

UK Data Archive
www.data-archive.ac.uk



Economic and Social Data Service

The UK Data Archive and the experience of digital preservation (the first 35 years)

K. Schürer

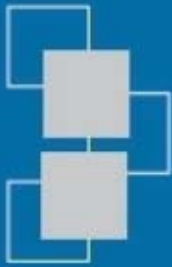
Director of UKDA and ESDS

Supported by:



Joint Information
Systems Committee

DPC Forum, TNA Kew
24 September 2003



Brief history & overview

- Archive established in 1968 (as 'Data Bank')
 - See EY2/34
- Funded by (then) SSRC to provide a service to UK HE sector
- Initial focus on government survey data
- New distributed service established 1 Jan. 2003
 - Economic and Social Data Service (ESDS)
- Mixed data types and formats
 - Specialist Qualidata unit and History Data Service
- Still predominately funded to provide service for HE/FE sectors
 - ESRC, JISC, University of Essex
 - Project funding (EC, JISC, MRC, AHRB, etc.)

Supported by:

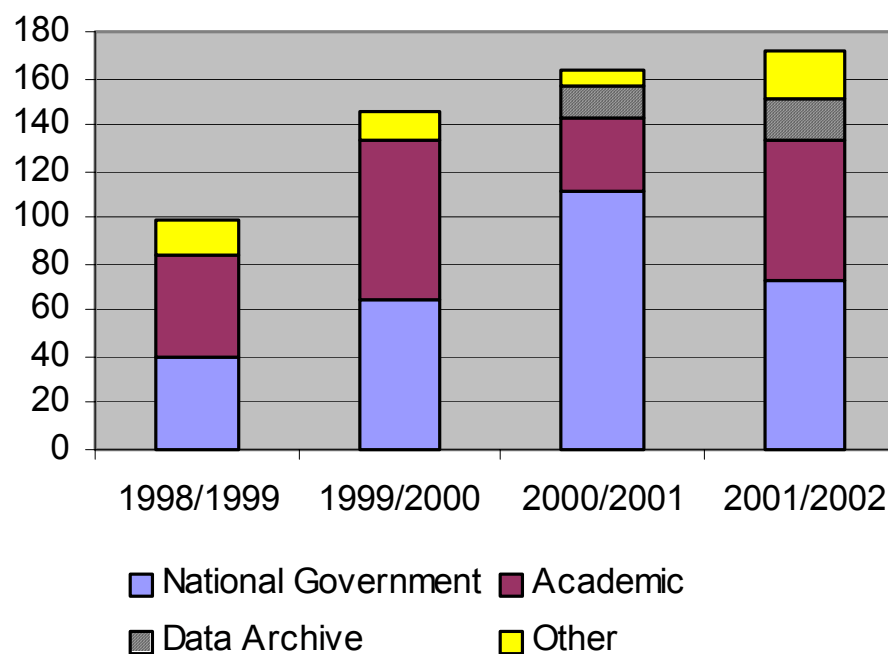


Joint Information
Systems Committee



Data In

Figure 3. New Acquisitions by Source



Supported by:

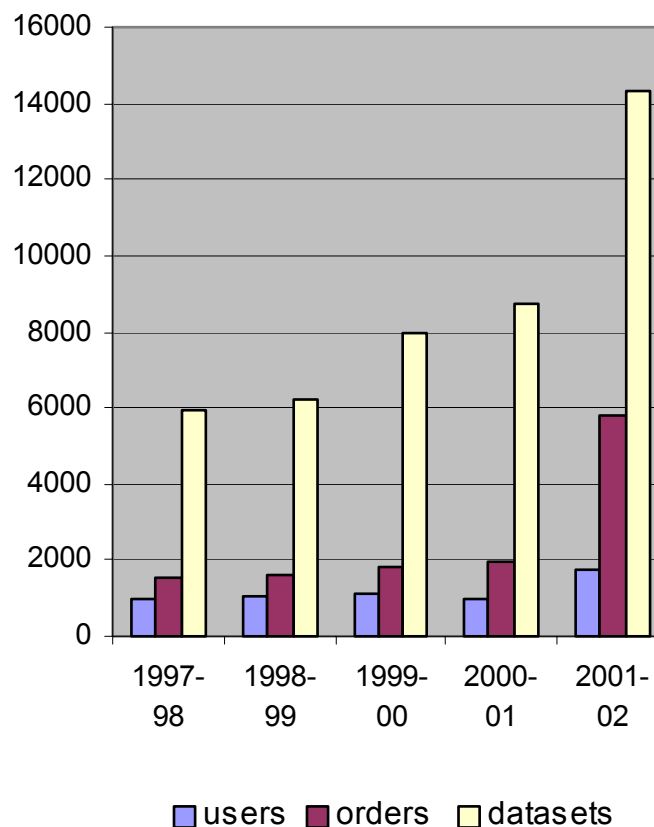


Joint Information
Systems Committee



Data Out

Figure 6. Users, orders and datasets



Supported by:



Joint Information
Systems Committee



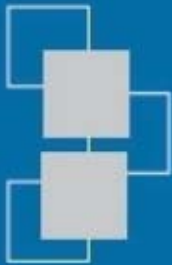
Top titles....

Rank	Title
1	Quarterly Labour Force Surveys
2	General Household Survey
3	Family Expenditure Survey
4	British Household Panel Survey
5	British Social Attitudes Survey
6	Eurobarometer Survey Series
7	1970 British Cohort Study (BCS70)
8	Health Survey for England
9	Workplace Employee Relations Survey
10	1981 Census
11	Family Resources Survey
12	British Election Studies
13	Youth Cohort Study of England and Wales
14	National Child Development Study and BCS70
15	National Child Development Study
16	Northern Ireland Family Expenditure Survey
17	Road Accident Data
18	Continuous Household Survey
19	British Crime Survey
20	ONS Omnibus Surveys

Supported by:



Joint Information
Systems Committee



Acquisition

- Data received (and disseminated) in early years by punch cards and magnetic tapes
- Magnetic tape still major media up to early/mid. 1990s
 - FTP
 - ZIP/JAZ cartridge
 - CD-ROM/DVD
 - DAT
 - Floppy disc
 - Exabyte
 - Email attachment
 - 9" reels
 - Punch cards

Supported by:



University of Essex



Joint Information
Systems Committee



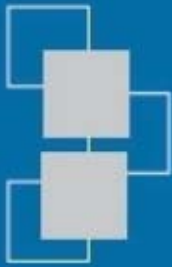
Ingest

Type of Data	Principal Ingest Formats
Alphanumeric data (held in a statistical package)	Delimited text, SPSS portable (.por), SPSS system (.sav), STATA (.dta), SAS (transport file)
Alphanumeric data (held in a database)	Delimited text, MS Access, dBase, FoxPro, SIR (export file), XML, Filemaker Pro, Paradox
Alphanumeric data (held in a spreadsheet)	Delimited text, MS Excel, Lotus, Quattro Pro
Textual data (e.g. transcripts)	MS Word, Rich Text Format (RTF), XML, SGML, HTML, WordPerfect, Plain (unformatted) Text, paper
Video data	MPEG-1, MPEG-4, Apple Quicktime
Audio data	MS Wave, MPEG-3
Raster (bitmap) images	TIFF, PNG, GIF, BMP
Vector images	DXF, SVG, Adobe Illustrator
Documentation/ metadata	MS Word, RTF, XML, SGML, HTML, WordPerfect, Adobe PDF, paper

Supported by:



Joint Information
Systems Committee



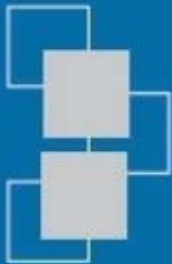
The Data Exchange Initiative

- an XML schema capable of storing any rectangular social survey type of dataset. This will be the first data preservation and interchange standard for social science datasets
- the UKDA plans to define, maintain and promote the schema, and provide software to import into XML from the major proprietary formats (SPSS, STATA, SAS, Excel and Access), and to export back out from the XML to any of these proprietary formats

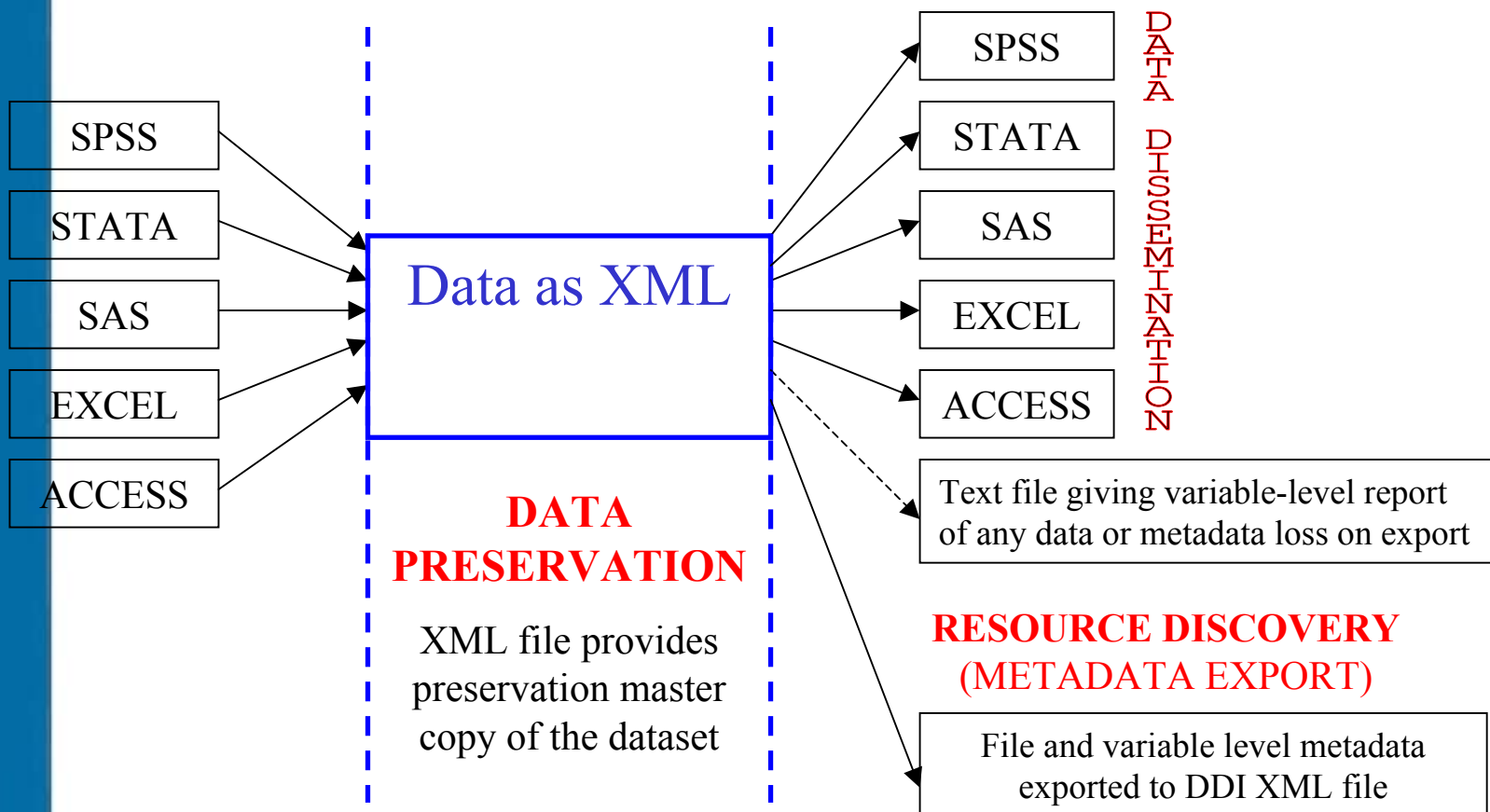
Supported by:



Joint Information
Systems Committee



DATA
AT
INQUEST



Supported by:



Joint Information
Systems Committee

Key functions of the Data Exchange software



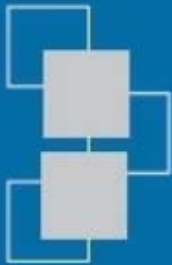
Metadata

Element Name	Comments
Identification	
Study number	<i>UKDA study number</i>
Study title	<i>As specified in the catalogue record</i>
Resource content	
Details of study received	<i>Author, depositor of the study and time of its deposit</i>
<i>Details of data files received:</i>	
File name	
File format	<i>Including version number</i>
Description of file content	
<i>Details of documentation files received:</i>	
File name	
File format	<i>Including version number</i>
Description of file content	
<i>Details of hard copy documentation received:</i>	
Content of hard copy texts	
Duplication with electronic files	<i>List of any file(s) of which hard copy is a printout</i>



Metadata II

Element Name	Comments
Processing history	
File name changes	<i>Files for issue to users as well as files not for no issue</i>
Level of processing	<i>One of the four levels of processing used at the UKDA</i>
Data file conversions:	
Conversion method	<i>Including what software was used for conversion</i>
File format created	<i>Including version number</i>
Validation	<i>Checks as required by each level of processing</i>
<i>Documentation file conversions:</i>	
Conversion method	<i>e.g. scanning of hard copy</i>
File format created	<i>e.g. into PDF</i>
<i>Dataset validation problems:</i>	
Problems encountered	<i>Note of any confidentiality problems, data errors, wild-codes, etc</i>
Solutions found	<i>Details of solutions to above problems</i>
Unsolved problems	
New edition information	
Delivery	
Notes to data delivery	
Notes from data delivery	
Name of the member of staff who processed the note file and date	



Metadata III

3500\dat@0@858166912@858166912@D@NOCRC
3500\dat\fr7172.dat@3209769@991959008@858166850@ 134a3c87faf275aa950a113faa82a40d
3500\dat\fr7172.sps@971@991959019@858166850@ @ce26d1a014925017627e051dfd55438b
3500\dp\fr7172.des@5565@991959178@858166914@ @a179162fcf2ce4a1fe21b763214c638f
3500\dp\fr7172.lst@131161@991959178@858166914@ @e00524d402faa3899acf2f9bd3817b24
3500\exp\fr7172.exp@2759508@991959192@858166934@ @992d4bace55a839ea016170be6f71d70
3500\mrdoc\image@0@865345296@865345296@D@NOCRC
3500\mrdoc\image\3500img.zip@2615815@991959318@865345296@
@3089d663a39dc781772adfdb5bae7ce
3500\mrdoc\pdf@0@915635038@915635038@D@NOCRC
3500\mrdoc\pdf\3500uab.pdf@5413472@991959319@915632856@
@0603e6fce7842f02a1ccffcc2ba39aab
3500\noissue@0@858166987@858166987@D@NOCRC
3500\noissue\dat@0@866023704@866023704@D@NOCRC
3500\noissue\dat\fr-7677@4510505@991959319@858166990@
@d51b6b5242ad3cdb5e37c7fd7dd261d7
3500\note3500.txt@4359@991959007@865431152@ @46f88e215c321110b5741709940f1e63
3500\read3500.txt@1082@991959008@865506394@ @26032b8fc90fa7db11d9257fb1a06979
3500\rf@0@867944847@867944847@D@NOCRC
3500\rf\3500rf.zip@189920@991959335@867944848@ @9016af95b2fd8ab01e0be04f4b18764d

Supported by:



Joint Information
Systems Committee



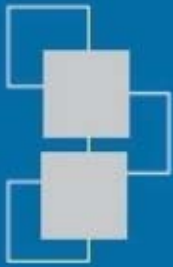
Preservations formats

Type of Data	Preservation Formats
Alphanumeric data (held in a statistical package, database, or spreadsheet)	Delimited ASCII with data definition statements, Tagged ASCII
Textual data (e.g. transcripts)	Rich Text Format (RTF), SGML, XML, Plain (unformatted) ASCII
Video data	MPEG-1, MPEG-4
Audio data	Currently holding MS Wave, MPEG-3 (preservation standard under review)
Raster (bitmap) images	TIFF
Vector images	DXF, SVG
Documentation/metadata	RTF, TIFF, XML, SGML, Plain (unformatted) ASCII

Supported by:



Joint Information
Systems Committee

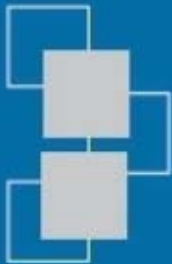


Multi-copy, multi-storage media, multi-site, multi version resilience

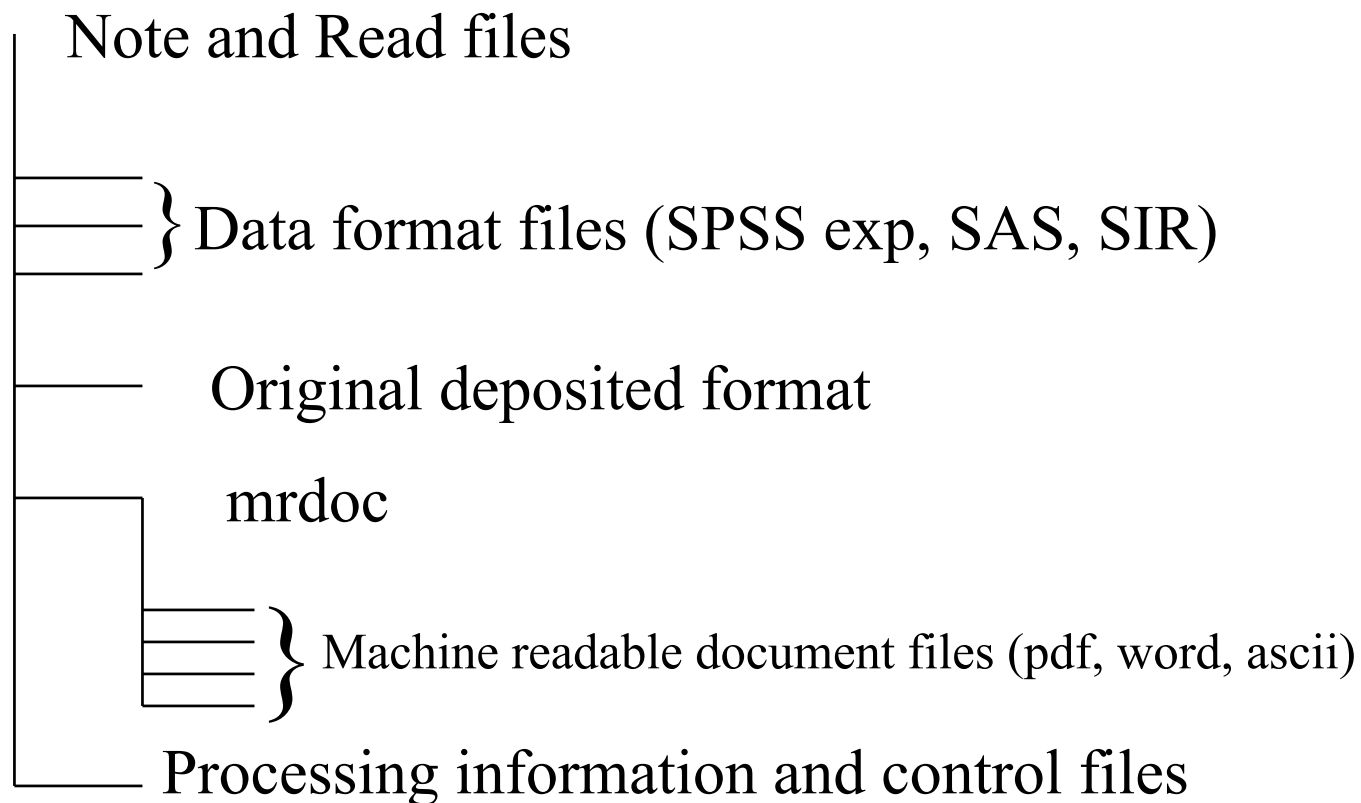
- ◆ Two copies on separate media in main system
- ◆ Up to 10 different versions of each individual file in the shadow area
- ◆ Read only CD-ROM copy with error checking
- ◆ Complete off-site near-line copy of all data with a high level security protection
- ◆ Tape monitoring and refresh strategy
- ◆ Front end copy to reduce load on main system

Supported by:





🌀 Study Number



Supported by:



Joint Information
Systems Committee

Consistent Directory Structure (simplified)