. 1
The entire polygon (light grey-colored) is missing (author: J. Dolleschall)

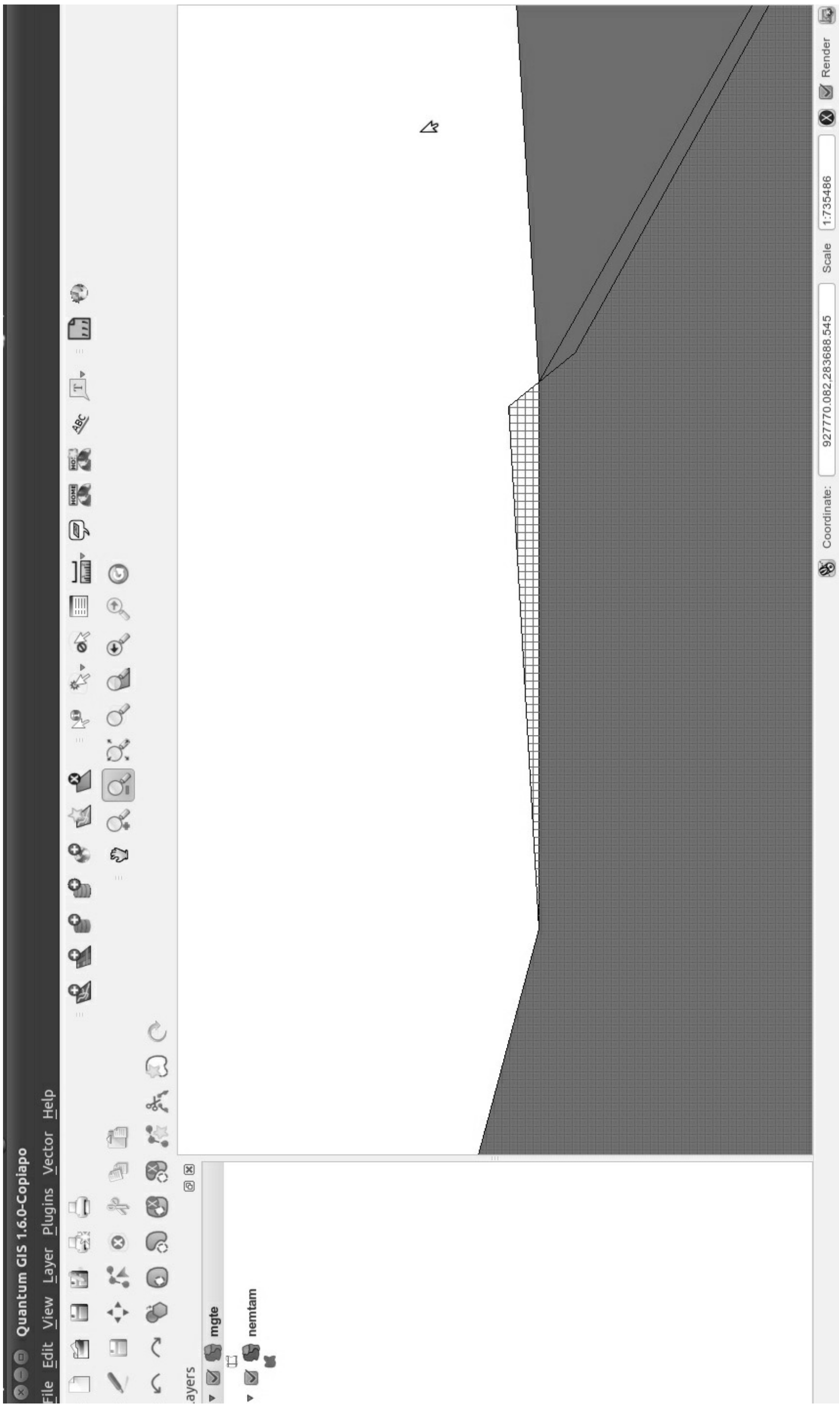

Fig. 2
A part of the polygon is missing (author: J. Dolleschall)

In these topological calculations the difference of a small (max. 12 Mb) and a huge (598 Mb) shape file was calculated at all times. During these calculations QGIS was significantly faster than ArcGIS. QGIS completed the difference calculation between areas affected with an forestation layer and non-eligible areas layer in one and half minutes, but ArcGIS only in 12 minutes.

(4) The topological quality or topological correctness was examined with automatic geometry validity check for the blocks, for the non-eligible areas and for the areas affected with forestation. A large number of very different types of topological errors can occur in the vector data. The most common topological error types in spatial vector data:

1. Floating or short lines
2. Overlapping lines
3. Overshoots and undershoots
4. Unclosed and weird polygons" (Maras et al. 2010).

However, sometimes it is difficult to decide whether a weird-looking polygon is topologically correct or not. Figure 3 is a good example of this: in some places the light grey-colored polygon is too thin (line-like). It can be a real and topologically correct object if the thin parts are thinner than the resolution of the source data (e.g. the aerial photo from which the polygons were digitized), but it can be a topological error as well. This geometry validity check was also carried out with both QGIS 1.6 and QGIS 1.7. QGIS 1.6 lists this problem as an error, but QGIS 1.7 and ArcGIS do not.

After the automatic topology test, the delineation of non-eligible areas was checked. The delineation of the non-eligible polygons must be within the blocks or the blocks and non-eligible areas must have the same border. Therefore the non-eligible polygons cannot intersect with the delineation of the block polygons. For this test a spatial query had to be set up. If all the polygons of the non-eligible layer are contained in the polygons of the blocks layer, there is no intersection. If two polygons have the same border they are contained in each other, but – as most of the GIS software interpret it – they also intersect with each other. So the form of the spatial query in QGIS is "select source features from blocks where the feature contains reference features of non-eligible". If there is no error in topology, all features in the blocks will be selected.

The geometry check was faster with QGIS (around 3 minutes), but it was very slow with ArcGIS (around 23 minutes). The spatial query for these large layers was around 27 minutes with QGIS and around 9 and a half minutes with ArcGIS.

The client's reaction

The investigation and the results were presented to the representatives of the client and the applied functionality of QGIS was demonstrated to the ones who had a GIS background. They easily understood the manner of operation,

Fig. 3
Weird polygon (light grey) (author: J. Dolleschall)

appreciated the simplicity of the user interface and expressed their interest in implementing related training to all the staff who work with LPIS.

Table 2
Run-time performances

GIS function	QGIS (minutes)	ArcGIS (minutes)
Basic data table functions	0.2–0.3	0.2–0.3
Area estimation for 373222 polygons	0.5	1.5
Topological difference of two large (500–600 MiB) files	55	22
Topological difference of large (600 MiB) and small (12 MiB) files	1.5	12
Geometry check	3	23
Spatial query	27	9.5

Lessons learnt

As described in the previous chapter, all the examinations were able to be carried out with Free and Open Source software. During the test QGIS showed good performance with no error. Table 2 indicates the run-time performance of QGIS and ArcGIS.

In some functions QGIS is significantly faster than ArcGIS, but in other functions ArcGIS is faster. So it cannot be fairly said that any one of these software packages showed better performance than the other. They seem to be equal tools for similar investigations. An important difference between the two, however, is the price. Openness is also an important difference, because if a software is open, one can understand how it works and the implementations of the functions could be examined. QGIS is a Free and Open Source software and ArcGIS is a proprietary one. Nevertheless it must be admitted that ArcGIS has far more functions than QGIS. But more is not always better, because the usage of a software with wide scale functionality is harder to learn. If one doesn't need all these functions, this can be a waste of time (and money in this case). For the investigations described above the functionality of QGIS was sufficient. QGIS functions can be easily found and there are many documentations and tutorials for this software.

References

Csornai, G., B. Csonka, Gy. Zelei, G. Nádor, L. Tikász, I. László, L. Martinovich, A. Kocsis, E. Bognár, Zs. Suba, Cs. Wirnhardt, P. Winkler 2005: Crop monitoring basis to the area based subsidy control and Land Parcel Identification System (LPIS-Hu) in Hungary. – New Strategies for European Remote Sensing, Proceedings of the 24th Symposium of the European Association of Remote Sensing Laboratories, Dubrovnik, pp. 507–514.

Csornai, G., E. Bognár, G. Mikus, Cs. Wirnhardt, Sz. Mihály 2009: Maintenance and Development of the Hungarian Land Parcel Identification System (LPIS-HU) for IACS. – Conference presentation. FIG Working Week 2009, Eilat, Israel.

Detrekői, Á., Gy. Szabó 2002: Térinformatika (GIS). – Nemzeti Tankönyvkiadó, Budapest, 380 p.

ESRI 1998: Shapefile Technical Description. – White paper, 34 p. (www.esri.com/library/whitepapers/pdfs/shapefile.pdf)

Grandgirard, D., R. Zielinski 2008: Land Parcel Identification System (LPIS) Anomalies' Sampling and Spatial Pattern. – JRC Scientific and Technical Reports, European Communities, Luxembourg, 84 p.

Maras, S.S., H.H. Maras, B. Aktug, E.E. Maras, F. Yildiz 2010: Topological error correction of GIS vector data. – International Journal of the Physical Sciences, 5/5, Academic Journals, pp. 476–483.

Mantovani, F., F.J. Gracia, P.D. de Cosmo, A. Suma 2010: A new approach to landslide geomorphological mapping using Open Source software in the Olvera area (Cadiz, Spain). – Landslides, 7, pp. 69–74.

Sagris, V., W. Devos 2008: LPIS Core Conceptual Model. – JRC Scientific and Technical Reports, European Communities, Luxembourg, 67 p.

Szatmári, J. 2010: A szabad szoftverek szerepe a geoinformatika oktatásban. (The role of open software in the education of GIS). – Conference presentation. Free Software Conference 2010, Szeged, Hungary.

Tsou, M.-H., J. Smith 2011: Free and Open Source Software for GIS education. – White paper, 18 p. (www.iapad.org/publications/ppgis/tsou_free-GIS-for-educators-whitepaper.pdf)

Vanmeulebrouk, B., U. Rivett, A. Ricketts, M. Loudon 2008: Open source GIS for HIV/AIDS management. – International Journal of Health Geographics, 7, 53 p.