A guide for digitising manuscript climate data

S. Brönnimann, J. Annis, W. Dann, T. Ewen, A. N. Grant, T. Griesser, S. Krähenmann, C. Mohr, M. Scherer, C. Vogler

Clim. Past Discuss., 2, 191–207, 2006
www.clim-past-discuss.net/2/191/2006/

Climate of the Past Discussions

Climate of the Past Discussions is the access reviewed discussion forum of *Climate of the Past*

A guide for digitising manuscript climate data

S. Brönnimann, J. Annis, W. Dann, T. Ewen, A. N. Grant, T. Griesser, S. Krähenmann, C. Mohr, M. Scherer, and C. Vogler

Institute for Atmospheric and Climate Science, ETH Zürich, Universitätstr. 16, CH-8092 Zürich, Switzerland

Received: 1 March 2006 – Accepted: 14 March 2006 – Published: 4 May 2006

Correspondence to: S. Brönnimann (broennimann@env.ethz.ch)

Abstract

Hand-written or printed manuscript data are an important source for paleoclimatological studies, but bringing them into a suitable format can be a time consuming adventure with uncertain success. Before starting the digitising work, it is worthwhile spending a few thoughts on the characteristics of the data, the scientific requirements with respect to quality and coverage, and on the different digitising techniques. Here we briefly discuss the most important considerations and report our own experience. We describe different methods for digitising numeric or text data, i.e., optical character recognition (OCR), speech recognition, and key entry. Each technique has its advantages and disadvantages that may become important for certain applications. It is therefore crucial to thoroughly investigate beforehand the characteristics of the manuscript data, define the quality targets and develop validation strategies.

1 Introduction

The age of digital computing and data storage has revolutionised data acquisition and administration. Since about the 1950s, climate data have been stored electronically or have been converted to electronic format. However, for centuries, climate data have been stored in the traditional way, i.e., hand written on paper. These data accumulate to hundreds of thousands of volumes in countless archives. While some of these data have been digitised in the past, this is not the case for the bulk of the data. The value of such data for climate research is nowadays highly esteemed with increasing demand from the paleoclimatological community and new numerical techniques becoming available (Brönnimann et al., 2005). However, digitising such data is a labour intensive undertaking that is often associated with a high risk of a "no result" (data quality does not meet scientific requirements). In order to reduce the risk and optimize the amount of labour it is important to spend a few thoughts beforehand on the characteristics of the data, the scientific requirements, and the quality tests and validation

CPD

2, 191–207, 2006

strategies. This can help choosing the optimal digitising technique.

In our own projects we have digitised historical upper-level data from various sources that differed in format, quality and layout (Brönnimann, 2003), total ozone from original observation sheets (Vogler et al., 2006), as well as meteorological observations from Mount Washington, USA (Grant et al., 2005). We used optical character recognition (OCR), speech recognition, and key entry to digitise the data. Following our publications we have repeatedly been contacted by other research groups concerning our experience with different techniques. In this paper we would like to share our experience.

2 Properties of manuscript data

Before describing digitising techniques, we would like to categorise the most important properties of manuscript data and their relation to the specific requirements of the planned research. In the following we distinguish between formal characteristics (format of the source and format of the information) and informational characteristics (information content in relation to the requirements, i.e., quality, redundancy, see Table 1).

The source can be available as an original (in any format), as photocopies, scanned images, or any other form. If originals are available, reproduction is sometimes necessary (see Sect. 4). Image processing or also photocopying may enhance the legibility of the source (e.g., in the case of faint pencil writing on yellowed paper) and is worth testing. Bound books often pose special problems. Photocopying is sometimes not possible, and even when photographing it can be difficult getting the bound books to lie flat. This is especially the case for old, fragile books. If OCR will later be applied, it can be advisable to make one-sided photocopies of the bound books as an intermediate step (rather than photographing or scanning directly). This preserves (most of) the information, while the actual scanning later on takes not much additional time, but can be optimised later on for speed and resolution.

The information type can be numeric, text, an alphanumeric code, or graphical. In this paper we mainly refer to numeric data; other considerations apply to other types of data. The format of the information can be a table, a text, a map (such as a weather map with station information on it), a graph, or a mixture of all these. This is probably the most important factor in deciding which technique to use. Thereby it should be kept in mind that the format and type of the information may frequently change within the same archival source over the period of time desired. This concerns not only the reporting (e.g., units, resolution), but also the layout (tables, weather maps). Another important issue is the typing of the data. Is it printed, typed, or hand-written? Finally, the legibility can be the most important constraint and is something that certainly needs consideration in advance.

A second set of criteria refers to the information content of the data (informational characteristics). The first question that often arises is: What part of the information is needed? Everything? Or just parts of it? Here one has to keep in mind that redundant information is valuable for quality checks. For instance, in our upper-air data project we were confronted with the problem of a large number of station records, from which we had to choose (due to limited resources) a small subset. This is a very common problem, and an obvious approach is to estimate the amount of additional information that can be gained in relation to the digitising costs, leading to a cost-benefit function (Jones and Trewin, 2002). However, in addition to the spatial and temporal coverage of the data series one has to take into account the varying quality as well as the ways of assuring the quality. Here, redundant information is important. We used historical literature research to obtain information on the quality and chose pairs of neighbouring stations wherever possible (especially when dealing with different networks). A second important question concerns the expected (based on theory and literature) quality of the data and its relation to the required accuracy and precision of the end product. Finally, it is important to think about the meta-information: What kinds of meta-data are available, what conclusions can possibly be drawn (what is the role of these data in the re-evaluation process), and how will the meta-data be archived? Answering

this question can be important, e.g., when the same data are available from different sources, one of which would be cheaper to digitise than others. For all questions related to the informational characteristics, thorough literature research is necessary.

3 Digitising techniques

In our project we have used three techniques for digitising numeric or text data, which are discussed in the following. Special techniques are necessary for digitising graphical data such printed figures or hand-drawn isolines on weather maps or for analogue data such as registering strips from barographs or meteographs, photographed spectra, or the like.

Optical character recognition (OCR) is a powerful technique to convert scanned or photographed documents into text. We used ScanSoft OmniPage Pro 14 for our work. The user can select the area of interest and choose between standard output formats (e.g., text, table, worksheet). We used OCR in conjunction with an Epson document scanner that allows scanning piles of sheets (in all cases, photocopies of the originals) to a series of files. We performed limited tests also with scanning pens, but decided not to use this method operationally in our project.

The second method discussed is speech recognition. We used Dragon NaturallySpeaking, Versions 5 and 7 Preferred (digitising languages German and English) in combination with an Excel spreadsheet. In this application, the speaker dictates numbers or text along with spoken commands (e.g., “new line”). There is a number mode that constrains the program to understanding only numbers and commands. Numbers can be spoken as numbers (e.g., 4267), sequences of ciphers (4-2-6-7), or mixed (42-67). The software must be trained by each speaker according to the specific needs.

The third method considered is key entry, which is self-explanatory. All software programmes are very inexpensive compared to the salaries and hardware and hence their price is not considered a factor in this paper.

CPD
2, 191–207, 2006

EGU

4 Procedure

In this section we describe the digitising procedure, thereby putting emphasis on those steps where decisions concerning the digitising techniques must be made. The originals are often not at hand. Sometimes the material can be loaned, or an archive is willing to scan the documents. But mostly a trip to the archive is required, which needs careful planning. How much time is needed? Can the digitising be made directly in the archive based on the originals? Or should one just photocopy everything, take the paper back home and start browsing through the material? Or should one bring a digital camera and a powerful laptop? It is very important to find people at the corresponding archives that are willing to provide sample photocopies of the data sheets in advance. In historical time periods, data reporting was less standardized, the layout of data sheets changed frequently, and it is advisable to ask for as many sample photocopies as possible.

Digitising directly in the archive using key entry or speech recognition is only rarely advisable (e.g., if there are just small pieces of information on a large number of oversized data sheets so that photocopying would take as much time as digitising). Having the data sheets at hand for later checks is very important, hence, it is mostly advisable to make photocopies or photographs (the latter requires careful testing, a good tripod or copy stand, and a fast connection to the computer). We normally photocopied all material. Per archive day, around 2000 copies can normally be made (make sure to discuss this with the archive beforehand).

Before deciding which method to use, it is worthwhile performing extensive tests. Following are the advantages and disadvantages we found for the three methods used in our project.

4.1 Optical Character Recognition (OCR)

OCR is usually the fastest way to digitise data, especially for printed or tape written, tabulated data. Combined with an automatic scanner (we usually used a resolution

of 300 dpi in greyscale), OCR is many times faster than the other two techniques. However, we found that the error rate is normally higher. Figure 1 gives a typical example of an OCR'ed data table. The right panel shows the uncorrected output. While the recognition of the numbers worked relatively well despite the somewhat blurred typewriting, there are still a lot of errors that have to be corrected: shifts in the columns, decimal separations (points or commas), strange characters, or tiny spots on the paper that became symbols. The correction is relatively time intensive. Many misrepresented characters for any sample may be repetitively represented as the same character, but automatic search algorithms can not easily be defined for all cases.

For one application (data were given in blocks of code rather than a table) we considered using a scanning pen and performed a few tests. The two tested models (MyPen by C-Channel and QuickLink pen by WizCom) both were slower and produced more errors than other methods. However, scanning pens should certainly be considered in special cases.

4.2 Speech recognition and key entry

Speech recognition and key entry share similar characteristics. They are normally used if OCR is not possible (e.g., for hand-written or hardly legible data) or would make too many errors, if only a small portion of the table or sheet is used, or if the data are scattered. Figures 2 and 3 give examples of data sheets where speech recognition is the most effective method. The first example is a weather map that includes station information, the second example is a data table that is printed in two parts, with columns in different order. Note that in both cases, the format of the resulting spreadsheet is much simpler than the original layout.

We found the error rate of both methods to be smaller than for OCR. If this difference means that a double-check or a double entry can be avoided (see below), speech recognition or key entry may turn out faster than OCR.

When dictating or typing directly into a spreadsheet, a template has to be created. This should be done in such a way that it allows fast digitising, but also minimizes

later reformatting (e.g. transpose rows into columns, skip lines, merge columns directly when speaking or typing, see Figs. 2 and 3). This can be an advantage over OCR, which reproduces the layout of the source (including all of the frequent changes of reporting). The range of the numbers accepted can be constrained in the worksheet settings, so that a large fraction of the errors can already be excluded when speaking or typing.

Whether speech recognition or key entry works better also depends on the person doing it. Some would get tired faster (and thus make more errors and be slower) when key punching the data. Speech recognition is probably faster and easier for persons not used to key entry because it allows you to fully concentrate on the manuscript sheet. In the cases shown in Figs. 2 and 3, speech recognition allows using the fingers of both hands to keep track. Also, the spoken commands (e.g., "seven lines down") have some advantages. A frequent error (when digitising in German) was that the software confounded 14 ("vierzehn") with 4 10 ("vier zehn"), which in the worksheet became 410. We found similar problems while digitizing in English, but these problems varied from person to person. Speaking the ciphers individually (2-3-1 instead of 231) reduces the error, but is slower.

Provided that the hardware is good (computer, sound card), the software can keep pace with any reasonable speed of speaking. The numbers are stored in a buffer and written to the spreadsheet during a breathing pause. We find that a trained speaker can digitise around 2400 5-digit numbers with speech recognition in a 2-h period. That includes the correction of visually (in the work sheet) apparent errors, but not a systematic error correction. We found, after two hours of digitising, attentiveness usually dropped and the error rate increased. One of us had problems with a sore throat.

Key entry has its own advantages and drawbacks. While for a trained, fast-typing person, the speed can be similar to speech recognition, someone who is merely a fast typist but not experienced in 10-key entry, the error rate can be high. Similar attentive issues occur as for speech recognition. Errors tend to include both keying mistakes and duplication or skipping of a line of data. The latter error is aggravated by having

paper sheets to work from (rather than a digital image which can often be lined up to match the key punch template on the computer screen). Some people develop repetitive stress injuries. Outsourcing to data entry professionals is also an option. Many firms offer guarantees of 99.9% accuracy or higher, generally achieved through double keying. In some cases using a professional, who has no information about what the data represents, can be a drawback. For example, if the data being keyed is temperature and dew point, someone familiar with atmospheric variables will know that dew point is lower than (or equal to) temperature and will be able to correctly decipher semi-illegible numbers more often than someone without that knowledge.

4.3 Correcting and formatting

After digitising, the data must normally be reformatted. In the case of OCR, a large number of individual tables must be concatenated or sorted. There are often layout changes, so that this step must be done carefully. In the case of key entry and speech recognition, this step may be mostly done during data entry simply by choosing an appropriate template beforehand (see Fig. 3). This has to be considered when determining the overall speed of the different methods.

In the next step the data need to be tested and errors corrected. Double entry (having two persons digitising the same data and then comparing the differences) or a double check (checking each number) are the best ways of avoiding digitising errors. However, resources for this step are often not available, or not justified due to a high risk of a "no result", and in the case of OCR, double "entry" may not offer any advantage since the software algorithm is static. If one decides for a double check (or double entry), then choosing the fastest method (regardless of the error rate) might give the best overall benefit. Otherwise choosing the method that produces the fewest errors may help avoiding a double check. In the case of our upper-air data (temperature and pressure from historical radio soundings; a high-risk data set with redundant information) we decided not to double check the data but used the redundancy within the measurements to find errors. We plotted correlated variables (e.g., temperature at

neighbouring levels) against each other, or the thickness between two layers against their mean temperature (hydrostatic check). This sequence of tests proved sufficient to detect even small errors (some digitising errors, some errors in the originally recorded data) with statistical techniques, but it took clearly more time for OCR'ed data than for those stemming from speech recognition or key entry. After this procedure, we periodically tested samples of 1000 randomly selected numbers. In total, around 25 samples were tested, and the number of errors was between 1 and 10 in all cases. Hence, the error rate after this step was clearly less than 1% (0.5% on average) and the errors mostly concerned the least significant digit. This was sufficient compared to our quality requirements.

In the case of the Mount Washington data (Grant et al., 2005), we found keying error rates of around 0.2% to 3% depending on the person doing it. After the quality assurance procedures the error rate was 0.2% or less, but the latter procedure included a manual check of almost all the keyed entries which was very time consuming and probably not worth the small increment in error rate.

5 Summary

In order to optimise the overall goals, i.e., to digitise the data correctly, quickly, and inexpensively and to preserve the meta data, considering the following questions might save time and trouble:

1. What fraction of the available data is needed? What are the available resources? Include the data quality, redundancy, and validation options in any cost-benefit analysis.

2. What is the expected error and what is the required quality? Is a double entry or double check possible or necessary? If yes, use fastest method. If no, use method with fewest possible errors (key entry or speech recognition better than OCR) or optimise quality assurance.

3. Is enough information available in order to prepare the stay in the archive (i.e., thorough literature research, enough sample copies or scans, assessment of legibility, test of reproduction methods, test of digitising methods)?

4. Are the data printed (OCR), type written (OCR) or hand written (speech recognition or key entry)?

5. Are the data organised in tables (OCR) or scattered (speech recognition or key entry, possibly scanning pen)?

6. Is the whole table needed (OCR) or just small excerpts (speech recognition or key entry)?

7. Are the numbers clearly legible (OCR) or faint (speech recognition or key entry)?

Acknowledgements. This work was funded by the Swiss National Science Foundation. We would like to thank all archive staff we had the pleasure to meet during our project.

Sources

Deutsche Seewarte: Täglicher Wetterbericht, Übersicht über die Höhenaufstiege, Hamburg, January, 1939.
Reichsamt für Wetterdienst: Aerologische Berichte, Zusammenstellungen von deutschen aerologischen Messungen, Monthly issues, Parts I and II, 1935.

References

Brönnimann, S.: A historical upper-air data set for the 1939–1944 period, Int. J. Climatol., 23, 769–791, 2003.
Brönnimann, S., Compo, G. P., Sardeshmukh, P. D., Jenne, R., and Sterin A.: New approaches for extending the 20th century climate record, Eos, Trans. AGU, 86, 2–7, 2005.

Grant, A. N., Pszenny, A. A. P., and Fischer E. V.: The 1935–2003 Air Temperature Record from the Summit of Mount Washington, New Hampshire, J. Clim., 18, 4445–4453, 2005.
Jones, D. A. and Trewin, B.: On the adequacy of digitised historical Australian daily temperature data for climate monitoring, Austral. Meteorol. Mag., 51, 237–250, 2002.
Vogler, C., Brönnimann, S., and Hansen, G.: Re-evaluation of the 1950–1962 total ozone record from Longyearbyen, Svalbard, Atmos. Chem. Phys. Discuss., in press, 2006.

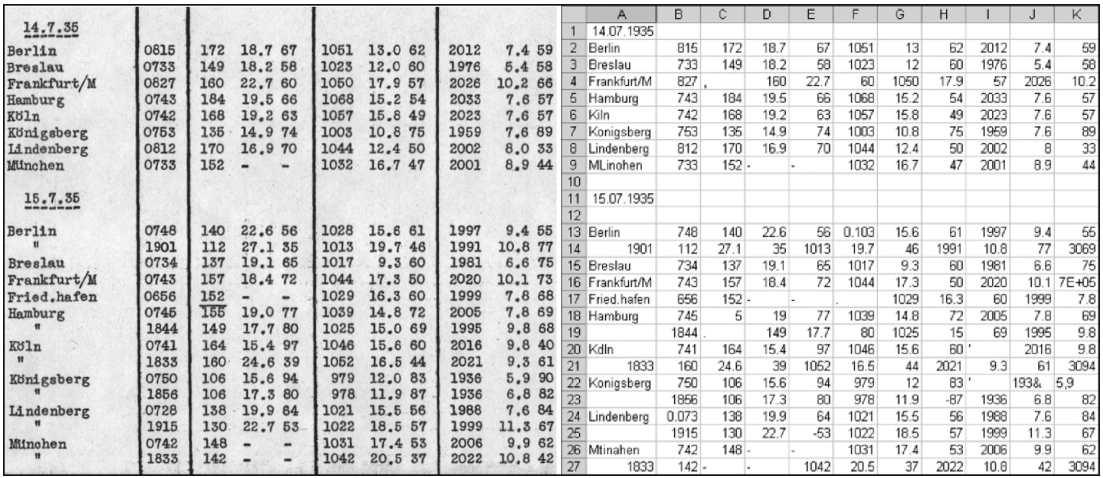
Table 1. Characteristics of the data to be digitised and their relation to the requirements of the planned scientific application.

Formal	Source format	Original (hardbound, loose sheets, etc.), carbon copy, photocopy, photograph, microfilm, image file
	Information type	Numeric, text, code, graphical
	Information format	Table, text, map, graph, mixture
	Typing	Printed, typewritten, hand written
	Legibility	Clear, faint, strike through, blurred, corrections on top of each other etc.
Informational	Data needed	Choice of stations, time periods, alternative sources
	Quality	Expected quality with respect to required accuracy/precision
	Redundancy	Possibilities to check quality and consistency, validation
	Meta information	What is available? How valuable? How archived?

Table 2. Characteristics of the data digitising techniques. Approximate speed is in 5-digit numbers per hour and refers to a trained person and well organised data. Note that these are rough approximations and that the actual speed may deviate considerably from these values. The qualitative assessment of error rate and post-processing is a subjective rating based on the experience of the authors (ten persons).

	Speed (num/h)	Error rate	Post-processing
OCR (scanner)	3000	High	High
Scanning Pen*	1200	Very high	High
Speech recognition	1200	Middle	Middle
Key entry	1000	Low	Middle

*no operational experience was gained, just limited testing.



14.7.35

Berlin	0815	172	18.7	67	1051	13.0	62	2012	7.4	59
Breslau	0733	149	18.2	58	1023	12.0	60	1976	5.4	58
Frankfurt/M	0827	160	22.7	60	1050	17.9	57	2026	10,2	66
Hamburg	0743	184	19.5	66	1068	15.2	54	2033	7.6	57
Köln	0742	168	19.2	63	1057	15.8	49	2023	7.6	57
Königsberg	0753	135	14.9	74	1003	10.8	75	1959	7.6	89
Lindenberg	0812	170	16.9	70	1044	12.4	50	2002	8.0	33
München	0733	152	-	-	1032	16.7	47	2001	8.9	44

15.7.35

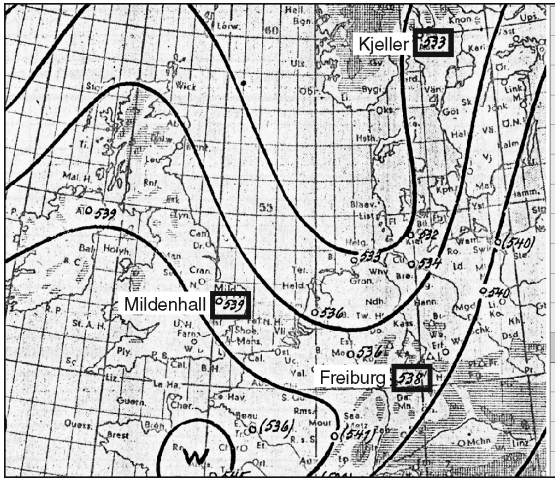
Berlin	0748	140	22.6	56	1028	15.6	61	1997	9.4	55
"	1901	112	27.1	35	1013	19.7	46	1991	10.8	77
Breslau	0734	137	19.1	65	1017	9.3	60	1981	6.6	75
Frankfurt/M	0743	157	18.4	72	1044	17.3	50	2020	10.1	73
Fried.hafen	0656	152	-	-	1029	16.3	60	1999	7.8	68
Hamburg	0745	155	19.0	77	1039	14.8	72	2005	7.8	69
"	1844	149	17.7	80	1025	15.0	69	1995	9.8	68
Köln	0741	164	15.4	97	1046	15.6	60	2016	9.8	40
"	1833	160	24.6	39	1052	16.5	44	2021	9.3	61
Königsberg	0750	106	15.6	94	979	12.0	83	1936	5.9	90
"	1856	106	17.3	80	978	11.9	87	1936	6.8	82
Lindenberg	0728	138	19.9	64	1021	15.5	56	1988	7.6	84
"	1915	130	22.7	53	1022	18.5	57	1999	11.3	67
München	0742	148	-	-	1031	17.4	53	2006	9.9	62
"	1833	142	-	-	1042	20.5	37	2022	10.8	42

	A	B	C	D	E	F	G	H	I	J	K
1	14.07.1935										
2	Berlin	815	172	18.7	67	1051	13	62	2012	7.4	59
3	Breslau	733	149	18.2	58	1023	12	60	1976	5.4	58
4	Frankfurt/M	827	,	160	22.7	60	1050	17.9	57	2026	10.2
5	Hamburg	743	184	19.5	66	1068	15.2	54	2033	7.6	57
6	Kiln	742	168	19.2	63	1057	15.8	49	2023	7.6	57
7	Konigsberg	753	135	14.9	74	1003	10.8	75	1959	7.6	89
8	Lindenberg	812	170	16.9	70	1044	12.4	50	2002	8	33
9	MLinohen	733	152	-	-	1032	16.7	47	2001	8.9	44
10											
11	15.07.1935										
12											
13	Berlin	748	140	22.6	56	0.103	15.6	61	1997	9.4	55
14	1901	112	27.1	35	1013	19.7	46	1991	10.8	77	3069
15	Breslau	734	137	19.1	65	1017	9.3	60	1981	6.6	75
16	Frankfurt/M	743	157	18.4	72	1044	17.3	50	2020	10.1	7E+05
17	Fried.hafen	656	152	-	-		1029	16.3	60	1999	7.8
18	Hamburg	745	5	19	77	1039	14.8	72	2005	7.8	69
19		1844		149	17.7	80	1025	15	69	1995	9.8
20	Kdln	741	164	15.4	97	1046	15.6	60		2016	9.8
21	1833	160	24.6	39	1052	16.5	44	2021	9.3	61	3094
22	Konigsberg	750	106	15.6	94	979	12	83		193&	5,9
23		1856	106	17.3	80	978	11.9	-87	1936	6.8	82
24	Lindenberg	0.073	138	19.9	64	1021	15.5	56	1988	7.6	84
25		1915	130	22.7	-53	1022	18.5	57	1999	11.3	67
26	Mtinahen	742	148	-	-	1031	17.4	53	2006	9.9	62
27	1833	142	-	-	1042	20.5	37	2022	10.8	42	3094

Fig. 1. (left) Excerpt from "Aerologische Berichte" as an example of a data source that easily undergoes OCR (Reichsamt für Wetterdienst, 1935). (right) Screen shot of the spreadsheet produced by OmniPage Pro 14.

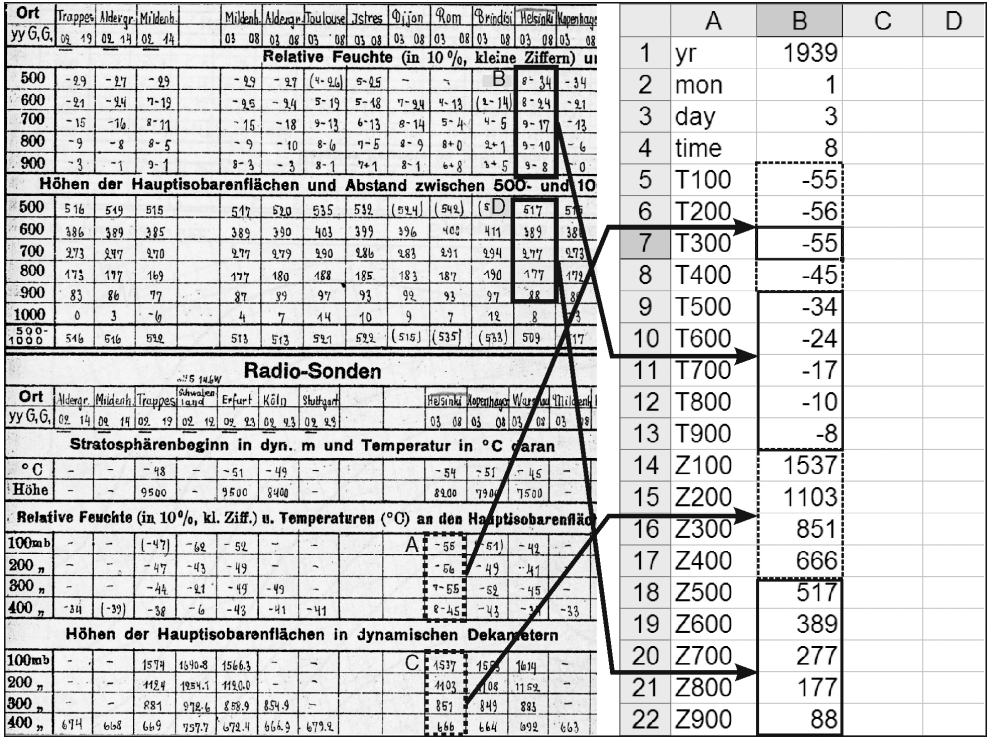
CPD

2, 191–207, 2006



	A	B	C	D	E	F	G
1	yr	mon	day	time	Kjeller	Mildenh.	Freiburg
2	1939	5	13	8	529	538	
3	1939	5	14	8	535	540	536
4	1939	5	15	8	534	535	
5	1939	5	16	8	532	537	537
6	1939	5	17	8	530	535	541
7	1939	5	18	8	528		534
8	1939	5	19	8	530	533	533
9	1939	5	20	8	533	533	532
10	1939	5	21	8	532	536	534
11	1939	5	22	8	533	539	538
12							
13							
14							
15							

Fig. 2. (left) Map of the 500/1000 hPa thickness that includes handwritten station data (from Täglicher Wetterbericht, Deutsche Seewarte, 22 May 1939). (right) Screen shot of the corresponding spreadsheet time series per station in columns. In this case, data from three stations are digitised. The layout is complex and only a fraction of the information is needed. Speech recognition allows using the fingers of both hands to track the data on the weather map while digitising and at the same time produces a suitable data format.



	A	B	C	D
1	yr	1939		
2	mon	1		
3	day	3		
4	time	8		
5	T100	-55		
6	T200	-56		
7	T300	-55		
8	T400	-45		
9	T500	-34		
10	T600	-24		
11	T700	-17		
12	T800	-10		
13	T900	-8		
14	Z100	1537		
15	Z200	1103		
16	Z300	851		
17	Z400	666		
18	Z500	517		
19	Z600	389		
20	Z700	277		
21	Z800	177		
22	Z900	88		

Fig. 3. (left) Table with handwritten aerological data in two parts, from Täglicher Wetterbericht (Deutsche Seewarte, 3 January 1939). (right) Screen shot of the corresponding spreadsheet. The data table is split into two parts and the columns are not in the same order in both tables. Speech recognition allows using the fingers of both hands to keep track on the paper sheet while digitising and thus allows reformatting the data into a suitable format in the same step. The speaker starts with field A in the lower part of the table, then moves up to B in the upper part of the table, then C and D. The time required for digitising one record in this way is not much longer than if it were in a well-organised format. Even if the numbers could be deciphered with OCR (which is not the case here), concatenating the different parts of the table would take a lot of time.

CPD
2, 191–207, 2006