EGU23-15072, updated on 20 Jun 2023
https://doi.org/10.5194/egusphere-egu23-15072
EGU General Assembly 2023

Automated Extraction of Bioclimatic Time Series from PDF Tables

Sabino Maggi, Silvana Fuina, and Saverio Vicario
CNR National Research Council, Institute of Atmospheric Research, Bari, Italy (sabino.maggi@cnr.it)

Since the development of the original specifications in the '90s the PDF document format has become the *de-facto* standard for the distribution and archival of documents in electronic form because of its ability to preserve the original layout of the documents, independently of the hardware, operating system and application software used to visualize them.

Unfortunately the PDF format does not contain explicit structural and semantic information, making it very difficult to extract structured information from them, in particular data presented in tabular form.
The automatic extraction of tabular data is a difficult and challenging task because tables can have extremely different formats and layouts, and involves several complex steps, from the proper recognition and conversion of printed text into machine-encoded characters, to the identification of logically coherent table constructs (headers, columns, rows, spanning elements), and to the breaking down of the data constructs into elemental objects.

Several tools have been developed to support the extraction process. In this work we survey the most interesting tools for the automatic detection and extraction of tabular data, analyzing their respective advantages and limitations. A particular emphasis is given on programmable open source tools because of their flexibility and long-term availability, together with the possibility to easily tweak them to meet the peculiar needs of the problem at hand.

As a practical application, we also present a workflow based on a set of R and AWK scripts that can automatically extract daily temperature and precipitation data from the official PDF documents made available each year by Regione Puglia, in Italy. The lessons learned from the development of this workflow and the possibility to generalize the approach to different kinds of PDF documents are also discussed.