Public use of travel surveys: The metadata perspective

Resource paper for the 2\textsuperscript{nd} International Conference on Transport Survey Quality and Information, Kruger National Park, August 2001

KW Axhausen

Arbeitsbericht
Verkehrs- und Raumplanung

April 2001
Öffentliche Nutzung von Befragungen zum Verkehrsverhalten: Die Metadaten-Perspektive

KW Axhausen
IVT
ETH
CH-8093 Zürich

Telefon: +41-1-633 39 43
Telefax: +41-1-633 10 57
eMail: axhausen@ivt.baug.ethz.ch

April 2001

Kurzfassung


Der Aufsatz diskutiert als zentrale Technologie, XML document type definitions (DTD) und stellt ein prominentes Beispiel, das DTD der Data Documentation Initiative (ddi) vor.

Schlagworte

The public use of travel surveys: The metadata perspective

KW Axhausen
IVT
ETH
CH-8093 Zürich

Telefon: +41-1-633 39 43
Telefax: +41-1-633 10 57
eMail: axhausen@ivt.baug.ethz.ch

April 2001

Abstract

The paper discusses the slow uptake of the data archiving and publication agenda discussed since the last conference of this series. It identifies the currently high start up costs, as the main reason for this. One-off solutions, such as www-cta.ornl.gov/npts or sturm.math.fundp.ac.be/~test, are too complex and too expensive for small and medium sized planning agencies and universities. A more promising approach, based on common standards and interfaces, www.nesstar.org is discussed and presented.

An example of the central technology for such common approaches, XML document type definitions (DTD) is presented and discussed: the DTD of the Data Documentation Initiative (ddi).

Keywords

1. The status-quo

The discussions at the 1998 Eibsee conference (Stopher and Jones, 2000 and Axhausen, 2000) highlighted the need to improve the documentation, access to and archiving of transport data, in particular travel behaviour data. Discussions at other venues, for example the meetings of the International Association of Travel Behaviour Research or of the relevant Transportation Research Board committees, have come to the same conclusions.

The action list implicit in those discussions has the following points:

- Systematic archiving of travel behaviour data with the relevant national data archive, certainly for large local, regional and national studies
- Stipulation of data archiving as part of research grants and consultancy contracts
- Development of uniform data set descriptors (variables and attribute values/vocabularies) (Metadata)
- Implementation of on-line tabulation software on the web
- Development of specialised transport data archives

Unfortunately, during the last three years the response to these action points has been rather muted within the transport policy and research community. Progress has been made, but mostly where individuals were willing to act. The range of activities mentioned has not yet been embraced as part of obvious and normal professional practise.

This judgement is not based on a comprehensive survey, but on a set of random checks: websites of planning agencies and ministries, catalogues of data archives, availability of archiving standards. Many planning agencies have websites, but few offer as much detail about their data and travel surveys, as the Bay Area\(^1\) and the New York/New Jersey\(^2\) metropolitan planning organisations, but even those two do not offer direct access to their data, neither as raw data or through a on-line tabulation system. Searches at various data archives (US, Germany, UK, Switzerland) reveal a very uneven coverage of transport related survey data: with the exception of the UK only elderly versions of national travel surveys, virtually no regional sur-
veys of note, even in the UK. Better represented are leisure and holiday travel surveys and the results from household income and expenditure surveys, maybe because both are normally conducted by sociologists and economists, which are closer to those disciplines’ tradition of data archiving.

In the US, the impact of the Bureau of Transportation Statistics (BTS) has been enormous and the Department of Transport has made large efforts to make its data available. While www.bts.gov gives direct access to the mayor surveys of the last decade, this seems to have interrupted the transfer of these data sets to the ICPSR Archive at Ann Arbour. There is no cross-referencing between the sites, which makes them difficult to find, unless one has a prior awareness of their existence, something which should not be taken for granted in today’s world of highly specialised disciplines and networks of knowledge.

On-line tabulation seems an obvious solution to many problems of members of the public, planners and policy makers: fast and easy access to the data and well produced and valid tables or graphs for the users, central control over the data, quality control of circulating numbers and a reduction in the workload for the planners. Given that the software tools required have now been available for a number of years, the number of implemented sites is disappointing: http://www-cta.ornl.gov/npts/ for the 1990 and 1995 US NPTS (National Personal Transportation Survey) and the European Union research programme funded site at http://sturm.math.fundp.ac.be/~test, which is based on some small scale research data sets on long-distance travel (see below for more detail) (Reginster and Toint, Forthcoming)³.

The advent of the institutional website as a preferred channel of communication for professional information has encouraged the publication of a great deal of information about travel surveys on the web, at least has made the relevant reports easier to obtain. Unfortunately, the absence of well understood and adopted patterns of how to report such surveys has led to a rather uneven quality in those available reports. The raw data is generally not available, although no alternatives are provided other then some tabulations in the reports. On-line tabulation seems to be clearly beyond the resources of most individual planning agencies, as long as the software has to be massively customised for each survey.

³ The design and technology of this site have been adapted for the UK National Travel Survey and the Belgian National Travel Survey. Neither site has been made available to the public as of April 2001, but those responsible predict that this will happen in June for the UK site and in autumn for the Belgian site (private communication with Barbara Noble, DETR (London) and Eric Cornelis, FUNDP (Namur).
Standardisation of the reporting requirements, standardisation in the formatting of the data and standardisation in the software tools would be a possible way forward to meet the requirements of the professional and general public at reasonable costs in terms of money and staff time.

After brief discussion of the current state of the transport on-line tabulations sites, the paper will discuss a number of developments outside transport which might make this goal achievable: xml-based document type definitions (DTD) and general tabulation systems based on the ddi DTD. The paper will continue with a discussion of the ddi proposal and of possible extensions from a travel survey perspective. The conclusions will discuss the next steps and a possible organisational frame for them.

2. Current on-line tabulation of travel behaviour data sets

The two sites identified above are instructive about the reasons for the slow uptake of this approach. Each site offers basic tabulation services based on pre-defined data sets, which bundle subsets of variables from the total set available in the respective survey. Oak Ridge National Laboratory developed the US site for the FHWA. The European site was the product of 4th Framework research project “Technologies for European Surveys of Travel Behaviour” (TEST). While the FHWA site is based on commercial and proprietary software (SAS, web32, Perl), the TEST site is built around shareware (Apache server, Perl, Xlispstat). Both sites require registration, but offer the user some storage of results and of user-defined-formats for variables in return.

The sites guide the user through the interactions in a series of steps (See Figures 1 to 4). The selection of available statistics is different, as are the ease with which subsets of the data can be analysed. The NTPS sites alerts the user to the need of a careful checking of the results, while the TEST site removes values, which are based on too small subsamples to be reliable.

From a user perspective the sites deliver their promise: tables and some graphs without the hassle of downloading the data. Still, the user has to learn about the chosen file structure and might find that the tabulations implemented are too limited for his or her need. The spreadsheet download available on the 1995 NPTS site allows the reformattting of the results to one’s own preferences, including the creation of graphs to ones own standards. Depending on
the level of sophistication of the user, he or she will soon outgrow the limitations of such sites.

One area, which is clearly missing on these two sites, is the mapping of the results. While national representative surveys cannot be used to generate maps at any level of spatial resolution, some higher levels can be supported; let's say NUTS2 or NUTS3 in the European nomenclature. For examples of the current-state-of-practise see http://www-atlas.usgs.gov/ or http://plue.sedac.ciesin.org/plue/ddviewer/index.html, which give a good idea of what is possible.

Figure 1   TEST website: First page

Source: Reginster and Toint (Forthcoming), Figure 2
Figure 2 1995 NPTS website: First page

![1995 NPTS website](image)

Choose The Analysis Tool

- Exploratory analysis (get a feel for the data)
- Table Wizard (your best source for tables and measures of travel)
- Create a customized variable value grouping (e.g., Low Income)

**Warning**: a new toolset has been installed. Please report any bugs.

Back Next Help

Figure 3 TEST website: Selection of variables for a two-way table

![TEST website: Selection of variables](image)

Crosstables

This page allows you to cross some variables. (Help)

1. Select a filter:

   There is no selected filter

2. Please, choose one or more countries:

   - France
   - United Kingdom

3. The current table is:

   ![Table Example]

To choose another row variable, click on the first radio button beside the variable.
To choose another column variable, click on the second radio button beside the variable.
A (*) beside the variable means that the counting is done on pre-defined classes.
A (cf) beside the variable means that the counting is done on user’s created classes.

Source: Reginster and Toint (Forthcoming), Figure 3
From the data owner perspectives, the sites deliver, but raise at the same time questions about the long term viability of such dedicated sites: Does one want to recreate all capabilities available in standard statistics software? How much control over the formatting does one want to give to the users? Is this one-off approach feasible and financially viable?

The slow speed of adoption seems to indicate that for most data owners and planning agencies the answer to these questions is “no”. While they appreciate the possible benefits, they have not been able to develop a standardised approach, which would spread the costs widely enough to make it acceptable. Even the existing sites have not yet been used as platforms for further data sets, the Belgian and British example notwithstanding.

3. Metadata and data archives

The need to reinvent the wheel has slowed down the development generally, if the set of sites compiled by the library at the University of Virginia is a fair representation of the current state of the art (see http://fisher.lib.virginia.edu/active_data/index.html). The sites listed there
offer tabulation and some graphing, selection of data subsets and data downloading, but rarely more.

The bespoke systems, discussed so far, know the structure of their data sets and the meaning of each element by design. They are therefore unable to be adapted quickly to new datasets. This missing separation of the data and of its description is not unique to statistical data, but applies to any data. It is therefore no surprise, that information science and all professionals concerned with data are searching for solutions to this problem, in particular solutions, which are independent of any particular software vendor. They are placing their hopes in metadata, the data about data and the standardisation of the metadata items and structures.

Metadata (see for example Wigan, 2001 or Ambur, 1999 among many others), the data about the data, has four main aspects in survey research:

- The description of the file, in which the data is stored: the variables and their relationships, their names, their labels, the way they were generated, their coding and the values of the codes. The relationships describe how the variables relate to the different units of the survey, say household, person, journey, trip, stage, vehicle in our context
- The description of the survey and of its conduct: location, timing, sampling procedure, protocol of the contacts with the respondents, survey materials, data editing and imputation methods, weighting approach, problems and deviations from the protocol.
- Supplementary materials: data used for weighting, networks and other external data used in conjunction with the survey
- The description of responsibilities for the survey: ownership, design, conduct, editing and imputation, maintenance and archiving

A fifth relevant aspect is the metadata about the metadata, in particular authorship and bibliographical details.

It is clear, that this task can be approached in any number of ways and that standardisation is therefore the basis for further products, such as on-line tabulation, but even more importantly efficient retrieval of data and information about the data from the archives or the www. It is therefore no surprise, that librarians and data archivist have taken a leading role in the development of metadata standards and of meta-metadata standards.
An interesting result of such early activities is the joint catalogue of the data sets held in the European social science data archives (see http://www.nsd.uib.no/cessda/europe.html) for details.

The diversity of available technologies and the high speed of their innovation and change makes all standardisation work perilous, as a standard might be technology outdated by the time of its adoption. Unfortunately, technological choices cannot be avoided, as a metadata standard needs to be implemented as software to become relevant for the potential users.

Central for travel behaviour research are the attempts to develop a common grammar and partial vocabulary, with which to describe data sets, so that other software can act on them (search, index, display, tabulate and analyse). Other efforts, which are either directed toward physical data exchange, such as those based on the mostly European EDIFACT standards, or those concerned with meta-metadata, for example the rules of how to define grammars and vocabularies and where to store and register them, are of lesser interest here.

The preferred technology chosen for the purposes of defining grammars (structures) and vocabularies at this time is XML (Extended Markup Language), a subset of the more general, more powerful, but also more difficult to learn SGML (Standard Generalised Markup Language) (Ray, 2001). SGML is a standard, which has thirty years of history, and is correspondingly powerful and sophisticated. It started with developments at IBM, which was faced early on with substantial problems in maintaining its documents, specifications and systems coherent. In a marked up documents or data set, the tags of the language describe the content or specify some action to be taken with it; for example that it should printed in bold or a larger font. But more importantly, the tags can also be used to structure the document and to specify meanings. The general public had generally forgotten about the strength of such approaches until HTML (Hypertext Markup Language) started its explosive growth on the web in the early 1990’s. HTML uses the conventions of SGML, but radically restricts its functionality to the needs of rendering text for the web and to the needs of linking web-hosted documents and services through hyperlinks (i.e. URL’s – uniform resource locators). Given the success of the web it was no surprise, that old functionality was added to various HTML dialects over the years. The governing body of the web was able to standardise it again, as XHTML, before the functionality of the web was seriously threatened, but the divergence had shown the need for a more flexible instrument consistent with HTML files, but less complex and expensive then SGML. XML was developed to fill this spot by World Wide Web Consortium (W3C) since
1996. It is now available as an official recommendation in version 1.0. XML documents are SGML compatible. The official recommendations define XML documents as

“made up of storage units called entities, which contain either parsed or unparsed data. Parsed data is made up of characters, some of which form character data, and some of which form markup. Markup encodes a description of the document’s storage layout and logical structure. XML provides a mechanism to impose constraints on the storage layout and logical structure” (W3C, 2000, Section 1).

The development goals were (W3C, 2000, Section 1.1):

- “XML shall be straightforwardly usable over the Internet.
- XML shall support a wide variety of applications.
- XML shall be compatible with SGML.
- It shall be easy to write programs which process XML documents.
- The number of optional features in XML is to be kept to the absolute minimum, ideally zero.
- XML documents should be human-legible and reasonably clear.
- The XML design should be prepared quickly.
- The design of XML shall be formal and concise.
- XML documents shall be easy to create.
- Terseness in XML markup is of minimal importance.”

The standard has been adopted very quickly and a substantial range of products are available to develop XML “languages” and to work with XML compliant documents. In particular, a number of associated technologies have been developed to render XML documents through style sheets, to convert them into new formats and to program for them.

At the core of XML is the ability to define specific application oriented “grammars” and “vocabularies” through document type definitions (DTD). If a DTD is associated with a document, that document has to conform to it fully, otherwise a XML parser will not declare it “fully formed”, i.e. correct. This ability allows groups of persons or firms to define standards for their own purposes, which can then be used for subsidiary technologies, such as statistical

---

analysis or tabulation packages. The depth of standardisation chosen is open and can be determined by the groups or institutions themselves.

It is unclear at this time, if XML will be the success, that many people predict that it will be. Given the speed of development of computing technologies any forecast is dangerous. Still, XML seems to positioned be right, addresses an obvious need and is accessible to non-specialist user for their data definition needs. Even if it is superseded quickly, the current investment in it will assure transition paths to the next technology.

4. ddi, NESSTAR and FASTER

The Data Documentation Initiative (ddi)\textsuperscript{5} started as an effort of the US Inter-University Consortium for Political and Social Research (ICPSR) to replace the previous OSIRIS codebooks it had used to document the files in its data archive at the University of Michigan (Ann Arbor). The ddi – Consortium grew quickly to include other interested parties from the US and Europe, including other data archives. The product of the work is a DTD, which specifies, which metadata are to be provided with each data set in the archive’s collection. It builds on the large and long experience of the data archives in collecting metadata about their holdings. It also reflects the wish of the participants to open up those holdings for comparative research and in particular survey method-oriented research.

The participating European data archives have quickly adopted the ddi DTD to further their own work. In particular, they adopted the ddi DTD as the basis for the EU-funded project NESSTAR (Networked Social Science Tools and Resources) (NESSTAR Consortium, 2000). NESSTAR combine the functionalities of a specialised search engine with those of an on-line tabulation tool (see www.nesstar.org)

\textsuperscript{5} See www.icpsr.umich.edu/ddi for details.
The client software, NESSTAR explorer 1.01, allows the user to search simultaneously all datasets published by institutions and archives using the NESSTAR system. The selected datasets can then be checked using the metadata provided by the ddi DTD, which is consistently used to describe the data sets (See Figure 5 for an example search result). In addition, the user can tabulate and graph any variable (See and ). The client software is implemented as a Java-applet, but the necessary virtual machine is provided by NESSTAR for installation with the relevant browsers.

On the server side, NESSTAR provides the tools to publish data sets using a mixture of web-based technologies (See Figure 8).

Figure 5  NESSTAR example search results
The NESSTAR system is available free, both client and server software. It allows basic access control and can therefore be used to publish data without losing control over it. It has been extensively beta-tested by various agencies around Europe. As the production version has only been available since last year, it is difficult to make predictions about its likely success and speed of adoption, but it is clear, that the archives themselves see this approach as a successor to their current joint CESSDA search engine. The follow-on project FASTER (Musgrave and Ryssevik, 2000; Ryssevik, 2000) wants to extend the reach of the approach to include time series, geo referenced data and aggregate data, e.g. tables from official statistical publications, plus increase the functionality of the client software further. In a parallel project, a multi-lingual thesaurus is being developed, which will be integrated into the client software to support search and data retrieval.
The inclusion of official tables is particularly important, as this would open the possibility of publishing tabular results, both new and old. The European statistical offices, in particular the Dutch CBS, is working on a ddi DTD compatible way to describe tables (CRISTAL).
5. The ddi DTD

The ddi DTD reflects the experiences of the data archives involved in describing data sets for publication, which results in a strong emphasis on issues of ownership and publication detail. It is less strong on some of procedural issues of survey research, which tend to get a summary treatment in written abstracts and summaries, but not as a set of detailed variables.

Part 1 of the DTD provides the bibliographical detail of the ddi codebook and of the codebook, if any, on which it is based. It clarifies the responsibilities for the various steps in the production of the codebook(s) (See Figure 9) (For the fourth level elements see the appendix). Parts1 and 2 provide the information required for the Dublin Core, a widely accepted metadata standard set of items for the description of documents (See www.dublincore.org).
Figure 9  ddi codebook outline (Part 1) (Without forth level elements)

0.0 codeBook (ATT == ID, xml:lang, source, version)
  -- 1.0 docDescr* (ATT == ID, xml:lang, source)  Description of document/codebook
     -- 1.1 citation? (ATT == ID, xml:lang, source, MARCURI)  Citation for XML/ddi codebook
         -- 1.1.1 titlStmt (ATT == ID, xml:lang, source)  Title of XML/ddi codebook
         -- 1.1.2 respStmt? (ATT == ID, xml:lang, source)  Responsible
         -- 1.1.3 prodStmt? (ATT == ID, xml:lang, source)  Producer
         -- 1.1.4 distStmt? (ATT == ID, xml:lang, source)  Distributer
         -- 1.1.5 serStmt? (ATT == ID, xml:lang, source, URI)  Series
         -- 1.1.6 verStmt* (ATT == ID, xml:lang, source)  Version
         ++ 1.1.7 biblCit?  (ATT == ID, xml:lang, source, format)  Bibliographic citation
         ++ 1.1.8 holdings* (ATT == ID, xml:lang, source, location, callno, URI)  Location

  -- 1.2 guide?  (ATT == ID, xml:lang, source)  Definitions of terms used
  -- 1.3 docStatus? (ATT == ID, xml:lang, source)  Status as draft or final
  ++ 1.4 docSrc* (ATT == ID, xml:lang, source, MARCURI)  Description of original codebook

* = Element is optional & repeatable    + = Element is mandatory & repeatable
? = Element is optional & non-repeatable

Source: http://www.icpsr.umich.edu/DDI/codebook.html

Part 2 describes the study in detail (Figure 10), in particular the data collection methods. At this time, it specifies the topics, which need to be covered, but it is lacking a fixed vocabulary and fixed set of numerical variables to describe a study. The author of the codebook is responsible here for covering all relevant issues. At the moment the following topics are requested:

- Time method, e.g. panel, cross section etc (No standard classification)
- Data collecting agency: name(s), but no standard classification
- Frequency of data collection, e.g. monthly, quarterly (No standard classification)
- Sampling procedure and major deviations from it (Free text, no standard classification)
- Mode of data collection, e.g. face-to-face, CATI (No standard classification)
- Type of instrument, e.g. structured (Proposed classification)
- Weighting (Free text, no requirement to store the external data used for weighting)
- Sampling error estimates (Free text)
- Response rate (Free text, no suggested standard for response rate calculation or specification of the required variables)
The subsection on availability covers the relevant issues comprehensively, as one could expect. Still, the individual archivist/user will need to interpret the information before he or she can take actions. Again, standard classifications are still missing, which would allow automated processing of permissions and access.

Part 3 provides the elements for the description of the data files and their internal structures (See Figure 11). It is possible to describe hierarchical structures of nested record groups with differing number of variables. Part 4 describes each individual variable (See Figure 12). Variables, which form a group because they derive from a common question frame, can be defined. The necessary elements (indices) to cross-link the variables to the files and to the questions of the original questionnaires are available. The original question text has to be included.

Source: http://www.icpsr.umich.edu/DDI/codebook.html
A final fifth section allows the inclusion of further related materials.

Figure 11  ddi codebook outline (Part 3) (No fourth level elements)

```
0.0 codeBook (ATT == ID, xml:lang, source, version)
 -- 3.0 fileDscr* (ATT == ID, xml:lang, source, URI, sdatrefs, methrefs, pubrefs, access)
  -- 3.1 fileTxt? (ATT == ID, xml:lang, source) Data file description
    -- 3.1.1 fileName? (ATT ==ID, xml:lang, source) Name
    -- 3.1.2 fileCont? (ATT == ID, xml:lang, source) Content description (free text)
    -- 3.1.3 fileStrc? (ATT == ID, xml:lang, source, type) File structure (flat, hierarchical)
    -- 3.1.4 dimensns? (ATT == ID, xml:lang, source) Dimensions of flat files
    -- 3.1.5 fileType? (ATT == ID, xml:lang, source, charset)Encoding standard (ASCII etc.)
    -- 3.1.6 format? (ATT == ID, xml:lang, source) Physical format
    -- 3.1.7 filePlac? (ATT == ID, xml:lang, source) Agency and location of file production
    -- 3.1.8 dataChck* (ATT == ID, xml:lang, source) Data checks (free texts)
    -- 3.1.9 ProcStat? (ATT == ID, xml:lang, source) Draft or final status
    -- 3.1.10 dataMsng? (ATT == ID, xml:lang, source) Missing data information (free text)
    -- 3.1.11 software* (ATT == ID, xml:lang, source, date, version) Software used for file
    --- 3.1.12 verStmt? (ATT == ID, xml:lang, source) Version of file
    +-- 3.2 notes* (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
```

* = Element is optional & repeatable  + = Element is mandatory & repeatable
? = Element is optional & non-repeatable

Source: http://www.icpsr.umich.edu/DDI/codebook.html

Each of the named elements above defines a tag for the XML file. XML enforces that the hierarchy of the tags is maintained and checks that all required tags are available. The attributes can be used to specify details about the elements, which are easily standardised. Figure 13 gives an example by defining a variable (See also Part 4 of the DTD below).

NESSTAR and other, commercial, providers have developed editors, which use a user specified DTD to guide the data entry and to enforce the completeness of the generated file.
Figure 12  ddi codebook outline (Part 4)

0.0 codeBook (ATT == ID, xml:lang, source, version)
  -- 4.0 dataDscr* (ATT == ID, xml:lang, source)
    -- 4.1 varGrp* (ATT == ID, xml:lang, source, type, var, varGrp, name, sdatrefs, methrefs, pubrefs, access)
      -- 4.1.1 labl*     (ATT == ID, xml:lang, source, level, vendor)
      -- 4.1.2 txt*      (ATT == ID, xml:lang, source, level)
      -- 4.1.3 defntn?   (ATT == ID, xml:lang, source)
      -- 4.1.4 universe? (ATT == ID, xml:lang, source, level, clusion)
      ++ 4.1.5 notes*    (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
    -- 4.2 var* (ATT == ID, xml:lang, source, name, wgt, wgt-var, qstn, files, vendor, dcml, intrvl, rectype, sdatrefs, methrefs, pubrefs, access)
      -- 4.2.1 location* (ATT == ID, xml:lang, source, StartPos, EndPos, width, RecSegNo, fileid)
      -- 4.2.2 labl*     (ATT == ID, xml:lang, source, level, vendor)
      -- 4.2.3 imputation? (ATT == ID, xml:lang, source)
      -- 4.2.4 security?  (ATT == ID, xml:lang, source, date)
      -- 4.2.5 embargo?   (ATT == ID, xml:lang, source, date, event, format)
      -- 4.2.6 respUnit?  (ATT == ID, xml:lang, source)
      -- 4.2.7 anlysUnit? (ATT == ID, xml:lang, source)
      -- 4.2.8 qstn*      (ATT == ID, xml:lang, source, qstn, var, seqNo, sdatrefs)
      -- 4.2.9 valrng*     (ATT == ID, xml:lang, source)
      -- 4.2.10 invalrng*  (ATT == ID, xml:lang, source)
      -- 4.2.11 undocCod* (ATT == ID, xml:lang, source)
      -- 4.2.12 universe* (ATT == ID, xml:lang, source, level, clusion)
      -- 4.2.13 TotlResp? (ATT == ID, xml:lang, source)
      -- 4.2.14 sumStat*  (ATT == ID, xml:lang, source, wgtd, weight, type)
      -- 4.2.15 txt*      (ATT == ID, xml:lang, source, level)
      -- 4.2.16 stdCatgry* (ATT == ID, xml:lang, source, URI)
      -- 4.2.17 catgryGrp* (ATT == ID, xml:lang, source, missing, missType, catgry, catGrp)
      -- 4.2.18 catgry*    (ATT == ID, xml:lang, source, missing, missType, country, sdatrefs)
      -- 4.2.19 codInstr* (ATT == ID, xml:lang, source)
      -- 4.2.20 verStmt*  (ATT == ID, xml:lang, source)
      -- 4.2.21 concept* (ATT == ID, xml:lang, source, vocab, vocabURI)
      -- 4.2.22 derivation? (ATT == ID, xml:lang, source, var)
      -- 4.2.23 varFormat? (ATT == ID, xml:lang, source, type, formatname, schema, category, URI)
      ++ 4.2.24 notes*    (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
    ++ 4.3 notes*      (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
  ++ 5.0 OthMat

* = Element is optional & repeatable    + = Element is mandatory & repeatable
? = Element is optional & non-repeatable

Source: http://www.icpsr.umich.edu/DDI/codebook.html
As mentioned above, this DTD is not yet specific enough to ensure a complete description of the survey process (See for example the checklist in Richardson, Ampt and Meyburg, 1995). The mostly free form elements can be used in this way, but they cannot enforce it. Topics, which would need to be covered, to reflect the current discussion in travel survey research, are:

- Protocol of the survey
  - Number of contacts (Minimum, maximum)
  - Contact (id, name, type and form, materials used, responsible unit, permissible time window, accepted response, rules for the selection of next contact, status as required or optional contact)
- Contact history of the household/person
  - Number of contacts
  - Contact (date, time, type, id, person making the contact, contact success, answer, person answering the contact)
  - Status (final status of household/person)
Equally, the DTD does not give much attention to the physical forms used, generally assuming the use of CATI or similar technologies. It would be helpful, if there were clear requests to document the forms, as part of a set of elements, which describe the printed forms (size, format, type of postage, type of envelope etc.)

6. Summary and outlook

The recent developments in computing technology and in the information sciences, in particular the development of languages or systems, such as XML, which make the development of common professional metadata standards feasible at reasonable cost and in reasonable time. The example combination of the ddi DTD and NESSTAR demonstrates the power of this approach. The travel survey community could use these developments as their starting point for their contributions to this growing set of standards. These contributions should reflect their particular concerns, especially in the analysis of survey non-response. The XML approach would allow the development of further DTD’s based on the parent ddi DTD, which address the needs of specific types of travel survey data, such as stated response and stated preference data, travel diaries or road side interviews.

This author hopes that the series of conference on “Transport Survey Quality and Innovation” in conjunction with the conferences of the International Association of Travel Behaviour Research, will become a forum for this work of definition and standardisation.

7. Acknowledgements

The author is grateful for comments of the referees and of the information provided by the colleagues at ICPSR and the Data Archive of the University of Essex.
8. Literature


Reginster, I. and Ph. Toint (Forthcoming), in K.W. Axhausen, J.L. Madre, J.W. Polak and Ph. Toint (Eds.) Capturing Long Distance Travel, Research Science Press, Baldock


---

6 Available at www.faster-data.org

7 Available at www.nesstar.org