Frédéric Falkenhagen – Carl-von-Ossietzky-University Oldenburg – frederic.falkenhagen@uni-oldenburg.de

Accounting for national tradition in data collection when creating an international database

Abstract:

This paper aims at giving an overview of different European traditions in questionnaire design and data documentation and their implications for international comparative research.
First the traditions of questionnaire development and revision will be assessed in four national contexts.
After this overview this paper will more specifically analyse three typical cases of international subjects rendered in national questionnaires from the most static to the freest.
Finally it will turn to the documentation of original data collection and their differing traditions, their usefulness and dangers in secondary use with a special focus on the risks to secondary analysts.
It will also briefly take an outlook into the long term trends and practices in documentation and compare the recent efforts undertaken in international projects like the ESS to improve documentation.

Introduction

This paper aims at giving an overview of different European traditions in questionnaire design and data documentation and their implications for international comparative research.

When trying to establish pool data on internationally observable political tendencies the national traditions of enquiry are one of the most lethal dangers to good databases and sound research.

First the traditions of questionnaire development and revision will be assessed in four national contexts: France, Belgium, Germany and the United Kingdom before the consequences of a breakdown in traditions will be briefly discussed. The focus will mainly be on the questionnaires of national election studies in the 1990s and early 2000s but the position of these developments in their long term traditions will be restated.

After this overview this paper will more specifically analyse three typical cases of international subjects rendered in national questionnaires from the most static (post-materialism) to the freest (authoritarian personality) without leaving out more mixed renderings (political efficacy). It will take a look at possible reasons for these developments and eschew strategies to ensure meaningful inclusion in international datasets.

Finally it will turn to the documentation of original data collection and their differing traditions, their usefulness and dangers in secondary use with a special focus on the occurrences when only extensive field knowledge can avert major disasters in data analysis. It will also briefly take an outlook into the long term trends and practices in documentation and compare the recent efforts undertaken in international projects like the ESS to improve documentation.

Origins of reflection

This analysis rests on the experience of creating a cumulative dataset for a PhD thesis on the structure of ethno-regional party votes in Western Europe over the period 1991 to 2003. The final cases investigated were Plaid Cymru (Wales, United Kingdom), the Scottish National Party (Scotland, United Kingdom), Vlaams Belang (Flanders, Belgium) and the Christlich Soziale Union in Bayern (Bavaria, Germany). The final case was initially designed as a negative case to add some contrast but turned out to fit much better into the pattern than anticipated. Its place as contrast party was taken by Scottish Labour (Scotland, United Kingdom). Apart from the countries concerned by these four cases France has been a significant reference in the work, as the research was done at a French institution in France and the researcher has enjoyed a French education in research methods and techniques. Prior work has also left the researcher with some knowledge of the Canadian traditions in empirical research and work during the PhD thesis has put the author into contact with the European Social Survey on various levels as user or assistant in collection and documentation and some minor research projects that have required documentation.

National tastes in questionnaire development

Before we can analyse the content of surveys it is necessary to restate the surveys in their context and their traditions. Just as the ethnographer has to get acquainted with the general practices of the society he is to live in to avoid major transgressions of code, the secondary analyst has to learn of the general habits and quirks the initial collectors show to avoid documentation despair and transgressions of codebook use. Therefore we will first try to roughly categorize the general types of questionnaire development over the medium and long time range (10 years or more) of exemplary European countries. All these countries were at some point connected to the construction of the final database.

The first question to ask is, whether there is such thing as a national tradition of questionnaire development. We firmly believe there is for a number of reasons. The first is the personal continuity in survey research: few senior researchers are involved in only one major survey as most tend to follow certain enquiries for a number of years. Even as these researchers retire change inside the collection team is generally a very gradual process. But the producing side is not the only one in minor flow, the data users tend to follow datasets for a long time and develop certain habits and tastes. The partial identity between both groups and their interaction lead to a consensus of what is a good questionnaire and its application to the development of such questionnaires. As most political survey questionnaires focused on essentially domestic topics, the natural core for the scientific community concerned with their results was national, albeit a few distinguished friends were included but did not fundamentally alter the nature of the community.

These semi-closed communities developed tastes for questionnaire evolution that can be characterized on a continuum ranging from almost complete absence of recurring elements to a stone-graved canon that suffers no alteration with a lot of space between these two extremes.

The case closest to the first extreme is the French case as Vincent Tiberj has eloquently shown in his analysis of the CEVIPOF-questionnaire from 1978 to 2007 (cf. Tiberj, Stimson & Thiebaud, 2009 and several other undocumented presentations). In a presentation for the debriefing of the 2007 survey the French tradition has been described as a state of permanent revolution. Only two items were present in all surveys and only a minority were used twice or more back to back[1]. The reasons for this tradition are manifold and while some refer to the historically uncertain and shaky funding of the CEVIPOF surveys (which have recently found a more stable base through partnership with the Ministry of Interior) the major reason may also be a preference for data perfectly adapted to the current situation. Thus every single survey represents a perfect fit for the research questions that were raised at the time of enquiry. The downside of this perfect punctual fit is the difficulty of long term comparisons as only few items are adopted once they leave immediate focus. Tiberj has shown that although virtually no item was carried through all surveys a system of lianas exists to enable comparison. Lianas are items that have been used in at least two successive surveys and offer thus a possibility to calibrate the relative scores between these two surveys. Multiple lianas provide a more accurate assessment of relative levels and a complete liana network needs to exist for the entire observation period in every subject field in which a composite measure is expected. The necessary analysis work is cumbersome as the individual coherence of every subject module has to be assessed before a pairwise system of value conversion can be established which will, in the end, require a form of harmonization to be used in ulterior analysis. Timeline analysis and cross-wave comparison remain possible, although at a cost.

The Belgian tradition is a bit more rigid than the French and represents a compromise that reduces the cost of cross-wave comparison without sacrificing the opportunity to have items right on the action in the survey.

The Belgian system is a form of carefully managed continuous change. Individual items are quite regularly dropped or added to the questionnaire but every module has a core set of items that only very gradually change. Thus over a period of twelve years and four waves from

[1] For detailed information refer to the questionnaire database available at http://bdq.reseau-quetelet.cnrs.fr/fr/Accueil

1991 to 2003 modules that were carried in all waves had a core of at least three to four items that did not change or were merely rephrased. At the same time a number of single shot modules were included and most permanent modules carried at least some items with more period than long term relevance. While relinquishing some flexibility compared to the French tradition the Belgian model adds some ease for medium to long term comparison. For permanent modules there are at least two possible strategies depending on the amount of detail required in the analysis. If few items are enough to characterize the subject an adoption of the stable items is the most straight-forward way to construct the dataset. In case more detail is sought to add more incremental variation (as more sophisticated analysis is projected) a puzzle strategy can be sought that adds to the stable items those that have more than period interest but that underwent modification or only appeared during the period. In this case the search and confirmation of functional equivalents is greatly aided by the presence of static items as calibration aids. If such equivalents cannot be found, depending on the ulterior use of the data, a treatment by systemic item non-response in case of single lapses may prove useful. In the specific case of research on the 1991 to 2003 period one interesting item for the authoritarianism scale had only entered the questionnaire in 1995. It was coded as a non-response for the 1991 wave alongside the stable items. The ensuing specific multiple correspondence analysis was able to cope with this situation without too much distortion, otherwise a very loosely fitting replacement item or imputed data might have had to be used as surrogate. The insular non-response has the benefit of not inferring data and thus limiting the artefact risk incurred in such operations.

A third position on this spectrum is taken by the German tradition which is a bit more static than the Belgian without being completely rigid. Where the Belgian tradition keeps the questionnaire in constant evolution while maintaining some anchor points, the German tradition is keeps a constant questionnaire allowing for regular evolution. The German tend to fix their questionnaire for a number of waves and then engage in a review process for the next series of waves. The waves that were relevant for the initial research were those for the German Federal Elections of 1994, 1998, 2002 that were united under the title "Voter Conduct in United Germany". They were followed by the "German Longitudinal Election Studies" (GLES) since the 2005 Federal Election. The questionnaire remained almost unchanged during the whole lifespan of the "Voter Conduct in United Germany" (VCUG) programme. The questionnaire was designed to address the specific issues that arose from

Reunification (cf. the title) and was thus able to offer strong period relevance for the 1990s on the one hand and on the other the opportunity for longer term comparison with the (almost) absolutely stable questionnaire. The period relevance influenced the questionnaire with large modules on the political system (relation to Nazism, Communism and Democracy) and values. It also partly influenced the sampling procedure as the *New Länder* were strongly oversampled. The following project is more concerned with individual long term evolution and seems to rely more heavily on a real Panel, while VCUG only contained a small panel that had to cope with a very high mortality.

The questionnaire of VCUG keeps strong traces of earlier German Election Studies (and other social science research) which enable comparison outside its cadre but contains a few very specific items on events in the early 1990s, especially in the modules on xenophobia and racism. Most of these items are inclusions from other research fields or expansions on the established short module form while a few try to directly tap into recent events. The latter generally show wear from the second wave on, and are of only minimal use beyond then.

The advantages of this Stop-and-Go development are the very smooth possibility to cumulate and compare inside the programme and the opportunity to tap into long term period effects. The disadvantage is the risk of major modifications between programmes due to changing focus and major questionnaire overhauls. Major overhauls always beg the question if every single change has been carefully enough monitored and prepared. In the German case the high quality ALLBUS research programme is a very useful tool for pretesting such overhauls as its continuous waves and changing modules accommodate innovation and testing over several cycles during the overhaul process but glitches happen and the bigger the project the higher the certainty for some minor glitch. From the outside the scope of questionnaire overhaul from VCUG to GLES remains uncertain, but considering the way VCUG was fitted and the noted absence of some standard modules that were peripheral in the specific scope of the project (especially the absence of a substantive module on economic policy) do not bode well.

The strategy as a secondary analyst of such projects is very contrasted depending on the timeframe. If the observation period is inside the scope of a single project all data is directly usable, if the observation period overlaps two project cycles or precedes an important cycle the situation may be worse than in case of a very liquid questionnaire as the system may even lack lianas to get indirect measures as the module may have been dropped or at least trounced in the big review process or so fundamentally changed that all postulates to functional equivalence are speculative and attackable.

The most rigid questionnaire tradition could be found in the United Kingdom up to the 1990s. The questionnaire used in the 1997 British Election Survey was in its modules on political values and ideologies virtually unchanged from the one used in October 1974. Especially economic policy and authoritarianism were not altered in these years although they were transferred from the personal interview to the mail-back questionnaire left at the end of the interview. Most items in these modules can be found as far back as the 1967 wave of Butler and Stokes seminal work on political change in Britain (and some date back to the initial 1963 version). While the British Election Studies have undergone a complete rebuilding prior to the 2001 elections, the Scottish Social Attitudes Surveys and other surveys still use parts of these modules which now can look back to forty years of service. This preservationist feat provides invaluable time series for those interested in long term evolution of values and attitudes. In this lies also its main weakness: the subjects that were not visible at the end of the 1960s are generally absent from the questionnaire and by tracking the evolution of public attitudes some results are simply the tracking of subjects that have completely fallen out of the attention of the general public.

This situation is generally the simplest for the secondary investigator as the stable questionnaire enables direct integration of observations for comparison or pooling for almost every time-frame. The large diffusion of some indicators into other surveys has opened up the opportunity to larger or more specific populations with significant case counts. Due to their long investigation most items are extremely well understood in their reception and their fit into scales or other indicators has been validated over and over. There are nevertheless two big downside risks to this form of questionnaire preservation. The first one is the potential absence of subjects that emerged recently and have now come to be of major importance to the investigation. A truly static questionnaire cannot account for such change or items that cater to specific periods. The other one lies in the orchestrated absence of adjustment. Like an old clockwork the model may spin out of synchronicity first only marginally than gradually more and more and without any revision the time (or results) given may not be of any use. In time, questions may be understood differently or options used in a different manner or simply the terms of the debate may have evolved. In case of such an ossification first a sharp decline in usefulness of the results may occur before a complete breakdown in tradition may come to readjust to the changing society.

Breakdown in tradition

From the case of the 1990s reshaping of the British Election Studies we would now like to investigate such a breakdown and the new start that has come out of this disaster.

In the late 1990s it became apparent that large parts of the old British Election Studies questionnaire were terribly outdated. The economic policy module still carried an item asking for an opinion on the opportunity to nationalize key industries just as if Margaret Thatcher's government had had no influence on the debate (a very similar item on the opportunity of privatizations might have been a useful replacement) and the whole module on mores needed an overhaul as it was still concerned with censorship to uphold moral standards.

While items were the first concern, the whole question format was also coming under scrutiny. True to the standards of their time the questions were almost exclusively worded as statements to be rated on a five-point Likert scale. Direct questioning was virtually inexistent although other traditions had introduced this format to change pace and reduce the risks of response sets. The Likert scales in general and the five point version in particular tend to be also very unreliable for any statistics beyond the classic cross-tabulation. Cross-tabulations were the most current form of analysis in times of the punch card, but regressions and factor analyses have since gained ground and raise issues of interval measurement and length of scale for their analysis.

Another evolution that the original questionnaire was ill prepared for was the shift from personal interviews to more modern forms like the telephone interview, which was still field tested and checked for results divergence in the 1997 Welsh Election Study. The traditional format of mainly post-electoral survey had not undergone any review to introduce in campaign or pre-campaign fieldwork.

With these and other issues piling one atop of the other the whole structure of the British Election Studies crumbled under the pressure and a complete renewal from top to bottom was undertaken before the 2001 elections.

From the outside the results are very mixed. On the one hand all benefits of long time series were lost as not a single item escaped the modernisation unaltered. Those elements that were tried and tested all were discarded and new ideas flew in. On the other many long overdue changes were finally introduced although many overshot the target. Breaking the reliance on Likert-Scales most items were fitted with a numeric ten point scale which theoretically resolves all problems of intervals and variance but always begs the question of participants' comprehension and usage (cf. Lebart, 1992 among others). In survey methods so many things were changed that while every single change bore huge promise the combination made that there was no longer any anchoring point to judge the progress made, as all old was wiped out or changed beyond recognition. Even some old items that had become international quasi-standards were dropped (most prominently an item on the main value of education that since the works of Adorno in the 1940s is often kept as shorthand for authoritarianism went silent).

This breakdown of tradition is characteristic in more than one way, as not only does it come after an old system has to most of its users outlived its usefulness but also that a considerable number of innovations were not to last. The 2005 questionnaire underwent another major overhaul that scrapped or modified many different items and led not only to a change of questionnaire but to the question whether the British tradition of extremely long lived questionnaires has come to an end. For the secondary user this period was the most disappointing as he was confined to being a bystander when this one thing that made the British Election Studies valuable even beyond their nominal scope was destroyed without being noticed by those who worked with. The 2001 British Election Study is in itself a very interesting study as it finally moves the survey closer to the concerns of the day, uses state of the art investigating techniques and has a rarely equalled quality of data collection (as most British surveys) but to the long term external user many decisions remain unclear in the detail and virtually no meaningful comparison can be done between the 1997 and 2001 editions (and even 2001 and 2005 are not simple to compare). Individually the reconstruction brought the British questionnaire closer to its international counterparts but at the same time it left some decisively British aspects behind.

While primary collectors and users have some influence on these developments secondary users simply suffer the results, thus the great risk of too ossified questionnaires for secondary users is that without established change management questionnaires can come down like derelict buildings without any reference points to continue the building of meaningful equivalents. In the British case luckily some modules have been carried on by other surveys which already had taken steps to gradually modernize individual items a few at a time.

Topic development

As we now have a general grasp of different life cycles of questionnaires and items we grasp one of the major questions in international medium or long term research: Not only do we have comparable items, but also what do they look like in the next period? There are potentially important differences between subjects with single established modules and subjects that spawned an infinite variety of possible items and modules. As a series of examples and their implications on our work we would like to discuss a case with clearly established modules, one that is completely free and a third that is somewhat intermediate.

Stability: post-materialism

One of the most stable subjects is post-materialism. Inglehart's work provided both the theoretical framework and the operationalisation in two different forms with his works of the 1970s (Inglehart 1977, 1997). Once the theory of post-materialism began to spread and the issue was seen as an interesting point to investigate the module for inclusion in questionnaires was already developed and extensively tested in several European countries. Investigators could use and have used one of the two modules off the shelf. Cases of such modules are the easiest for international comparison. Not only do the modules match, they are also tested for local differences and dimensional problems. There is also a standardized form of analysing the results which may be the initial interpretations or streaks that the scientific community has come to develop. In the case of post-materialism there are two different modules. A brief (and more widespread) one consisting in having respondents choose the two most important goals for the society from a list of four. Two of these items are associated with materialist and two with post-materialist priorities. The respondents who chose one of either type are classified as mixed. The longer version (that is largely in disuse) asks to choose five items out of a twelve item list. Those who choose four or five items of one type are classified as materialists or post-materialists respectively while the remaining are classified as mixed. The Belgium Election Studies used the long form while the German (VCUG) and British surveys used the second, with not every British survey carrying the module. In case of absent module there is no other opportunity than to opt for a systemic module non-response (unless one removes the survey) and seek a treatment for this issue afterwards. We thus limited our effective database to the classification materialist, post-materialist or mixed but maintained a coding of the priorities given for further analysis especially for analysis of national cross-sections which were coherent. We deemed them directly comparable on behalf of other studies and the international testing that has been done in recent years. In case of a theory only tested in one or very few national contexts we would have had to establish the comparability or question the initial collectors to get to their arguments for having accepted the module as functionally identical in their context to the initial context. Post-materialism opens up to such a debate as the issues it tries to tap have left the immediate political space. Especially fighting inflation which was a big subject in the 1970s and most of the 1980s has become unknown to a generation that grew up in the 1990s and 2000s that saw remarkably stable prices. In this the question of ageing questionnaires will have to be raised on the initial collectors level, but not by secondary investigators as long as they can trust the initial collectors to apply due diligence, which is implied in using their data.

Free development: authoritarianism

The counterpoint to the situation of post-materialism is the investigation into authoritarian personality which is characterized by a multitude of items and methods. This development can be traced back to the original development in the 1940s. Adorno's et al. (1950) work was never meant to be applied to an international audience and thus lacked any international dimension in the development and testing of the questionnaire and not only the international audience was left out, almost all testing occurred on white middle class Americans living in Pacific Coast states with the exception of the psychic patients and some excursions to workers' social clubs still in Pacific states. The questionnaire was thus clearly situated in this context and was tailor made to their issues. Triggers for prejudice included Filipinos and Oklahomans, that were very specific to this time, region and social stratum. A second issue is the design of the questionnaire that has cumulated problematic decisions. Most importantly every item that has been tested consisted in a statement that showed the trait investigated (authoritarianism at large) and asked for an evaluation of this statement. The single orientation meant that higher scores given to statements automatically were higher scores on the authoritarianism scale, regardless of real content. This leads to the immense risk of absolute set response as general computation of results was only a sum index without weighing or hierarchy. This unsatisfactory situation had to be rapidly addressed in questionnaire development and was in different ways (generally by inverting at least one of the items but more profound changes have also occurred). But the scale used was also an issue of fashion. Adorno opted for a six point Likert scale without neutral position on the questionnaire but with a coded neutral position as non-response as the median point of the scale. This scale is widespread in psychology and the neurosciences but has fallen into disuse in the social sciences for several reasons: first it has been established at least since the 1960s that non-response may not be neutral and especially not in case of issues exerting high social control like authoritarianism (Michelat & Thomas, 1966) and second the distinction between two low-intensity answers has proven problematic to even formulate expressively (an issue Adorno et al. circumvented by stating on the forms that "the intermediate numbers serve to qualify your answer") and have thus been dropped in favour of Likert scales with either four (without neutral position) or five positions (with neutral position). As the arguments for and against neutral positions have their respective merits both have been used depending on national tradition and individual collectors.

The issue of limited scope of testing has spiralled even further due to the age of the initial development. The United States of Adorno's initial research were still a racially segregated society and thus the issue of racial prejudice was much more prominent than it has been in recent years. Evolution in other moral standards (especially with relevance to sexual habits but also dress codes) have also taken place and displaced the base of most triggers Adorno used. The scale was in need of renovation and shortening as in the final version Adorno had still kept thirteen items which is much more than the usual space devoted to a module in current surveys. Unless a module taps into the heart of the inquiry around four to six items seems to be the maximum space allotted. The necessary overhaul of the measure of authoritarian values and their internationalisation have not been coordinated centrally and have thus taken any number of different developments and through the almost complete obsolescence of initial development work, few common characteristics remain.

The general picture is that of distinct national corpuses of established items that may or may not undergo evolution depending on the national traditions.

For the secondary researcher the strategy to follow implies a lot of trust in the national collectors. One has to assume that they have tested their items for coherence and applicability in the national context. While the Belgians have adapted a few items from the initial questionnaire to their national context, the British have kept the general themes but developed specific items and the Germans have travelled a different road by developing major modules that tap either into the xenophobia/ authoritarian dimension or into the fascism/ chauvinism dimension with clear references to different stages of Adorno’s questionnaire development. There is no practical way for direct comparison of items and it is thus very difficult to judge differences in national levels of authoritarian values. This is a weakness in the analysis that we will have to accept as all standard comparative strategies would potentially cause more harm through artefacts or random choices than good in giving numbers to compare. Nevertheless one can still compute a relative comparison by creating or using nationally tested composite indicators and segment the population by relative level. This may permit to include authoritarian values as an independent variable in the analysis but not to make the value achieved a dependent variable in conventional analysis for fear of lacking class homogeneity (no one can be sure that a certain level on the British scale corresponds to a certain level on the Belgian scale, although relative positions can be ascertained). The step to code relative groups instead of the standardized values the respondents achieve on their national composite indicator reduces the analytical value of the module by limiting variance, but not to do so would imply that the secondary collector can be sure that the extreme points of the scales are the same across nations. To clarify this problem another example may be useful: the Belgian questionnaire has a module on acceptable forms of political participation that runs from signing petitions to political violence while its British counterpart runs from writing to the MP to participating in a lawful demonstration. A full score on the Belgian scale would denote a terrorist while in Britain anyone who held a sign under the windows of a council meeting would achieve this and in case of direct correspondence they would be considered equal. The relative correspondence assumes that in the top group of respondents those the most open to unconventional forms gather, regardless of what the national culture considers to be these unusual forms of participation. Through these devices we can achieve a degree of assurance that the data has been treated truthfully and still produce results (especially when analysing which groups attract those that score the highest or lowest on authoritarian values and consider national differences and similarities).

The middling grounds: political efficacy

Between these extremes which lead to very clear cut and drastic strategies there is a vast midfield that calls for much more complex treatments that in the end may yield very useful and complete results but only at the price of major effort in putting together the dataset. One of such cases is political efficacy a widespread notion that has spawned a lot of different operationalisations. There is no single canonical module in widespread use but the developments follow a path that suggests mutual exchange and learning. This web of discussion and evaluation has lead to a certain similarities in modules used. They generally investigate a number of shared sub-fields alongside specifically local questions, but they do not share single formulas of wording. Sometimes these are not completely fixed nationally and show some slight local variation, too. The format of these questions is relatively homogenous as most use balanced 5 point scales either in Likert format or as responses to questions. There remains thus a large variation of wordings for closely similar concerns.

In such a case some detective work is necessary in digging into documentation and national testing of wordings. The aim is to identify functional equivalents on an item for item basis. There are potentially a lot of “false friends” around, that look like identical questions but in national acceptance have very different meanings. A known example is an economic policy item that asks whether the respondent is in favour of nationalisations. In Britain and France this item is part of the left-right economic scale (although with diminishing efficiency over time) while in Germany “nationalisation” is associated with the nazi expropriations in the 1930s and thus the item ranks quite high on the fascism scale. In our database we opted to work with tools of analysis that are able to account for insular non-response to include as many items with even partial presence possible. This was especially important as the Belgian questionnaire only carried a very short module in 1991.

After long research we were able to identify two items that were almost universally used (one on satisfaction in democracy and one on general interest in politics) and six subject matters that were included in most questionnaires (on the sense of voting, if parties only want votes, if it matters who is in power and three on the characteristics of politicians: being out of touch, trying to do the best, and listen to ordinary people). Apart from the 1991 Belgian Election Study the 2003 Wales Life and Times Survey was the one with the least exhaustive module and several British studies had some of the items only included in their mail back form which led to a relatively high rate of system non response. At the bottom line the only subject every questionnaire carried was on the sense of voting, but at the same time except for the 1991 Belgian Election Study with only three items no survey tapped into less than five of the eight subfields. And the lowest performing item was satisfaction with democracy that was only carried in 11 out of 16 questionnaires (with the others scoring between 13 and 16). The final roundup enabled sensible comparison as well globally as on single subfields in national and temporal space. This result is encouraging, but the amount of effort and work needed to accomplish the conversion table depends largely on acute knowledge of the national cases studied to try to evade false friends and also on the quality and accessibility of the documentation that accompanies the datasets.

In establishing the database not everything is possible and a several points the secondary investigator has to recognize the weakness inherent in his data and act on it. The price to pay is either more complex data analysis routines or less impressive results, but the reward is results which are based on data that one is confident of being not stretched over its resistance. A strategy of only assuming what one knows will not be overstretching the foundations can open the opportunity to be brave in the data analysis as the base is firm. This data collection depends to large parts on acute knowledge of the debates surrounding a subject in the relevant countries, knowing how the questionnaires tend to evolve nationally and individually but also all types of national traditions and history and as it is rare to have this knowledge on more than a limited number of subjects and countries there is one thing that can make or break such a dataset: it's the documentation in content, quality and usefulness that are nor always positively linked. We will thus turn to the way data is or has been documented and its incidence on the work of secondary analysts.

Traditions in documentation

Traditions in documentation can be summarized into three distinct groups. The first can be called traditional as it at least assumes that the documentation is a book with information that is written and printed to be understood linearly. The second is the off-line digital documentation that ties documentation closely to the computer terminal used to produce the dataset and used in its analysis. The third is the on-line digital that integrates the opportunities and the webbed structure of international communications.

Traditional documentation

The traditional form of documentation can be studied on the basis of the codebooks and other documentation from the early days of international circulation of survey data but is still present in current Belgian documentation. The situation that this tradition envisions for the use of its dataset is the reception of a bulky package of punch-cards accompanied by a short letter from the original collector and a linen-covered book to guide the researcher in his or her work with the dataset. For easy reference it consists of a number of chapters of explication of data and context before a long appendix with the codebook proper and generally a series of frequency tables. These codebooks generally contained a short summary of the political system and a brief introduction to the campaign, its main issues and its most important events. These really brief introductions, generally only a few pages, were not meant to give a full lecture on the subjects but hand some background on main tendencies (for political system) or give the scholar of the country that did not witness the events (the campaign information) an inside account. The next chapters were short accounts of the questionnaire development, which items were dropped, which were modified and what new modules entered the questionnaire (in case of moderated change) or how the questions were chosen (in case of less stable questionnaires). The final chapters of such a typical codebook were the data collection notes that detailed the technicalities of the survey (sampling, response rate etc.). The codebook proper was in the appendix where items were transformed into variables and these variables introduced with technical information (format, coding, name and/or position in the dataset) and content information on the coded values of different responses or non-responses. Historically very useful were the inclusion of a frequency table alongside the coding information, as in punch card times every single frequency analysis could take several hours. Today their usefulness is much reduced as with the dataset ready this information is accessible in split-seconds. They have thus generally disappeared in recent surveys, although sometimes a quick look at the distribution of a variable when reading through the codes can point to variables to include or exclude in further research.

Documentation of this type can currently still be found in Belgian Election Studies (including frequency tables) and at least up to 2000 Canadian National Election Study. Historically the

British followed this tradition up to the mid 1990s and virtually all research prior to the 1990s included this form of documentation.

This form has two major disadvantages: as an edited volume of essays and technical information it is quite cumbersome to use. A book is by definition mono-dimensional and depending on the format and precision in indexing may be a nightmare to search for the really useful bits of information, especially for those with excellent knowledge of the case and general patterns of questionnaire development. Especially older codebooks can be a true detective story when the only available documentation is a collection of typed sheets with variable description with manuscript frequency annotations that have been Xeroxed a generation or more ago and were laid on an early scanning device to be photographed with a slight tilt, as I experienced with a 1968 Canadian National Election Study. The second is the necessary selection and edition of information on campaign events and questionnaire development. As the idea is that the codebook is an edited and printed book there is just a set amount of space in this format. Summarizing six weeks (or several months) of campaign into five or ten pages implies an important selection work and introduces an observer bias that may be problematic even without intention. The omission of seemingly unimportant aspects may destroy another inquiry with a different approach.

The selection and effort to write a coherent introduction to the data is also the greatest strength of this tradition. The concentration on a few key events in the campaign and the few pages of thematic introduction make this form of codebook the most easily accessible especially for those unacquainted with the dataset and potentially its context. The ideal-type would be the international researcher that tries to get a comparison sample for another research or the student that for his term paper is fed up with the dataset used in class and wants some hands-on experience. This form of codebook is really useful as an introduction to the dataset and if frequencies are still noted may be enough for a short reference on general attitudes or practical issues (e.g. check the quality of response rates or compare regional fieldwork to national results on standard items). It may even represent a useful volume by itself as it contains results and accounts of situation.

Offline digital documentation

The first step towards a better disclosure of all technical details is the offline digital model of documentation. This model was developed in the 1990s when datasets were no longer punch cards or weighty magnetic bands but a few floppy disks or a CD-ROM and desktop computers not only suited for data analysis but also capable of multitasking and displaying the codebook alongside the statistics programme. At this time the hard copy documentation was no longer a constraint and information selection could be largely done away with. Digital indexing and searching provided the key to unlock the mono-dimensionality that made editing a necessary feature of documentation. In practice some technical limitations still applied (especially RAM but also processor power) that forced some decisions. The chapters that were most edited either disappeared (introduction to political system) or severely reduced (campaign events). Even the account on fieldwork was often reduced to a simple series of relevant numbers.

The most complete example of this model is the British documentation since the late 1990s. The British attained a level of disclosure that is as close to absolute as one can get. The codebook proper was in fact nothing less than the unedited source code of the questionnaire for their preferred investigation programme. As in the age of CAPI and CATI the interviewer interface effectively produces the dataset all information for the codebook is inside the source code. Every single phrase the interviewer is to read out, every single coding advice displayed on his screen is there, laid bare to those who can read it. There is in theory no way to hide the devices that distort data in a favourable way or put respondents on certain rails. Every bit of information that has been collected can be traced to its origin.

What seems to be best of all documentation worlds turns into a nightmare for most users. Those who are not thoroughly acquainted with the programme used and/ or reading code face an insurmountable challenge in getting to grasp essential information in this ocean of data and command ciphers. There is virtually no simple introduction to the data but only data. Those who do not possess extensive field knowledge not only on the field investigated but on the field of investigating this field by this team have no simple access. The problems of information selection have lead in this model to the absence of editorial work and incapacity of decisions to restrain the flow of information to an amount the outsider might read. There remains doubt whether this reduction of editing was only a matter of giving better unfiltered access to data or a matter of reducing workload by dumping data on the user instead of editing a volume. The scholar from abroad who only looked for some comparative data could retrieve the data with traditional documentation but risks getting swamped with off-line digital information and the student will have to invest a lot of hard work into his term paper. Hard work that will pay in the long term, as once he has acquired the secret knowledge of code reading, he will have complete access to the data documentation. The high entry barrier remains the main problem of this form of documentation that once this barrier has been overcome is almost complete.

Online digital documentation

The tendency to dematerialize was taken to a new level at the turn of the century by the model of online digital documentation. This model not only takes advantage of progress in computing but of the progress in usage. If offline digital documentation still was a linear thread of one or at most a few files, online digital is a collection of short documents on any particular aspect. Offline digital relied on the index, online relies on links either for cross-reference or for individual access to a subject. The exploration technique used also evolves: if users of traditional documentation did linear reading, offline digital users worked through the index to cater to the needs of their research, while online digital users can browse through the documentation creating a personal path to the data by being guided by links giving potential direction to their working. The main structure of an online digital documentation is webbed and divided into a great number of individual files which are accessible either through main overview lists or through cross-references in specific other files. Two examples are the German archives at Gesis and the European Social Survey (ESS). While offline systems dump the data on the user online systems can internally link information (for example coding information on a particular question to campaign events or fieldwork remarks that may have had an influence on it) without adding bulk to documentation as a single information file may be linked to any number of other information without creating redundancies. In this the webbed structure enables a truly three-dimensional documentation where information can be accessed through any chain of links. This linking gives the documentation more structure and may help to manage huge data and documentation reservoirs without restricting access.

Apart from linking (which would have been possible offline) another great feature is the possibility to outsource certain parts of the documentation to other readily accessible sources. The general introduction to the political system may be dropped in favour of a link to a good introduction by official services (e.g. ministry of foreign affairs), scholarly work (e.g. an open access publication or lecture) or vulgarisation (e.g. online encyclopaedia) which spares the effort to write those while still providing the service for those who need it. The same applies to general campaign introduction or explanations on generally known investigation techniques or sampling routines.

The crucial advantage though is the opportunity to keep the documentation up to date. While in theory this was possible with other models through errata and addenda it was too cumbersome to be done regularly. For online documentation this is no longer a problem as only the data on the server needs to be altered. Especially if certain effects (news, sampling, response set, fieldwork etc.) prove to be much stronger than anticipated they can be put to the fore through specific linking or alteration of their entry. It also opens up the back channel from data users to data collectors. If users make public their experience and their results the collectors can more readily engage a dialogue and amend the documentation. These contacts can be through standard online feedbacks or through online publication of results but in theory a user forum would be possible. This forum could also serve to connect users and create a community of users that might highlight frequently encountered issues that standard test rounds do not detect. Even if we discount community and cloud euphoria, the user as content generator is an asset online documentation has.

The true downsides of this model are the dependency on functioning access to the depository server. Network outages are rare currently, but bad access and low availability always threaten as we are reminded any time when major cables go silent. There is also the long term perspective. A set of punch cards and an edited volume may lie forgotten in the archives for a decade and more and still be usable whether the data distributor is still in activity. Once the distributor shuts down online documentation will no longer be available, even if the dataset survives on offline media. This is no theoretical issue as the disappearance of the *Belgium Archives* for *Social Sciences* (BASS) has shown us in recent years. The connected issue is the need for external documentation that bears the same risks.

Another important issue originates in application of the model. Most examples have trouble acknowledging that the workload involved in good online documentation is considerably higher than in traditional documentation. Instead they try to introduce automation into the tedious work of documentation. A prime example of this is the English version of the 1994-2002 *Voter Conduct in United Germany* surveys whose codebook was a trap for anyone without good knowledge of German and the German political system. Translation was generally limited to the question asked and did not extend to the responses, which was especially dangerous as the German tradition of Likert scales is coded inversely to most other countries (Germans assigning code “1” to “completely agree” and “5” to “completely disagree”). A true nightmare was the translation of administrative terms, as the plural “Länder” was generally correctly translated to “states” while the singular “Land” was systematically translated as “country” although it often referred to a single state rather than Germany the final straw was the confusion created by a translation of “Land” as “federal state” (confusing federal for federated). Without the ability to recur to the original questionnaire or good knowledge of German standard items this documentation would have been a major source of trouble. This has been addressed at the latest software overhaul through a workaround in question description (the frequency tables still are only in German).

The ESS a potential best practice?

The European Social Survey as one of the biggest surveys to date has tried to build upon the experiences and fulfil the promise of online documentation. It has been willing to do the work needed beyond dumping data on the user but has not been able to quell the worst enemy of digital documentation in general and online in particular: drowning the user.

The ESS has sought to have very careful documentation for the questionnaire development in every single language and particularly strict regulations on local deviation from questionnaire or methodology. It even succeeded in bringing together collectors from different countries sharing a language to discuss and remove differences in question wording. These documentations and advisories to the dataset have reached the point where only those that have been involved throughout the process have any chance of knowing every step and the long warnings on non-standard administration of the questionnaire are often ignored as they would exclude a number of countries ex officio, because some routines are not applicable in certain contexts. As an example the manual asks that once an address has been drawn at random the interviewer first takes the names of all mailboxes (alternatively doorbells) and then proceeds to choose the household interviewed. This method is strictly impossible in Paris and some other larger French cities as to access the mailboxes virtually all buildings in these cities require the entry of a code to discourage begging and housebreaking.

ESS also aims at the best possible coverage of the news context before, during and immediately after the fieldwork. The general indication calls for about 70 to 100 entries into this database per wave and country. If one multiplies five waves with 34 participating countries, this comes to an endless news stream. In fact the number of events noted for the first 4 waves ranges from 0 (Lithuania and Latvia) to 700 (Spain) totalling more than 6000 short descriptions of events. Even though possible questionnaire links are included, the necessary work to piece it all together would be tremendous.
This information overkill is a major issue and is partially addressed through summary reports but will still need some improvement.
A field where ESS puts emphasis is user relation as feedback is aggressively sought, a lot of training offered and publications on ESS widely published on the site. This may be the first step towards a user community it assuredly is a step towards dynamic documentation.

Conclusion: going for a best practice

If true online documentation seems the way to go, preparing some old fashioned single volume thoughtfully edited documentation may prove useful as an introduction to the dataset. Those who will thoroughly use the dataset will carry on to the web of knowledge online documentation offers, but those just taking a peek will not be shut out or be endangered through ignorance of major facts.
This way may unite all the advantages in exhaustiveness of online documentation while at the same time lowering the usage barrier that the information flood creates for the casual user. To create useful documentation an important amount of work will be needed to edit the introductory texts, although they are generally reusable with slight modifications for several consecutive issues. To neglect the quality control for online documentation would at the same time be foolish, as sloppy translations or inaccurate labelling as a result of a lack of proofreading would risk completely devaluating the dataset for secondary analysis or long period analysis.
Still secondary analysis requires general knowledge of traditions of questionnaire development to evaluate whether using a dropped item in the analysis is acceptable or implies using data that was judged sub-standard by national collectors (the first being generally the case in France’s flexible questionnaire, the second in English style fixed questionnaires) and

respect and thoughtful use of data analysis techniques when several surveys are cumulated into a single larger dataset. All this depends on useful documentation that is reasonably complete, up to date and succinct to be absorbed by a single individual.

Appendix

Bibliography

Adorno Theodor Wiesengrund et al. *The Authoritarian Personality*. New York, Harper and Row, 1950. 989 p.

Butler, David, Stokes, Donald (1974). *Political Change in Britain – the evolution of electoral choice*, 2nd ed. Basingstoke, Macmillan.

Falkenhagen, Frédéric (2010). Les *électorats ethno-régionalistes en Europe Occidentale*. PhD thesis, Paris. http://ecoledoctorale.sciences-po.fr/theses/theses_en_ligne/f_falkenhagen_scpo_2010/thesef_f_falkenhagen.pdf

Inglehart, Ronald. *The Silent revolution : changing values and political styles among Western publics*. Princeton (NJ) : Princeton University Press, 1977. 482 p.

Inglehart, Ronald. *Modernization and Postmodernization : cultural, economic and political change in 43 societies*. Princeton (NJ) : Princeton University Press, 1997. 453 p.

Lebart, Ludovic. *La qualité de l'information dans les enquêtes = Quality of information in sample surveys*. Paris, Dunod, 1992. 549 p.

Michelat, Guy, Thomas, Jean-Pierre H.(1966) Dimensions du Nationalisme. Paris, Collin.

Tiberj, Vincent, Stimson, James, Thiébaut, Cyrille. *Structure et structuration d'un public policy mood à la Française*. Paper presented at the 2009 AFSP Congress.

Data References:

Scotland:

Jowell, R. et al. , Scottish and Welsh Referendum Studies, 1997 [computer file]. Colchester, Essex: UK Data Archive [distributor], December 1998. SN: 3952.
National Centre for Social Research et al. , Scottish Social Attitudes Survey, 1999 [computer file]. Colchester, Essex: UK Data Archive [distributor], June 2001. SN: 4346.
National Centre for Social Research, Scottish Social Attitudes Survey, 2001 [computer file]. Colchester, Essex: UK Data Archive [distributor], February 2004. SN: 4804.
Scottish Centre for Social Research, Scottish Social Attitudes Survey, 2003 [computer file]. Colchester, Essex: UK Data Archive [distributor], January 2005. SN: 5076.

Flanders:

ISPO, PIOP, Belgian General Election Study, 1991. Leuven/ Louvain, Belgium. Niwi (Netherlands Institute for Scientific Information Services)[distributor], 1997. p :1228.
ISPO, PIOP, Belgian General Election Study, 1995. Leuven/ Louvain, Belgium. Niwi (Netherlands Institute for Scientific Information Services)[distributor], 1998. p :1422.

ISPO, PIOP, Belgian General Election Study, 1999. Leuven/ Louvain, Belgium. Niwi (Netherlands Institute for Scientific Information Services)[distributor], 2002. p :1693.
ISPO, PIOP, Belgian General Election Study, 2003. Leuven/ Louvain, Belgium. 2004 [distributed by original collection centre]

Bavaria:

Falter, J. et al., Political Attitudes, Political Participation and Voter Conduct in United Germany 1994. Mannheim, Germany. GESIS (German Social Science Infrastructure Services) [distributor],2004. ZA3065
Falter, J. et al., Political Attitudes, Political Participation and Voter Conduct in United Germany 1998. Mannheim, Germany. GESIS (German Social Science Infrastructure Services) [distributor],2004. ZA3066
Falter, J. et al., Political Attitudes, Political Participation and Voter Conduct in United Germany 2002. Mannheim, Germany. GESIS (German Social Science Infrastructure Services) [distributor],2005. ZA3861

Wales:

Jowell, R. et al. , Scottish and Welsh Referendum Studies, 1997 [computer file]. Colchester, Essex: UK Data Archive [distributor], December 1998. SN: 3952.
Jones, R. Wyn et al., Welsh Assembly Election Study, 1999 [computer file]. Colchester, Essex: UK Data Archive [distributor], 14 December 2000. SN: 4180.
Jones, R. Wyn et al. , Welsh Election Study (Wales Life and Times Study), 2001 [computer file]. Colchester, Essex: UK Data Archive [distributor], August 2002. SN: 4546.
Jones, R. Wyn, Heath, A. and National Centre for Social Research, Wales Life and Times Study (Welsh Assembly Election Study), 2003 [computer file]. Colchester, Essex: UK Data Archive [distributor], December 2004. SN: 5052.