|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **DEXt Bid**  **Copyright, University of Essex**  Name of Capital Programme: JISC Repositories and Preservation Name of Lead Institution: University of Essex Name of Proposed Project: Data Exchange Tools and Conversion Utilities (DExT) Name of Project Partners: UK Data Archive  **Full Contact Details for Primary Contact:**  Name: Louise Corti Position: Associate Director, UK Data Archive Email: [Louise Corti](mailto:corti@essex.ac.uk) Address: University of Essex Wivenhoe Park, Colchester CO4 3SQ Tel No: +44 (0)1206 872145 Fax No: +44 (0)1206 872003  **Project Start and End Dates:**  Length of Project: 12 months Start date: 11 December 2006 End date: 31 March 2008 (extended)  **Total Funding Requested from JISC**  £139,421  **Funding Broken Down over Project Years:**  Year 1 (1 Nov 06 - 31 March 07) £57,640 Year 2 (1 April 07 - 31 March 2008) £81,781  **Outline Project Description**  This project is exploring the feasibility of developing data exchange models and data conversion tools for primary research data collected in the course of empirical research. It is developing, refining and testing models for data exchange for both survey data and qualitative research data based on XML/RDF schema and developing tools for data import and export. The data formats 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 is being undertaken to inform JISC of the most viable options for future development in this area.  **1. Background**  1.1 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.  1.2 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 UK Data Archive 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. This proposal argues that open data exchange formats are necessary for maximising the opportunities for data sharing and long-term archiving.  1.3 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. 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.  1.4. Metadata Encoding and Transmission Standard (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 should therefore be relevant to the development of data conversion standards and tools. The TNA and UK Data Archive recently undertook a JISC-supported project to report on Open Archival Information System (OAIS) Reference Model and METS.  1.5 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.  **2. Overview**  2.1 This bid is for a project under the Tools and Innovation strand, which aims to develop and pilot innovative approaches to repository use and digital preservation through the development of new software and tools. The outputs will also be 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.  2.2 The proposal supports 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 will thus 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 will be an increase in productivity for preservation and data sharing services and the enhancement of both the reprocessing of legacy datasets and the data refreshment element vital to good data preservation practice.  2.3 The proposal builds on development work arising from the existing JISC programmes: Digital Repository Programme and the Supporting Digital Preservation and Asset Management in Institutions. Findings from the JISC projects set out in the Background section 1.0 will be utilised where appropriate. It should be noted that the 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 this proposal. The proposal also builds 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'.  2.4 The UK Data Archive 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 authentication, 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, and will be built upon for user guides and documentation purposes for the conversion tools to be built.  **3. Aims and Objectives**  3.1 This proposal aims 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.  3.2 The specific objectives are 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   **4. Overall Approach**  4 The kinds of data to be dealt with under the remit of this proposal are 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, as set out in Appendix B. The UK Data Archive 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 can be easily read.  **4.1 Numeric data exchange and conversion tools (WP2)**  4.1.1. 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   4.1.2. 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.  4.1.3. The main aim of this strand is to develop an XML standard for data exchange and curation. This standard will be defined and expressed as an XML schema. It will store ALL information in a statistical dataset including the internal metadata (variable and code labels, missing value definitions, variable level notes, variable formatting, etc.). This would be a logical extension to the Data Documentation Initiative (DDI) that is used globally to describe social research data collections, with the DDI acting as the metadata standard for survey datasets and this new standard providing the standard for the data themselves - and thereby providing a much needed open standard for curating statistical data resulting from surveys.  4.1.3.1 Specifically the standard would need to take account of:   * 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   Also the standard would need to take account of any other feature (data type) which might be found in a spreadsheet, statistical package, or relational database.  4.1.4 The resulting standard would be DDI 3 compatible and would need to be able to be used to generate DDI 3 XML files via XSLT stylesheets. The XML schema would be published on the project and ESDS web sites. 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 some pilot import and export utilities will be developed to accompany the new standard.  The basic work plan for this development work is set out in Section 11.  **4.2 Non-numeric data exchange and conversion tools (WP3)**  4.2.1. 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.7 The softwares have similar basic functionality that include:   * 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   4.2.2. 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.  4.2.3. A standard format for representing richly encoded qualitative data is necessary because it: ensures consistency across datasets; supports the development of common web-based publishing and search tools; and facilitates data interchange and comparison among datasets. Importantly, it could also enable data and linked 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.  4.2.4. The work to be undertaken for this strand is a refinement and joining of two new and roughly specified models that have been developed by ESDS Qualidata at Essex and ANU in Australia. 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 has been prepared by ESDS Qualidata. The Universities of Melbourne and Queensland have developed a draft Qualitative Data Interchange Format (QDIF) for e-Social Science (QDIF)10. Both centres have been working closely together in this very early development phase, but as yet, neither have any dedicated funding to work further on realistic development or testing, so the specification work remains on the back-burner.  4.2.5. In essence, the model is aiming at specifying in XML or RDF, a data exchange format that can represent a data collection and retain links between data, annotations and related objects. The model makes use of the DDI and TEI for metadata, and is compliant with Dublin Core and OAI. The model will be tested on various data types including output to a generic archival format and export from (and possibly import to) 2-3 of the brand leaders of the CAQDAS software. Some basic import/export tools will be developed. Agreed testers of the models ad tools are listed in Appendix C.  The workplan for this development work is very similar to WP2, as set out in Section 11.  **5. Project outputs**  5.1 The outputs of the project will comprise a suite of tools for achieving best practice in data format conversions:   * a model for a data exchange format for qualitative data (first instance text - of NVIVO, Atlas ti, MaxQDA, Qualrus, QDAMiner ) and a limited set of import/export tools * import from SPSS and Excel   + these would be using visual basic scripts run from within SPSS and Excel   + these utilities would be available for people to download and execute locally from within the relevant software   .   * export to SPSS, STATA, and SAS   + this would be via XSLT stylesheets and would provide delimited text datafile and accompanying command file(s) to recreate the data in SPSS/Stata/SAS   + these stylesheets would be made available for people to download and execute locally (i.e. no software dependence beyond downloading one of the freely available XSL compilers) * export to DDI 3 metadata files   + this would be via XSLT stylesheets which would be made available for people to download and execute locally, again, with no software dependencies   5.2. The experience of the UK Data Archive/ESDS in this line of work and the synergies this proposal has with existing projects and staff expertise within the organisation can offer great value in terms of promoting and subsequently maintaining the standard and import/export utilities - not least because it is a standard that UK Data Archive will adopt for their own preservation and data exchange work.  **6. Project Outcomes**  6.1 The following outcomes will be sought from the project:   * specification of data exchange models and partner consultation * pilot conversion tools and testing with partners * evaluation of performance and results   6.2. Once these demonstrator tools can be shown to work it should draw attention to the potential for a fully specified statistical data curation/exchange web service. Users could upload data that had been extracted from a software package and return a chosen converted format. This is a logical extension to the work that would be funded within this bid and it would be a highly beneficial low investment tool for institutional repositories. The short time scale and limited budget for this proposal does not allow for full exploration of this kind of facility, but the potential will be discussed in the final report.  **7. Stakeholder Analysis**  *Group*: Researchers across all domains who undertake data collection and analysis using digital formats *Interest*: More research data that is tied up in proprietary or legacy formats will become available. Data and analysis, rather than simply a single dataset will be able to be long-term preserved. Incorrect conversion of survey data will no longer be necessary. Appreciation of metadata will be improved *Value*: Very high  *Group*: Post graduate students *Interest*: As above *Value*: Very high  *Group*: Undergraduate students *Interest*: Low use of primary data in teaching is an issue of concern for some of the Research Councils. More availability of data will be an asset *Value*: Medium  *Group*: Research funding agencies *Interest*: Greater value for money from grants awarded through an increase in the archiving and availability of primary data. Increased re-use and further analysis of datasets whose creation they have funded. Greater accountability for public funds through increase in data usage and citation. *Value*: High  *Group*: Library community *Interest*: Increased recognition by academic and research communities of their libraries' new role in promoting and pointing users to raw data sources for research *Value*: Medium  *Group*: General public *Interest*: Typically the general public do not create or analyse research data *Value*: Low  **8. Risk Analysis**  *Type*: Staff recruitment *Probability*: Medium *Action to minimise*: Immediate widescale advertisement of positions subject to award letter. Head-hunting initiated prior to this *Response to worst case*: Development timeline reassessed if time permits  *Type*: Financial mismanagement *Probability*: Low *Action to minimise*: Part-time Project Manager and PIs with financial responsibility and project meetings Institution liable to under-write losses. Project halted  *Type*: Legal problems *Probability*: Low *Action to minimise*: Project Manager and PIs well-briefed and sensitive to legal environment *Response to worst case*: Advice sought from JISC Legal and ultimately from legal personnel at the University of Essex  *Type*: Technical problems *Probability*: Medium *Action to minimise*: Established track record in technical development. PIs to apply rigorous monitoring *Response to worst case*: Technical development strategy reassessed and new staff recruited, if time and resource permits  *Type*: Staff retention *Probability*: Low *Action to minimise*: PIs to monitor *Response to worst case*: Development timeline reassessed and new staff recruited, if time and resource permits  *Type*: Tester drop out *Probability*: Low *Action to minimise*: PIs to monitor *Response to worst case*: New testers recruited  **9. Standards and Technical Developments**  9.1 The Project will adhere to the JISC Information Environment Architecture Standards Framework. Standards and protocols employed will be open wherever possible. These are likely to include Java, OpenURL, OAI-PMH, SRU/W, XML and SQL. Metadata will conform to Qualified Dublin Core. UK Data Archive uses the Data Documentation Initiative (DDI) and the Text Encoding Initiative (TEI), which both map onto Dublin Core but provide more relevant detailed description.  **10. Project Partnership**  10.1 This is a small project and as such is limited to a single institution. A number of organisations in England and Wales have agreed to be user testers, as set out in Appendix C.  10.2 The UK Data Archive (UK Data Archive) has been providing high quality social science data to the UK academic community for over thirty years. In order to continue improving its ability to satisfy the requirements of the social sciences HE/FE communities, the UK Data Archive has engaged in a sustained, long-term effort to define, design and implement consistent and wide-ranging curation and preservation strategies, procedures and formalised services. UK Data Archive is a centre of expertise in data acquisition, preservation, dissemination and promotion and is curator of the largest collection of digital data in the social sciences and humanities in the UK. Founded in 1967, it now houses several thousand datasets of interest to researchers in all sectors and from many different disciplines. UK Data Archive provides resource discovery and support for secondary use of quantitative and qualitative data in research, teaching and learning. As a lead partner of the Economic and Social Data Service (ESDS) funded by the ESRC, the JISC and the University of Essex. The UK Data Archive also hosts AHDS History11, one of the five Centres of the Arts and Humanities Data Service, and the Census Registration Service (CRS)12, facilitating access to the census data resources for UK higher and further education.13  10.3 UK Data Archive contributes to the development of internationally recognised metadata standards (the DDI standard for statistical metadata) and has led the development of: state-of-the-art distributed statistical dissemination software that has been widely adopted by the Data Archives community (the NESSTAR system); a European-wide data portal (the Madiera portal) and the multilingual social science thesaurus, ELSST (European Language Social Science Thesaurus).  10.4 UK Data Archive is recognised as a legal Recognised Place of Deposit for TNA for public records and provides preservation services for other data organisations, supports the National Centre for e-Social Science (NCeSS)14 and facilitates international data exchange through agreements with other national archives.  10.5 UK Data Archive is involved in Research and Development projects making a significant contribution to new developments in data preservation and dissemination, metadata standards, software for web browsing, data discovery and data delivery.15 UK Data Archive has led the initial development of data exchange formats for both survey and qualitative data and is a project partner in the is JISC StORe (Source To Output Repositories) project16 and the JISC-funded OHPR (Online Historical Reports Project).  **11. Work packages**  WP1: Project Management Objective: to ensure that all the workpackages of the project are managed coherently and that all the project outputs are delivered within the agreed deadlines and budget. Timeline: 12 months  Project Administrator   * contract management * financial management   Project Director   * Detailed workplan * Running project meetings * Production of bi-monthly progress reports and key reporting to JISC * Creation and maintenance of project web site * Management of dissemination in collaboration with WP staff   WP2: Develop technical specification of data exchange models - survey data Timeline: 9 months  WP3 Develop technical specification of data exchange models - qualitative data Timeline: 12 months   * Developing the standard- 6 months * Import/export utilities - 4 months * Finalising demonstrator tools and promoting the model - 2 months   WP4: Testing and Evaluation Objective: to ensure that formative evaluation takes place throughout the project and is reported to JISC, and a summative evaluation is performed to consider recommendations of the project. Timeline: 12 months (formative); 2 months (summative)   * Limited evaluation of demonstrator system based on high quality user testing * Report to Project Committee and JISC on the value of the project, with recommendations for future development work   **12. Project Management**  12.1 The Principal Investigator, Louise Corti will lead the project and manage the two work packages WP3 and 4. They will liaise regularly with staff and meet fortnightly. The project will be overseen and run by the Project Director, Louise Corti. Her time is being offered at no cost to this project. Expert consultancy for digital preservation has been built in for Matthew Woollard, Head of Digital Preservation and Systems for the UK Data Archive. He will lead on WP2 which is to be redefined in scope given the recent death of the fomer PI Alasdair Crockett.  12.2 A small Project Committee will be formed to steer and advise the project. The Project Committee will be composed of people with data curation or data handling expertise in the UK who will meet 3-4 times a year (face-to-face, VC or AGN) to consider progress, advise on quality assurance and provide advice based on leading edge ideas in their fields. They will also help to disseminate the work of the project, particularly among senior staff in UK universities and research agencies, to government, and overseas. The Board will consist of 'research community champions' from the domains represented in our project, together with senior figures from the worlds of digital libraries and data curation. The members are:   * UK Data Archive, Ken Miller, Head of Information Development and e-Social Science * UK Data Archive, Matthew Woollard, Head of Digital Preservation and Systems * Data Curation Centre (DCC), Chris Rusbridge * A representative from the statistical community (tba) * University of Queensland - Andrew Smith   **13. Evaluation**  13.1 Formative evaluation will be the responsibility of the PIs. Summative evaluation will be undertaken in the final two months of the project to assess performance and achievements against the initial objectives.  **14. Quality Assurance**  14.1 Quality assurance of the technical workpackages will also be the responsibility of the PIs, who will set requirements for documentation, usability and accessibility. Advice on all aspects of quality assurance will be provided by the Project Committee.  **15. Dissemination**  15.1 Project personnel will look for opportunities to present the work of the project at conferences and to publish papers in appropriate professional journals - including the academic journals of our researcher groupings. Wherever possible, academic colleagues will be encouraged to publicise the work of the project to their peers. The Project Committee will provide strategic-level dissemination. UK Data Archive will host a project web site that will provide copies of all project publications, reports and presentations.  **16. Exit/Sustainability and IPR**  16.1 This project seeks to undertake groundwork preparatory to further infrastructure development. It will conclude with recommendations for a greater promotional effort towards uptake of the data standards and a full-scale data conversion facility/service, and the UK Data Archive will archive the project web site for as long as JISC requires, up to a maximum of five years after completion of the project.  16.2 Software developed will be made available to the UK HE/FE community on an open source basis. Where appropriate, all outputs will be made available via SourceForge for further development.  **17. Budget**  17.1 Staffing The main costed staffing element of the project covers the programmers and consultants at the University of Essex. Specifically, Louise Corti will act as Project Director, and be ultimately responsible for all the elements of the project, and will lead WP3. Her time and consultancy is offered at no cost. Matthew Woollard will be used as overall consultant to the project contributing digital preservation expertise. One new post will be sought to undertake the research and programming work for WP3 and 4. WP2 staffing is being reconsidered to match the scope currently under revision, but will most likely continue to comprise a consultant and a full-time research officer. |  | |  | | --- | |  | | . | | |
|  |