Clim. Past, 2, 137–144, 2006
www.clim-past.net/2/137/2006/

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 Zurich, Universitätstr. 16, CH-8092 Zürich, Switzerland

Received: 1 March 2006 – Published in Clim. Past Discuss.: 4 May 2006
Revised: 19 September 2006 – Accepted: 12 October 2006 – Published: 19 October 2006

Abstract. Hand-written or printed manuscript data are an important source for paleo-climatological studies, but bringing them into a suitable format can be a time consuming adventure with uncertain success. Before digitising such data (e.g., in the context a specific research project), it is worthwhile spending a few thoughts on the characteristics of the data, the scientific requirements with respect to quality and coverage, the metadata, and technical aspects such as reproduction techniques, digitising techniques, and quality control strategies. Here we briefly discuss the most important considerations according to our own experience and describe different methods for digitising numeric or text data (optical character recognition, speech recognition, and key entry). We present a tentative guide that is intended to help others compiling the necessary information and making the right decisions.

1 Introduction

The age of digital computing and data storage has revolutionised data acquisition and administration. Starting around the 1950s, climate data have been stored electronically or on machine-readable media in digital 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).

Manuscript climate data can be digitised for the development of a multitask database or, in the context of a research project, to solve a specific problem (see comment by I. Smolyar, http://www.cosis.net/copernicus/EGU/cpd/2/S108/cpd-2-S108.pdf). In the latter case, which is the topic of this paper, questions concerning the quality of the data become important and affect the digitising process.

Digitising manuscript 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 better assess 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 with respect to quality and coverage, the metadata, and technical aspects such as reproduction techniques, digitising techniques, and quality control strategies. In this paper we present a tentative guide that can be followed in this process. We hope that this guide may eventually contribute towards the development of a set of commonly accepted formal procedures for digitising and processing qualitative and quantitative climate data.

The paper is based on our own experience. We have digitised historical climate data from various sources that differed in format, quality and layout. We digitised historical upper-air data from many different sources (Brönnimann, 2003), a project which carried an unknown (presumably high) risk as nothing was known about the quality of pre-1948 upper air data. In other projects some of us digitised total ozone observations from Longyearbyen, Svalbard (Vogler et al., 2006), as well as meteorological observations from Mount Washington, USA (Grant et al., 2005). Here we report the experience we gained from these projects.

The structure of the paper follows the procedure of digitising historical manuscript climate data, which consists of several steps: defining the requirements for data quality and coverage and compiling the relevant information (Sect. 2), assessing the properties of the manuscript data and choosing the data source and archive (Sect. 3), preparing the archive visit and reproducing the data (Sect. 4), digitising, formatting and correcting the data (Sect. 5), and finally assessing and describing the data quality (Sect. 6). As a summary, a list of guidelines is given in Sect. 7.

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

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 coverage	Available stations/time periods with respect to the required coverage
	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?

2 Defining requirements and compiling information

As a first step one has to describe the data requirements based on the scientific objectives of the project. The product is a list of quality targets (quantitative or qualitative). For instance, in our upper-air data project, we specified the following targets beforehand (Brönnimann, 2003): the accuracy (bias) of the historical data should be within ±0.75°C for temperature and within ±15 to 30 gpm (depending on the pressure level, 850 hPa to 100 hPa) for geopotential height. The predefined targets for the precision were ±1.6 (±30 to 80 gpm) for monthly mean values and ±4°C (±70 to 160 gpm) for individual profiles (the numbers represent a 90% range). For the historical total ozone data the goal was to obtain a data series that is suitable for deriving a climatology for the 1950s (prior to the era of chlorofluorocarbons) and allows addressing interannual variability. For the Mt. Washington data it was envisaged to obtain a data set suitable for trend analysis. These data quality targets are revisited in Sect. 6 in context of the data validation.

In addition to the data quality, data requirements also concern the coverage. How many stations are needed, and what is the desired temporal resolution and coverage? Here one has to keep in mind that redundant information is valuable for quality checks (see below) and it is helpful, at this stage, to plan the validation. The result of this process is a list of qualitative or quantitative criteria that concern both the coverage of the data and the quality.

In the second step, one has to find out what kind of data is available, where, and in what form. The starting point must be a thorough study of the historical literature, including journal articles and technical reports. This is also extremely helpful with respect to meta-information. Following is an example from our own projects: For a number of upper-air stations, historical journal articles helped us to determine the time of observation, which was incorrect in the original data source. Of course, for a description of the instruments and associated errors one often has to rely on the historical literature.

In order to locate the data, it is worthwhile searching the internet. In the context of data imaging projects, historical manuscript climate data have been photographed and archived. Examples for such projects are the NOAA Central Library Climate Data Imaging Project (http://docs.lib.noaa.gov/rescue/data_rescue_home.html) or the International Environmental Data Rescue Organization, Ltd (http://www.iedro.com/). Climate data were sometimes also published in journals that can be accessed online at the publisher's website (e.g., the Monthly Weather Review, http://ams.allenpress.com/) or via JSTOR (http://www.jstor.org/). It is also important to be informed about the activities of others in order to avoid duplicating the work.

In many cases that data can only be found on paper in a meteorological archive. 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. In any case it is very important to find people at the archives that are willing to provide sample photocopies (or scans) of the data sheets in advance. In historical time periods, data reporting was less standardised, the layout of data sheets changed frequently, and it is advisable to ask for as many sample photocopies as possible.

Table 2. Questionnaire facilitating the choice of the appropriate digitising method.

1.	What is the expected error and what is the required quality (Table 1)? 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.
2.	Are the data printed (OCR), type written (OCR) or hand written (speech recognition or key entry)?
3.	Are the data organised in tables (OCR) or scattered (speech recognition or key entry, possibly scanning pen)?
4.	Is the whole table needed (OCR) or just small excerpts (speech recognition or key entry)?
5.	Are the numbers clearly legible (OCR) or faint (speech recognition or key entry)?

3 The manuscript data and its relation to the scientific project

The information that is prepared as outlined in Sect. 2 must now be compared with the manuscript data. We do this in the form of a table that describes the properties and information content of the manuscript climate data. 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., coverage, quality, redundancy, and meta-information; see Table 1). It is recommended to fill out a similar table before starting a project in order to make sure that no important piece of information is missing. The table may be important in order to choose the appropriate reproduction and digitising techniques or the quality assurance procedure.

The first manuscript property is the source format. 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 often necessary or advisable (see Sect. 4). 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. 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. Note that the legibility depends on the reproduction (see Sect. 4) and should be assessed based on the same source format that will later be used for digitising. The formal characteristics of the manuscript data mainly affect the choice of the digitising technique (see Sect. 5 and Table 2), which is arguably the single most important decision.

A second set of criteria refers to the information content of the data (informational characteristics). After having compiled a list of requirements with respect to the spatial coverage, it should be assessed how these are matched by the available data. This leads to a decision concerning which data series should be digitised (any a priori information on the data quality, e.g., from the literature, is very helpful at this point). 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, having thought about ways of assuring the quality (Sect. 2), redundant information may be judged more valuable than good spatial coverage. In our case, for instance, we chose pairs of neighbouring stations wherever possible.

A second important question concerns the expected quality of the data and its relation to the predefined accuracy and precision of the end product. The quality can be estimated based on theory, historical literature, and sometimes also based on the reporting apparent on the sample copies. Here it should be kept in mind that often a large amount of processing is necessary to obtain the final product, which may affect the quality more than the uncertainty of the actual measurements. In our cases, the upper-air data need to be corrected for radiation and lag errors, which are not well known and hence add uncertainty. In the case of total ozone, one needs the air mass, ozone slant path, and absorption and scattering coefficients to derive total ozone from the digitised instrument reading. There are a number of interfering factors that affect the calculated total ozone value, and there are a number of assumptions behind this approach.

Finally, it is important to think about the meta-information: What kinds of meta-information are available (see Sect. 2) and what is the role of this information in the re-evaluation process? Answering this question can be important, e.g., when the same data are available from different sources, one of which must be chosen. For all questions

Table 3. 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 (correction of errors, formatting) 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.

related to the informational characteristics, thorough literature research is necessary.

4 Archive visit and reproduction

After filling out Table 1 and deciding what fraction of the data is needed, a visit to the archive can be envisaged. This poses important logistical questions. 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?

Digitising directly in the archive 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). 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 Optical Character Recognition (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 for speed and resolution.

During our projects, we normally photocopied all material. Per archive day, around 2000 copies can normally be made (make sure to discuss this with the archive beforehand). Travelling well-prepared to an archive also is important with respect to the meta-information. The better one knows the data and the scientific problem, the better one can search for specific information, which probably is available at the archive.

5 Digitising and formatting

The next step is to actually digitise the data. 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.

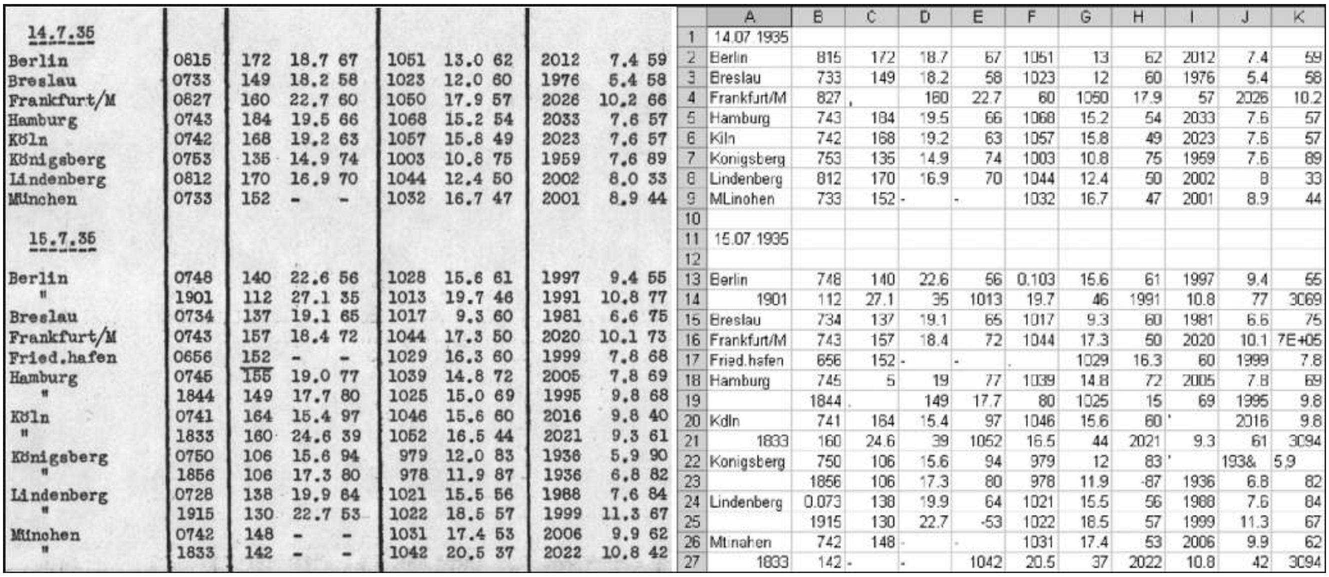
Optical character recognition (OCR) is a powerful technique to convert scanned or photographed documents into text. We used ScanSoft OmniPage Pro 14 for our work. The user can select the area of interest and choose between standard output formats (e.g., text, table, worksheet). We used OCR in conjunction 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. 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. A list of questions that is designed to help choosing the appropriate method is given in Table 2. Table 3 lists information about the performance of the three techniques in a qualitative and quantitative way.

5.1 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



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.

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.

5.2 Speech recognition and key entry

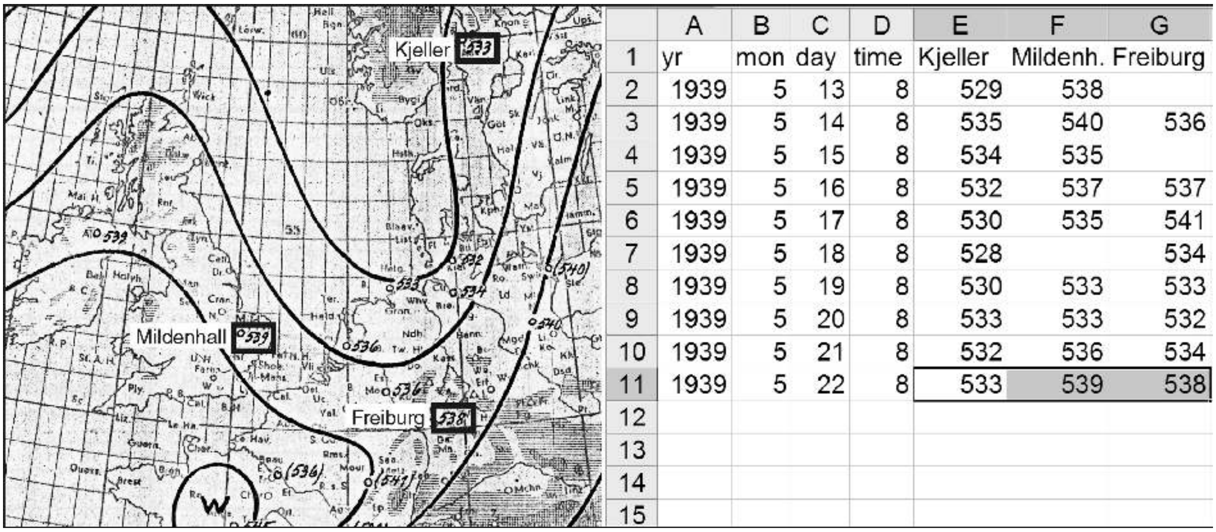
Speech recognition and key entry share similar characteristics. They are normally used if OCR is not possible (e.g., for hand-written or hardly legible data) or would make too many errors, if only a small portion of the table or sheet is used, or if the data are scattered. Figures 2 and 3 give examples 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 shifted columns. 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



	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.

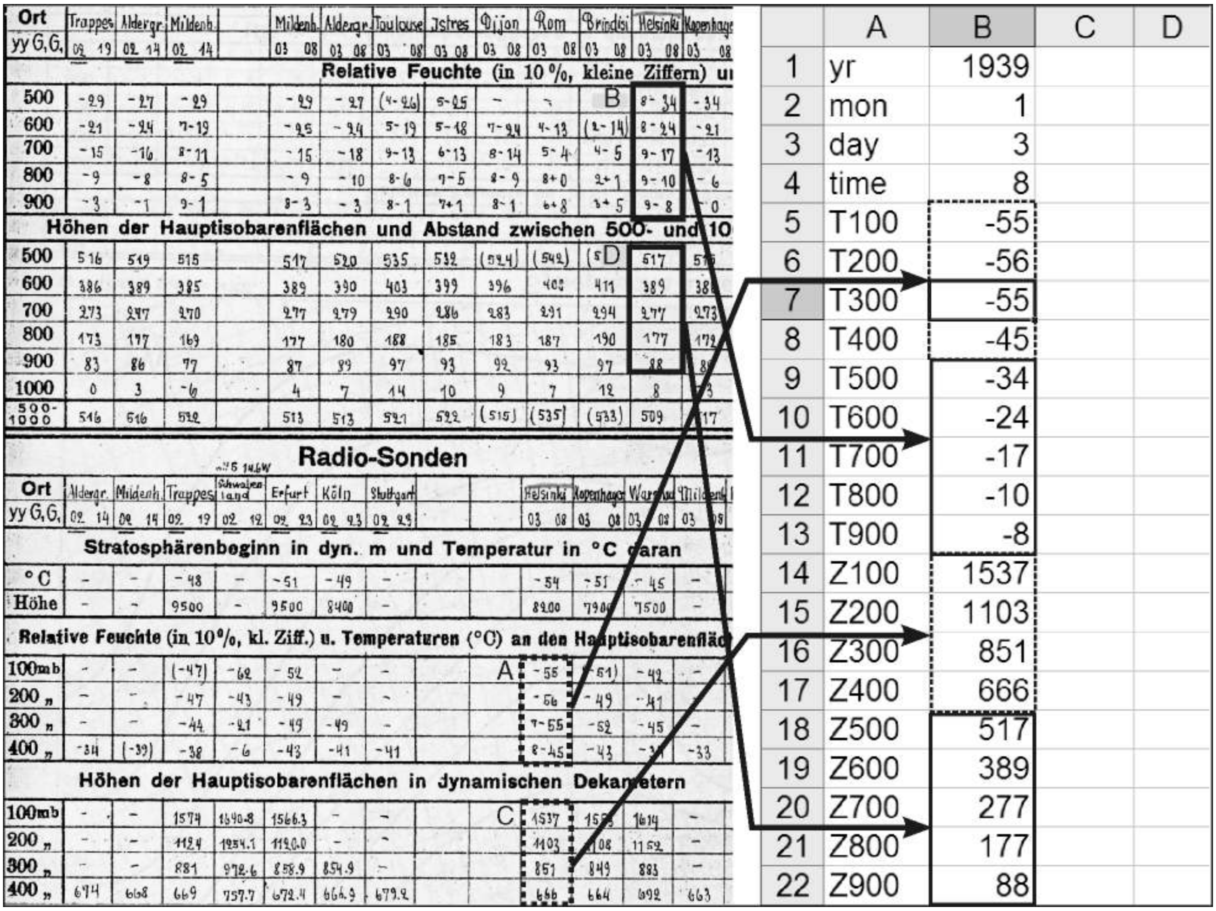
of speaking. The numbers are stored in a buffer and written to the spreadsheet during a breathing pause. We find that a trained speaker can digitise around 2400 5-digit numbers with speech recognition in a 2-h period. That includes the correction of visually (in the work sheet) apparent errors, but not 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 number more often than someone without that knowledge.

5.3 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 double checks (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



Ort	Trappes	Aldergr.	Mildenh.		Mildenh.	Aldergr.	Toulouse	Istres	Dijon	Rom	Brindisi	Helsinki	Kopenhage
yy G,G,	02 19	02 14	02 14		03 08	03 08	03 08	03 08	03 08	03 08	03 08	03 08	03 08

Relative Feuchte (in 10 %, kleine Ziffern) u

500	-29	-27	-29		-29	-27	(4-26)	5-25	-	-	B	8-34	-34
600	-21	-24	7-19		-25	-24	5-19	5-18	7-24	4-13	(2-14)	8-24	-21
700	-15	-16	8-11		-15	-18	9-13	6-13	8-14	5-4	4-5	9-17	-13
800	-9	-8	8-5		-9	-10	8-6	7-5	8-9	8+0	2+1	9-10	-6
900	-3	-1	9-1		8-3	-3	8-1	7+1	8-1	6+8	3+5	9-8	-0

Höhen der Hauptisobarenflächen und Abstand zwischen 500- und 10

500	516	519	515		517	520	535	532	(524)	(542)	(5D	517	5[illegible]
600	386	389	385		389	390	403	399	396	408	411	389	38[illegible]
700	273	247	270		277	279	290	286	283	291	294	277	273
800	173	177	169		177	180	188	185	183	187	190	177	172
900	83	86	77		87	89	97	93	92	93	97	88	8[illegible]
1000	0	3	-6		4	7	14	10	9	7	12	8	[illegible]3
500-1000	516	516	522		513	513	521	522	(515)	(535)	(533)	509	[illegible]17

Radio-Sonden

5 14.6W

Ort	Aldergr.	Mildenh.	Trappes	Schwaben-land	Erfurt	Köln	Stuttgart			Helsinki	Kopenhagen	War[illegible]	Mildenh.
yy G,G,	02 14	02 14	02 19	02 12	02 23	02 23	02 23			03 08	03 08	03 08	03 08

Stratosphärenbeginn in dyn. m und Temperatur in °C daran

°C	-	-	-48	-	-51	-49	-			-54	-51	-45	-
Höhe	-	-	9500	-	9500	8400	-			8200	7900	7500	-

Relative Feuchte (in 10%, kl. Ziff.) u. Temperaturen (°C) an den Hauptisobarenfläc

100mb	-	-	(-47)	-62	-52	-	-		A	-55	(-61)	-42	-
200 „	-	-	-47	-43	-49	-	-			-56	-49	-41	
300 „	-	-	-44	-21	-49	-49	-			7-55	-52	-45	-
400 „	-34	(-39)	-38	-6	-43	-41	-41			8-45	-43	-3[illegible]	-33

Höhen der Hauptisobarenflächen in dynamischen Dekametern

100mb	-	-	1574	1640.8	1566.3	-	-		C	1537	15[illegible]	1614	-
200 „	-	-	1124	1254.1	1120.0	-	-			1103	1108	1152	-
300 „	-	-	881	972.6	858.9	854.9	-			851	849	883	-
400 „	674	668	669	757.7	672.4	666.9	679.2			666	664	692	663

	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.

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.

In summary, the formatting and correction is an important part and often takes as much or more time than the actual digitising. At the same time, the formatting and correction depends on the digitising technique. Therefore, considerations concerning formatting and correction should be considered in the decision concerning the digitising method. Table 2 is designed to facilitate this decision.

6 Validation and description of the quality

The last step is the validation of the data and the description of the quality. This step depends very much on the specific data type and project. There often is a lot of processing between the formatting and the validation such as the correction for known systematic errors, or the desired variable (e.g. total ozone) must first be derived from the digitised data using complex equations or even models (see Sect. 3). Therefore, it is difficult to give general rules. In our upper-air data work (Brönnimann, 2003), we developed statistical tests based on comparing neighbouring stations and based on comparisons with independent reference series. These methods allowed testing whether or not the predefined targets were met. The result of this process, i.e., a quantitative or qualitative assessment of the final data product, should be described (including the assumptions that were necessary in the context of the assessment) and published together with the data products.

There are various ways how this can be achieved, including error bars, flags, summary statistics or assessments in the form of a text. The description should be accurate enough for another user, with different requirements, to decide whether or not the data are useful.

7 Recommendations

The following steps are recommended for digitising historical manuscript climate data:

- Define quality targets (qualitative or quantitative)
- Define requirements with respect to spatio-temporal coverage (include requirements for quality assessment such as redundancy)
- Compile and study historical literature
- Be informed about the work of others, check document imaging projects
- Check data availability at various archives
- Ask archive staff to provide as many sample copies as possible
- Using all the above information, fill out Table 1.
- Based on Table 1, choose data source and archive, choose fraction of data to be digitised
- Test reproduction methods and discuss with archive staff beforehand
- Go well prepared to an archive visit (e.g., in order to locate meta information)
- Use appropriate reproduction technique
- Choose appropriate digitising technique (see Tables 2 and 3)
- Digitise and format the data
- Assess the digitising error, correct errors
- After the necessary processing, validate the final data product
- Provide description of the quality of the final data product

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.

Edited by: H. Goosse

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 20^{th} 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. Disc., 6, 3913-1943, 2006.