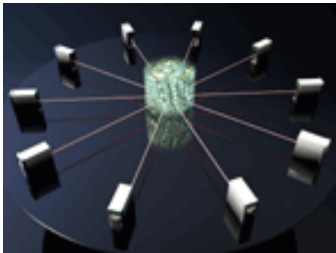




# Dexterity: Data Exchange Tools and Standards for Social Sciences

Louise Corti, Herve L'Hours,  
Matthew Woollard (UKDA)  
Arofan Gregory, Pascal Heus (ODaF)

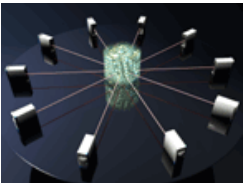


I-Pres, 29-30 September 2008, London



# Introduction to DExT

- data exchange models and data conversion tools for primary research data
- two aims:
  - a standard format for representing richly encoded qualitative data
  - conversion tools from proprietary statistical software to open formats

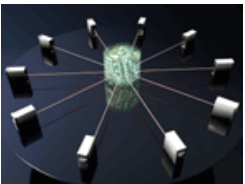




# Project Environment

## JISC funded

- funded by JISC (Joint Information Systems Committee)
- Capital Programme: Repositories and Preservation (*Tools and Innovations strand*)
- small budget for one year – proof of concept

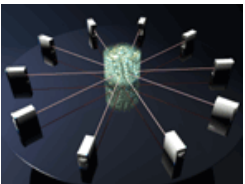




# Project Environment

## UKDA

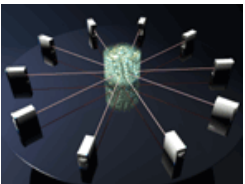
- the leading UK social science data archive
- documenting, preserving and disseminating data for over 40 years
- pioneered the archiving and sharing of qualitative data at national level
- commitment to open standards and formats
- support to DExT project through testing, maintaining and embedding conversion tools and standards





# 1. DExT: qualitative data

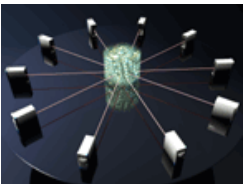
- data selected were from the social sciences
- 'qualitative' data: annotated and coded text, multimedia, linked sources
  - in-depth and semi-structured interviews
  - focus groups, observations, field notes, case study notes
  - diaries, open-ended survey questions
  - personal documents and photographs
- the data formats are typically found across all domains of primary research





# ESDS Qualidata

- national service providing access to and support for a range of social science qualitative datasets
- part of the ESRC funded Economic and Social Data Service (ESDS)
- promotes and facilitates effective use of data in research, learning and teaching
- offers a resource hub via the [www.esds.ac.uk](http://www.esds.ac.uk)
- committed to creating value-added data resources through enriched data and description



# ESDS Qualidata

About

Data

Create/deposit

Online

Support

News

Events

Which service?

Select service

 [Print-friendly page](#)

About ESDS Qualidata

[Brochures](#)

[FAQ](#)

[Contact](#)

[Advisory committee](#)

[About this web site](#)

## About ESDS Qualidata

ESDS Qualidata is a specialist service of the ESDS led by the [UK Data Archive](#) (UKDA) at the University of Essex. The service provides access and support for a range of social science qualitative datasets, promoting and facilitating increased and more effective use of data in research, learning and teaching. The work builds on Qualidata's expertise and international reputation in this area.

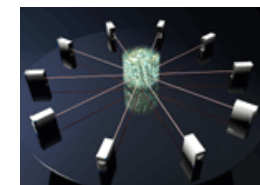
The service's focus is on acquiring digital data collections from purely qualitative and mixed methods contemporary research and from UK-based 'classic studies'. Data are typically acquired via the ESRC Datasets Policy that requires all research grant award holders to offer data collected during the course of their research for preservation and sharing through the ESDS. ESDS Qualidata works closely with data creators to ensure that high quality and well-documented qualitative data are produced. General guidance and a dedicated advisory service for data creators and depositors on research project management, issues of confidentiality and consent, and data documentation of data for archiving are provided.

ESDS Qualidata offers a resource discovery hub via the [Data Catalogue](#), enabling users to locate accessible sources of qualitative data across the UK. ESDS Qualidata facilitates the preservation of important large paper collections, and where appropriate, digitises samples of these collections. The service is committed to creating value-added data resources through enriched data context as well as developing authenticated online access to qualitative data. Related research and development work undertaken by ESDS Qualidata staff have provided some important methodological and technical developments for sharing qualitative data.

Finally, ESDS Qualidata offers [user support](#) and [training](#) to encourage professional researchers and research students alike to make full use of the rich sources of archived qualitative data.

[More about ESDS Qualidata...](#)

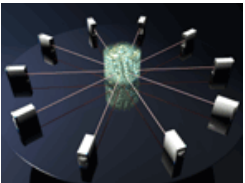
The ESDS web site will be unavailable between 08.00 and 18.00 on Saturday 25 August 2007 due to [essential maintenance](#) at the University of Essex.





# Data ingest

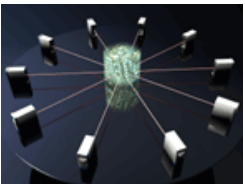
- text ingested as WORD, RTF, TXT
- CAQDAS software formats, text extracted as RTF or TXT (...)
- audio ingested as WAV and MP3
- paper digitised to TIFF and where possible to RTF, XML





# Data delivery

- authenticated web download
  - text delivered via as RTF or PDF, depending on level of digitisation
  - audio as MP3
- online data browsing for selected interview text collections in XML (marked-up using TEI).  
Extending to multimedia
- audio streaming coming (MP3)



# Basic XML mark-up in TEI

- document level
- content-level structural mark-up

- header
- interview attributes
- utterances
- selected interviewee
- turn-taking

```
</teiHeader>
<text>
  <body>
<U who="Interviewer" id="1">I'd like to start, if
  <U who="Subject" id="2">November 9th, <date>1902<
<U who="Interviewer" id="3">Could you tell me how
  <U who="Subject" id="4">There were 11 of us. I wa
<U who="Interviewer" id="5">Could you tell me, if
or girls.</U>
  <U who="Subject" id="6">Well, the first 3 of us w
<U who="Interviewer" id="7">So you were approximat
  <U who="Subject" id="8">That's right, yes.</U>
<U who="Interviewer" id="9">And do you know approx
  <U who="Subject" id="10">Oh, maybe 21, 22.</U>
```

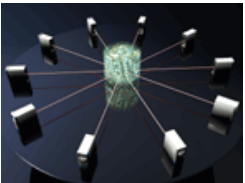
- but linear structural mark-up (e.g. TEI) not suitable for coding as codes may overlap



# Enhanced content-level mark-up

- defining start and end points for segments within a file and assigning values to those segments or to entire files.
- assigned values may be further arranged in a hierarchical structure
  - text
  - images
  - audio
  - video
  - links to external sources, URIs etc

Some examples →



# Value Hierarchy (e.g codes arranged in a coherent hierarchical structure)

LP: There's just one or two factual things first of all do you mind my asking how old you are?

G24: 49.

LP: And what schools did you go to?

G24: King Street, Woodside and Hilton.

School

LP: Uh-huh ... and how old were you when you left the school?

G24: 14.

LP: And you work at the moment? What sort of work do you do?

G24: Well I've gone back to get shorter hours, I've went back to domestic, which I dinna really care for. But then I used to be in the pharmacy department at ARI ... just pharmacy assistant. At least it was better than cleanin'! But then they've nae part-time workers there so...

LP: And did you work in the pharmacy long?

G24: I was there for eleven years.

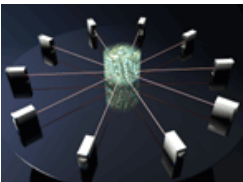
LP: And did you have any other sort of jobs?

G24: Where? Since I left school, like? Well, when I first left school I was just a shop assistant in a number of shops like Reid and Pearsons, which is... we hinna got it ony mair.

Current Work

JOB

First Job



# MEMOS: Assign Notes or Comments (e.g. to a segment or a code)

0008 Date of birth : 1902  
0009 Gender : M  
0010 Marital status : Married  
0011 Occupation : Postman  
0012 Geographic region : Colchester, Essex  
0013  
0014 I : I'd like to start, if I may, by asking you your birth date.  
0015 K : November 9th, 1902.  
0016 I : Could you tell me how many children there were in your family?  
0017 K : There were 11 of us. I was the eldest.  
0018 I : Could you tell me, if you remember, how they went after that and roughly the  
0019 space between them and whether they were boys or girls.  
0020 K : Well, the first 3 of us were boys, then I had a sister, another brother, three more  
0021 sisters and twin brothers at the end.

**ME:ME - 30/07/07 {1-Me} - Super**  
this date might be significant

ME - 30/07/07



## File Hierarchy/file classification (e.g. files arranged in a coherent hierarchical structure)







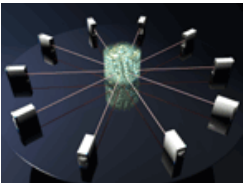
Male Family

-  Interview 1
-  Interview 2
-  Interview 3
-  Interview 4



Female Family

-  Interview 1
-  Interview 3
-  Interview 2
-  Interview 4



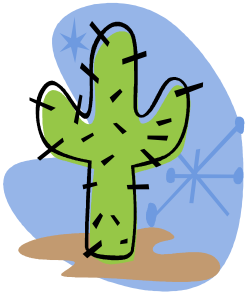
# CAQDAS: What does the software do?

- popular analysis softwares used by researchers. They support a common range of functions:

- coding
- searching
- memoing
- variables/attributes
- grouping codes and
- documents

- but they are all proprietary
- presents a problem for sharing

- Atlas-ti
- HyperResearch
- Max-QDA
- NU\*DIST 6
- N\*VIVO 2
- QDA Miner
- QUALRUS
- Weft QDA



Computer Assisted Qualitative Data Analysis

# Atlas-ti text coding

The screenshot shows the ATLAS.ti software interface. The main window displays a text document with the following content:

Study Name: SN 2000 Family Life and Work Experience Before 1918  
Interview number: 2000int001  
Interview ID: Mr. Keble  
Depositor: Paul Thompson  
Version: rekeyed from PDF scan of paper; not anonymised

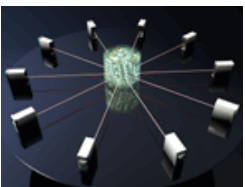
Date of birth : 1902  
Gender : M  
Marital status : Married  
Occupation : Postman  
Geographic region : Colchester, Essex

I : I'd like to start, if I may, by asking you your birth date.  
K : November 9th, 1902.  
I : Could you tell me how many children there were in your family?  
K : There were 11 of us. I was the eldest.  
I : Could you tell me, if you remember, how they went after that and roughly the space between them and whether they were boys or girls.  
K : Well, the first 3 of us were boys, then I had a sister, another brother, three more sisters and twin brothers at the end.  
I : So you were approximately 7 boys, is that right, and 4 girls?  
K : That's right, yes.  
I : And do you know approximately how old your parents were when you were born?  
K : Oh, maybe 21, 22.  
I : And when the last child was born?  
K : Oh, I suppose they were 45.  
I : Did they lose any?  
K : Yes, we lost a sister at school leaving age. Well, she was about 12. And I lost my fourth brother from sclerosis about 3 or 4 years ago.  
I : Your mother didn't lose any as babies?  
K : No, not to my knowledge, no.  
I : Can you tell me what your father's job was?  
K : Well, my father was a farm worker. He worked on a farm at St. Osyth Wick for thirty-odd years.  
I : Can you describe to me what particular kind of farm work he did?  
K : He was a horseman. You see, all the farms then were ... all the big machines were horse-drawn. There was no mechanical machines, not in those days, so it was all horsedrawn traffic, you see, and he was employed as a horseman.  
I : Do you have any idea, roughly speaking, what he was paid?  
K : Well, I can remember him, in my early days, 12, 6d. a week, and that went up to about 14/-, but that was the earliest recollections.  
I : That was when you were a small child that you remember that?  
K : Yes, yes, when I can first start remembering things, you see, I only remember him earning about 12/- a week.  
I : Did your mother ever have any work outside the house?  
K : Oh, very occasionally. You see occasionally she would do a little bit of occasional work on a farm, you know, like picking up potatoes or things like that. Just occasional ... very occasional, because she had plenty to do without that. But there were occasions like the potato season they would go, most of the women went out odd days digging up potatoes.  
I : Did she do any other kind of ... you know, fruit picking, or pea picking?  
K : Well, pea picking, and of course, she did take in a little bit of washing for the moneylender's wife.  
I : And the potato picking, how many days at a time would she go out?  
K : Oh, 3 or 4 during the season. Very seasonal, of course.  
I : What I'm trying to get at is would she be out at work for several weeks at a time?  
K : Oh, no, only odd days. Only odd days.  
I : And the washing she took in, how regular was that?  
K : Once a week. These people lived about 5 miles from where we lived, you see

On the right side, a coding scheme is visible with the following codes:

- ME - 30/07/07
- Father's employment
- mother's employment~

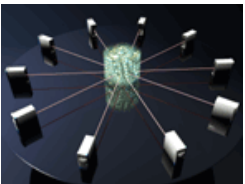
The status bar at the bottom indicates: Loaded PT: P 1: 2000int001.txt, G:\Quads\MembersOnly\DEXT\_sample\_data\sn2000\_v4\ingested\data\txt\2000int001.txt





# The problem with CAQDAS

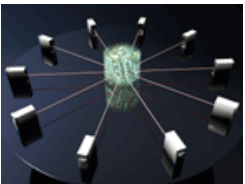
- CAQDAS software use different methods to store links between annotated data and annotations
- in the past, general unwillingness to build import or export functions. No interoperability. You buy into one software:
  - expensive or dependent upon limited University site licences (typically one product only)
  - steep learning curve for most
- VERY recent efforts by vendors to export in XML





# UKDA's basic needs

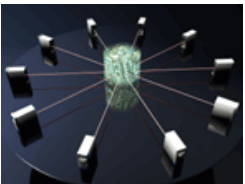
- robust pointing system to relate segments of text, and audio-visual to codes, researcher annotations, keywords and other linked sources
- XML model for data exchange
- vendor-neutral format
- open standard and format for preservation
- back-end system for the management of:
  - study and case files
  - associated documentation
  - metadata enrichment





# The solution: our basic needs

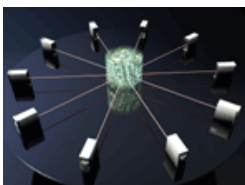
- vendor-neutral format in XML: **QuDEx**
- standard for the management of **METS**
  - study and case files
  - associated documentation
  - metadata enrichment
- project utilised XML consultants who have built and advised on many international schemas and related tools, eg DDI, EBML, SDMX





# the QuDEx Schema

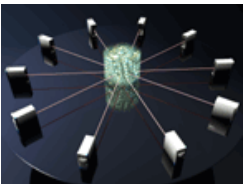
- initially working with XML output from 2 CAQDAS vendors: Atlas-ti and QDAMiner. Keep simple!
- methodology uses embedded segment identifiers pointing to external files
- existing solutions considered:
  - SMIL (Synchronized Multimedia Integration Language)
  - QDIF (Qualitative Data Interchange Format)
  - TEI (Text Encoding Initiative)
  - MPEG – 21 (Moving Picture Experts Group)

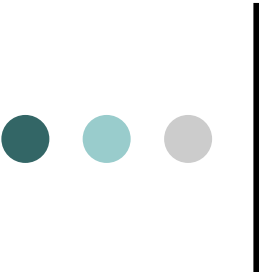




# QuDEX structure: elements

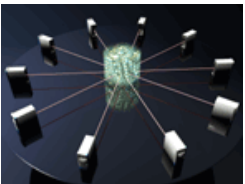
- **resourceCollection**: lists and collates all content
- **segmentCollection**: subset of a document (text, AV etc). May overlap
- **codeCollection**: codes that can be assigned to a segment or document. Can be controlled vocabulary
- **memoCollection**: internal or external notes as a text string. Can be assigned to document, segment, code or category
- **CategoryCollection**: a string assigned to one or more documents. May be nested
- **RelationCollection**: relationship between two objects (all)





# Archival file management: metadata for a whole study

- deal with complex qualitative studies which may consist of multiple data files of different types:
  - interview texts
  - audio recordings
  - photographs
  - textual field notes
  - video capture
  - survey data
- only selected parts may have been analysed in a CAQDAS package, and the rest remains in its raw format
- need a way to represent the whole collection for longer term preservation in addition to QuDEX
- and document how each part is related to other parts e.g. how a single case may have text, audio and image data associated

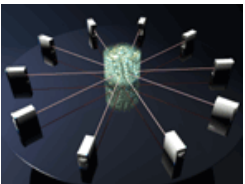




# METS

- METS was chosen to describe the structure and to package all the files relating to a study
- XML standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library
- is 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 can point to other XML schema already in use for the study, e.g. DDI, TEI, DC and MODS

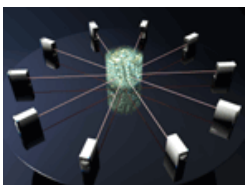
<http://www.loc.gov/standards/mets/>





# Progress

- Version 3 of QuDeX Schema released  
[www.data-archive.ac.uk/dext/schema](http://www.data-archive.ac.uk/dext/schema)
- XML schema, full documentation, UML model and examples of XML and transformations
- feedback from vendors
- basic Java viewer for QuDEX
- final report to JISC



UKDA\_DEXT\_VIEWER - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://localhost:8080/viewer/qudex.jsp#

ESDS Lucene http://oai.esds.ac.uk... TinyURL! Local DEXT GUI

Google Actstone Blueone download Search Bookmarks PageRank Check AutoLink AutoFill Send to Actsone Blueone download Settings

Google Mail OpenXML Developer: Creating an Op... Have french characters in web applica... UKDA\_DEXT\_VIEWER Study Files

File Edit View Tools Help

C:\Documents and Settings\abhat\Desktop\Sample\_files\_Qudex\SAMPLE QUDEX\QuDEX\_v03\_00.xml

Expand All Contact All

- segments
  - segment 1
  - segment 2
  - segment 3
  - segment 4
  - segment 5
  - segment 6
  - segment 7
- codes
  - part-time work
  - full-time work
- memos
  - memo 1
  - memo 2
- categories
  - female category memo
  - pre 1900
  - male pre 1900 category
- relations
  - codeCode
  - segmentSegment
  - memoSegment

Treetview: Click QuDEX Elements to view details

code : A short alphanumeric string; usually a single word which may be assigned to a segment, document or memo, though it is optional.

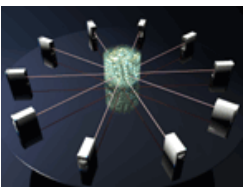
Attribute Name	Attribute Value
id	code_1
cdate	2007-11-15T11:56:38
mdate	2007-11-15T11:56:38
creator	Angad
label	part-time work
displayLabel	part-time work code
language	en
authority	UKDA

For detail information about QuDEX model click here [Schema](#)

Find: xp Next Previous Highlight all Match case Phrase not found

Done zotero

start Inbox - Microsoft Of... UKDA\_DEXT\_VIEWE... Interviews SpecFunct J2EE - CodeFamilies... Document2 - Micros... shortcuts EN 10:36



UKDA\_DEXT\_VIEWER - Mozilla Firefox

File Edit View History Bookmarks Tools Help del\_jcio.us

http://localhost:8080/viewer/xml.jsp

ESDS Lucene http://oai.esds.ac.uk... TinyURL! Local DEXT GUI

Google Actstone Blueone download Search Bookmarks PageRank Check AutoLink AutoFill Send to Actstone Blueone download Settings

Google Mail OpenXML Developer: Creating an Op... Have french characters in web applica... UKDA\_DEXT\_VIEWER Study Files

File Edit View Tools Help

C:\Documents and Settings\labhat\Desktop\Sample\_files\_Qudex\SAMPLE QUDEX\QuDEx\_v03\_00.xml [Text view](#) | [Expand All](#) | [Contact All](#)

```

qudex
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.data-archive.ac.uk/dext/schema/draft http://www.data-archive.ac.uk/dext/schema/draft/QuDEx_v03_00.xsd"
  xmlns="http://www.data-archive.ac.uk/dext/schema/draft"
  cdate="2007-11-15T11:56:38"
  mdate="2007-11-15T11:56:38"
  creator="Angad"
  label="Qudex Example"
  displayLabel="Qudex based xml example showing all the elements and attributes present in schema model "
  status="testing"
  language="en"
  id="qudex_02_00_full"
  <!-- resources -->
  resourceCollection
    id="rc_1"
  <!-- segments -->
  segmentCollection
    id="scol_1"
  <!-- Codes -->
  codeCollection
    id="ccol_1"
  <!-- Memos -->
  memoCollection
    id="mcol_1"
  <!-- Categories -->
  categoryCollection
    id="catcol_1"
  <!-- Object relations -->
  relationCollection
    id="relcol_1"

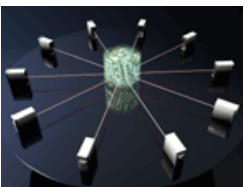
```

[For detail information about QuDEx model click here Schema](#)

Find: xp Next Previous Highlight all Match case Phrase not found

Done zotero

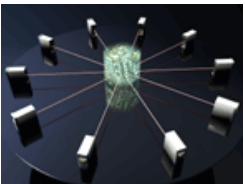
start Inbox - Microsoft Of... UKDA\_DEXT\_VIEWE... interviews SpecPunct J2EE - CodeFamilies... Document2 - Micros... shortcuts EN 10:37





# A home for the standard

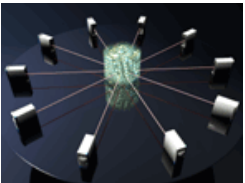
- want other data producers/archives to take up the standard
- need mechanism for greater feedback on model and technical possibilities
- need a well respected home for the standard and associated tools
- and the capacity for refining/nurturing of the standard
- need vendors to buy and build XML export (and import) tools
- now a **DDI Specialist Working Group**





## 2. DExT: statistical survey data

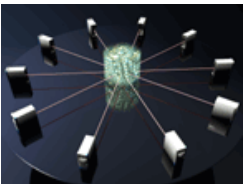
- similar problem: need non-proprietary format
- greater interoperability than CAQDAS as existing exchange format (SPSS.POR) but this is not truly open nor vendor neutral
- need to capture existing data conversion errors (truncation, rounding, corruption etc. between softwares e.g. SPSS to STATA)
- need to easily create study, file and variable level XML metadata (DDI) from common analysis format
- no existing reliable tools available





# Examples of errors

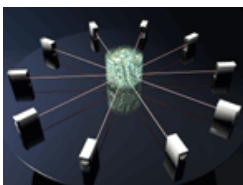
- SPSS: the command 'PRINT FORMATS' often used to perform data typing upon conversion
- but rarely matches the actual data: can lead to coarsening of data upon conversion or inflation of file size
- MS Access: export precision can be controlled by the number of decimal places in the 'Regional Options' of the Windows Control Panel..little known
- MS Access: embedded characters such as 'tabs' or 'carriage returns' in fields: unless these are stripped out prior to conversion to delimited text, the data will lose its rectangular structure





# Progress

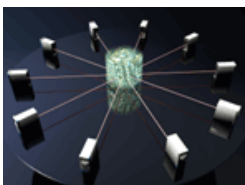
- worked with same consultants (Open Data Foundation (ODaF) to create a java-based data conversion tool
- proof-of-concept:
  - Input: SPSS format
  - archive-neutral format: fixed ASCII for data and the Data Documentation Initiative (DDI) 3.0 XML specification for documentation
  - output: SAS, Stata and SPSS formats
  - conversion error report
- tool built so that it is completely extensible to conversion to and from other data formats e.g. setup file formats for R, Excel, and SQL databases
- been tested by some of the national data archives





# Programming

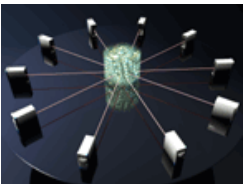
- package developed in Java programming language using the Eclipse Integrated Development Environment (IDE)
- conversion of the DDI-XML metadata into setup files for the statistical packages were developed using the XSLT v2.0 language and processed by Saxon XSL v8.9
- tested on various platforms – Windows, Linux, and Solaris
- utilised ODaF's Guidelines for Tools Development and Recommendations for Operating Environment"  
<http://www.opendatafoundation.org/?lvl1=resources&lvl2=papers>

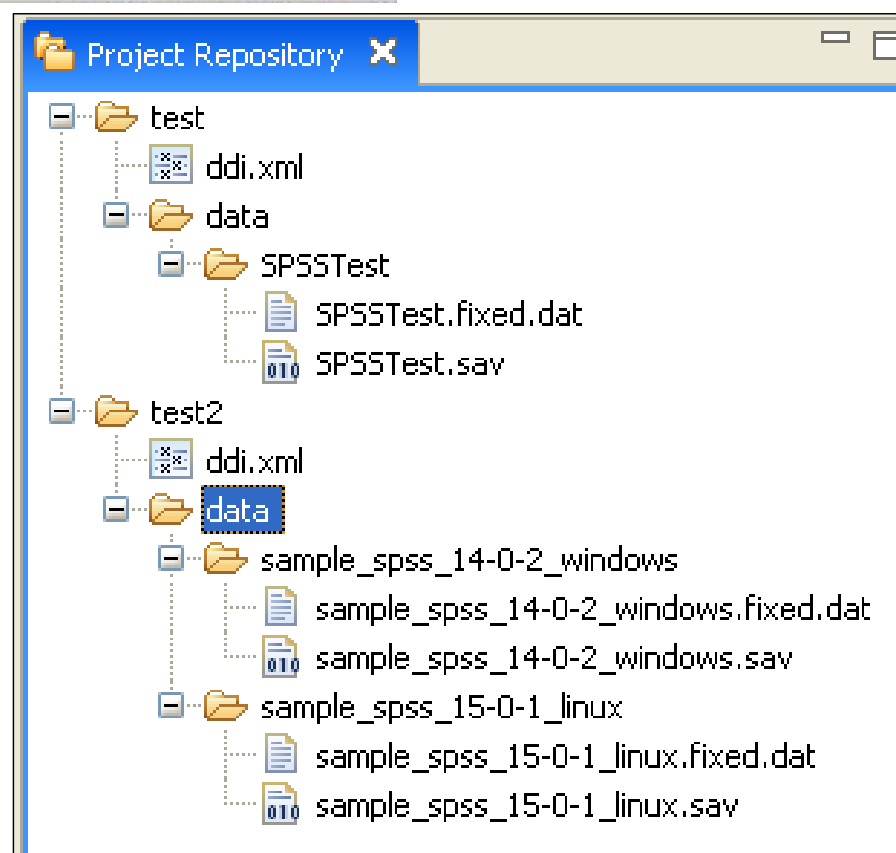
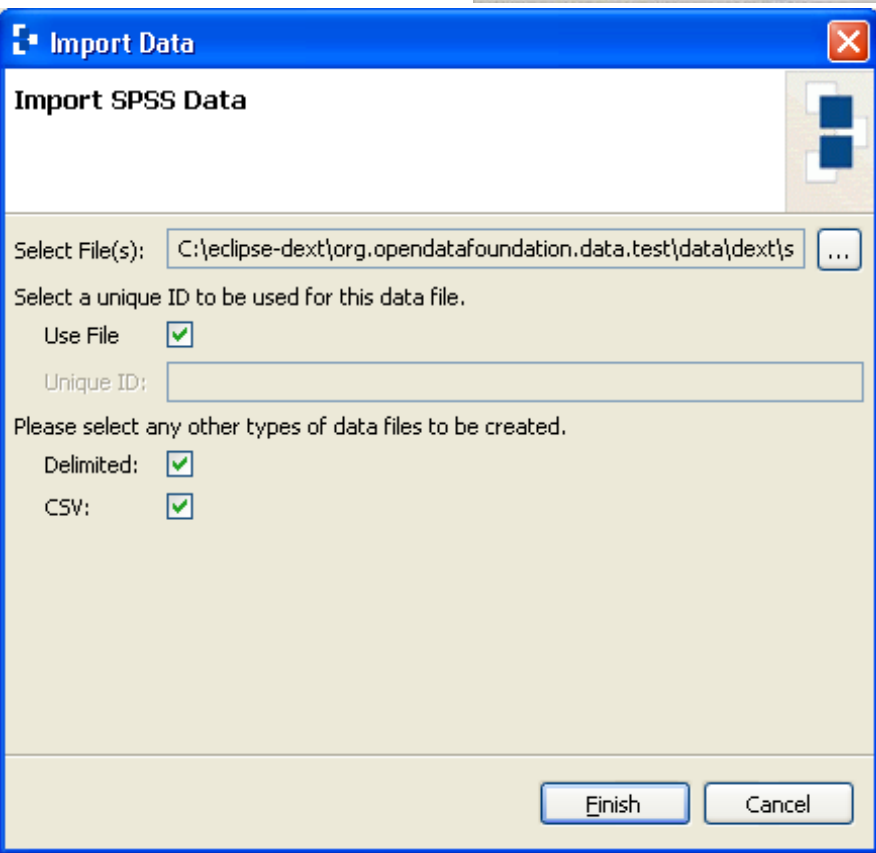




# Availability

- DDI-DExT installer and SPSS Reader can be downloaded from the DDI Tools web site  
<http://tools.ddialliance.org/?lvl1=product&lvl2=dext>
- fully documented
- source code is published and maintained in the ODaF Forge public repository  
<http://forge.opendatafoundation.org>
- distributed under GNU Lesser General Public License





## Export Wizard

### Export Wizard

Select the export format for the setup scripts.

#### Data / Metadata Export

- Fixed  Comma Separated (CSV)  Delimited  Tab  Other
- Include original data file  DDI 3.0  DDI 2.0

#### Setup scripts

Stata/IC (standard) ▼

7 ▼

Add

SAS(9)  
STATA(7)

Remove

#### Output options

- Compress output  Save report

Save

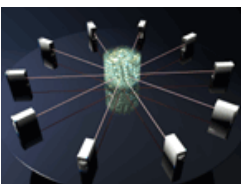
Load

< Back

Next >

Finish

Cancel



test SPSSTest ASCII Data    test SPSSTest SPSS Data

### Physical Instance

ID	ID_1ea650fb-6c4e-4635-b89f-1bf4b2084d6a_Phylins_SPSS
Data File	file:/C:/eclipse-dext-runtime/test/data/SPSSTest/SPSSTest.sav
# of cases	12

### Proprietary Record Layout

SPSS10.1 [@(#) SPSS DATA FILE MS Windows Release 11.0 spssio32.dll]

Property	Value
Compression	1
CompressionBias	100.0
MachineCode	720
FloatingPointRepresentation	1 [IEEE]
Endianness	2 [Big endian]
CharacterSet	2 [7-bit ASCII]
Sysmiss	-1.7976931348623157E308
HighestSysmissRecode	1.7976931348623157E308
LowsetSysmissRecode	-1.7976931348623155E308

### Record Layout

ID	Name	Label	Representation	Type	Format	Properties
V1	NUMERIC	Numeric variable	<a href="#">Code Scheme [3]</a>	numeric	F8.2	Width=8, Decimals=2, MissingFormat=3, MissingValue0=1.0, MissingValue1=2.0, MissingValue2=3.0, DisplayWidth=8, Alignment=Center, Measure=Scale
V2	NUMER16	Numeric 16.2	<a href="#">Code Scheme [5]</a>	numeric	F16.2	Width=16, Decimals=2, MissingFormat=-2, MissingValue0=1.0, MissingValue1=5.0, DisplayWidth=8, Alignment=Center, Measure=Scale
V3	NUMER16B	Numeric 16.0	<a href="#">Code Scheme [6]</a>	numeric	F16.0	Width=16, Decimals=0, MissingFormat=-3, MissingValue0=1.0, MissingValue1=5.0

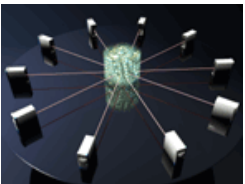
Data Dictionary    Data View

Metadata for a converted fixed ASCII file



# Next steps

- more testing in-house
- integrate into data preservation and dissemination workflow (OAIS)
- other partner archives testing
- seek opportunities to extend development to include other data formats
- could well be an On-the-fly data conversion web service
- may be taken up by international DDI Tools initiative
- working with Dutch team to look at open formats for databases and spreadsheets





## UKDA contacts

- Louise Corti [corti@essex.ac.uk](mailto:corti@essex.ac.uk)  
Qualitative data work
- Matthew Woollard [matthew@essex.ac.uk](mailto:matthew@essex.ac.uk)  
Structured data work

