Public use of travel surveys: 
The metadata perspective

Resource paper presented at the 2nd International Conference on Transport Survey Quality and Information, Kruger National Park, August 2001

KW Axhausen and MR Wigan

Arbeitsbericht
Verkehrs- und Raumplanung

Titel: (IVTd.eps) 
Erstellt von: Adobe Illustrator(TM) 6.0

Titel: (ETH.eps) 
Erstellt von: Adobe Illustrator(TM) 6.0

May 2002
Öffentliche Nutzung von Befragungen zum Verkehrsverhalten: Die Metadaten-Perspektive

KW Axhausen
IVT
ETH
CH-8093 Zürich

MR Wigan
Oxford Systematics
GPO 126
AUS – Heidelberg, Vic 3084

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

Mai 2002

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

Verkehrsbefragungen; Archiv; Zweitnutzung; Öffentlicher Zugriff; Metadata; SGML; Datenbanken

Bevorzugter Zitierstil:

The public use of travel surveys: The metadata perspective

KW Axhausen  MR Wigan
IVT  Oxford Systematics
ETH  GPO 126
CH-8093 Zürich  AUS – Heidelberg, Vic 3084

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

May 2002

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.

Recent progress has provided new tools for data specification and data access. The crucial technology is the XML specification language, including XML document type definitions (DTD). These are presented and discussed using the DTD of the Data Documentation Initiative (ddi) as an example. The European social science data archive have demonstrated the power of this approach with their NESSTAR system (see www.nesstar.org).

The key benefit is that automated access to reliably specified and structured metadata becomes possible thereby simultaneously reducing the costs and improving the quality of data access and use. To take advantage of these developments the transport research community needs to develop specialised versions of the basic DTD and devote significant effort in developing suitable metadata (text, structures and vocabularies).

Keywords

Metadata; XML; Travel surveys; Data archive; Secondary use; SGML; database

Preferred citation style

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 and of the relevant Transportation Research Board committees, have come to the same conclusion.

The action list implicit in those discussions had and 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.

At the core of the problems causing the delay is the missing separation of the data and of their description, which is not a problem unique to travel and transport data, but applies to any data – including text. 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, 2001a b ,c or Ambur, 1999 among many others), the data about the data, has five main aspects in survey research:

- The description of the file(s), in which the data are 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
- The qualification of the quality and applicability of the data and the processes by which they were generated; the equivalent of the peer review of academic publication may be needed for this.

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

This set of items is not unique to transport or travel research, but applies to all survey researchers. A description covering these items permits the researchers (originators) of the data to document their work fully and allows data users access to the data and insight into its quality. It preserves the data for later secondary analysis, which is becoming increasing important in transport research, as we starting to address the issues of long-term behavioural change using surveys from 70’s onwards. As an aside, most of the earlier surveys have been lost due to insufficient or missing archiving.

In the US, the impact of the Bureau of Transportation Statistics (BTS) has been enormous and the Department of Transportation has made large efforts to make its data available. While www.itdb.bts.gov gives direct access to the major surveys of the last decade, this seems to have interrupted the transfer of these data sets to the ICPSR Archive at Ann Arbor. 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. In addition, many US planning agencies have websites, but few offer as much detail about their data and travel surveys, as the Bay Area and the New York/New Jersey metropolitan planning organisations,

1 See especially the mapping tool provided for a number of data sets
2 www.mtc.ca.gov/datamart
but even those two do not offer direct access to their data, either as raw data or through an on-line tabulation system.

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 remaining 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)\(^4\).

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 and 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 are generally not available, although no alternatives are provided other than 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, in particular to handle issues such the calculation of confidence intervals and complex weighting schemes.

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. While standardisation might lead to the loss of some information specific to some surveys, it would support the documentation of the vast number of currently undocumented surveys through the provision of standard tools and check lists. The individual survey e-
searcher could concentrate on the main tasks and would not need to laboriously reinvent the wheel each 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.

A practical advantage of developing such formal descriptors is that they enable apparently disparate data to be brought together automatically from many different sources with verifiably identical items. This not only increases the scope and effectiveness of data integration for specific purposes, but also reduces the costs and enables at least some forms of analysis to be carried on such dynamically collated data.

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 earlier 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 is the ease with which subsets of the data can be analysed. The NPTS site alerts the user to the need of a careful checking of the results, while the TEST site removes values that are based on subsamples too small to be reliable.
From a user perspective both sites deliver their promise: tables and some graphs without the hassle of downloading the data. The user still 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 reformatting of the results to one’s own preferences, including the creation of graphs to one’s 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; for example 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 by using web-enabled GIS tools.

Figure 1 TEST website: First page

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

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

FHWA  CTA  ORNL

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

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](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.

However there are examples where the continuity of data collection systems with a continuing remit are in place, examples include VATS (Victoria Activity and Travel Survey) in Melbourne (Transport Research Centre, 1999) and the SHTS (Sydney Household Travel Survey) (Battelino and Peachman, 2001) and the systems being developed for LATS (London Area Travel Survey) (Neffendorf, Ashley, Bates, Jones, Collop, Capell, McCaig, Gunn, Fearon, Ramsey and Wigan, 2000). Each of these have developed greater or lesser levels of data base archiving, and have a cumulative commitment to each of these areas.

There are some real issues raised by simply attempting to describe such systems. The SHTS data base is held in a relational database, but which is not fully normalised, and thus requires care in interpretation as the results must be obtained from well-formed SQL (Structured Query Language) with a knowledge of the data base structure. This is a good example where direct access by users is probably inappropriate, and there is a need to re-examine the design principles used for such systems. It is significant that the managers of both VATS and the SHTS have not as yet addressed metadata, although both have spent a great deal of effort to ensure that the data itself is internally consistent within their internal definitions and formats.

The SHTS has a formal database design, but this is not sufficient to allow end users direct access to the data base, irrespective of the commercial or privacy issues involved. The VATS data is held in a simpler format and is usually delivered as SPSS files to clients.

Notably both continuing committed programs do not yet have formal metadata standards or access systems. One can therefore see some interesting constraints on current operational thinking and practise: the design of data holding systems maintains past best practice.

Past and current best practise in the social sciences has matched data holding and analysis systems. Examples are OSIRIS data holding system with SAS, or the SIR hierarchical database system designed to work with SPSS. Both these and other systems are oriented towards hierarchical structures that tend to suit transport data. There are still substantial questions about data access possibilities when these are restricted to hierarchical data structures, just as relational databases can have some problems in dealing with the hierarchical structures used in travel behaviour records.
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.

It is clear that this task can be approached in any number of ways and that standardisation is therefore the basis for further analysis tools, 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 archivists 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 any particular implementation of a standard might be outdated by the time of its adoption. Unfortunately, technological choices cannot be avoided, as a metadata standard needs to be implemented formally to become relevant for the potential users. There is therefore a clear need for a technology-neutral specification of any such standard.

Central to 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 data message exchange, such as those based on the EDIFACT standards, or those concerned with meta-metadata are of lesser interest here (for example the rules of how to define grammars and vocabularies and where to store and register them).

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; Goldfarb, 1990). 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 coherently. In a marked up document 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 technical 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 additional 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.

The confusion between data and formatting inherent in HTML had led to these major differences in the extensions. But this divergence had shown the need for a further instrument, consistent with HTML, but distinguishing between data and structure; preferably less complex and expensive then SGML. XML was developed to meet this requirement by the 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 used as the foundation for a substantial range of further layers, standards and tools. Examples are tools to develop XML “languages” and to create, transform and extend 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) (See Appendix A for more detail). If a DTD is associated with a document, then that document has to conform to it fully (verification); otherwise, an XML parser will not declare it “fully formed”, i.e. correct. This ability ensure the completeness of a description, allowing in turn automatic processing of such files.

The ability to define extended grammars and vocabularies makes it straightforward for groups of persons or firms to define standards for their own purposes. These can then be used to communicate to different tools, such as those for statistical analysis and tabulation. The detail of standardisation chosen is open and can be determined by the groups or institutions themselves.

XML has already been widely adopted as a data specification tool. It is unclear at this time if the many standards and draft standards based on XML will be the successes that many people predict them to be. Given the speed of development of computing technologies any forecast is dangerous. Still, XML seems to be positioned correctly, addresses an obvious need and is accessible to non-specialist users for their data definition needs. Even if many of the subsequent standards are not adopted, the current investment in data definitions in XML will assure transition paths to the next technology.
4. ddi, NESSTAR and FASTER

The Data Documentation Initiative (ddi)\(^6\) 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 (See Appendix B for the content of the ddi DTD). 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. Some US agencies are also considering the use of it or are considering the interoperability of their own systems with it. The European archives adopted the ddi DTD as the basis for the EU-funded project NESSTAR (Networked Social Science Tools and Resources) (NESSTAR Consortium, 2000). NESSTAR combines the functionalities of a specialised search engine with those of an on-line tabulation tool (see www.nesstar.org).

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 Figure 6 and Figure 7). 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).

---

\(^6\) See [www.icpsr.umich.edu/ddi](http://www.icpsr.umich.edu/ddi) for details.
The NESSTAR system is available free, both for 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 intensively beta-tested by various agencies around Europe. As the production version has only been available since mid 2000, 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, mentioned above. The follow-on project FASTER (Musgrave and Ryssevik, 2000; Ryssevik, 2000) aims to extend the reach of the approach by including time series, geo-referenced data and aggregate data, e.g. tables from official statistical publications. Also, the functionality of the client software is to be extended 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).

Figure 5 NESSTAR example search results
Figure 6  NESSTAR example tabulation results

Figure 7  NESSTAR example chart
5. 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, that 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 intercept surveys.

As discussed in the Appendix B, the ddi 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). These are mostly described in free form elements, but there is no formal validation through an XML editor or browser. Topics required to reflect current discussions in travel survey research include:

- Protocol of the survey\(^7\)
  - 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 ddi 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 to be clear requests to document the forms as part of a set of elements that describe the printed forms (size, format, type of postage, type of envelope, etc.).

The advent of XML provides transport researchers with a great incentive to develop clear and unambiguous specifications for their data, and to include sound and informative metadata in a formal framework. The incentive is that if this is done, then not only will the survey data become better documented, communicable and stable for the future, but also the substantial efforts across the world in developing XML-base standards and tools will become accessible.

Improving survey data specifications is only part of the story, although these will require professional effort, as the precision required will also demand that specifications of individual items (‘trip’ for example) become clearly specified in all their variants. The ‘eXtendable’ aspect of XML is an important asset here. Current specifications in public transport tend to be

---

\(^7\) The method repositories for the new ebXML – based standards for electronic business, which allow the exchange of business methods might be applicable to the exchange of survey protocols. The ebXML standards provide the tools to specify the messages involved in a series of interactions between different parties. This is an area for future work. See also www.ebxml.org.
drawn from GIS sources (e.g. FTA, 2000) and a more widely-based framework is clearly neces-

International collaboration and cooperation will therefore be highly desirable when specifying and refining suitable vocabularies and document definitions. Existing bodies such as IATBR, ASCE, ITE, IHT, IATUR, AET and TRB (particularly the TRB Subcommittee on Urban Metadata A1B08-2) could provide appropriate mechanisms for this to occur.

The need to agree exactly what various items mean in a formal manner (an XML vocabulary) is essentially a task that has been initiated many times. The key difference on this occasion is that this effort can realistically be expected to open the way for easier access, and more reliable usage across both the national and international stages – and the ability to automate much of the work now requiring significant manpower and effort for each study or studies where results may need to be compared or combined. This is coordination effort that needs to be doen progressively and in cooperation and is not technically demanding.

DTDs are more complex and do demand some technical skills. But, once created, they then allow documents to be formally verified and permit reliable and rapid computer access and integration. The combination of the agreements on standardised vocabularies in each area of transport and a set of DTDs is as far as survey workers and data collectors need be concerned, as the later usage of these newly validated formats is then enabled.

While the pioneering work already done on survey DTDs by the US ICSPR and the European social science data archives offers an excellent starting point from which to build, it is essential to involve the transport community in developing high quality metadata for their own existing and future datasets. No number of technical tools can replace the process of developing such material.

The GIS community has addressed at least one aspect (spatial accuracy) of data quality in spatial data sets, but transport data offers many more grounds – and needs - for commentary on ‘quality’. This will probably require the development of a peer review mechanism to de-fuse the sensitivities raised by this process, but the effort will be well worthwhile.

Taking metadata and formal (XML –based) formats seriously will allow a number of benefits to be realised. The most important of these are:

- Distributed queries over multiple databases
• Higher quality and less ambiguous results from such queries
• Automated tools to locate, extract, secure, format and protect such data
• Better control of privacy, and commercial aspects of survey data in more efficient searches

The two stages, metadata development and formal (XML) data specifications, are not dependent on any particular technology. Extensions in terms of transport-specific vocabularies will allow specialised transport survey tools to be developed, and for users to switch from one to another quite easily, without losing access to many of the database, analysis and presentation tools that are emerging based on XML document formats.

The next steps are to build on existing initiatives (such as the ddi discussed here) through the various transport research and practice networks and forums, build up transport specific vocabularies, and to clarify what tools are likely to be of greatest value to the transport survey community.

As distributed transport survey database holding and searches will then be possible, the privacy and commercial issues involved will have to be addressed. There are already numerous general SGML\(^8\) and XML-based tools and databases which can handle security and transaction control tied right down to XML fragments, so the task of specification is the key. The tasks of implementation and delivery will therefore be far less burdensome for the transport sector, as they are already being addressed by a broader IT community with many common interests in developing that infrastructure.

The discussion so far has assumed that the data owners in the transport domain want their data properly archived to protect it, to simplify its use, to make it accessible to the general and professional public. The discussion has also assumed that the transport community would be best advised to join the social sciences in their drive for standardised tools for data archiving, at this time XML and standards based thereon. Clearly, the professional archiving of a data set is not cost free, but in this case one can argue, that a “data archiving” step in the survey process only collects a wide variety of activities, which are undertaken anyway, into one cost heading. The costs are being incurred currently anyway hidden among other cost headings. They are now becoming visible. Tools such as the ddi DTD, Nesstar and subsequent associated tech-

\(^8\) Such as www.mds.rmit.edu.au
nologies simplify the tasks and replace traditional reporting requirements. In balance, the authors would expect professional archiving to be cost-neutral, especially when one considers later data retrieval and reconstruction costs. This will be especially true, when the organisations have learned the tools and have adopted the technologies. Clearly, there will be learning costs, but they will be repaid in the interest accruing from the protected investment in the organisations data and knowledge. The entry point requires no more than broad agreements on vocabularies on which to start to create validatable formal XML documents. This first basic step is simply a necessary 21st century advance on collections of definitions created by librarians (thesauri). Taking simply this step will allow later workers to secure all the potential benefits over time.

6. Acknowledgements

The authors are 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.

7. Literature


NESSTAR consortium (2000) NESSTAR: Networked Social Science Tools and Resources, final report to the European Commission, EPSRC Data Archive, University of Essex, Colchester\(^10\).


Transport Research Centre (1999) The VATS process: 1998 release2, TRC, RMIT University, Melbourne\(^11\).


---

\(^9\) Available at [www.faster-data.org](http://www.faster-data.org)

\(^10\) Available at [www.nesstar.org](http://www.nesstar.org).

Appendix A: XML documents and Document Type Definition (DTD)

A XML document is made up of elements and attributes - neither of which are defined in the XML standard. A document comprises both data (content) and markup (the commands that define its structure). This can be viewed as a form of HTML with the data between the tags removed and undefined. The W3C resource description framework (RDF) specification\(^\text{12}\) is easily interpreted by XML, as RDF is a\(^\text{12}\) vocabulary in XML and this vocabulary defines all the entities and elements required for RDF.

This process gives a window into the vocabulary functions required of the transport professions (be they travel behaviour, safety, traffic, transport, environmental, demography or physical planning). A vocabulary is not just terms, but also rules for parsing statements in those terms, and requires to be consistent with the other structures needed for an XML document to be a valid and verifiable document.

There are two key structures that are required to progress towards metadata specification, both with the initials DTD.

- Document Type Definitions (DTD)

These contain rules on elements and rules on attributes and a DTD (Document Type Declaration)

- Document Type Declarations (DTDs)

A short example may contain statements such as:

```xml
<!ELEMENT company (name, address, phone*)>
<!ELEMENT address (road, city, state)>
...
<!ELEMENT name (#PCDATA)>
...```

\(^{12}\) See: http://www.w3.org/RDF/
<!ATTLIST person key ID #REQUIRED>
<!ATTLIST phone location CDATA #IMPLIED>

are the starting point, and make it straightforward to define XML documents and to allow their automatic validation. Any XML document needs both an XML declaration and a document type declaration as in the Prologue and Document instance to follow. This is a very clearly defined form for a document.

An XML document using the previous elements and attributes would be as follows:

```xml
<?xml version="1.0 ">
<!DOCTYPE company SYSTEM "company.dtd ">
<company key="cocac ">
  <name>Coca Cola</name>
  ...
</person>.
```

Here the structure and the content are brought together.

The standards for XML fall in two categories. Vocabularies (which are simply applications of XML for a specific application field) and data modelling. The really interesting stage introduces the standards for XML enhancements and extensions – which in turn can be XML vocabularies themselves. One of these is XML Schemas, which are sets of rules to constrain the structure and to articulate the information datasets that comprise XML documents.
Appendix B: The ddi DTD

This section presents and discusses a DTD developed for the documentation social science data. This particular ddi (data documentation initiative) DTD reflects the experiences of the staff at 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. There are other aspects, in particular those of spatial accuracy and data quality, which would need further work for transport application.

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 www.icpsr.umich.edu/ddi/codebook.htm. Parts 1 and 2 provide the information required for the Dublin Core, a widely accepted metadata standard developed by librarians primarily for the description of text documents (See www.dublincore.org).

Figure 9  ddi codebook outline (Part 1) (Without fourth level elements)

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Authors’ comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0 codeBook (ATT == ID, xml:lang, source, version)</td>
<td></td>
</tr>
<tr>
<td>-- 1.0 docDescr* (ATT == ID, xml:lang, source)</td>
<td>Description of document/codebook</td>
</tr>
<tr>
<td>-- 1.1 citation? (ATT == ID, xml:lang, source, MARCURI)</td>
<td>Citation for XML/ddi codebook</td>
</tr>
<tr>
<td>-- 1.1.1 titlStmt (ATT == ID, xml:lang, source)</td>
<td>Title of XML/ddi codebook</td>
</tr>
<tr>
<td>-- 1.1.2 rspStmt? (ATT == ID, xml:lang, source)</td>
<td>Responsible</td>
</tr>
<tr>
<td>-- 1.1.3 prodStmt? (ATT == ID, xml:lang, source)</td>
<td>Producer</td>
</tr>
<tr>
<td>-- 1.1.4 distStmt? (ATT == ID, xml:lang, source)</td>
<td>Distributer</td>
</tr>
<tr>
<td>-- 1.1.5 serStmt? (ATT == ID, xml:lang, source, URI)</td>
<td>Series</td>
</tr>
<tr>
<td>-- 1.1.6 versStmt* (ATT == ID, xml:lang, source)</td>
<td>Version</td>
</tr>
<tr>
<td>-- 1.1.7 bibliCpt? (ATT == ID, xml:lang, source, format)</td>
<td>Bibliographic citation</td>
</tr>
<tr>
<td>++ 1.1.8 holdings* (ATT == ID, xml:lang, source, location, callno, URI)</td>
<td>Location</td>
</tr>
<tr>
<td>-- 1.2 guide? (ATT == ID, xml:lang, source)</td>
<td>Definitions of terms used</td>
</tr>
<tr>
<td>-- 1.3 docStatus? (ATT == ID, xml:lang, source)</td>
<td>Status as draft or final</td>
</tr>
<tr>
<td>++ 1.4 docSrc* (ATT == ID, xml:lang, source, MARCURI)</td>
<td>Description of original codebook</td>
</tr>
<tr>
<td></td>
<td>Elements as in 1.1.</td>
</tr>
</tbody>
</table>

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

Source: http://www.icpsr.umich.edu/DDI/codebook.html
Part 2 specifies the items to describe in detail the survey to be documented (Figure 10), in particular the data collection methods. At this point it specifies the topics that need to be covered, but lacks both a fixed vocabulary and a fixed set of numerical variables to describe a study adequately.

Figure 10  ddi codebook outline (Part 2) (Without fourth level elements)

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Authors’ comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0 codeBook (ATT == ID, xml:lang, source, version)</td>
<td></td>
</tr>
<tr>
<td>-- 2.0 stdyDescr+ (ATT == ID, xml:lang, source, access)</td>
<td>Study description</td>
</tr>
<tr>
<td>-- 2.1 citation+ (ATT == ID, xml:lang, source, MARCURI)</td>
<td>Citation for the data collection</td>
</tr>
<tr>
<td></td>
<td>+-- Elements as in 1.1.</td>
</tr>
<tr>
<td>-- 2.2 stdyInfo* (ATT == ID, xml:lang, source)</td>
<td>Description of the survey</td>
</tr>
<tr>
<td></td>
<td>-- 2.2.1 subject? (ATT == ID, xml:lang, source)</td>
</tr>
<tr>
<td></td>
<td>-- 2.2.2 abstract* (ATT == ID, xml:lang, source, date)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>-- 2.3 method* (ATT == ID, xml:lang, source)</td>
<td>Method</td>
</tr>
<tr>
<td></td>
<td>-- 2.3.1 dataColl* (ATT == ID, xml:lang, source)</td>
</tr>
<tr>
<td></td>
<td>-- 2.3.2 analyInfo? (ATT == ID, xml:lang, source)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>-- 2.4 dataAccs* (ATT == ID, xml:lang, source)</td>
<td>Access rules and status</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

The author of a specific survey codebook is fully responsible for covering all relevant issues within the specification set out. At the moment information about 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)
• Response rate (Free text, no suggested standard for response rate calculation or specification of the required variables)
• Sampling error estimates (Free text)

The subsection on availability comprehensively covers the relevant issues. Still, the individual archivist/user would need to interpret the information before he or she could give access to the data. However, a computer programme could interpret the specification automatically under a yet to be specified set of rules. This is one of the benefits of this general approach.

Part 3 specifies 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 numbers of variables. Part 4 describes each individual variable (See Figure 12). Variables that 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.
### Figure 11  ddi codebook outline (Part 3) (No fourth level elements)

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Authors’ comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0 codeBook (ATT == ID, xml:lang, source, version)</td>
<td></td>
</tr>
<tr>
<td>-- 3.0 fileDescr* (ATT == ID, xml:lang, source, URI, sdatrefs, methrefs, pubrefs, access)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-- 3.1 fileTxt? (ATT == ID, xml:lang, source) Data file description</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>--- 3.2 notes* (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)</td>
</tr>
</tbody>
</table>

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

Source: [http://www.icpsr.umich.edu/DDI/codebook.html](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 research and 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 defn? (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, dcm1, 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
Figure 13  ddi DTD example of the coding of a variable

<fileName ID='Household'>Household file</fileName>

<var files='Household' name='hhinc' qstn='A-54' >
  <labl>Household Income</labl>
  <location StartPos='55' EndPos='57' width='3'></location>
  <imputation>Hotdesk imputation using size, hours worked, car owned</imputation>
  <respUnit>Head of household</respUnit>
  <anlysUnit>Household</anlysUnit>
  <qstn>What is your total household income [kUS$] after taxes and social security</qstn>
  <valrng>
    <range Units='INT' maxExclusive='250' min='5' max='240'></range>
    <key>250 250k and more</key>
  </valrng>
  <invalrng>
    <range Units='INT' minExclusive='0' min='998' max='999'></range>
    <key>0 Refused 998 Dont know 999 Not applicable</key>
  </invalrng>
  <TotlResp>450 valid responses</TotlResp>
  <sumStat type='Min'>5</sumStat>
  <sumStat type='Max'>220</sumStat>
  <sumStat type='Min'>65</sumStat>
</var>