MTA CSFK GGI, Csatkai Endre u. 6-8, Sopron, Hungary

The digitisation of historical telluric recordings

T. Nagy, V. Wesztergom, István Lemperger, Árpád Kis, Ernő Prácser, Károl

Abstract

As a contribution to the European Seventh Framework Programme project European Risk from Geomagnetically Induced Currents (EURISGIC), MTA CSFK GGI has offered to digitize its historical telluric recordings. Altogether 4539 film rolls (with a total length of approximately 8 km) has been digitized, covering data from the year 1957 to 1997. During the process a best practice has been developed about how to split the project into multiple stages allowing collaborative work of a group of people, how to store data in a secure and easily accessible way, how to control the quality of the resulting data product and how to make it available to the public.

1 Introduction

The Earth's magnetic field is influenced by the solar activities e.g. coronal mass ejection (CME) which may affect the electrical network. For instance, in 1989 a geomagnetic storm in Montreal, Canada resulted in the malfunction of Hydro-Quebec's power grid which has caused nine hours of total power outage affecting at least six million people. For these reasons, it is essential to better understand the processes originating from solar activities.

Based on Faraday's law of induction an electric field is related to the time-varying geomagnetic field. According to Ohm's law terrestrial (telluric) currents are formed in the finite conductivity of subsoil due to the electric field. The nearly harmonic, small amplitude, exogenous changes is the geomagnetic field are called pulsations, whereas the intense changes associated with strengthened solar winds are called geomagnetic storms.

A typical duration of a geomagnetic storm is a few hours but some of them may even last for several days. The interactions between the Earth's magnetic field and the plasma clouds ejected from the Sun during flares yields to a typical pulse-like beginning of the storm (Storm Sudden Commencement –SSC). The intensity of the storm is determined by the combined effect of solar wind velocity and interplanetary magnetic field (IMF). Their amplitudes are a few hundred nT at mid-latitudes and a few thousand nT in the auroral zone. The energy input process of the particles flowing from the Sun in magnetic field is the field line interconnection. This process determines the level of magnetic disturbance, i.e. the general geomagnetic activity.

Earth current measurement with high time resolution started at the Nagycenk Observatory in late 1957. This is a representative, homogeneous and unique data

set for statistical analysis of the long-term variation of the geomagnetic induction effect. Global trends are also influenced by the solar activity through feedback, therefore our goal is to examine how the solar activity changes in the Earth's plasma environment. Towards this the first step is the digitization of our uniquely long analog telluric recordings.

2 Discussion

Our digitization workflow has been built on the follwing pillars:

Reproducibility. It is an essential requirement to be able to reproduce any step of the workflow at any time. For this reason, we have chosen to rely only on open source software. This way we have full control over the developed system: everyone is free to check that what we have calculated is identical to what we claim to have calculated. Moreover, if for some reasons (such as correcting an error or making our sampling more accurate) in a later time we want to update our database then we can avoid the problem of not being able to do so caused by f.i. expired software licences.

Transparency. If someone notices an error in the end product it is very important to be able to identify easily which level of the process it originates from, what was the modification within that level which caused the error and who is responsible for that modification. This way we know immediately who to turn to when we are about to figure out the actions need to be taken in order to correct the source of the error. We ensure these requirements by designing the different stages of the workflow in a way that each stage depends exclusively on the previous stage. Different people are responsible for each stage and files created during a stage are under version control, so it is possible to explore not only the current but former versions of a file as well. Another aspect of tranparency is to make regular progress reports about the completeness of each stages.

Collaboration. Because there are a lot of materials to process, it requires the collaboration of many people. As such, it can lead to conflicts when multiple people are trying to work on the same file. Having the files we work on under version control eliminates such issues. Furthermore, team members are sometimes working at separate locations from each other so it must be ensured that they can exchange files with each other efficiently over the world wide web.

Data protection. To avoid loosing data due to an extraordinary event such as a breakdown of a hard drive or an accidental file remove operartion, it is essential to create backups. First of all, to each intermediate file belongs at least one team member who already worked on that file, so the team members' computers serve as a form of backup. The version controlled repository associated to the project is a form of backup in itself too. On top of that, there are further automatisms in the institute which create regular backups.

Intelligence. Due to the large number of materials we developed the workflow so that it is as automated as possible. An unexpected input (such as a typo

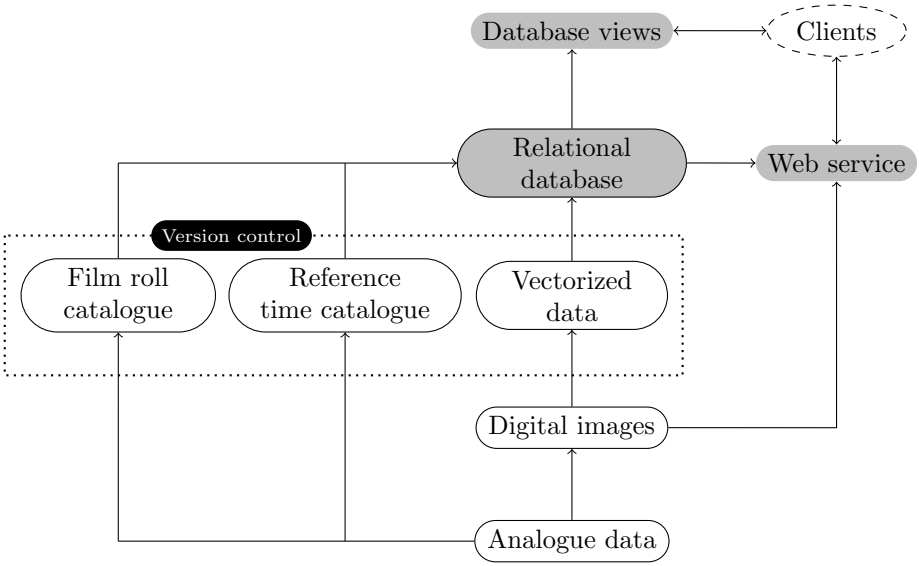

Figure 1: Flowchart of the digitization process. The arrows connecting the nodes illustrate the dependencies between them.

in a filename) can break such automatisms very easily, therefore our scripts are trying to guess human intentions rather than give up processing.

There are essentially five different layers of the digitization process (Figure 1). The first layer holds the analogue data stored on a transparent material comparable to film rolls. This serves as the initial input to the entire process. The second layer holds static (or rarely changing) files which are the digitized (scanned) versions of the film rolls in the first layer. The result of vectorizing the files on the second layer gives the files on the third layer. During the process of vectorization these files get updated very frequently, therefore the files associated to this layer are under version control. Additional to the vectorized data, on this layer an another type of file collection, the so-called reference time catalogue exists also in the form of version controlled files. On top of that, there is also a film roll catalogue here which holds textual data collected from the film rolls. All information contained on this layer is then loaded into a relational database, which serves as our fourth layer. The fifth layer consists of collection of server side programs and sql views which serves data to the clients from the relational database (fourth layer) or from the image repositories (second layer).

Analogue film rolls were digitized using an ordinary long-format paper scanner in the resolution of 118.18 pixels/cm, which corresponds to 300dpi. 5cm on the film roll covers 2 hours of registration. A typical length of a film roll varied somewhere between 0.5 and 1m, but some of them where as long as 4.5m. These latter ones were too large to scan them in a single shot due to the limitations of our device, leading to the need of further post-processing on the resulted digital

images. Altogether 4539 files were created by a group of 5 people, but to reduce their dimensions we cutted them into smaller segments, therefore the total number of files in our final image repository was 22055. The film roll catalogue (which is a single plain text file containing tabular information) of layer 3 were created during the scanning process and the the reference time catalog (which is a hierarchical collection of plain text files containing tabular information) was created during the cutting process.

We used the software called Engauge Digitizer (ED) to vectorize the curves which were present on our digital images. The vectorization involved 2 steps: first we defined a coordinate system on each image by specifying the actual coordinates of three reference points (this step defines the transformation between pixel coordinates and actual coordinates), then we traced the point of the curve with a digitizing tablet using either the semi-automatic or the manual sampling mode offered by ED. This step also resulted in a hierarchical collection of plain text files containing information about the vectorized curves in a tabular format.

To ease up data access we loaded all of the information contained in the files of layer 3 into a PostgreSQL database. In this database there is a separate table for the film roll catalogue and another one for the reference time catalogue. In order to speed up queries, the vectorized pointsets are sorted into smaller tables each of which containing nearly one year of registration. The end users communicate only with this relational database, either directly via a database view or via a web service wrapped around a certain view. F.i. the url

`http://geodata.ggki.hu/tellurics/api/view.php?basename=19600312_0000`

gives back a plot of the vectorized data overlaid above the original image with id `19600312_0000`. A graphical user interface for exploring our data is available at

`http://geodata.ggki.hu/tellurics`

3 Conclusions

We presented the method we developed for digitizing our historical telluric recordings. Besides opening the door towards analyzing a uniquely long time series of its kind, the developed process is general enough to serve as a tool for managing other digitization projects as well.

Acknowledgements

The following persons were permanent or occasional members of the digitizing group: Benke, Erzsébet; Boros, Eszter Agnes; Bódis, Virág Bereniké; Guttmann, Eszter; Holler, Gáborné, Holler, Ildikó; Király, Zsuzsanna; Kurucz, Gergő; Meditz, Andrea; Meditz, Júlia; Nagy, Annamária, Pusztai, Annamária; Szabó, Henriett; Szita, Renáta; Novák, Attila; Pálla, Gyula; Szokoli, Kitti. We took advantage of the following software products: Engauge Digitizer, bash/R/C++/PHP, ImageMagick, GIMP, Subversion, PostgreSQL, Apache2, gdal2tyles.py, Debian.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant

agreement 260330. This study was supported by the TAMOP-4.2.2.C–11/1/KONV-2012-0015 (Earth-system) project sponsored by the EU and European Social Foundation.