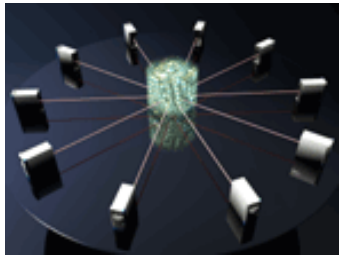
Data Exchange and Conversion Utilities and Tools (DExT)

Angad Bhat, Louise Corti, Herve L'Hours

DExT project, UK Data Archive



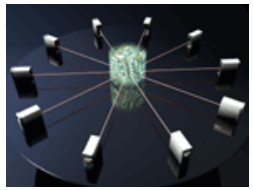
ODaF Meeting, 13-15 September 2007, USA

ACS Meeting, 12-13 September 2007, UK

UK Data Archive

Introduction to DExT

- data exchange models & Data conversion tools for primary research data collected in the course of qualitative research
- a standard format for representing richly encoded qualitative data
- small budget for one year – proof of concept
- developing, refining and testing models for data exchange for qualitative research data based on a combination of existing and internationally recognised schema
- test data selected are from the social sciences (multimedia, linked, annotated data etc.), but these formats are typically found across all domains of primary research

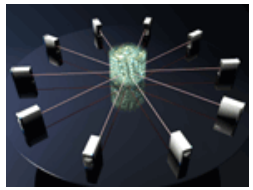


UK Data Archive

Project Environment

JISC

- funded by JISC (Joint Information Systems Committee) under the Repositories Programme
- “provides world-class leadership in the innovative use of ICT to support education and research”
- funds UK national services, programmes & projects

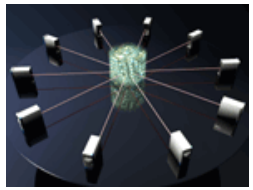


UK Data Archive

Project Environment

UKDA

- the leading UK social science data archive
- pioneered the archiving and sharing of qualitative data
- preserving and disseminating data for 40 years
- offers a robust data service on a national scale with a dedicated infrastructure – ESDS Qualidata

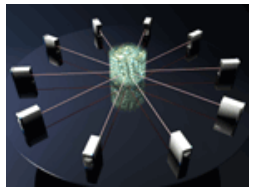


UK Data Archive

Project Environment

ESDS Qualidata

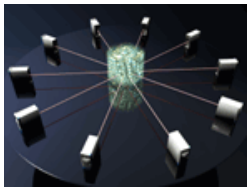
- provides access and support for a range of social science qualitative datasets
- promotes and facilitates effective use of data in research, learning and teaching
- offers a resource hub via the www.esds.ac.uk delivering support and training in:
  - research project management
  - issues of confidentiality and consent
  - documentation of data for archiving
- committed to creating & disseminating value-added data resources through enriched data context



UK Data Archive

Defining qualitative data

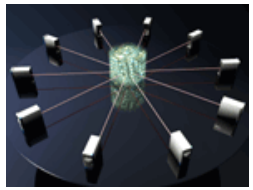
- audio/video tape recordings
- in-depth and semi-structured interview transcripts
- focus groups
- observations and field notes
- unstructured/ semi-structured diaries
- open-ended survey questions
- personal documents and photographs
- records of meetings and case study notes
- collections of press cuttings



UK Data Archive

ESDS Qualidata Technical Environment

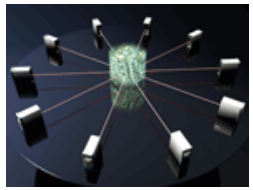
- authenticated data download via web
- online data search and browse facility for selected textual collections



UK Data Archive

Standard data delivery

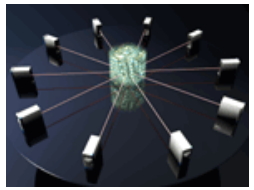
- text delivered via web download as rtf or pdf, depending on level of digitisation
- audio as mp3, or streaming of examples
- video as mpeg4
- behind authentication system



UK Data Archive

Online data browsing system

- enables more precise searching/browsing of archived qualitative data beyond the standard summary record
- allows querying and display of full interview texts across data collections through a standard web browser
- XSL Style sheets to display XML textual documents
- XML texts are currently interviews based on basic TEI mark up
- extending to display audio visual content

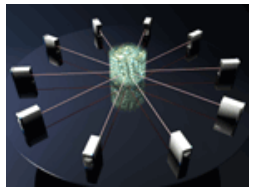


UK Data Archive

ESDS qualitative collections

already utilise known XML schema

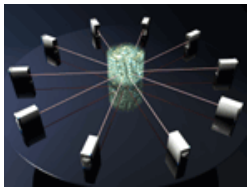
- DC/OAI – basic bibliographic and study description
- DDI2 – study level description
- TEI – content level structural mark-up
  - header
  - interview attributes
  - utterences
  - selected interviewee
  - turn taking
-
- Some fixed vocabulary for qualitative data types, data formats and data collections methods



UK Data Archive

An exchange format for qualitative data

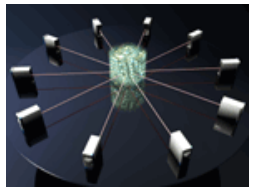
- data exchange models and data conversion tools for primary research data collected in the course of qualitative research.
- a standard format for representing richly encoded qualitative data



UK Data Archive

Qualitative Data Mark up

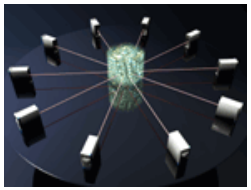
- the process of defining start and end points for segments within a file and assigning values to those segments or to entire files. Assigned values may be further arranged in a hierarchical structure
- initially the mark up (aka coding or annotation) and analysis of qualitative data
- originally textual e.g. interview transcripts
- information technology has been used to facilitate this process
- now expanded to incorporate images, audio and video



UK Data Archive

What is CAQDAS?

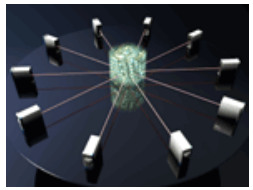
- CAQDAS, Computer Assisted Qualitative Data AnalysiS is a term, introduced by Fielding and Lee in 1991
- refers to the wide range of software now available that supports a variety of analytic styles in qualitative work
- most have been under development for many years



UK Data Archive

CAQDAS: What does the software do?

- most of the popular programs now support a common range of functions:
  - coding
  - searching
  - memoing
  - variables/attributes
  - grouping codes and documents
  - see: http://onlineqda.hud.ac.uk/ for details

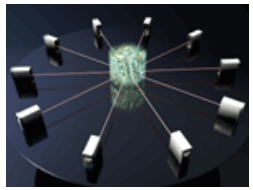


UK Data Archive

CAQDAS: Key functions

- segment: A subset of a file (text, audio, video, image) EXAMPLE 1
- code: A short alphanumeric string (usually a single word) assigned to a segment or file EXAMPLE 2
- hiCode: The top level in coherent hierarchical structure of codes EXAMPLE 3
- fileClass: A short alphanumeric string assigned to one or more files EXAMPLE 4
- memo: A variable length (from a word to a detailed document) alphanumeric string assigned to a a segment or code EXAMPLE 5

or file



UK Data Archive

SEGMENTS: Identify Subsets of the study (e.g. text or line selections)

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.

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?

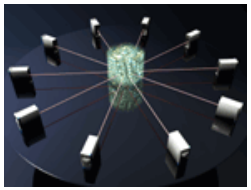
G24: Well I've gone back to get shorter hours, I've went back to domestic, which I dinna really care for. But then I used to be in the pharmacy department at ARI ... just pharmacy assistant. At least it was better than cleanin'! But then they've nae part-time workers there so...

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.



UK Data Archive

CODES: Assign Values to a Subset of a study (e.g a segment)

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?

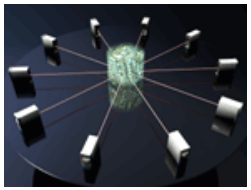
G24: Well I've gone back to get shorter hours, I've went back to domestic, which I dinna really care for. But then I used to be in the pharmacy department at ARI ... just pharmacy assistant. At least it was better than cleanin'! But then they've nae part-time workers there so... — Current Work

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. — First Job



HiCODES: Create a 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.

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?

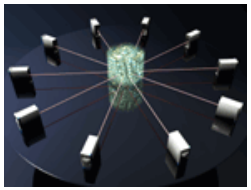
G24: Well I've gone back to get shorter hours, I've went back to domestic, which I dinna really care for. But then I used to be in the pharmacy department at ARI ... just pharmacy assistant. At least it was better than cleanin'! But then they've nae part-time workers there so...

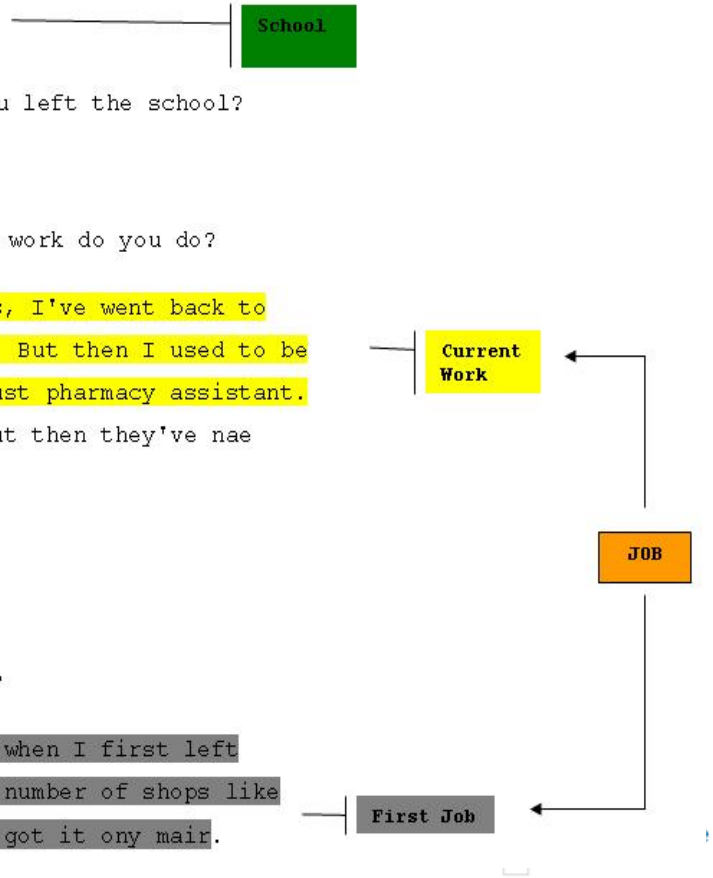
LP: And did you work in the pharmacy long?

G24: I was there for eleven years.

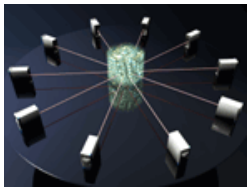
LP: And did you have any other sort of jobs?

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.





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



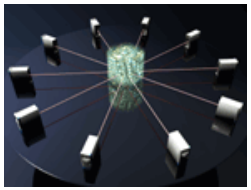
UK Data Archive

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

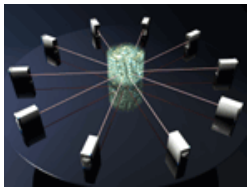
ME - 30/07/07



UK Data Archive

The problem with CAQDAS

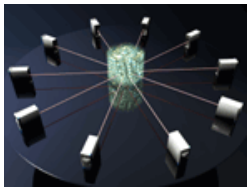
- Large number of programs
  - Atlas-ti
  - HyperResearch
  - Max-QDA
  - NU*DIST 6
  - N*VIVO 2
  - QDA Miner
  - QUALRUS
  - Weft QDA



UK Data Archive

The problem with CAQDAS

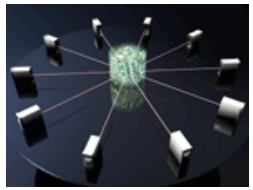
- linear structural mark-up (e.g. TEI) not suitable for coding as codes may overlap
- need robust pointing system to relate segments of text/audio-visual to codes/researcher annotations/keywords
- CAQDAS software use different methods to store links between annotated data and annotations



UK Data Archive

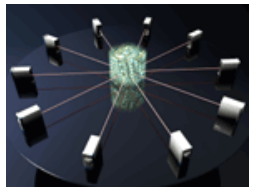
The problem with CAQDAS: For example

- Atlas –ti – links codes to identified segments from the text being analysed
- QDAMiner embeds the XML in the text being analysed
- ‘value-added’ work (mark-up /coding/annotation) that is carried out within the package typically cannot be exported
- neither can previously annotated data from another software be imported
- recent efforts by vendors to export in XML



The solution: our wish list

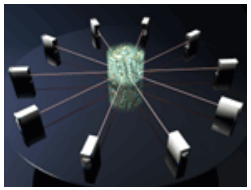
- long-term preservation requirements (software and platform independent formats)
- in-house toolsets for preparing qualitative data for multiple forms of dissemination
- enable ‘added-value’ data to be retained and exchanged e.g CAQDAS-specific functionality
- offers a standard for data creators to store and publish data in multiple formats e.g. web-based publishing
- more precise searching/browsing of archived qualitative data beyond a summary record
- facilitates annotated data exchange and data sharing across dispersed collections and repositories (comparative analysis and e-science)



UK Data Archive

The solution: our basic needs

- ESDS: Vendor-neutral format
- UKDA : System for the management of
  - all study & case files
  - associated documentation
  - metadata enrichment



UK Data Archive

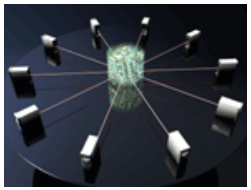
The solution: our basic needs

- ESDS: Vendor-neutral format

QuDEx

- UKDA : System for the management of
  - all study & case files
  - associated documentation
  - metadata enrichment

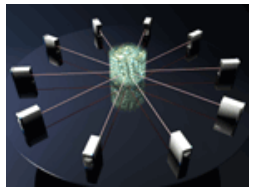
METS



UK Data Archive

Vendor Neutral Format: the QuDEx Schema

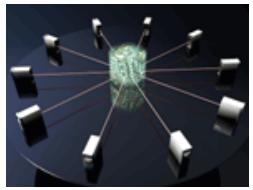
- initially working with XML output from 2 CAQDAS Vendors: Atlas-ti and QDAMiner
- methodology uses embedded segment identifiers pointing to external files



UK Data Archive

QuDEx: Solutions considered

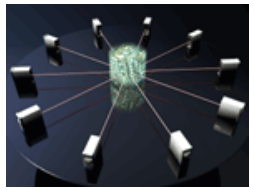
- SMIL (Synchronized Multimedia Integration Language)
- QDIF (Qualitative Data Interchange Format)
- MPEG – 21 (Moving Picture Experts Group)
- TEI (Text Encoding Initiative)



UK Data Archive

QuDEx: Solutions rejected

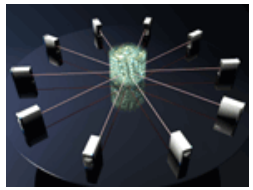
- SMIL
  - no descriptive relationship
  - Flexible but can be complex sometime
- QDIF
  - abstract way of identifying and linking fragments
  - not a good interchange and long term preservation method
- MPEG -21
  - continuous media (audio/video) only & no discrete media
  - hard to identify image and text fragments
- TEI
  - no relationship scheme
  - does not provide line “offsetting”



UK Data Archive

QuDEx: Decisions

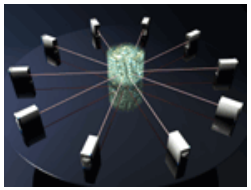
- stand alone, independent schema holding all the concepts with descriptive nature
- simplified XML format for vendors
- contains all key constructs
  - Segment(s)
  - Code(s)
  - Hicode(s)
  - Memo(s)
  - File(s)
- easily interchangeable



UK Data Archive

QuDEx structure

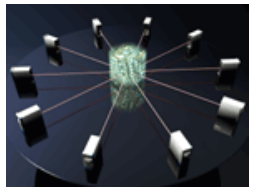
- segmentCollection: contains segments that hold the pieces of text and memo information
- codeCollection: contains codes which can have “segmentRef” to related segments plus a “codeRef” to other low-level related codes (nesting concept)
- hiCodeCollection: contains hicodes which can have “childCodeRef” to subordinate codes (which might or might not have low-level codes)
- memoCollection: contains memos which can have a “memoRef” that could be linked to file, segment, code, hiCode and memo.
- fileClassCollection: contains all the files



UK Data Archive

Archival File Management: Metadata for a whole study

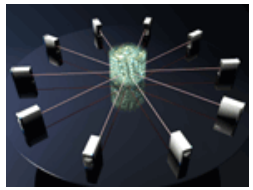
- a qualitative study may consist of multiple data files of different types:
  - interview texts
  - audio recordings
  - photographs
  - textual field notes
  - video capture
  - survey data
- only selected parts may have been analysed in a CAQDAS package, and the rest remains in its raw format
- we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



UK Data Archive

METS

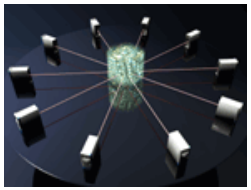
- METS has been chosen to describe the structure and to package all the files relating to a study
- METS is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language
- the standard is maintained in 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 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/



UK Data Archive

Structural Maps

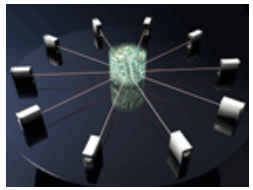
- these are used to split a study in any way; the usual example is by chapter and page. Each split is identified by a tag
- **CAQDAS:** are constructed for Values, Value Hierarchies and File Hierarchies
- **Logical and Physical:** Logical (by section) and Physical (file by file). Structural maps provide a mechanism by which 3rd party programs can access the whole of the original study as well as the vendor-specific markup



UK Data Archive

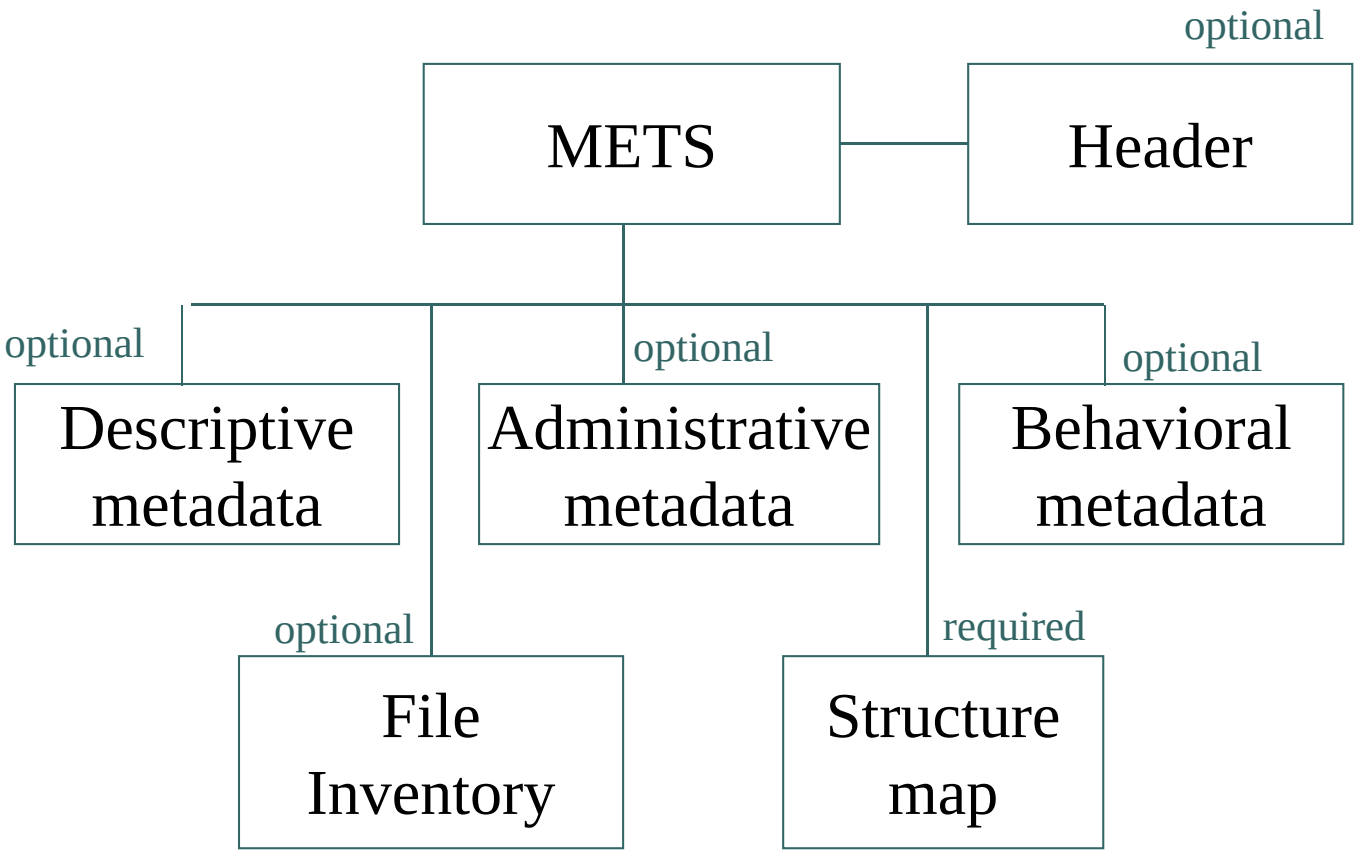
Content Packaging

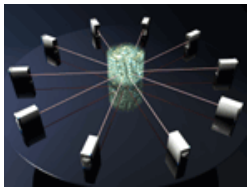
- in addition to a DExT-METS version of the core data concepts the METS file (*METS File Section*) may also retain
  - original files from the study
  - any rtf format versions created for analysis
  - original vendor-specific xml file describing the resource
  - any report output from the vendors program
  - any supporting documentation, notes or content delivered with the study but not part of the core deliverables



UK Data Archive

METS





UK Data Archive

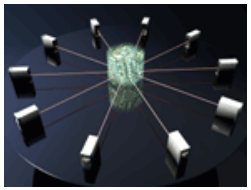
Linking in METS Documents

DescMD
mods
relatedItem
relatedItem

AdminMD
techMD
sourceMD
digiprovMD
rightsMD

fileGrp
file
file

StructMap
div
div
fptr
div
fptr



UK Data Archive

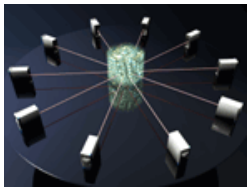
Linking in METS Documents

DescMD
mods
relatedItem
relatedItem

AdminMD
techMD
sourceMD
digiprovMD
rightsMD

fileGrp
file
file

StructMap
div
div
fptr
div
fptr



UK Data Archive

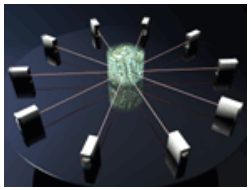
Linking in METS Documents

DescMD
mods
relatedItem
relatedItem

AdminMD
techMD
sourceMD
digiprovMD
rightsMD

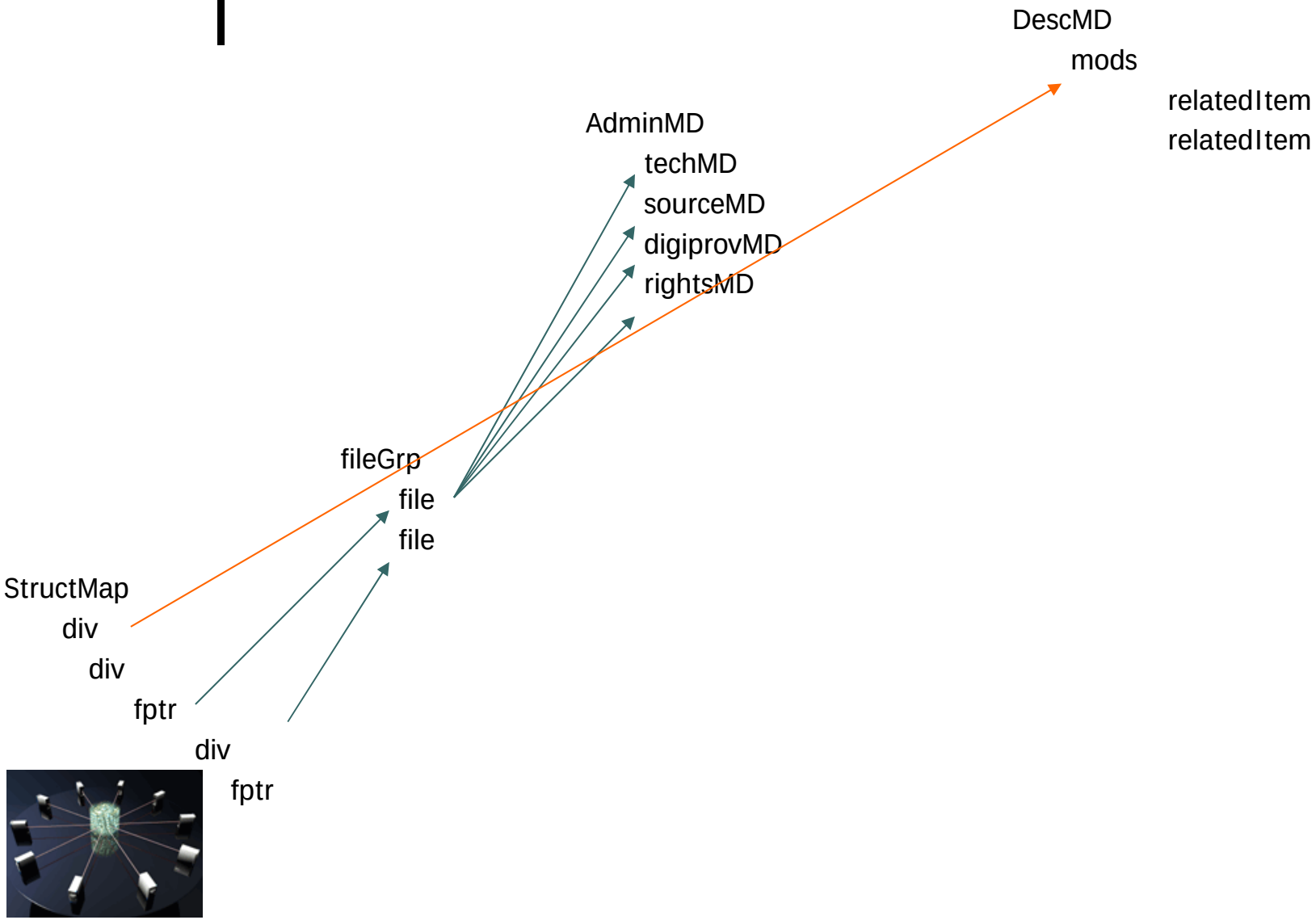
fileGrp
file
file

StructMap
div
div
fptr
div
fptr



UK Data Archive

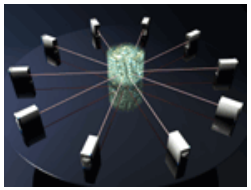
Linking in METS Documents



UK Data Archive

How far have we got?

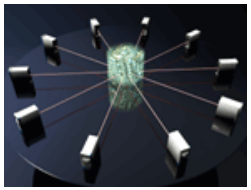
- representative sample dataset
- schema
- sample METS
- UML model
- import GUI plan
- viewer plan under review
- initial meeting with software vendors



UK Data Archive

Next steps

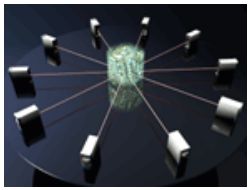
- proof of concept
- import GUI
- review existing tools
- stand alone METS reader
- initial METS profile
- review with vendors
- future of the standard



UK Data Archive

A home for the standard

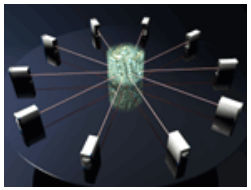
- want other data producers/archives to take up the standard
- need mechanism for feedback on model and technical possibilities
- need a well respected home for the standard and associated tools
- and the capacity for refining/nurturing of the standard



UK Data Archive

Options

- UKDA DExT project extension
- UKDA
- An existing standards body e.g. DDI, OASIS

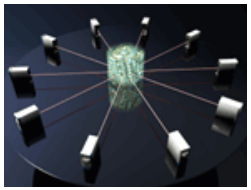


UK Data Archive

Contact

DexT team at Essex

- corti@essex.ac.uk
- herve@essex.ac.uk
- abhat@essex.ac.uk



UK Data Archive