CO-268

AUTOMATIC PROCESSING OF THE FIRST RELEASE OF DERIVED STATE MAPS SERIES FOR WEB PUBLICATION

TALICH M., ANTO F., BÖHM O.
Research Institute of Geodesy, Cartography and Topography, ZDIBY, CZECH REPUBLIC

1. Digitizing center ODIS VÚGTK

In order for the digitizing of old cartographic products in the Research Institute of Geodesy, Topography and Cartography (VÚGTK) to be on a high level, a new digitizing center was founded in year 2007. This center is a part of the Branch Information Center (ODIS) which also operates the Surveying library. In this library are stored some old and valuable publications from the field of geodesy and cartography. Important part of the digitizing center is large flat bed scanner Trias Vidar on which it is possible to digitize artworks up to format A0+ without risk if damage to the original. Parts of the center are also workstations for digital images processing and several-terabyte data storage for archiving and for backing up digital maps and publications. The Digitizing Center also runs a web server mapy.vugtk.cz, where old digital maps are published. On digitizing of map collections we cooperate with The Institute of History of Academy of Sciences of the Czech Republic and with Central Archive of Surveying, Mapping and Cadastre.

Fig. 1 – New digitizing center in VÚGTK

2. Derived state map 1 : 5 000 (SMO5)

After the Second World War new mapping has begun in Czechoslovakia and thanks to this mapping a new national map series should have been created for the needs of urban planning and for public use. For these purposes the scale of 1 : 5 000 was chosen as the most suitable scale and the maps should have consisted of detailed planimetry and altimetry. Mapping in this scale was slower than necessary and that is why a decision was taken that as a provisional solution a new uniform map series in continuous and uniform map projection of the whole territory of former Czechoslovakia will be created from maps which were already available. This national map series was named Derived State Map 1 : 5 000 (SMO5).

The planimetric component of the map is drawn by black color and was derived from cadastral maps. Cadastral maps covered in the form of maps of selected areas the whole territory, but in different scales and map projections. That is why the planimetry had to be generalized first, than modified and adjusted and after that translated by photoreproduction methods into the map section of SMO5. From planimetry of cadastral map land parcel numbers were left out and building parcels were marked by a dot inside build-up areas.

The altimetry was taken mainly from topographic maps of a scale 1 : 10 000 or 1 : 20 000, but in most cases contours were interpolated from a scale 1 : 25 000. Because the altimetry component of map were transformed from different kinds of large scale maps, the accuracy of contours is only approximate but sufficient for all purposes of its time. Contours are drawn in map by brown color and their equidistance is varied.

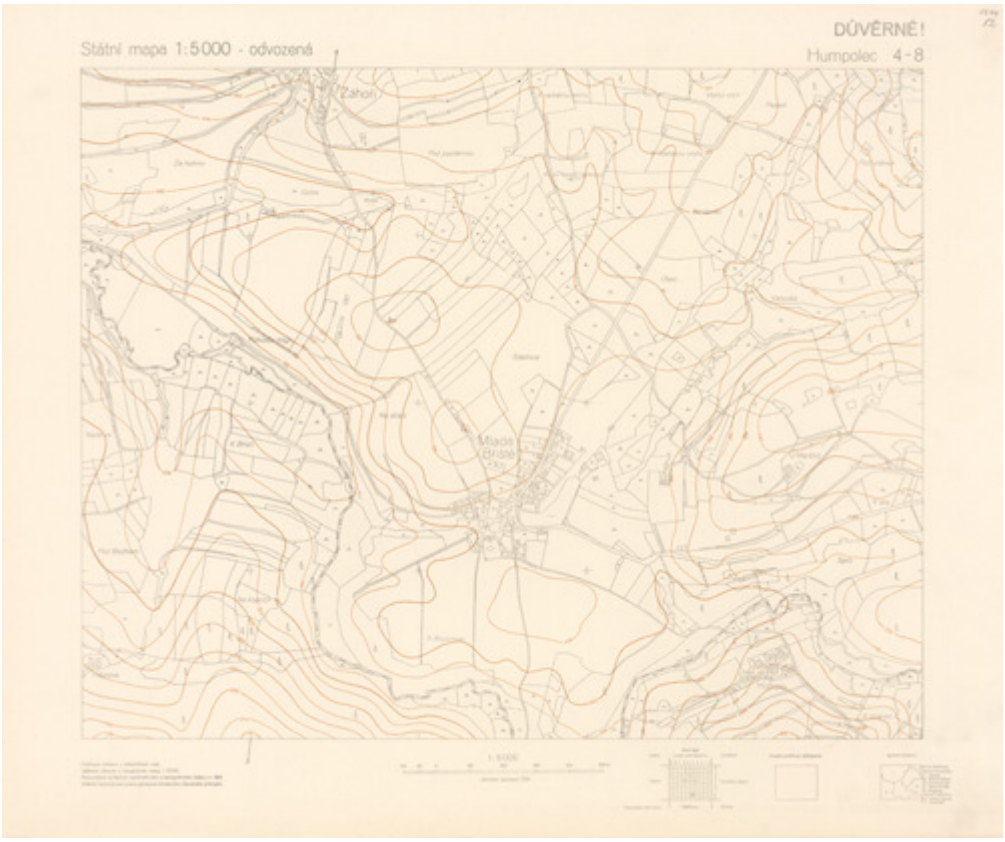

Fig. 2 – A map sheet of the Derived State

The first release of Derived State Map for the whole territory of former Czechoslovakia has taken almost 10 years and although it was intended as a temporary solution, until a new mapping for whole territory will be done, this national map series has been maintained until to this day. This map series is the most detailed national map series of large scale and in numbers of 16 193 map sheets covers the whole territory of the Czech Republic. The first release is interesting for display of boundaries of plots that show the state of land before land consolidation during the communism era.

The Derived State map is nowadays a binding state map series which is maintained and released by governmental office. The series is maintained in geodetic reference system of the Datum of Uniform Trigonometric Cadastral System (S-JTSK) and the map layout is coming from orthogonal coordinate system in a plane of Křovák´s universe conform conic projection where parallels with axes X and Y form orthogonal sections with proportions of 50x40 cm. The map sheet names are derived from fictive map of the scale 1 : 50 000 which is delimited by parallels by 20 kilometers with axis Y and by parallels by 25 kilometers with axis X. This fictive map is divided again by parallels into 10 columns and 10 rows so now the fictive map consist of 100 map sheets of SMO5 in the scale 1 : 5 000. The map sheet is named after the fictive map which is named after an important city in the map and the name continues by a number of column and row. Columns and rows are enumerated from 0 to 9 in a west east and north-south direction. For example, the SMO5 map of Prague territory is named like Prague 4-6. In ideal case the fictive map has all 100 map sheets, but in border areas the maps out of the Czech territory do not exist or in some places map sheets are missing. The whole process of digitalizing and image processing proceeds in a batches of sets which are formed by just one fictive map.

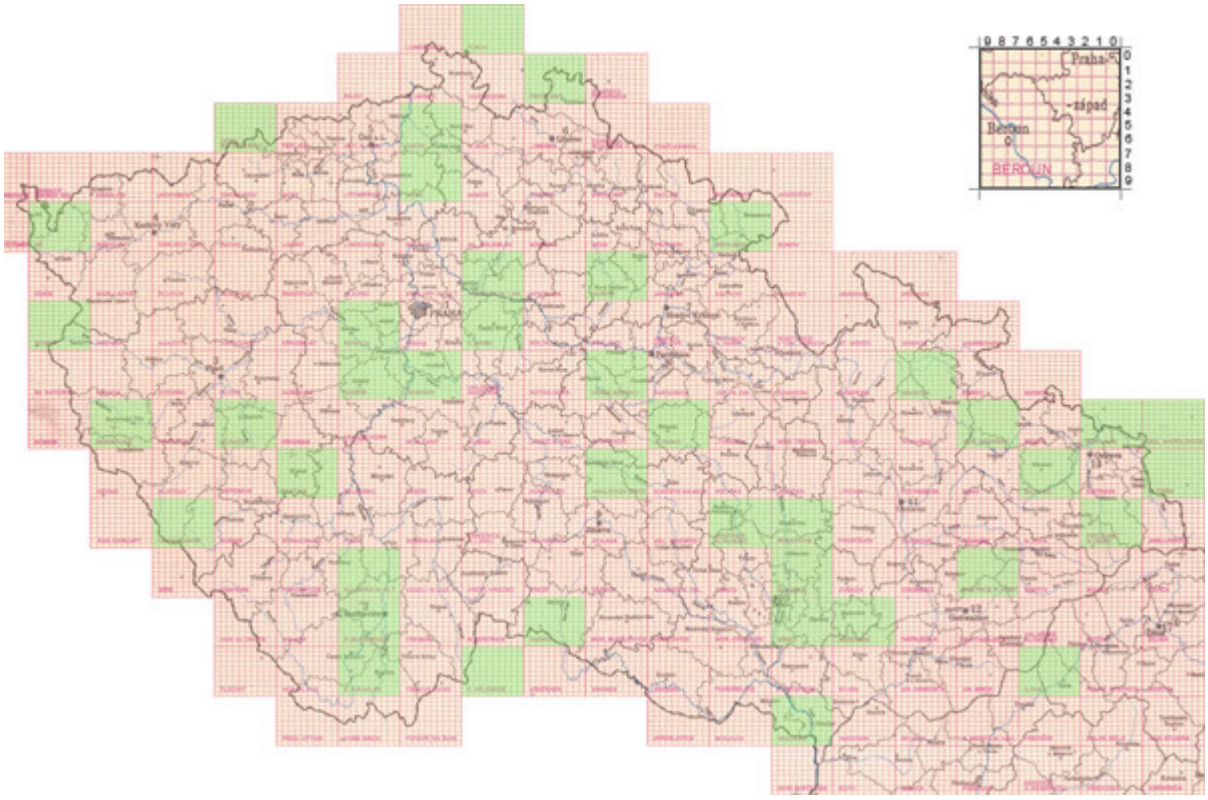

Fig. 3 – The map layout of the fictive maps (the green are processed) and upper-right the map layout of SMO5 maps for fictive map Beroun

3. The process of map SMO5

3.1 Digitizing

As previously stated, the Derived State Map covers the whole territory of the Czech Republic in the 16 193 map sheets. Approximately 90% of all map sheets from the first release was preserved to this day, which still presents a lot of maps. Because we planned from the beginning of digitizing to publish the whole map series on internet by using different web technologies, we have chosen a suitable workflow, which would allow the routine parts of processing of the digital copies of maps to be automated. The digitizing works are done in the digitizing center of VÚGTK by a large format flat bed scanner Trias Vidar which has a certification for cartometry scanning. A flat bed scanner captures only an area which is determined in computer that controlls the scanner. This is an advantage, because a scan of map has always the same size in pixels. This is used in the first step of automatic processing. A disadvantage of this type of scanner is that scanner captures not only a map but also a pad as a background which pushes the map to scanner´s glass. The background has to be cut out from the image in the second step of automatic processing.

We use the following configuration of the scanner's parameters. The optical resolution of the scanner is set to 400 dpi, the color depth to 24 bits and images are stored in Tiff file format. This format is suitable for subsequent automatic processing of raster images.

Fig. 4 – The flat bed scanner Trias Vidar

The proportions of map sheets of SMO5 maps are not uniform. Basically there three different sizes - 70x50 cm, 60x50 cm and a specific size. The specific size is used for border regions where a map drawing is displayed only in a small part of a map sheet. This is solved either by reduction of a map frame or by joining the map drawing to the neighbor map sheet and drawing over the map frame. These specific cases are automatically processed only in part and they need to i a large part manual processing.

Both digitizing and digital maps processing runs in the sets of the fictive maps, i.e. in the maximum amount of 100 map sheets. Thanks to the format of the scanner two maps of SMO5 can be digitized at the same time. This makes digitizing more efficient, but it also demands more attention to an operating staff. Maps must be precisely placed on a scanner´s pad marks which guarantee us that digital maps do not have to be straightened. The digitizing of the map set starts from a map named 0-0 and a filename is set to reflect this. The filename is automatically generated by the scanner after capturing the image. If the numbering is not continuous, because of missing maps sheets, then the operating staff follows a set protocol. Either digitize only one map sheet or change the filename. This schema ensures us a correct automatic naming in the first step of processing when the scan image is being split into the individual map sheets.

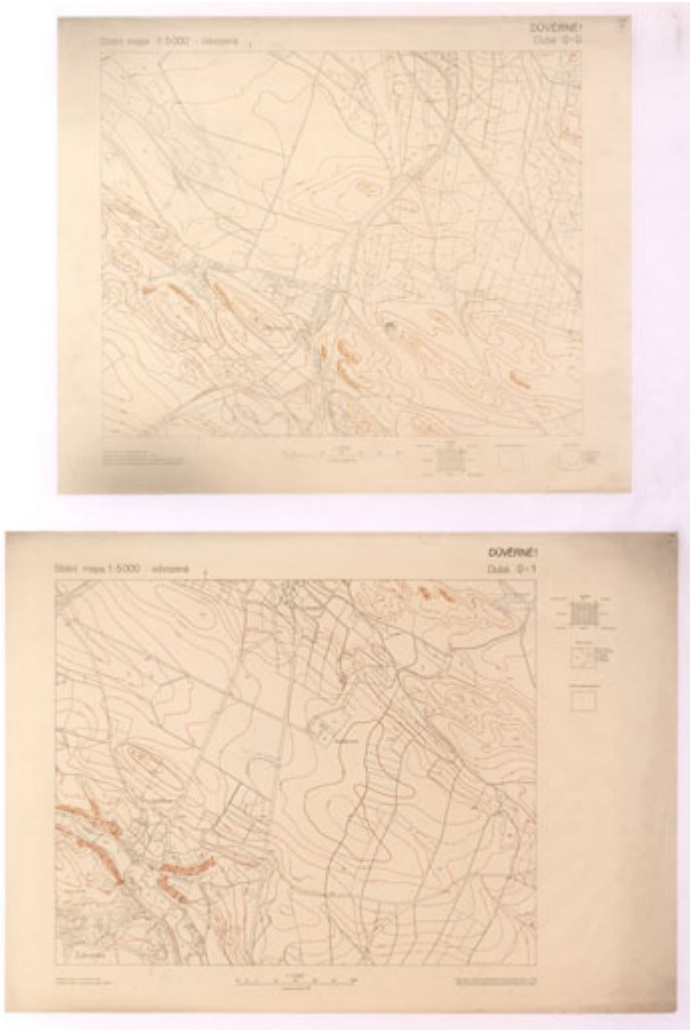

Fig. 5 – The scan with two different sizes of map sheets

When the scanner is scanning maps, the operating stuff fills metadata about maps into a database. The database is used for archiving of digital copies and for generating an overview of maps available on our map server.

3.2 Automatic raster image processing

The process of preparing maps for internet publication is relatively straightforward, but it comprises quite a lot of routine tasks especially if you need to prepare the output in several formats. The flowchart (fig. 6) below shows schematically our workflow when processing maps for online publication.

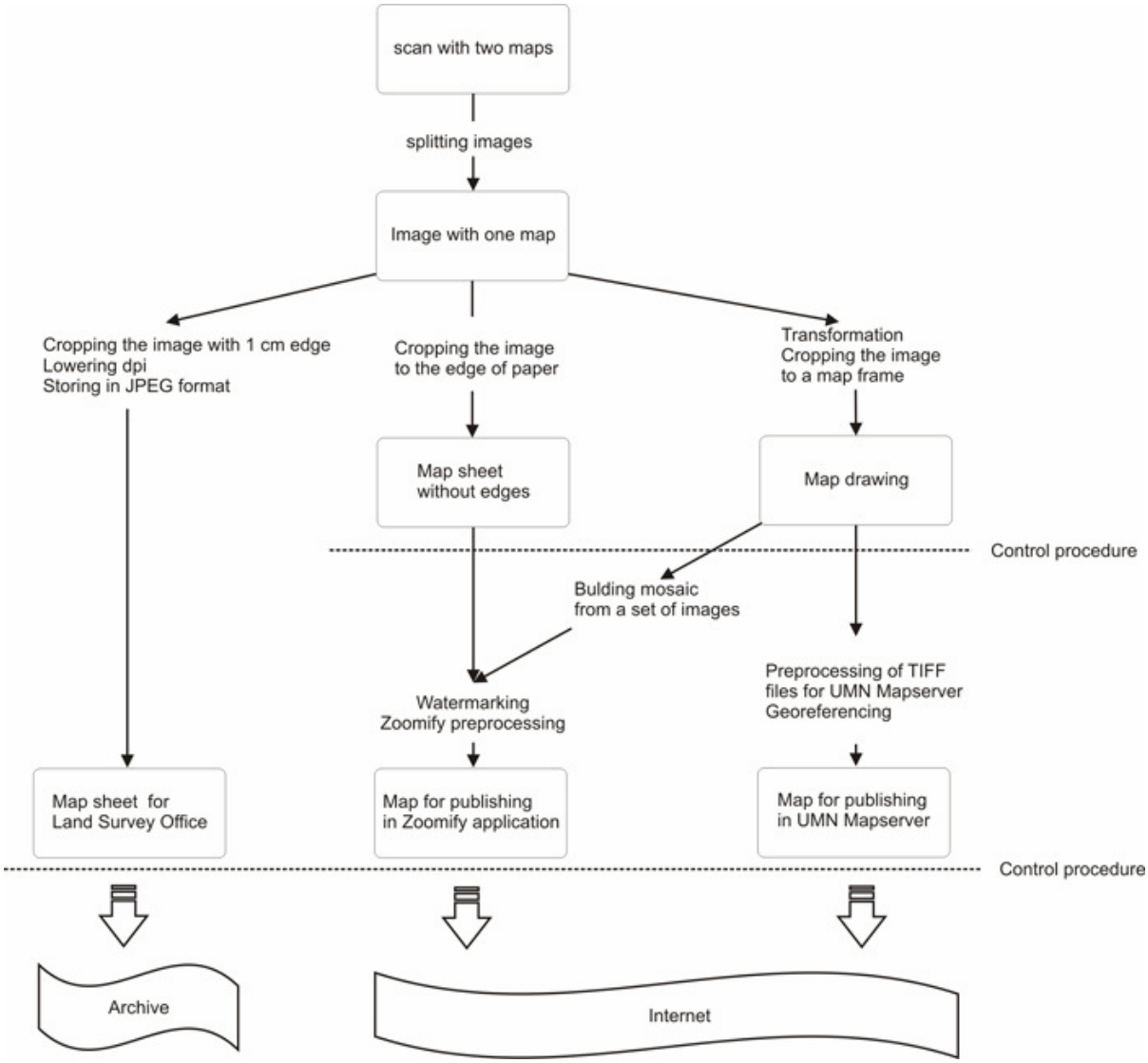

Fig. 6 – The flowchart of our workflow

It's easy, if tedious, to do this manually for individual maps or even small map sets. But in our case the SMO5 map set contains over 16 000 maps, so doing all of the processing manually would take a ridiculous amount of time and repetitive work. So we started to look for ways to automate the process as much as possible.

We used the Python programming language, the Python Imaging Library (PIL) and exiftool and exiv2 command line utilities to develop Merlin - a command line tool that takes the raw scanned images and produces images in various formats and resolutions. The results are either used directly as archive images or further processed by tools like Zoomify or FWTools for online publications (via Zoomify or MapServer).

Merlin works on data sets of arbitrary size. It takes a directory as its input parameter and processes all images found in this directory. The resulting images are named according to the original filenames. That means that if we keep a sensible naming scheme of the scanned images, the results are named in accordance with the map set's nomenclature.

To accomplish its work Merlin must be able to:

• Split the images
• Crop images
• Watermark images
• Read and write metadata in the form of exif header.

Each of these tasks is performed by Merlin automatically. The tasks are discussed in details in the following paragraphs.

3.2.1 Splitting images

Because of the scanner's large format, two map sheets can be scanned simultaneously. The map sheets are placed symmetrically, so obtaining the individual map sheets is done simply by splitting the original image in half.

However we need to account for the cases where there is only one map sheet. The script discerns this from the scanned image's size in pixels.

3.2.2 Cropping

Cropping is an operation that removes from the image excess areas around its edges. Excess area is a vague term though. It depends on the output's purpose and desired properties. For example both we and the map provider - the Land Survey Office - want to archive the scanned images. While we cut away only the empty area around the map sheet and keep the image with original dpi resolution, the Land Survey Office wants to cut away also the map sheet edge while still retaining all the outside-the-frame

information and they want to lower the resolution to 300dpi. And for other than archiving purposes we want to crop the image only to the map frame.

Therefore cropping is the main processing stage and the most difficult and most computing intensive one in the whole workflow. The following paragraphs describe the algorithms and tools used to accomplish it.

3.2.2.1 Cropping to map sheet edge

The purpose of cropping to map sheet edge is to get rid of the excess "blank" area around the map sheet edge. That means we need to find this edge in the image and then use it as a crop mask. In theory this should be easy if we use contrasting background for the scanning. Unfortunately we found out that some of the map sheets were too thin and colored background would shine through affecting the scanned map's colours so we were forced to use white background. Of course for white background the contrast isn't so good but it's sufficient for most cases. The basic idea behind finding the papers edge is quite simple.

We use thresholding to convert the image to a binary black and white image. Ideally the white background stays white, while the darker map sheet and especially its edge should be turned to black. Then we start the search on the image's edge and continue to the image's center. Since the background is empty, the paper edge must be the first homogenous black area encountered along the way. So whenever a black pixel is encountered a black colour frequency is computed for its surrounding area. This frequency divided by the area size gives us a measure of the area's black colour homogeneity. When the homogeneity is high enough (meaning there are almost no white pixels) we have found the paper edge.

There are some problems with this simple algorithm. The biggest challenge is to find the right threshold value so we don't lose information. If we choose the threshold too low the paper edge can degrade into only a few scattered black pixels. If we choose a too high value even the background can be colored black and the edge will again be indiscernible (see fig 7). The contrast between the map and the background simply isn't as high as it seems and more importantly it's not constant, because the map sheets are in various state of damage from age, dust etc.

Fig.7 - wrong choice of threshold value (too low on the left and too high on the right)

However the contrast becomes more pronounced if only the blue channel of the image is used. Using only the blue channel for thresholding we managed to find a good value for the threshold that allows us to detect the edge correctly in most cases. Furthermore the program tries to detect errors based on several criteria (mainly the size in pixels of the cropped image and the amount of pixels cropped). If it fails it automatically tries again with different threshold value. If even this doesn't help the program prints a warning message.

Although this method manages to crop correctly the majority of pictures it isn't completely error free and the output images still need to be checked.

3.2.2.2 Cropping to map frame

We could use the analogous algorithm to the previous case to crop the image to the map frame. We would however run into problems caused by imperfect alignment of the map with the scanner sensor - we would cut away a little piece of the map on one side and have left a bit of outside-the -map-frame area on the other side as shown in the picture below.

Furthermore we would like every resulting image to have the same size, as that would greatly simplify putting the images together to produce mosaics and tiles for online publication. The solution to this problem is looking for corners of the map frame instead of edges. These corners are then used as identical points for transforming the map frame to a new image of predefined size.

In reality the image is only slightly misaligned. We use the same algorithm as when looking for paper edges to find the map frame edges. This gives us four coordinates: $X_{min\neg}$, $X_{max\neg}$, $Y_{min\neg}$, $Y_{max\neg}$. These coordinates define approximate position of the map frame corners:

$c_1^{approx.} = [X_{min}, Y_{min\neg}]$

$c_2^{approx.} = [X_{max}, Y_{min\neg}]$

$c_3^{approx.} = [X_{max}, Y_{max\neg}]$

$c_4^{approx.} = [X_{min}, Y_{max\neg}]$

Then we look for the precise position in the vicinity of these approximate points by using pattern matching. For every pixel in the area we take a small surrounding area of the pixel (similarly to image filtering) and compare it to a template of an "ideal corner". We then choose the pixel with the best match ratio as the precise position of the corner. To avoid accepting even really bad matches only match ratios of 0.7 and higher are considered.

Fig. 8 - Loss of map drawing when cropping simply to map frame edges

The results of this script are less reliable than the cropping to paper edge. The reason is the actual map drawing, which can prevent the pattern matching from finding a good match. Even so this script can greatly reduce the amount human work needed to obtain the map tiles.

3.2.3 Watermarking

Watermarking is done by a script which simply inserts tiled watermark image into the map image and saves the result. This is the simplest part of the whole processing chain.

3.2.4 Metadata

The original scans contain various properties in the form of exif header. We wanted to keep these data for the processed images and the Land Survey Office wanted them present too. Unfortunately the Python Imaging Library doesn't write exif headers, so we had to look for other tools. There are various tools to read and write the exif headers, but the support for various properties and the read write operations is sometimes less than ideal. That's why we use two tools - exiftool and exiv2 command line utilities - to manipulate the exif header and get the desired result.

3.3 Automatic processing outputs

Each map sheet of SMO5 is after automatic processing stored into three different types of format. Each type has its own goal – archiving and web publishing.

3.3.1 The map sheet in JPEG format

The owner of the Derived State Map collection archives the digital map copies in JPEG (The Joint Photographics Experts Group) in 300 dpi resolution. Their priority is small volume of data. The JPEG is raster format using a lossy compression for reducing digital image data. In our case the volume of one digital map sheet is around 3 MB.

3.3.2 The map sheet prepared for Zoomify application

Zoomify is an internet application for interactive web publication of a huge raster data likemaps etc. We use this application for web publication of maps as a simple images.

The principle of publication is relatively easy. The digital raster maps have to be preprocessed by Zoomify software before they're uploaded on the server. This program makes a directory structure in which one big image is stored as several tens of thousands tiny files in the JPEG format. Digital map preprocessed by Zoomify software can be published on web sites and can be interactively operated, i.e it can be zoomed smoothly and panned in all directions. The interaction is fast and smooth because the image in the web browser is put together from tiles corresponding to the displayed area. Maps we published by this application are watermarked to protect them against downloading and unauthorized use.

The map image ready to be published by Zoomify is another output from the automatic processing.Maps are published in two variants. The first is publishing of individual map sheets with information on the map frame and outside the frame. The second variant is building a mosaic from the set of one fictive map. For creation of one big seamless digital map each map sheet needs to be georeferenced. More about georeferencing will be in the next chapter.

3.3.3 Map as a GeoTiff

The raster format Tiff (Tagged Image File Format) is one of the few image formats that can be georeferenced. It means the information about position in a coordinate reference system and information about a size of pixel can be joined to the digital map. The georeferenced Tiff (GeoTiff) can be published on internet by the WMS (Web Map Service).

The first part of georeferencing is cropping the image to the map frame and transforming it to a set rectangular size. This process was described in the chapter 3.2.2.2 and the outputs are maps that contain only the map drawing and have the same size in pixels.

The second part is only about automatic generating of a TFW file for each map. This file contains coordinates of a upper-left corner and the size of pixel. The size is still the same and coordinates are thanks to the regular map layout derived from the upper-left corner of the fictive map. The coordinate system of paper maps and digital maps is the same, so georeferencing is easy. In the end we are able to build the mosaic from the set of georeferenced maps and create seamless map of the fictive map.

Before uploading georeferenced maps to the mapserver the maps are saved in several resolutions as so-called pyramids. This makes serving the maps by WMS technology sufficiently fast.

3.4 Quality of outputs and accuracy of the automatic processing

The quality and the accuracy of outputs are explicitly affected by the quality of map sheet. Problems appear when the paper is not uniformly coloured (smeary, yellowed etc.) or is damaged. If it happens the command line tool detects as the edge the map frame instead of map paper and crop the map badly. The script checks the proportions of cropped maps and alerts to wrong outputs. This error occurs approximately in 2 % of all map sheets.

The accuracy of automatic georeferencing can be divided to an absolute and a relative one. The absolute accuracy can be described by comparing a real objects position with a position of objects in the map. But this accuracy is affected by technology used for the map creation when the planimetry from different scales of cadastral maps was photomechanicaly transformed into the SMO5 maps. The relative accuracy, describing how the neighbouring map sheets fit together, tells us more about the accuracy of automatic georeferencing. But the map drawing is not our interest because it was deformed by photomechanic transformation. What is not deformed in the map drawing is a grid consisting of small crosses in a density of 10 x10 cm. These crosses have their own coordinates and can be compared with real coordinates and on neighbouring map sheets half-crosses should be precisely connected. These exact connections are more affected by a paper shrink, which should be cancelled out by transformation to corners of the map frame. The accuracy of the automatic georeferencing could be better if all crosses in the map would be used in a transformation as identical points. But crosses inside the map very often collide with map drawing and automatic detection of all crosses would be almost impossible.

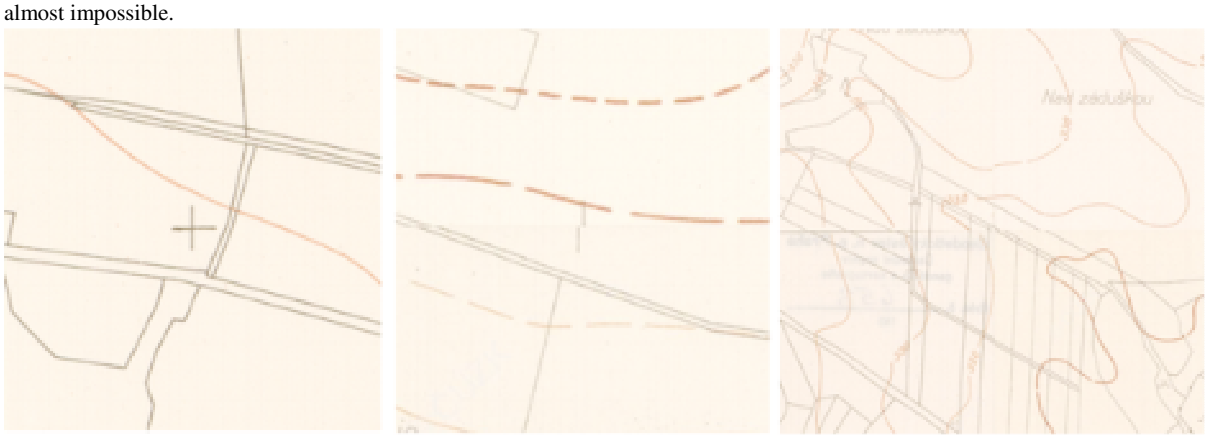

Fig. 9 – The cross in the map, half-crosses on a border of neighbouring maps, map drawing in corners of neighbouring maps

By statistical measuring deviations of crosses positions we have counted maximum deviation up to 13 meters in the scale of the map and the mean deviation in position is 4 meters. We can make a conclusion that the accuracy of georeferencing is fully suitable for archive works and for ordinary works in GIS applications.

Thanks to the automation of raw scans processing we are able to digitize and publish the whole set of 16 000 maps in two person in four months. By our estimation it is approximately ten times faster that to do it manually. Our workflow can be used not only for the first release of the Derived State Map 1 : 5 000 maps, but also for map sets of following releases.

4. Map portal http://mapy.vugtk.cz

We publish the maps we process on our map portal http://mapy.vugtk.cz. The majority of these maps are comprised of old maps of the area of today Czech Republic. All the published maps and map sets are provided with brief description. More interesting details about the maps' origins, contents and purpose are given in accompanying text.

We use two technologies to publish our maps: Zoomify and Web Map Services (WMS). Zoomify is our primary means of publishing and practically all our maps are made available via it. But we aim to make as many maps as possible available via WMS. We use the open source UMN MapServer to provide the service. As a complement we created our own WMS viewing web application.

Fig. 10 – SMO5 map in Zoomify application

The SMO5 maps are published on this map portal too. They can be viewed in two ways: Zoomify and in our WMS viewing application. Zoomify shows the SMO5 maps either as a fictitious concatenated whole (a mosaic of all the map sheets) or as individual map sheets complete with out-of-frame information. Users can select desired area in a clickable map and the application then displays the appropriate map sheet which can be zoomed and panned as desired. Alternatively users can select specific map sheets in a clickable map sheet layout overview and have them displayed individually.

More interesting is map publishing as a web map service. This has great advantage in that it keeps the cartographic properties of the maps. Therefore we can determine coordinates, azimuths, distances and areas without the need for the original paper map. On top of that we can easily compare the content of the old maps with new ones. Because of that, we think this is the most effective way of using old maps on the internet.

We connected the SMO5 maps to our WMS viewer along with our other maps: Müller's map of Bohemia (1720) and Moravia (1716) and Special Maps from III. Military Mapping (published 1923 - 1928). We connected also several other free WMS data sources: maps from II. Military Mapping, former Cadaster of Lands maps, and present day ortophoto map.

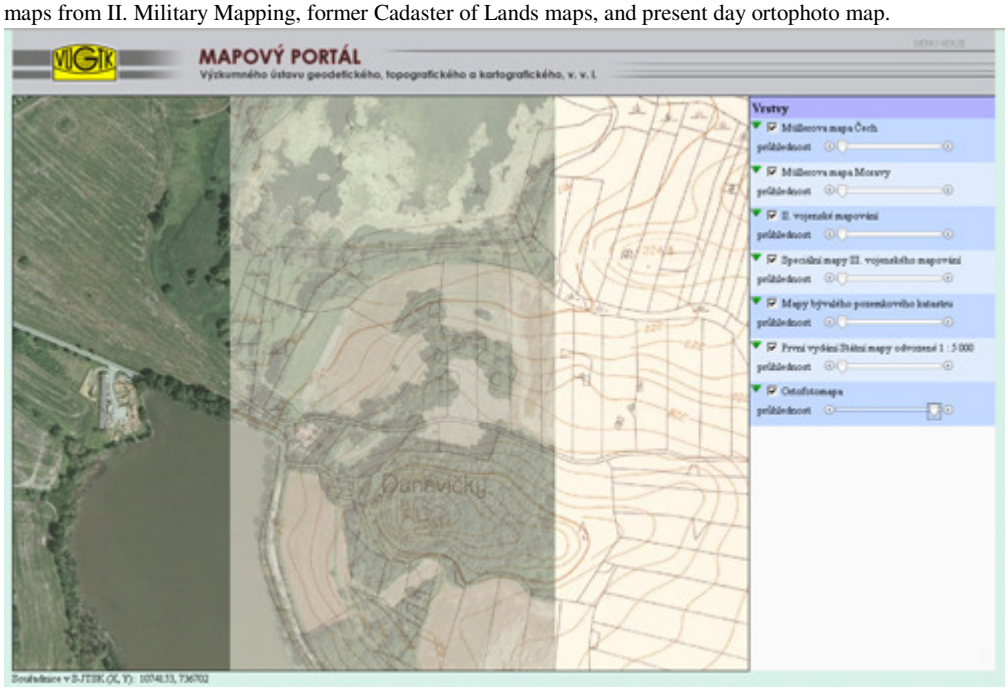

Fig. 11 – Derived State maps published via WMS and an example of the layer´s transparency

The application shows each data source as an individual layer. Users can turn these layers on and off, order them arbitrarily and set the layers' transparency. This allows to compare the contents of several map sources - for example find out the differences between II. and III. Military Mapping or compare the first edition of SMO5 to present day ortophoto.

The application is controlled with mouse and keyboard in a manner similar to commercial map portals. The status bar shows information about the current mouse position, giving the position in S-JTSK coordinate system.

5. Conclusion

Based on our experience in digitization and online publishing of old maps we can advise to concentrate on specific steps in the whole process. The first is the actual digitization – scanning, which should by done with a precise certified cartometric scanner with resolution of at least 400dpi and 24b colour depth. The following data processing can be automated to a high degree. The automation becomes a necessity when dealing with large maps sets containing thousands of map sheets. Even when the map set is relatively smaller, hundreds of map sheets, the productivity increase can be quite substantial.

When publishing maps online as whole map sheets it's important to select a technology that allows sufficiently short load times at common connection bandwidths. From our experience such a technology is Zoomify and in the future the JPEG2000 format.

However georeferencing is necessary to fully harness the cartographic properties of maps. Here the best publishing option seems to be the web map service, because it makes the maps available in a widely recognized standard and so enables many and varied uses of the maps from simple viewers to various mash-ups, overlays etc. combining several map sources. One of the more interesting uses of this is the comparison of content of various maps from various time periods.

Literature:

BOGUSZAK F.: Vývoj a problémy mapy 1 : 5 000, Zeměměřický Obzor SIA, ročník 11/38, 1950)

PTÁK J.: Vznik a účel Státní mapy 1 : 5000 – odvozené, Zeměměřictví roč. 2/40, 1952, č.10