WP3: Historic data – tellurics

EURISGIC Final Meeting

Tamás Nagy

✉ nattomi@ggki.hu

HAS RCAES GGI

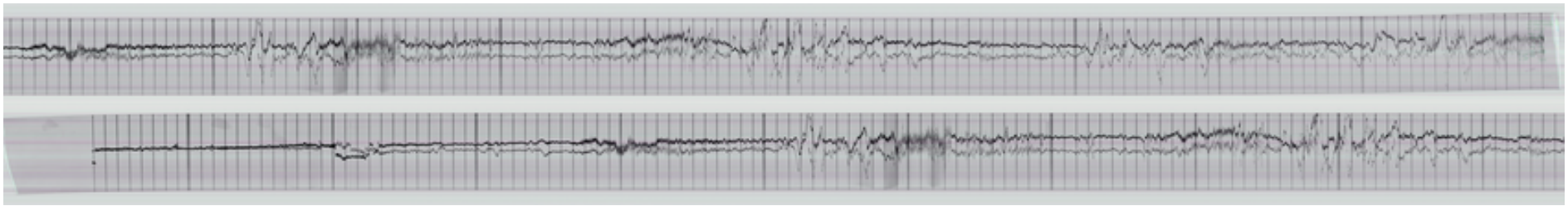
16-17 January 2014

Helsinki, Finland

SEVENTH FRAMEWORK PROGRAMME

EURISGIC

Introduction



- What do we have?
  - Exceptionally long recording of the induced geoelectric field a.k.a. *telluric field* is available from the data archive of HAS RCAES GGI in the form of film rolls.
  - Started in 1957, lasted till 1997
  - The recording goes on up until today, but from 1994 also (and from 1997, exclusively) in digital form.
- How is it related to the objectives of the project?
  - Direct measurement of the electric field allows us to assess worst case scenarios
  - May serve as an excellent possibility for the validation of the overall GIC model
- A desire for digitizing our analogue recordings naturally arises

Formulating the problem of digitization

- **Input:** 4539 film rolls (total length of approximately 8 km).
- **Desired output:** Adequatly dense discrete resolutions

$$t_x^1 < t_x^2 < \ldots < t_x^n \quad \text{and} \quad t_y^1 < t_y^2 < \ldots < t_y^m$$

of the time interval $[-378741600, 854953200]$ and the associated irregularly spaced time series

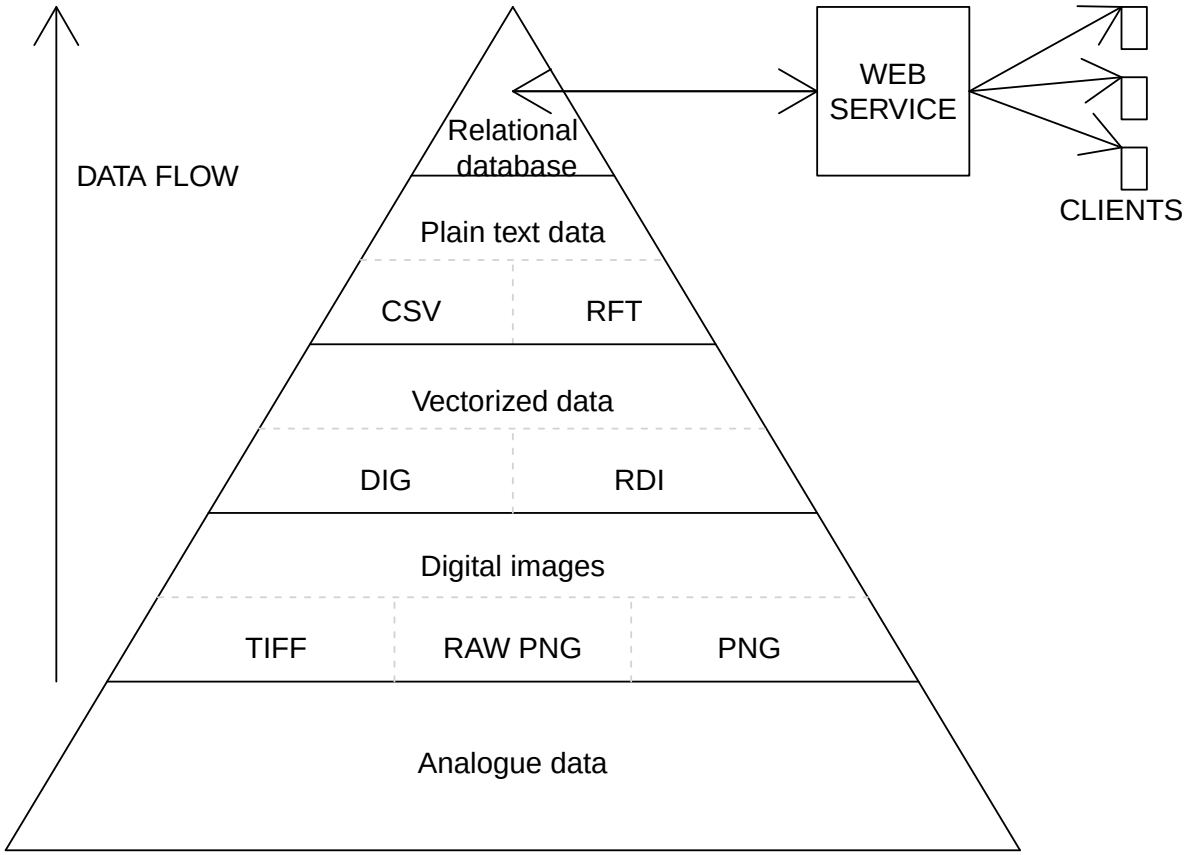
$$(t_x^i, E_x^i) \quad \text{and} \quad (t_y^i, E_y^i)$$

  - Maximal possible resolution (1min sampling means $n, m \geq 20561580$)
  - Ninimizing information loss
- **Tools required:**
  - Scanner
  - Curve tracing software
  - Curve tracing ability of the human brain

Key aspects

- Reproducibility
  - Ability to update it at any time
  - Workflow depends only on free software
  - Track file history with version controlled file system
- Transparency
  - Mistakes, bugs must be easy to spot
- Collaboration
  - Easy way to transfer/exchange files
  - Avoid conflicts caused by multiple persons working on the same file
- Data protection
  - Manual backups
  - Dedicated data servers are within the scope of the automatic backup service of the institute
- Intelligence
  - As automated as possible
  - Try to guess human intentions and correct human mistakes (f.i. a typo)

The digitization pyramid



Analogue film rolls to digital images

film rolls

Panasonic KV S1025C
long format paper scanner

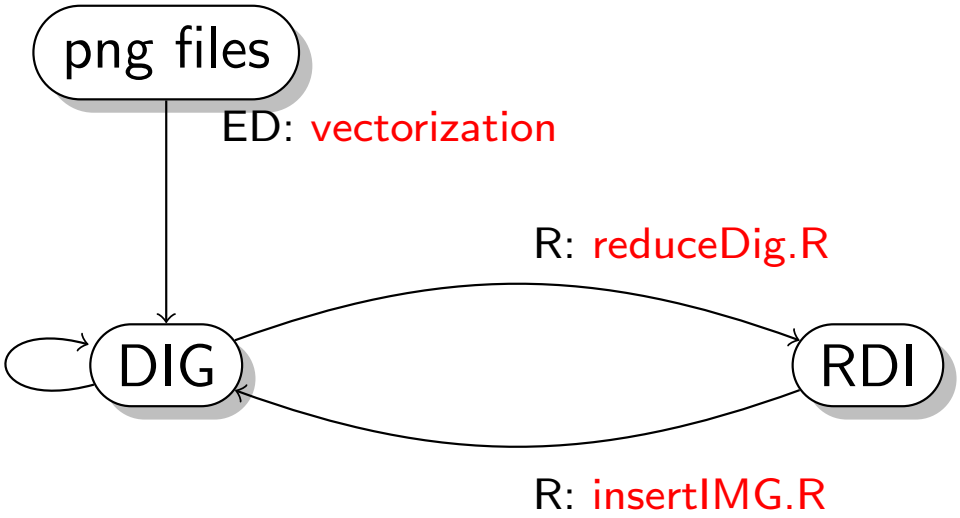
tiff files

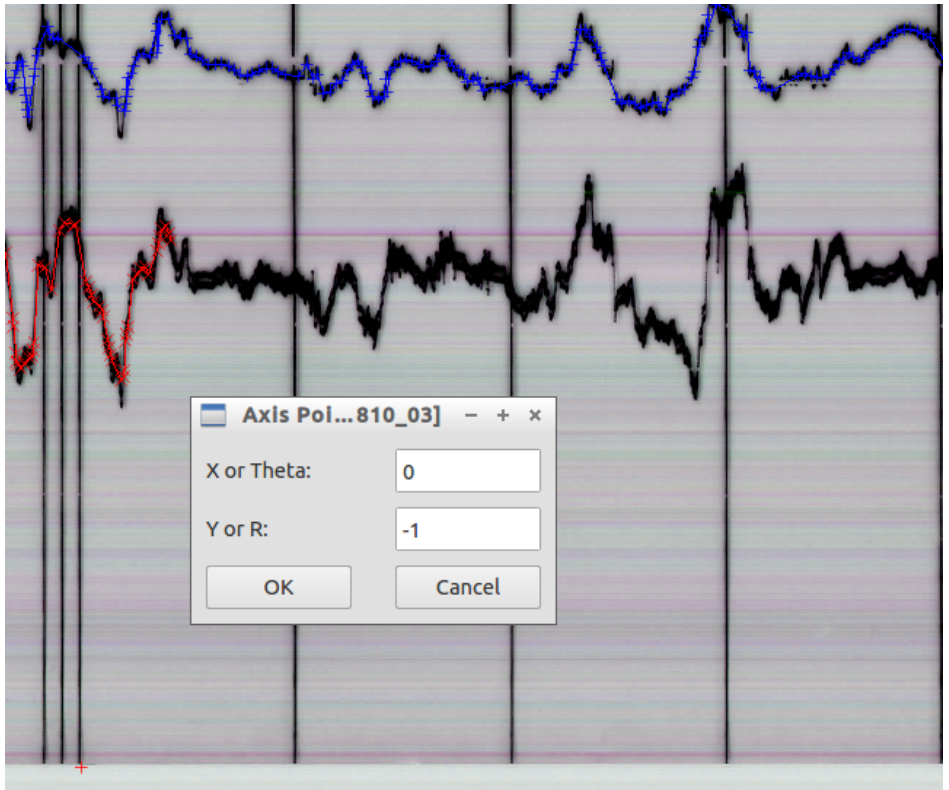
R (via ImageMagick): Tiff2Png.R
manual: rotating/splitting

png files

Resolution: 118.18 pixels/cm (300dpi)
5 cm on the film roll covers 2 hours
Maximal length of film rolls: 4.5 m
Number of film rolls: 4539
Number of tiff files: 5850
Number of png files: 22055
Number of people: 5
`http://eurisgic.ggki.hu/tiff`
`http://eurisgic.ggki.hu/png_raw`
`http://eurisgic.ggki.hu/png`

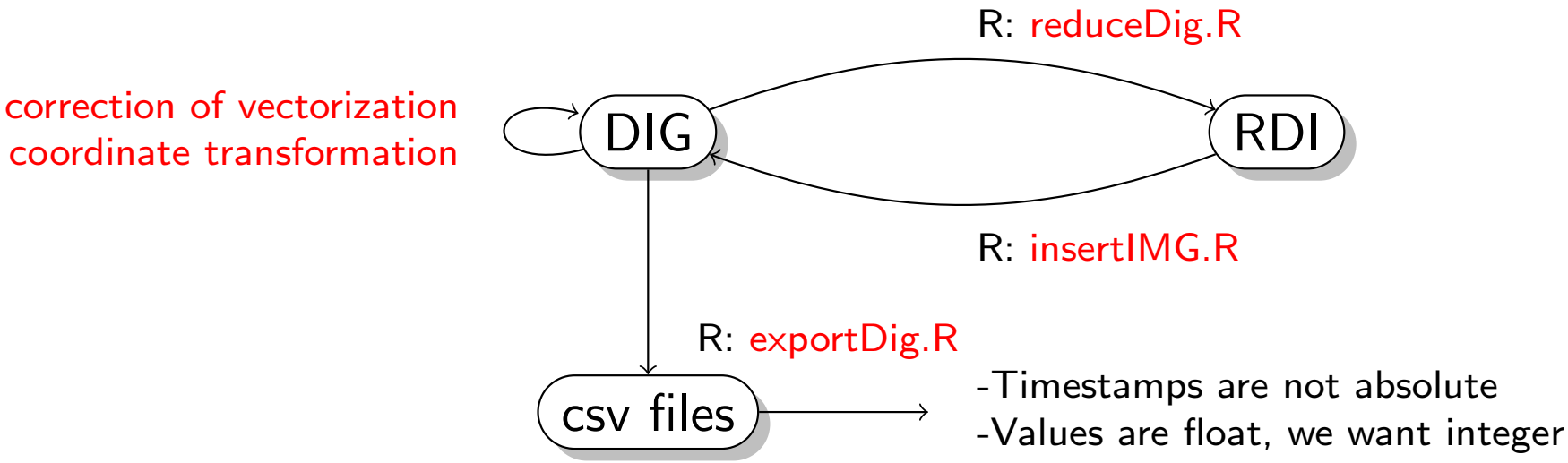
Digital images to vectorized data





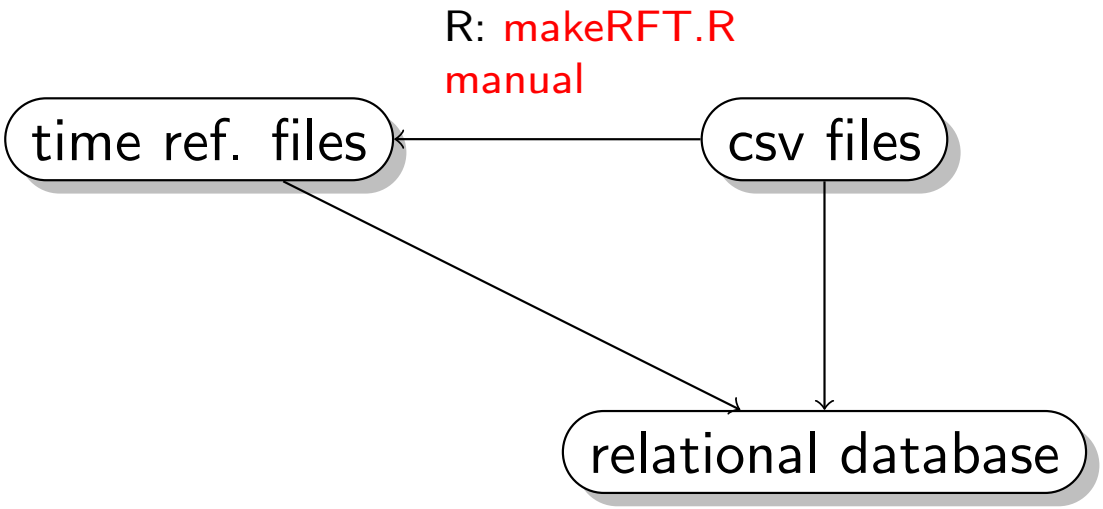
- Number of files: 20330
- Total size: 711MB
- Number of people: 14
- `svn://geodata.ggki.hu/digit/dig`

Vectorized data to plain text data



- Each csv file need to be time-referenced
- Number of files: 24614
- Total size: 487MB
- Time consuming (brute-force regeneration takes $\approx$ 1 day
- `svn://geodata.ggki.hu/digit/csv`

Reaching the top: plain text data to relational database



- Stored as one table in the database
- Current size is about 2GB
- Database engine is PostgreSQL
- Has a dedicated user for the web service which allows select queries
- Importing is fast, "brute force"-style update
- Update of portions of the database is also possible
- "Suspicious" data is not uploaded

Quality control

- Automatic checking for missing files based on naming convention
- Automatic checking wether coordinate trandformation was carried out
- Checking by manual exploration

Web service

- Communicates with the Postgres database (top of the pyramid)
- Serves data queries via http protocol
- Password protected
- Can be queried for 1 day of data
- Can return original (raw) or resampled time series
- Can return graphical display
- Examples:
  `http://tellurika.ggki.hu/api/get?date=19580220`
  `http://tellurika.ggki.hu/api/get?date=19580220format=original`
  `http://tellurika.ggki.hu/api/get?date=19580220format=resampled`
  `http://tellurika.ggki.hu/api/get?date=19580220format=png`
  `http://tellurika.ggki.hu/api/get?date=19580220format=p`

Acknowledgements

- Core members of the digitizing group: *Benke, Erzsébet; Boros, Eszter Ágnes; 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*
- Occasional members of the digitizing group: *Novák, Attila; Pálla, Gyula; Szokoli, Kitti*
- Collaborators of the following great open source softwares:
  - Engauge Digitizer
  - R Statistical Language and Environment
  - ImageMagick image manipulation library
  - Subversion
  - PostgreSQL database engine
  - Apache web server
  - rApache - R module for Apache
  - Debian