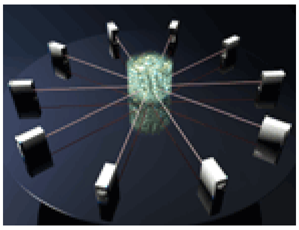
JISC

JISC Final Report

Data Exchange Tools and Utilities (DExT)
Repositories and Preservation Tools Programme



Louise Corti
March 2008

Contact Details for Primary Contact:

Louise Corti
UK Data Archive
University of Essex
Colchester CO4 3SQ
corti@essex.ac.uk
01206 872145

Table of Contents

Acknowledgements

The Data Exchange Tools and Conversion Utilities (DExT) project was funded under the JISC Capital Programme: Repositories and Preservation. The UK Data Archive (UKDA) based at the University of Essex led the project with major input from key consultants – Herve L'hours and Pascal Heus and Arofan Gregory and Chris Nelson of the Open Data Foundation (ODaF). Pascal Heus authored the text for WP2 that is incorporated into this Final report and also appended. Acknowledgments go to members of the UKDA team - Head of Preservation, Dr. Matthew Woollard and DExT Programmer, Angad Bhat for technical support. Thanks go to Claire Flaxton, Anne Etheridge and Sharon Jack or their work in promotion and publicity and the website. Our great thanks also go to the late Dr. Alasdair Crockett who helped formulate and write the bid, particularly for WP2.

Executive Summary

Data conversion and proprietary data entry and analysis are particularly important and problematic aspects of data management and curation. The Data Exchange Tools and Conversion Utilities (DExT) project aimed to provide researchers and support staff working with primary research data with a suite of tools that will enable data to be long-term curated and exchangeable.

Much important primary research data is created every day in the course of academic and policy research. While Data Sharing Policies are encouraging sharing and formalised archiving of data, the ideal life cycle for data creation to re-use remains beset by obstacles. The main issues involve the buying-in to a dedicated analytic strategy and typically a particular software package. Over the years the UKDA has seen a number of such softwares quickly become obsolete. To address the problem of incompatibility between software various data conversion tools have come of the market. One example for numeric and statistical data is StatTransfer and DBMSCopy that enable conversion from say SPSS to Stata. Equally the development of the format SPSS.por in the 80s enabled import and export between the major statistical analytic packages. However in the qualitative data analysis software field there are no such inter-software conversion tools.

The rationale for the DExT work is that open data exchange formats are necessary for maximising the opportunities for data sharing and long-term archiving.

The project led by the UK Data Archive researched, developed, refined and tested models for data exchange for both survey data and qualitative research data based on XML/RDF schema and developed some demonstrator tools for data import and export. The key objective was to develop an intermediate data exchange format that is as vendor neutral as possible. Software and platform neutral formats clearly have huge advantages for long-term archiving and for portability – passing between data analysis softwares. In order for take up of and sustainability of a format/standard, ideally they need support from software vendors to enable import and export functionality. The project worked closely with a number of key software vendors to loath these possibilities.

The data formats included in the demonstrator project are those that are commonly used in research predominantly, but not exclusively in the social sciences – for survey data: SPSS, STATA, CSV and qualitative data: word, Atlas-ti, MaxQDA and Nvivo. These formats are also typically found across all domains of primary research.

A small scale evaluation of the XML schema and data conversion tools was carried out in order to inform JISC of the most viable options for future development in this area. A longer-term aim of this work might be to build a fully functional and scalable facility or service where data formats can be submitted and seamlessly returned in a chosen, desired format – via a neutral exchange format– a true data exchange service.

The project was divided into two separate and quite distinct work packages: The first worked with qualitative social interview-based multimedia data and produced and tested a formal XML schema to support data exchange. This work and the schema was termed QuDex (for qualitative data exchange). A version 3 of the QuDEx schema was released for public comment plus its accompanying documentation, and XML instance files. These are published on the DExT website and will be available on the Open Data Foundation (ODaF) website. A METS file was also produced.

The second worked with survey data, namely in SPSS format and created a java-based data conversion tool ((Survey-DExT), which exported multiple version of SPSS to an 'open' format plus associated formal metadata, and to other major data analysis formats (STATA, SAS). The tool is built so that it is completely extensible to conversion from other data formats such as STATA and SAS and so on. The open-source conversion tool is fully documented and downloadable from the Open Data Foundation website.

The work under the small scale DExT project has contributed to the Repositories Programme by investigating and developing a formalized XML schema and some open source demonstrator tools. These form as solid grounding on which to develop internationally agreed standards and tools that could be widely used to support ingest, long-term preservation, and dissemination of social science research data. This project was formally part of the Tools and Innovation strand, which aimed to develop and pilot innovative approaches to repository use and digital preservation through the development of new software and tools. The outputs are relevant to i) the Discovery to Delivery stream, as the proposed service is based upon common standards for data interoperability, and (ii) Shared Infrastructure Services for resource discovery, repositories and curation - machine to machine services that support rights, profiling, terminologies, registries, file format and representation etc.

The project supported the Programme's desire to improve the efficiency and quality of repository functions, by helping automate the processes of data conversion, and by providing SMART data and tools - in the form of a universal data exchange format. The work set out has thus contributed to the need for refinement of the application of standards and specifications for digital repositories and preservation by building software and tools for both digital repository use and digital preservation. An immediate and indeed longer-term benefit should be increase in productivity for preservation and data sharing services and enhance both the reprocessing of legacy datasets and the data refreshment element vital to good data preservation practice.

Background

Data conversion and proprietary data entry and analysis are particularly important and problematic aspects of data management and curation. The DExT project aimed to provide researchers and support staff working with primary research data with standards and tools that will enable data to be long-term curated and exchangeable. The project researched, developed and tested tools for contemporary quantitative data or statistical data and qualitative data typically used by social researchers.

Much important primary research data is created every day in the course of academic and policy research. While Data Sharing Policies are encouraging sharing and formalised archiving of data, the ideal life cycle for data creation to re-use remains beset by obstacles. The main issues involve the buying-in to a dedicated analytic strategy and typically a particular software package. Over the years the UKDA has seen a number of such softwares quickly become obsolete. To address the problem of incompatibility between software various data conversion tools have come of the market. One example for numeric and statistical data is StatTransfer and DBMSCopy that enable conversion from say SPSS to Stata. Equally the development of the format SPSS.por in the 80s enabled import and export between the major statistical analytic packages. However in the qualitative data analysis software field there are no such inter-software conversion tools.

Legacy formats do and will present problems, a matter recognised for some time by the data archiving community and by some commercial suppliers. A JISC/NPO study on the preservation of electronic materials in 1997 gauged the extent of legacy materials sitting in our institutions - significant.1 Bennett developed a framework of data types and formats and looked at issues affecting the long-term preservation of digital material and the DCC is playing some part in this endeavour, but we are still a long way short of actually dealing with legacy on a managed or efficient scale. Universal exchange formats will help to alleviate the build up of yet more unreadable digital data in our institutions.

Outputs from primary research are also linked, either implicitly or explicitly, to associated research data - for example statistical data, qualitative data, data encoded in a scientific language, or even graphical data. Links between the output research findings and the source data upon which they are based are highly desirable for virtual research environments which are under construction in many institutions at the present time. Such links scarcely exist as yet, however, and are certainly not currently made from the relatively embryonic institutional output repositories which are currently under development. The STORE project is addressing some of these matters. Linking to truly exchange data formats that are professionally curated will be of greater benefit to the research community. The SHERPA project is also using OAIS and METS to harness institutional repository systems with the AHDS preservation repository to create an environment that fully addresses all the requirements of the different phases within the life cycle of digital information.

This project was part of the Tools and Innovation strand, which aimed to develop and pilot innovative approaches to repository use and digital preservation through the development of new software and tools.

Aims and Objectives

The rationale for the DExT work was that open data exchange formats are necessary for maximising the opportunities for data sharing and long-term archiving. The project thus **aimed** to provide researchers and support staff with standards and associated tools for data (and metadata) format conversion. Central to these aims are the recognition and use of best practice in longer-term data management and curation.

The **specific objectives** were to:

- research and develop a numeric data exchange standard and conversion tools
- research and develop a qualitative data exchange standard and conversion tools
- test and evaluate these standards and tools
- assess the feasibility of developing a web-based service for data conversion based on these tools sets

The kinds of data dealt with under the remit of this proposal were data arising from primary research using typical social research methods and techniques based on fieldwork. These are primarily: structured social surveys and more in-depth interviews or focus groups. The UKDA has preferred formats, and also distinguishes between acquisition, dissemination and preservation formats. Users make choices about which software to use in their research and require formats that be easily read.

The project was divided into two separate and quite distinct work packages: The first worked with qualitative social interview-based multimedia data and produced and tested a formal XML schema to support data exchange. This work and the schema was termed QuDex (for qualitative data exchange). A version 3 of the QuDEx schema was released for public comment plus its accompanying documentation, and XML instance files. These are published on the DExT website and will be available on the Open Data Foundation (ODaF) website. A METS file was also produced.

The second worked with survey data , namely in SPSS format and created a java-based data conversion tool ((Survey-DExT), which exported multiple version of SPSS to an ‘open’ format plus associated formal metadata, and to other major data analysis formats (STATA, SAS). The tool is built so that it is completely extensible to conversion from other data formats such as STATA and SAS and so on. The open-source conversion tool is fully documented and downloadable from the Open Data Foundation website.

All of the aim and objectives were met despite a major set back prior to the start of the project. The UKDA suffered the dreadful loss of a colleague and the DExT co investigator who died unexpectedly prior to the project starting and the project even being submitted . Dr. Alasdair Crockett, pretty much single handedly formulated the thinking and work behind WP2. The content of WP2 was drawn up to meet his original ideas but very few people had the background, inspiration and depth of knowledge that Alasdair possessed to undertake the work. The start for WP2 was significantly delayed as a new project leader had to be arranged and a new detailed project work plan drawn up. Instead of recruiting a single in-house programmer to undertake WP2, a known consultant were bought in to take forward the tool research and development. This proved to be highly efficacious and successful, with the work being delivered within the set time frame.

The methodology, results, outputs and outcomes are reported in two parts to make it easier to follow for the reader.

WP2: Survey data (DDI-DExT)

WP2 Methodology

The DDI-DExT project's primary objective was to produce utility software for the conversion of statistical data files into an archive neutral format for long term preservation along with the option to re-package the data for dissemination to end users for use with various statistical packages. This initial effort was intended as a “proof-of-concept” tool to demonstrate the feasibility and sustainability of such an approach.

Software applications are increasingly designed to provide the researcher with a view of their data and its 'internal metadata' (variable and code labels, variable formats, etc.) that is divorced from the software's internal (i.e. underlying) representation, used in conversion. Data creators, data centres and researchers often need to translate datasets between formats, but lack the tools to do so in an accurate and automated way. Data/research centres may use proprietary translation software for certain types of conversion, while individual research projects often rely on the inbuilt import and export functions of a given software package. Both options tend to be poorly documented and operate on the software's internal representations of data. A range of subtle but significant conversion errors often occurs as a result. The major problems that affect data format conversion are:

- rounding/truncation of numeric data
- truncation of textual data
- differences in handling 'internal' metadata (differential label lengths, missing value handling etc.)
- corruption of specially formatted variables (especially date/time variables)
- embedded special characters

For example, The SPSS command 'PRINT FORMATS' is often used to perform data typing upon conversion, yet it seldom matches the actual data and this can lead to catastrophic coarsening of data upon conversion or, conversely, unnecessary inflation of file size. Similarly, one of the ill-documented features of MS Access is that the export precision can be controlled by the number of decimal places in the 'Regional Options' of the Windows Control Panel. Lastly, and an example of point 5 above, embedded characters are an issue in MS Access, wherein fields may contain characters like 'tabs' or 'carriage returns'. Unless these characters are stripped out prior to conversion to delimited text, the data will lose its rectangular structure.

In WP2 the main issues that need to be confronted and dealt relate to survey data ***conversion*** methods and include:

- numeric data in all relevant storage modes, i.e. integer, real and complex
- numerical data in specific formats, e.g. date/time
- textual (character) data, at a UTF-8 standard
- categorical data (ordered or unordered)
- logical data
- 'null' and undefined data

The tool constructed should enable data conversion, a resulting metadata schema that captures these specifities, plus a reporting of all 'errors in translation'. The resulting standard should be DDI 3 compatible and required generation of DDI 3 XML files via XSLT stylesheets. T

The development was conducted from June 2007 to February 2008 as a collaborative effort between the UK Data Archive (UKDA) and the Open Data Foundation (ODaF). For the purpose of this project phase, the input data file format was limited to the most commonly used survey data analysis package format, SPSS[1]. The *archive neutral* formats selected were fixed ASCII for data and the Data Documentation Initiative (DDI) 3.0 XML specification[2] for the metadata. In addition to fixed ASCII, delimited and comma separated formats are also supported. ASCII remains the most commonly used and recommended format for long term preservation. DDI 3.0 was selected for its metadata richness as well as being the leading specification for the documentation of microdata. Direct compatibility with DDI ensures that the tools developed under DExT are compatible with other products and that the generated metadata can be imported by other packages or exchange with other institutions.

Statistical packages targeted as *output formats* were SAS, Stata and SPSS. This included different flavors of these products (such as Small Stata, Stata/CI, etc.) as well various versions (SPSS 11, Stata 7, etc.). The export wizard attempts as much as possible to take into account the differences or incompatibilities between the software packages (like variable name lengths, string and date formats, etc.). It also has the ability to simultaneously generate multiple output formats for multiple files and save the resulting files in either a folder or a compressed ZIP file. The output can also contain DDI 3.0 and DDI 2.1 XML metadata for the generated files.

WP2 Implementation

The project was managed by Matthew Woollard from the UKDA and under the technical coordination of Pascal Heus (ODaF). The development team consisted of Pascal Heus (OdaF) and Jack Gager (Metadata Technology). Joachim Wackerow from GESIS/ZUMA was also brought in as a technical expert for the conversion of SPSS to SAS. Mary Vardigan from ICPSR and Erwin Werkers from the CentERdata in the Netherlands (http://www.uvt.nl/centerdata/nl/) also provided valuable inputs for the testing of the SPSS Reader component.

The project team worked part time on the development and the first phase was completed in mid-October 2007 with the release of the alpha-test version of the DExT Tools application supporting the SPSS file import functionalities. The second phase development phase concluded end of January 2008 with the final version of the product that included the SAS, SPSS, and Stata export functions. Activity and progress report were regularly exchange by email and a face to face meeting took place at the UKDA mid-January 2008 to present the product to the project manager and discuss lessons learned and next steps.

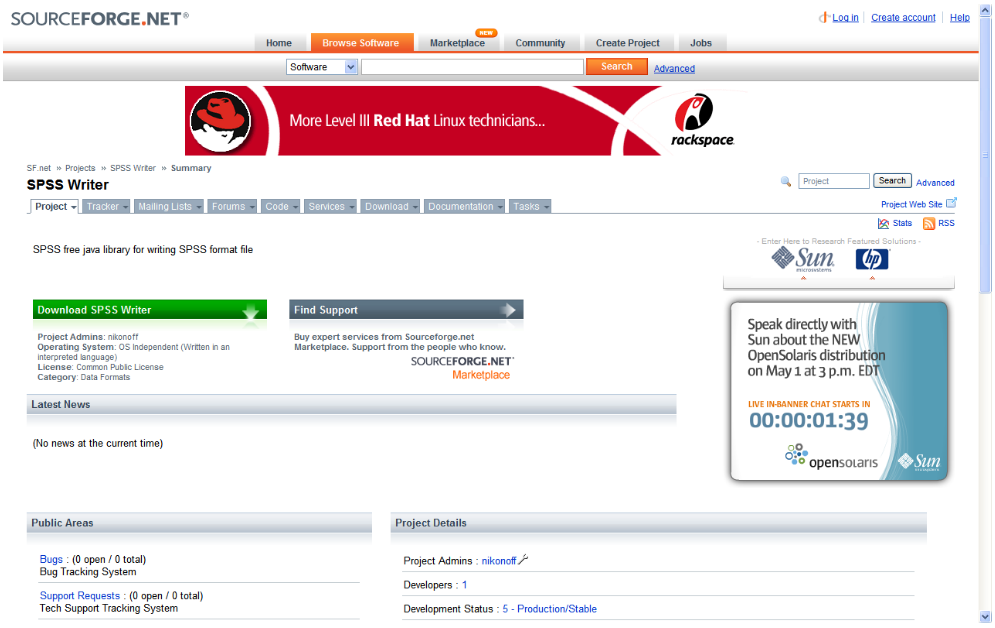
The project team followed the general recommendations outlined in the Open Data Foundation "Managing social, behavioral and economic data and metadata: Guidelines for Tools Development and Recommendations for Operating Environment"[3] document. The packages were developed in the Java programming language using the Eclipse Integrated Development Environment (IDE). The source code is published and maintained in the ODaF Forge public repository[4].

The end user DDI-DExT product was developed as an Eclipse Rich Client Platform[5] (RCP) application to maximize openness and portability across-platform. Conversion of the DDI-XML metadata into setup files for the various statistical packages were developed using the XSLT v2.0 language and processed by Saxon XSL v8.9[6].

In terms of managing a contractor for this WP (and one based in the US) we were surprised how straightforward the management, development and delivery process was. In addition to the initial formalised contract and agreed programme of work, the UKDA asked for monthly reports and had 3 face to face meetings and a couple of tele conferences. Communication was mostly by email. The subcontract worked extremely well and was considered to be excellent value for money. It is unlikely that UKDA could have recruited or afforded a developer/programmer of this caliber to produce the tools in the timeframe that was achieved by Pascal Heus of ODaF. The UKDA is now considering contracting out some of its more discrete and development-oriented technical projects to consultants such as those working for ODaF.

WP2 Outputs and Results

The primary outputs of the project are the DDI-DExT Tools application and the underlying SPSS Reader component. Both products have been released under the GNU Lesser General Public License[7] and donated to the DDI Foundation Tools Program[8] as a contribution from the UKDA and the Open Data Foundation. The DDI-DExT installer and SPSS Reader can be downloaded from the DDI Tools web site[9] and the source code is publicly available on the ODaF forge repository.


SOURCEFORGE.NET®
Log in | Create account | Help
Home
Browse Software
NEW
Marketplace
Community
Create Project
Jobs
Software
Search
Advanced
More Level III Red Hat Linux technicians...
rackspace
SF.net » Projects » SPSS Writer » Summary
Project
Search
Advanced
Project
Tracker
Mailing Lists
Forums
Code
Services
Download
Documentation
Tasks
Project Web Site
Stats
RSS
- Enter Here to Research Featured Solutions -
Sun microsystems
hp
Speak directly with Sun about the NEW OpenSolaris distribution on May 1 at 3 p.m. EDT
LIVE IN-BANNER CHAT STARTS IN
00:00:01:39
opensolaris
Sun

SPSS Writer

SPSS free java library for writing SPSS format file

Download SPSS Writer

Project Admins: nikonoff
Operating System: OS Independent (Written in an interpreted language)
License: Common Public License
Category: Data Formats

Find Support

Buy expert services from Sourceforge.net Marketplace. Support from the people who know.
SOURCEFORGE.NET® Marketplace

Latest News

(No news at the current time)

Public Areas

Bugs : (0 open / 0 total)
Bug Tracking System

Support Requests : (0 open / 0 total)
Tech Support Tracking System

Project Details

Project Admins : nikonoff

Developers : 1

Development Status : 5 - Production/Stable

Screen shots of the tool can be found in the WP2 Technical Report in Appendix 1.

A wide range of SPSS files were collected from various sources to test the import capabilities of the product. As initial test cases, a simple SPSS 11 file containing all possible variable formats was created. The UK Data Archive also provided a set of 9 files as use cases. In addition, the following individuals/organizations contributes test SPSS files: Mary Vardigan (ICPSR, USA), Erwin Werker (CenterData, Tilburg University, Netherlands), Joachim Wackerow (GESIS/ZUMA, Germany), Dan Kristiaensen (DDA, Denmark) and Guido Gay (IReR, Italy).

This provided for a wide range of test cases varying from SPSS version 4 to 16 and produced on different platforms (Windows, Linux, Solaris). With very few exceptions, the DExT SPSS Reader is able to properly read and export the data and metadata. A known issue exists for files produced on the Solaris platform (SPSS 4) or with the open source SPSS Writer Java package distributed through SourceForge[10]. A fix is trivial and will be implemented in upcoming maintenance release of the reader.

The export functionalities of DexT-DDI were tested internally with SPSS 11, Stata 8 and SAS 9.1. External feedback has been limited at this time given the small number of testers. While some adjustments are expected to be necessary, such changes are based on an XSL transformation and do not require the recompilation of the entire application. The Open Data Foundation and other contributors are expected to integrate such changes in future releases.

Three levels of documentations were maintained:

- Source code level: documentation is present in the Java files in the form of comments captured in the standard JavaDocs[11] format based on Sun Microsystems recommendations (see ODaF guidelines).
- Application level: user level documentation is provided in the DDI-DExT application help system

- Project level: technical documentation and administrative activities. These documents are available upon request.

The technical documentation includes a detailed paper on the SPSS and other file formats. As this information was collected from various sources on the Internet whose author could not always be identified or may be copyrighted, this paper cannot be released in the public domain. Several public sources documenting the SPSS format are however available. We want in particular to thank the members of the PSPP project[12] community who have been particularly helpful and supportive during the development of the DexT SPSS Reader package. The full technical report for WP2 by the contractor, Pascal Heus can be found in Appendix 1 and the URL for the final DExT tools is http://sourceforge.net/projects/spss-writer:

To advocate the availability of the tools, presentations and demos are planned at the ODaF Europe 2008 meeting and the IASSIST 2008 conference in Stanford, CA in May. Announcement will also be made through relevant channels such as the DDI Users, IASSIST and Open Foundation mailing lists.

It was agreed at the end of the project to move the products into maintenance mode and leave a period of 6-months for further testing by the UKDA, the DDI community and other interested users. Bug fixes and minor enhancements will be contributed by the Open Data Foundation, the UKDA development team and potentially the open source community.

Short term planned enhancements include full alignment of the DDI 3.0 specification once officially released end of May (the tool are currently using a candidate release version) as well as improvements to the script generator and other know minor issues.

Potential future enhancements include:

- Support for additional input formats such as Stata, SAS, SPSS Portable or Nesstar
- Support for additional setup file formats such as R, Excel or SQL Databases.
- Data export to a “generic” ASCII format that uses ISO standard dates, general number formats.
- Addition of summary statistics to the DDI metadata

WP2 Conclusions

The DDI-DExT project successfully demonstrated the feasibility of using open source solutions, ASCII text files and standards based metadata for the conversion, preservation and dissemination of microdata. While not a production grade product, the DDI-DExT application can already be used as a free stand alone import/export utility for SPSS files and the automated creation of DDI metadata. The availability of an open source Java library for reading SPSS data is also an important contribution as it greatly facilitates access to data and metadata stored in the proprietary SPSS format by any applications and web based services. The library as already been integrated in the DANS MIXED project[13] and has drawn interest from several other individuals and open source product developers.

The project also provided major contributions to the DDI community: first, it was one of the main use case for the testing and enhancements of the candidate recommendation versions of the DDI 3.0 XML metadata specification. As such, it directly leads to several improvements to the upcoming new standard. The package is also one of the first known implementation of the DDI 3.0 and will is likely to support and foster the adoption of the new specification.

Finally, the work performed on the setup files script generator highlighted many of the differences and incompatibilities between the SPSS, Stata and SAS software. These lessons learned could potentially be formalized into a paper that outlines the caveats and provides general recommendations to minimize information loss when converting data across statistical packages.

Overall, this initial effort was successful in more way than initially anticipated. We hope to have the opportunity to further develop the package to turn it into a versatile set of utilities for the management of microdata.

WP3: Qualitative data (QuDEx)

WP3 Methodology

The majority of social researchers undertaking qualitative research methods are making use of some form of data management software. This can be MS Word or MS Access but in the past 15 years a number of dedicated packages have come on the market. These are called generically termed Computer Assisted Qualitative Data Analysis Software (CAQDAS) packages and include the market leaders NVivo, Nudist, Atlas-ti and MaxQDA. CAQDAS packages were developed in the late 80s typically by keen qualitative researchers and the resulting software thus embodies different methodological and analytical approaches. The past decade has seen a huge take up of the use of these packages in research and in teaching and in the UK the CAQDAS Networking group has provided an invaluable information portal, forum and outreach program for helping get users started. While a basic common denominator set of functions can be seen across the software, various new functions have been added to some and not others. Thus each has its own flavour, and also terminology.

The softwares have similar basic functionality that includes:

- structuring work - ability to access to all parts of a project immediately
- staying 'close to data' - instant access to source data files (e.g. transcripts)
- exploring data - tools to search text for one word or a phrase
- code and retrieve functionality - create codes and retrieve the coded sections of text
- project management and data organisation
- searching and interrogating the database - search for relationships between codes
- writing tools - memos, comments and annotations
- outputs - reports to view a hard copy or export to another package

The key problem for a data archive interested in acquiring and disseminating data from qualitative research studies is that CAQDAS formats are proprietary and there exist no export or import formats. Thus once a researcher uploads data into a particular package, s/he is locked into that particular software and format. The work carried out within these packages, which is primarily coding and annotating data, is stored as an integral part of the softwares 'project' which can typically be exported as whole unit. However it is not possible to share the added value undertaken within the package. A list of annotations or codes can be exported but the links to the underlying data cannot. As yet, there are no open source products which can compete with the functionality of the leading softwares, Atlas-ti and QSR NVivo.

A standard format for representing richly encoded qualitative data is necessary because it: it ensures consistency across datasets; supports the development of common web-based publishing and search tools; and it facilitates data interchange and comparison among datasets. Importantly, it could also enable data and l inked products to be imported and exported directly into and out of CAQDAS packages, avoiding the reliance on just a single product, and offering the opportunity to share analytic workings outside the confines of any particular software.

In essence, the model we were aiming at was a data exchange format that can represent data collections and retain links between data, annotations and related objects. The model should be aware of complimentary metadata standards (DC, MODS, OAI, DDI and TEI).

The work undertaken in this WP took as its starting point two recent and quite roughly specified data models that have been developed. The first has been an ongoing work programme of ESDS Qualidata[14] at the UK Data Archive (Louise Corti is head of this service) who in the early 2000s developed a draft but limited formal definition of a common XML vocabulary and Document Type Definition (DTD) based on the Text Encoding Initiative (TEI) for describing these structures[15]. The Universities of Melbourne and Queensland have further developed a draft Qualitative Data Interchange Format (QDIF) for e-Social Science (QDIF)[16]. Both centres have been working closely together in this very early development phase, but as yet, neither has any dedicated funding to work further on realistic development or testing, so the specification work remains on the back-burner.

WP3 was directed and managed by Louise Corti from the UKDA with a technical/ metadata coordinator consultant, Herve L'hours bought in to support the full-time programmer Angad Bhat. XML consultants were bought in in the summer of 2007 to formally evaluate the initial schema and have major input into its subsequent development.

QuDex is the name given in this project to the qualitative data exchange model for the archiving and interchange of data and metadata between CAQDAS packages. The draft QuDEx standard/schema is essentially a software-neutral format for qualitative data that preserves ***annotations*** of and ***relationships*** between data and other related objects. The QuDEx XML schema is based on a small number of key concepts and elements that can represent: coding, classifying, memoing, and relating.

All of the key vendors were consulted throughout the QuDex development process. These were:

- ATLAS.ti
- Nvivo
- NU*DIST
- Max QDA
- QDA Miner
- Qualrus
- HyperResearch
- Tinderbox
- Transana
- WeftQDA

In the first phase of WP2, the key functionalities for the market leading software packages were compared. This comparison helped to distinguish the baseline, what may be thought of as common denominator functions possessed by all the softwares: annotation of data through codes; memos; classifications and relationships. It helped to establish the core concepts on which to build the schema. The report is appended in Appendix 2 and available at: http://www.data-archive.ac.uk/dext/Software2.xls

A comparison of existing and possibly relevant metadata schema was also undertaken. This can be found in Appendix 3 and at http://www.data-archive.ac.uk/dext/comparison3.doc

The initial schema for the baseline concepts was presented to a significant number of the CAQDAS vendors at a captive-audience conference on qualitative computing in April 2007. This is the first time the vendors had an opportunity (if only though curiosity) to come together to discuss what potentially might represent competition between them. None currently provide import and export to each others software formats, though a couple do export some basic XML. At this meeting an overview of a potential basic model for data exchange was provided by the DExT team. While not all in agreement, the majority decided that a discussion of which basic aspects of functionality might be common across packages would be worthwhile. A WIKI was been set up for discussion on this issue, and the feedback utilised in developing Version 1 of the QuDEx schema.

Transforming data from a proprietary data structure into the Qudex Schema is impossible without having access to the underlying data structure. DExT was not able to do this within the time frame, and a single leading software was chosen as the test case. This exported data into a basic, schemaless XML. This enabled a mapping from this format (Atlas-XML) to the QuDex generic XML and back again to demonstrate proof of concept.

The project only had enough funds to support one relatively junior member of staff as a programmer and what was clearly lacking was ***developer*** input. The real impact of this void was noted and in late spring 2007, a local consultant was bought in to help manage the progress of schema development by providing communication between the Investigator and the project's programmer. Herve l'Hours bought with him experience of metadata schema development and implementation for multimedia data and in particular, a theoretical ***and*** working knowledge of Metadata Encoding and Transmission Standard METS (see Outputs and Results for discussion).

The schema was developed with input from XML consultants from the Open Data Foundation (ODaF), who have extensive experience in the formulation and development of formalised XML specifications in this area (Electronic Business using eXtensible Markup Language: ebXML[17] and the survey data

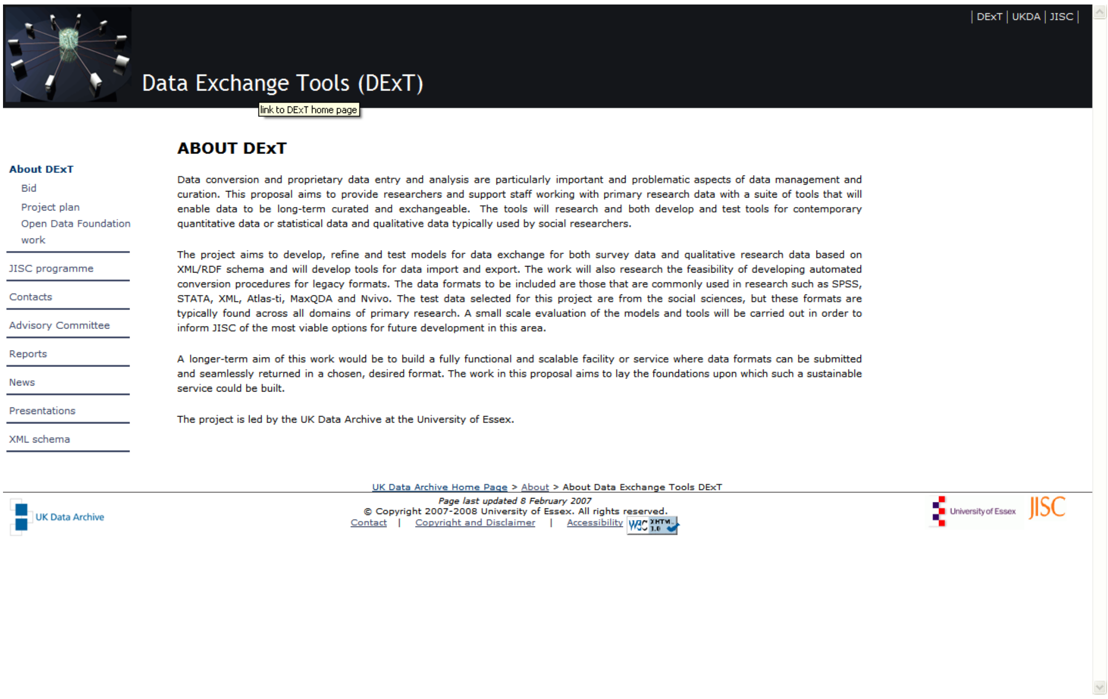
XML schemas: SDMX[18] and DDI). A visit by Arofan Gregory of ODaf (as DExT consultant) proved to be a significant milestone in formulating a more robust first draft schema. His expertise in dissecting complex and competing data structures and models aided the team's progress enormously.

Following agreement of the last version under the DExT funding, a QuDex viewer and QuDex transformation tool were hastily built, to show proof of concept, i.e. that the schema represented CAQDAS functionality.

In September 2007, two DExT staff attended the invitation only ODaF meeting in California *to* discuss the DExT project and tools. The team was invited to propose a working group of the Data Documentation Initiative (DDI) alliance in an attempt to progress the standard. This was a major breakthrough in getting such a standard for qualitative data recognized. A presentation at the Association of Survey Computing (ASC) in September 2007 also gave an opportunity for feedback by social survey tools developers. In April 2008 after the project finished another ODaF meeting was held, hosted by the UKDA, at which the DExT tools were discussed with unanimous support for the development under the ODaF umbrella of visualisation tools based on the QuDex schema.

WP3 Outputs and Results

Three versions of the schema (V1 to V3) were released, focussing on a generalise standard to represent the core concepts. The schema enables import from one of the CAQDAS packages Atlas to the intermediate schema (QuDEX0 and back out. It captures the key functions of CAQDAS packages identified. The DExT web site and a WIKI enabled opportunities for promotion, communication and feedback.



Data Exchange Tools (DExT)

ABOUT DExT

Data conversion and proprietary data entry and analysis are particularly important and problematic aspects of data management and curation. This proposal aims to provide researchers and support staff working with primary research data with a suite of tools that will enable data to be long-term curated and exchangeable. The tools will research and both develop and test tools for contemporary quantitative data or statistical data and qualitative data typically used by social researchers.

The project aims to develop, refine and test models for data exchange for both survey data and qualitative research data based on XML/RDF schema and will develop tools for data import and export. The work will also research the feasibility of developing automated conversion procedures for legacy formats. The data formats to be included are those that are commonly used in research such as SPSS, STATA, XML, Atlas-ti, MaxQDA and Nvivo. The test data selected for this project are from the social sciences, but these formats are typically found across all domains of primary research. A small scale evaluation of the models and tools will be carried out in order to inform JISC of the most viable options for future development in this area.

A longer-term aim of this work would be to build a fully functional and scalable facility or service where data formats can be submitted and seamlessly returned in a chosen, desired format. The work in this proposal aims to lay the foundations upon which such a sustainable service could be built.

The project is led by the UK Data Archive at the University of Essex.

UK Data Archive Home Page > About > About Data Exchange Tools DExT

Page last updated 8 February 2007
© Copyright 2007-2008 University of Essex. All rights reserved.
Contact | Copyright and Disclaimer | Accessibility

The schema has been built on best practice for text and audio-visual annotation and makes use of Xpointer and some elements from the audiovisual standard SMIL. This is fully documented in the QuDex Reference document and other documents can be found on the DExT site (see URL refs later) but this key document is also appended (Appendix 5) to this report. A basic outline follows.

Key ***elements*** used in the QuDex standard together with their definitions are shown in the Table 1. below:

Table 1 QuDex elements and definitions

Top level Elements	**Sub elements**	**Definition**
<qudex>	resourceCollection segmentCollection codeCollection memoCollection categoryCollection relationCollection	The root element; a 'wrapper' for all other elements of the QuDEx Schema. Each top level element in QuDEx is defined as a 'collection' and must appear in the order outlined below
<resourceCollection>	sources memoSources documents	The *resourceCollection* section lists and locates all content available to the QuDEx file. A *source* points to the original location of the resource while each author working on the QuDEx file is assigned a surrogate *document* which points to the relevant *source*. The child elements *sources* and *memoSources* contain direct references to the files under analysis; the *documents* section contains their surrogates
<segmentCollection>	Segment (sub elements text, audio, video, xml, image)	The parent element for all *segment*s, which is a subset of a *document* (text, audio, video or image) under analysis defined in a manner appropriate to the format (text, audio, video, image or xml). *Segment*s may overlap and multiple *memo*s and *code*s may be assigned to a *segment.* Start and end points can be formally assigned to segments of text, and audio visual materials an other document
<codeCollection>	code	The parent element for all *code*s. A code is a short alphanumeric string, usually a single word; may be assigned to a segment or document though assignment is not required. A code may optionally be taken from a controlled vocabulary defined under @ authority
<memoCollection>	memo (sub elements memoDocumentRef, memoText)	The parent element for all *memo*s; these may be pure text and embedded in the QuDEx file (inline memo) or may refer to external files. A memo is a text string internal to the document (inline memo) or an externally held document (external memo) which may be assigned to a segment, code, document, category or to another

<categoryCollection>	category	The parent element for all categories. A category is an alphanumeric string (stored in @label) assigned to one or more documents. Categories may be hierarchically nested. Documents contained within a category are referenced using @documentRefs. Nested categories are referenced using @categoryRefs
<relationCollection>	objectRelation	The parent element for all relationships between objects. For the purposes of a *relation* all of the following are considered to be 'objects' ▪ A *document*: surrogate of a *source* or *memoSource* ▪ A *segment* within a *document* ▪ An assigned value: *code*, *memo*, *category*, *relation* A relation is a link between two objects in a QuDEx file. Each object is either the start or end point of a relation (source vs target). Every relation may, optionally, have a name

A number of **attributes** are commonly used within the QuDEx standard, with standard attribute groups assigned designed to support the management of complex layers of analysis by multiple authors within a single QuDEx instance. These are:

@ cdate: the date and time the instance of the element was created
@ mdate: the last date and time the instance of the element was modified
@ creator: the original creator of the instance of the element or the author of the relevant resource
@ label: a human readable string for the element in general or its specific contents
@ displayLabel: a version of the label text appropriate for display, for example in a user interface
@ language: this caters for describing the overall language of the study while permitting element level variations such as defining a segment, memo or code as being in a different language

Appendix 4 contains a glossary of acronyms.

Each successive version of the schema was mounted on the DExT WIKI and the CAQDAS vendors and key stakeholders emailed (see Appendix 5 for stakeholders, organisations and individuals who were targeted and who contributed):

- **October 2007: Version 1.0 of QuDEx Schema released**

- **December 2007: Version 2.2 of QuDEx Schema released**

- **February 2008: Version 3.0 of QuDEx Schema released**

In December 2007, version 2.0 of the QuDEx draft schema was added to the DExT WIKI for comment. Following feedback from vendors, some key stakeholders and XML consultants from the Open Data Foundation (ODaF0 some critical amendments were made for Version 3.0. Version 3.0 is the final version that has been accomplished under the JISC-supported DExT project funding and is considered by the international XML consultant to be a very solid grounding for which to test more use cases and build both conversion scripts and data and visualisation tools.

The Version 3.0 draft schema, its UML model, accompanying documentation and XML instance files are listed below and can be viewed on the DExT Website (www.data-archive.ac.uk/dext/schema). Change notes also accompany each successive schema version to reflect the changes made and rationales for making these changes. The Reference Documents is the only document of these Schema-related documents to be appended to this report (Appendix 6).

Full Documentation of QuDex schema:

- *Release notes* for the Schema version 3.0

- *Open issues* A collection of the open QuDEx issues to be addressed in successive schema versions

- *Reference* A reference guide to QuDEx elements and attributes which we hope may be expanded to provide further implementation support in future versions. A list of relevant acronyms is provided at the end of this document providing basic definitions (not technical details) of QuDEx and other related concepts

- *QuDEx mapping from Atlas-ti to QuDex XML*

Schema and XML instances

- *QuDEx Schema* The latest version of the QuDEx Schema v3.0

- *QuDEx Object Model* The latest version of the QuDEx Object Model v3.0

- *Full instance* A full example instance of the xml containing most elements and attributes

- *Sample instance* An Atlas-ti XML example based on the sample data prepared for the DExT project

- *Sample METS* Supporting document for archiving of QuDEx and other material

Metadata Encoding and Transmission Standard (METS) is a standard for encoding descriptive, administrative, and structural metadata regarding digital objects, expressed using the XML schema language. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. METS is becoming an increasingly attractive standard for dealing with preservation metadata and confirming validation processes and is JISC's chosen stand for interoperability. It was therefore considered to be relevant to the development of data conversion standards and tools.

In 2006 the TNA and UKDA recently undertook a JISC-supported project to report on Open Archival Information System (OAIS) Reference Model and METS. The Quali DExT work evaluated this report and other METS related work undertaken in-house for the JISC Digitisation HISTPop project and developed a ***basic*** METS schema to complement the QuDex schema. That is, rather than attempting to perform every function required by researchers to describe a complex research data collection, QuDEx aimed to deliver ***core*** functionality while other metadata standards can be used as appropriate. In DExT all materials and metadata relating to a collection of objects pertaining to a single research study (data files, documentation such as methods and protocols and outputs) were packaged as ‘complex objects’ using METS. In other words, METS has been used to define a broader single 'collection' and to specify the relationships between parts of the collection.

Screen shots for the **QuDex Viewer** and instruction for installation plus the **QuDex Transformation** are shown in Appendix 7. The Viewer is an open source tool whose objective is to facilitate the use and understanding of QuDEx XML files and transform one of the CAQDAS vendor packages XML (in this case Atlas.ti) to its native XML format. It has been designed for simple browsing, transforming and viewing of core constructs such as code, segments, memo and their relationships. It requires the installation of tomcat server on to a local directory to run the application into a standard web browser such as Firefox or Internet Explorer. The QuDEx Viewer leverages the open sources packages, Yahoo! UI Library and Bubbling Library[19].

To use the software, any valid QuDEx XML file on a computer can be opened by using the application's menu at the top of the page. For all QuDEx XML documents, it provides basic functionalities such as a listing of all the codes, memos, segments, categories and their relationships and the ability to browse the document in either text or xml view. The sample QuDEx documents can be found on the DExT web site at www.data-archive/dext/schema/QuDex Viewer.zip.

The application also allows transforming Atlas.ti xml to QuDEx format by simply selecting the files for transformation. This feature can also be accessed using the top menu of the application. The sample Atlas.ti xml file and a copy of schema could be accessed from the URL as above. The files should be stored on a local drive and the application pointed to these files for transformation. These tools are very much demo tools, but could be used as the basis for a import/export and transformation tools and also a basic open source CAQDAS tool working on data in the open XML structure.

While the XML schema is published on the project web site, and will be linked to from other key sites (e.g. the DDI), it is important to realise that the standard is only of use to the potential user community if freely available tools are made available to get data into the curation/interchange standard from proprietary data formats and also export back out into proprietary formats. Hence a demonstrator of import and export utilities was developed in the last month of the project. This is only a rough tool and user friendly conversion tools would need to accompany the new standard if take up is to be encouraged. The ODaF are keen to work to support some basic utilities to support the standard, but it is the software vendors who may need to take some initiative in developing their own import and export tools if they wish to cater for open format translation.

WP3 Conclusions

Following a final round of comments to this latest release, we anticipate that the schema work will be transferred under a working group of the DDI committee (http://www.icpsr.umich.edu/DDI/). The schema will also be supported by the Open Data Foundation (ODaF) and any further versions and tools will be hosted on their web site (http://www.opendatafoundation.org/). We welcome anyone interested in participating in further developments to contact the team at UKDA to get involved.

The project's lifespan was short to undertake a fairly major task - a full blown schema development for complex objects primarily to suit the needs of the social science research community. The team were happy that so much has been achieved within such a relatively short period. As with the WP2, much of the progress can be attributed to the two consultants, Herve L'hours and Arofan Gregory who had the knowledge and skills to make quite critical decisions in haste.

Outcomes and Implications

In addition to the DExT website, the DExT project was also promoted at a number of key events: the UK e-science conference; national research methods conferences; all ESDS and UKDA events (some 50 per year); the two annual IASSIST conferences; the 07 CAQDAS conference; two years at the Association of Survey Computing (ACS); and at two of the Open Data Foundation meetings. It is hoped that by targeting these audiences, news has filtered out into the spheres which are most likely to consider further development of tools and take up.

The DExT outputs are relevant to a) the Discovery to Delivery stream of the Repositories Programme, as the proposed service is based upon common standards for data interoperability, and (b) Shared Infrastructure Services for resource discovery, repositories and curation - machine to machine services that support rights, profiling, terminologies, registries, file format and representation etc.

The work supported the Programme's desire to improve the efficiency and quality of repository functions, by helping automate the processes of data conversion, and by providing SMART data and tools - in the form of a universal data exchange format. The work aimed to contribute to the need for refinement of the application of standards and specifications for digital repositories and preservation by building software and tools for both digital repository use and digital preservation. An immediate benefit should be increase in productivity for preservation and data sharing services and enhance both the reprocessing of legacy datasets and the data refreshment element vital to good data preservation practice.

The project built on development work arising from the existing JISC programmes: Digital Repository Programme and the Supporting Digital Preservation and Asset Management in Institutions. Existing current projects from these JISC programmes are addressing publications and theses, learning objects, experimental and geospatial data, but largely do not cover the types of (widely used) data central to the DExT work. The project also built upon other contemporary reports and investigatory papers into preservation and dissemination issues at academic institutions. The 2005 report on 'Digital Repositories Roadmap: looking forward' called for the provision of a 'solid environment within which a wide variety of software tools (open source and commercial) and added value services can be developed' and 'functionality and services that support curation, migration and preservation'.

The UKDA has already produced three Best Practice Guides for the MRC and a joint ESRC/NERC/BBSRC RELU Programme in: Data Management - covering data format selection, metadata and documentation standards and content, version control, access, and authentication6 and in Data Format Conversion - covering the issues involved in data conversion. These provide solid background advice for researchers and centres with data sharing or archiving commitments. As yet while they suggest preferred formats, they have not attempted to cover vendor neutral data formats. It is anticipated that the new formats created in the DExT work will be integrated into advice if they are chosen to become the UKDA s preferred preservation format.

In conclusion, open source standards and utilities for the conversion of data into a standard archiving format and for exchange were developed as a proof of concept that focused on a limited number (yet high usage) software formats. The DExT work has laid the foundation for more advanced tools that can support additional input and output formats and provides enhanced functionalities. Ongoing support for such tools within our own community will ensure that they will further tested, refined and our hope is for them to be fully embedded in everyday repository practice.

Recommendations

It is recommended that all future JISC, HEFCE and RCUK initiatives that look at data and metadata standards for ***research data*** and longer term repository storage consult this project, as the underlying standards and tools are robust ad also quite generic. They are also supported by a significant international community – through the association of social science data archives - IASSIST[20].

Over the next year the UK Data Archive will work will work to further test and embed these standards and tools into everyday practice, and are more than happy to discuss use and adaptation of tools by other groups – they are open source and very much built for the community. The use of METS and MODS in the project has also created a useful and fairly simplistic way of describing a complex collection of research data and associated context and outputs, which we will gladly share and discuss. We would encourage JISC to include the project’s work on its list of repository tools and standards, and also after the Repositories Programme has ended.

Following further review of the tools there is scope for exploring the feasibility of a fully specified data curation/exchange service that is web-based. Users could upload data that have been extracted from a software package and return a chosen converted format. This is a logical extension to the ground work that has been accomplished under DExT and would be highly beneficial low investment tool for institutional repositories. The short time scale and limited budget for this project did not allow for exploration of this kind of facility, but the potential is certainly worth following up.

References

[1] SPSS created by SPSS Inc. is a leading worldwide provider of predictive analytics software and solutions., http://www.spss.com/

2 The DDI is the internationly agreed standard for describing survey micro data http://www.ddialliance.org

3 http://www.opendatafoundation.org/?lvl1=resources&lvl2=papers

4 http://forge.opendatafoundation.org

5 http://wiki.eclipse.org/index.php/Rich_Client_Platform

6 http://saxon.sourceforge.net/

7 http://www.gnu.org/copyleft/lesser.html

8 http://tools.ddialliance.org/?lvl1=ftp

9 http://tools.ddialliance.org/?lvl1=product&lvl2=dext

10 http://sourceforge.net/projects/spss-writer/

11 http://java.sun.com/j2se/javadoc/

12 http://www.gnu.org/software/pspp/

13 http://www.dans.knaw.nl/en/projects/mixed/

14 ESDS Qualidata www.esds.ac.uk/qualidata

15 ESDS Qualidata XML application for qualitative data www.esds.ac.uk/qualidata/online/about/xmlapplication.asp

16 Qualitative Data Interchange Format (QDIF) http://sourceforge.net/projects/qdif

17 ebXML (Electronic Business using eXtensible Markup Language) is a modular suite of specifications that enables enterprises of any size and in any geographical location to conduct business over the Internet. Using ebXML, companies now have a standard method to exchange business messages, conduct trading relationships, communicate data in common terms and define and register business processes. http://www.ebxml.org/

18 SDMX , SDMX is an initiative to foster standards for the exchange of statistical information. http://www.sdmx.org/

19 Yahoo! UI Library http://developer.yahoo.com/yui and Bubbling Library http://bubbling-library.com/

20 IASSIST The International Association for Social Science Information and Service and Technology http://www.iassistdata.org/

DExT Website with reports and presentations: http://www.data-archive.ac.uk/dext/

WP2 tools: http://sourceforge.net/projects/spss-writer

WP3: QuDEx comparison of CAQDAS packages: http://www.data-archive.ac.uk/dext/Software2.xls

WP3: QuDEx comparison of possible contending metadata schema for qualitative data : http://www.data-archive.ac.uk/dext/comparison3.doc

WP3 QuDEx sSchema and documentation: http://www.data-archive.ac.uk/dext/schema