Use of Optical Character Recognition for Digitizing Climate Data

Alexey Kaplan (LDEO of Columbia University)
alexeyk@ldeo.columbia.edu

Abstract

A pilot project is proposed to prove the concept of the use of optical character recognition (OCR) for the automatic digitization of large volumes of published climate data from the 19th and early 20th centuries. These data are effectively excluded from the mainstream climate research because they are not available in digital form. Funds to employ enough typists to digitize vast amounts of such data generally are not available. However, huge collections of such data are already publicly available in the form of scanned images. Fast development of the OCR technology makes its use attractive at least for some parts of the collection. First attempts to use it for this purpose have given encouraging results, when applied to a specific data table formats. Development of a flexible system for post-processing OCR results, which would be applicable for a wide variety of table formats is a crucial next step. Proposals suggesting effective ways to do this are especially encouraged.

Scientific Background. Importance of historical climate data for the field of climate research is self-obvious. Yet our records of basic meteorological parameters taken on the land weather stations around the world are very incomplete before 1950s, even though a large number of meteorological stations existing today were taking a few observations a day during the 19th century as well. Most of preserved records are hand-written and stored in various countries' weather services, archives of public records, etc. Monthly summaries of station observations brought into a digital form constitute the backbone of the land climate data sets available for the climate research [*Peterson and Vose*, 1997]. However, it is becoming increasingly clear that the temporal resolution of monthly averages is not adequate for addressing many important questions: study of weather extremes, frequencies of precipitation and occurrence of certain weather types, interpreting data on the harvests and tree growth of the past, etc., require daily resolution of weather records. Even judging the quality of the monthly data (which is essential for our ability to make responsible inferences about climate change) requires at least daily resolution. With inevitable changes in measurement techniques and observation times affecting almost any station record longer than a few decades, the original measurements taken a few times a day are in fact necessary for producing homogeneous daily and monthly records. Yet there are few daily temperature or pressure records which extend to more than two centuries [*Yan et al.*, 2001).

NOAA collection of scanned data images. Handwritten archives of the weather records are being digitized by the historical climatologists or by typists under their supervision; these records require individual human attention. There is, however, an interesting class of weather records which are in between handwritten archives and digital databases. These are published books of typeset weather records which are often available in the form of scanned images. In fact, large collection of such images recently became publicly available via NOAA Central Library Climate Data Imaging Project as the part of their data rescue (i.e. from deteriorating paper medium) project. Countries whose old data collections are presently available on their website (`http://docs.lib.noaa.gov/rescue/data_rescue_home.html`) are indicated in Figure 1. If digitized, these collections will make many useful additions to existing digital databases even at monthly temporal resolution; for daily resolution it will provide vast amounts of so far unavailable data (e.g. Phil Jones, pers. comm., regarding

Russian 1837-1913 collection). Digitization of the climate data contained in the images, however, is not in the mandate of this NOAA project, and is not being attempted by any other NOAA division (Doria Grimes, pers. comm.).For example, NCDC's Climate Database Modernization Program, in its third year now, focuses on the U.S. data and spends a few million dollars annually on imaging and keying in of 1895-1948 weather station bulletins and 1820-1890 weather observations from military forts. Even if electronic images are available, keying the data in is still a labor-consuming, expensive enterprise.

Utility of the optical character recognition (OCR) technology. Fast development of the OCR technology makes an accurate digitization of high-quality images a plausible task. Informational scientists are already attempting to use OCR for automatic cataloging and indexing of book images [*Tseng*, 2001]. On the other hand, as the technology develops, the top-line OCR software (e.g. ABBYY's FineReader) becomes available for a trial download and affordable for purchasing. It seems that with some post-processing of the OCR output, a climatologist can set-up a system of an automatic (or with a minimal human intervention) digitization. However, this was never done before.

Pre-pilot results. With Merissa Sakuda (CC' 02) and Allison Lim (current LDEO summer intern), helped by the advice of Benno Blumenthal (on a possible set-up of the post-processing system) and Lisa Fish (on historical data sources and current uses of the OCR) we did initial evaluation of the NOAA image collection. Our trial version of the OCR post-processing system for the Russian 1837-1846 pressure tables (see Figure 2 for an imaged table from [*von Kupffer*, 1837] was quite successful. Automatically digitized daily Catherinenbourg (called Sverdlovsk during Soviet era) pressures for the year of 1837 shows variations similar to those in the record taken 160 years later (Figure 3). Comparison done for St.Petersburg, Russia, made with already existing database of monthly pressures proved the validity of the implemented digitization procedure (Figure 4).

Required work. More work is needed before the concept is convincing enough to seek government funding. We need to develop **a flexible pilot system which uses typical check and correction tools from the quality control for climate observations, estimates the uncertainty in results of digitization, and includes an option for a human intervention**. The usefulness of this system will be demonstrated by its application to a few data sets (e.g. Russian pressures – good image quality, Chilean temperatures – ENSO relevance). Two kinds of labor are necessary in this project: an algorithmic and programming expertise for developing the post-processing system, and an operator to run the system on the actual climate images, check, correct, evaluate results, etc.

References

Peterson,T.C. and Vose,R.S., 1997: An overview of the Global Historical Climatology Network temperature data base, *Bull. Amer. Met. Soc.*, **78**, 2837-2849.

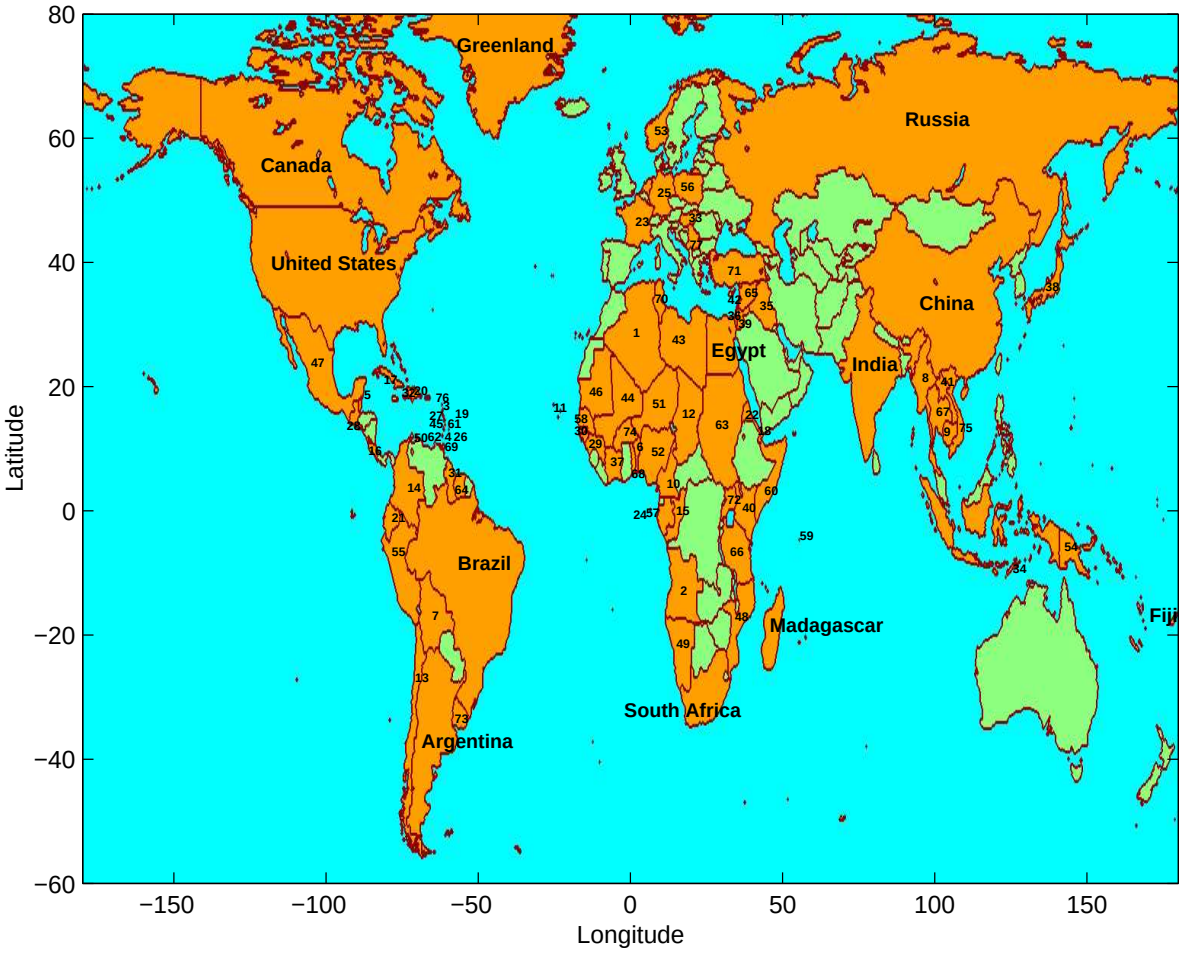
Tseng, Y.H., 2001: Automatic cataloging and searching for retrospective data by use of OCR text. *J. Amer. Information Sci and Tech. Soc.*, **52**, 378-390.

von Kupffer, A.T., 1837: *Annuaire magnétique et météorologique.* Corps des ingénieurs des mines de Russie, St.Petersbourgh. 244p.

Yan,Z., P.D.Jones, A.Moberg, et al., 2001: Recent trend in weather and seasonal cycles: An analysis of daily data from Europe and China. *J. Geophys. Res*, **106**, 5123-5138.

Scanned Images with Climate Data in NOAA Central Library

(reflects images available on the server as of January 2004)



Numbers on the map indicate the following countries:

1 - Algeria	21 - Ecuador	41 - Laos	61 - St. Lucia
2 - Angola	22 - Eritrea	42 - Lebanon	62 - St. Vincent and the Grenadines
3 - Antigua and Barbuda	23 - France	43 - Libya	63 - Sudan
4 - Barbados	24 - Gabon	44 - Mali	64 - Suriname
5 - Belize	25 - Germany	45 - Martinique	65 - Syria
6 - Benin	26 - Grenada	46 - Mauritania	66 - Tanzania
7 - Bolivia	27 - Guadeloupe	47 - Mexico	67 - Thailand
8 - Burma	28 - Guatemala	48 - Mozambique	68 - Togo
9 - Cambodia	29 - Guinea	49 - Namibia	69 - Trinidad and Tobago
10 - Cameroon	30 - Guinea-Bissau	50 - Netherlands	70 - Tunisia
11 - Cape Verde	31 - Guyana	51 - Niger	71 - Turkey
12 - Chad	32 - Haiti	52 - Nigeria	72 - Uganda
13 - Chile	33 - Hungary	53 - Norway	73 - Uruguay
14 - Colombia	34 - Indonesia	54 - Papua New Guinea	74 - Upper Volta - Burkina Faso
15 - Congo	35 - Iraq	55 - Peru	75 - Vietnam
16 - Costa Rica	36 - Israel	56 - Poland	76 - Virgin Islands
17 - Cuba	37 - Ivory Coast	57 - Sao Tome and Principe	77 - Yugoslavia
18 - Djibouti	38 - Japan	58 - Senegal	
19 - Dominica	39 - Jordan	59 - Seychelles	
20 - Dominican Republic	40 - Kenya	60 - Somalia	

[The map was produced in the LDEO of Columbia University;
Contact: A.Kaplan `alexeyk@ldeo.columbia.edu`**]**

Figure 1: Geographical distribution of climate data in image form.

CATHERINENBOURG. JANVIER 1837.

BAROMÈTRE À 13°½ R.

DATE	8 HEURES DU MATIN	10 HEURES DU MATIN.	MIDI.	2 HEURES APRÈS MIDI.	4 HEURES APRÈS MIDI.	6 HEURES DU SOIR.	8 HEURES DU SOIR.	10 HEURES DU SOIR.
1	595,13	594,66	594,09	593,53	593,40	592,84	592,25	591,30
2	587,75	586,85	585,77	584,92	584,60	584,02	583,20	582,88
3	574,03	579,64	578,08	577,88	577,24	575,79	575,29	574,39
4	577,67	578,17	578,69	578,81	578,72	578,10	576,33	574,50
5	571,50	572,27	573,23	574,86	577,11	578,43	580,36	582,10
6	588,70	589,89	590,38	590,63	591,14	591,16	590,86	590,40
7	589,96	589,98	589,33	588,37	588,97	588,82	588,77	589,43
8	592,16	593,13	593,74	594,05	594,69	594,81	594,69	594,70
9	593,29	593,39	593,03	592,54	592,54	592,49	592,01	591,71
10	589,49	589,09	588,39	587,45	586,92	586,56	586,14	585,52
11	582,31	581,84	580,92	580,42	580,41	580,45	580,22	580,22
12	580,38	580,67	580,69	580,74	581,49	581,79	582,37	582,99
13	585,29	585,83	585,93	585,96	586,63	586,71	586,56	586,78
14	584,54	584,10	583,40	582,54	582,37	581,89	581,85	581,96
15	580,62	580,65	580,50	580,03	580,30	580,21	579,84	579,80
16	580,74	581,07	580,58	580,04	579,54	579,37	578,90	577,76
17	570,15	568,09	566,23	565,04	565,00	565,17	565,39	565,72
18	570,91	571,41	572,15	572,69	572,53	573,36	573,89	574,93
19	572,90	571,49	569,18	567,66	566,50	565,32	564,32	562,46
20	564,23	566,10	567,34	568,63	571,18	571,81	572,37	572,96
21	575,73	576,17	576,95	577,58	579,18	580,98	582,12	582,92
22	581,68	579,12	577,42	576,72	575,25	575,07	574,99	574,85
23	584,34	585,69	586,99	586,99	587,43	587,59	587,38	586,80
24	583,04	582,28	581,78	581,40	581,24	581,12	580,95	580,72
25	578,37	578,03	577,19	576,54	576,13	575,54	575,02	574,17
26	570,19	569,79	569,64	569,72	569,58	569,68	570,57	571,63
27	573,61	573,51	573,36	573,00	573,15	573,43	573,69	573,94
28	574,57	574,37	574,19	573,94	574,09	574,99	573,95	574,23
29	574,43	574,55	573,79	572,71	572,63	572,25	572,51	572,73
30	574,43	574,64	574,87	575,13	575,34	576,09	576,43	576,65
31	577,28	577,10	577,68	577,71	578,04	578,77	579,20	579,90
MOYENNES	579,98	580,12	579,85	579,63	579,79	579,79	579,75	579,71

1

Figure 2: A typical table of Russian pressure data in 1836-1840s [*von Kupffer*, 1837].

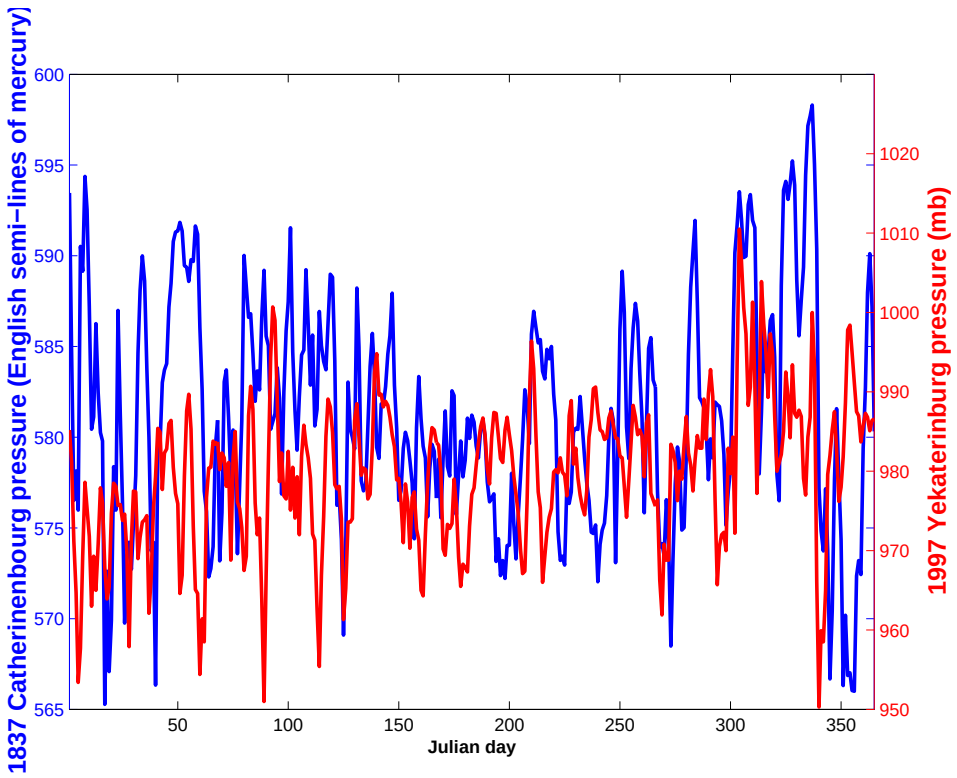


Figure 3: Two daily records for Catherinenbourg, 1837 and 1997

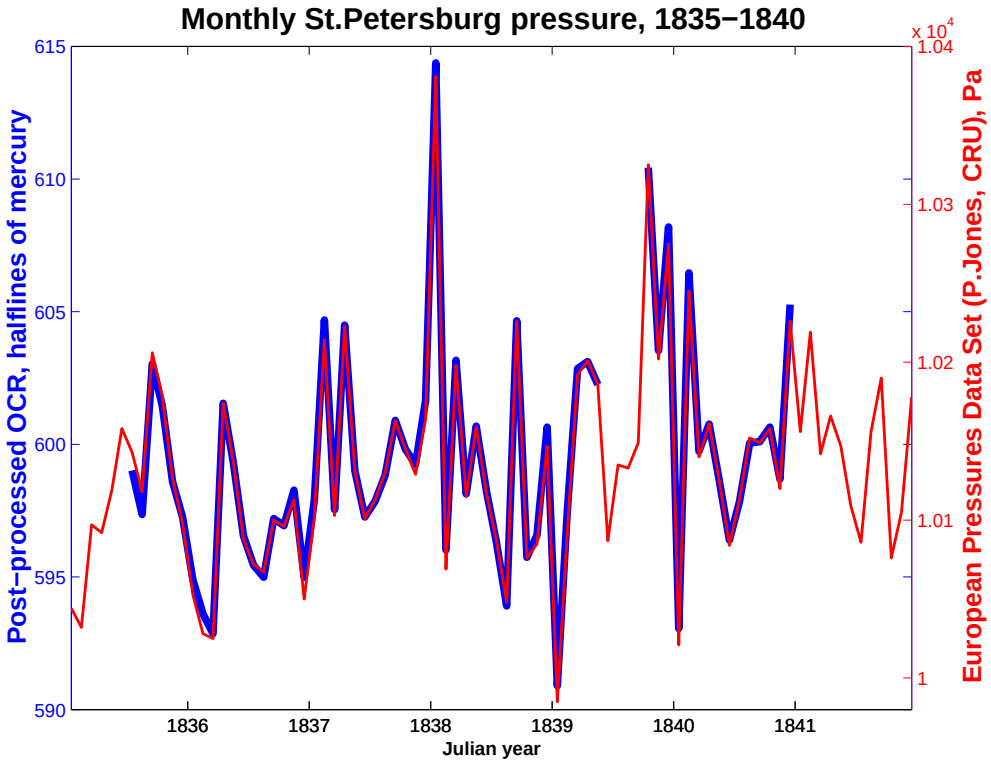


Figure 4: Comparison of monthly averages from newly digitized data with those from an existing monthly database