*** DRAFT – DO NOT QUOTE / DISSEMINATE ***

APPENDIX 6:

DDI-DExT Tools Project
Technical Report

*Pascal Heus (*pheus@opendatafoundation.org*)*

Version [2010-05-13]

Table of Contents

Overview

The DDI-DExT Tools project's primary objective was to produced a utility software for the conversion of statistical data files into an archive neutral format for long term preservation along with the option to re-package the data for dissemination to end users for use with various statistical packages. This initial effort was intended as a "proof-of-concept" tool to demonstrate the feasibility and sustainability of such approach. The development was conducted from June 2007 to February 2008 as a collaborative effort between the UK Data Archive (UKDA) and the Open Data Foundation (ODaF). This document summarizes the technical aspects and activities of the project.

Methods

Supported file formats

For the purpose of this project phase, the input data file format was limited to SPSS.

The *archive neutral* formats selected were fixed ASCII for data and the Data Documentation Initiative 3.0 XML specification[1] for the metadata.Note that in addition to fixed ASCII, delimited and comma separated formats are also supported. ASCII remains the most commonly used and recommended format for long term preservation. DDI 3.0 was selected for its metadata richness as well as being the leading specification for the documentation of microdata. Direct compatibility with DDI ensures that the tools developed under DExT are compatible with other products and that the generated metadata can be imported by other packages or exchange with other institutions.

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

Environment

The project team followed the general recommendations outlined in the Open Data Foundation "Managing social, behavioral and economic data and metadata: Guidelines for Tools Development and Recommendations for Operating Environment"[2] document.

The packages were developed in the Java programming language using the Eclipse Integrated Development Environment (IDE). The source code is published and maintained in the ODaF Forge

1 http://www.ddialliance.org
2 http://www.opendatafoundation.org/?lvl1=resources&lvl2=papers

public repository[3].

The end user DDI-DExT product was developed as an Eclipse Rich Client Platform[4] (RCP) application to maximize openness and portability across-platform.

Conversion of the DDI-XML metadata into setup files for the various statistical packages were developed using the XSLT v2.0 language and processed by Saxon XSL v8.9[5].

Documentation

Three level of documentations have been maintained:

- Source code level: documentation is present in the Java files in the form of comments captured in the standard JavaDocs[6] format based on Sun Microsystems recommendations (see ODaF guidelines).
- Application level: user level documentation is provided in the DDI-DExT application help system
- Project level: technical documentation and administrative activities. These documents are available upon request.

Note that the technical documentation includes a detailed paper on the SPSS and other file formats. As this information was collected from various sources on the Internet whose author could not always be identified or may be copyrighted, this paper cannot be released in the public domain. Several public sources documenting the SPSS format are however available.

We want in particular to thank the members of the PSPP project[7] community who have been particularly helpful and supportive during the development of the DexT SPSS Reader package.

Activities

Development

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

The project team worked part time on the development and the first phase was completed in mid-October 2007 with the release of the alpha-test version of the DExT Tools application supporting the SPSS file import functionalities. The second phase development phase concluded end of January 2008

3 http://forge.opendatafoundation.org
4 http://wiki.eclipse.org/index.php/Rich_Client_Platform
5 http://saxon.sourceforge.net/
6 http://java.sun.com/j2se/javadoc/
7 http://www.gnu.org/software/pspp/

with the final version of the product that included the SAS, SPSS, and Stata export functions.

Activity and progress report were regularly exchange by email and a face to face meeting took place at the UKDA mid-January 2008 to present the product to the project manager and discuss lessons learned and next steps.

Testing

Import

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

This provided for a wide range of test cases varying from SPSS version 4 to 16 and produced on different platforms (Windows, Linux, Solaris). With very few exceptions, the DExT SPSS Reader is able to properly read and export the data and metadata. A known issue exists for files produced on the Solaris platform (SPSS 4) or with the open source SPSS Writer Java package distributed through SourceForge[8]. A fix is trivial and will be implemented in upcoming maintenance release of the reader.

Export

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

Outputs

The primary outputs of the project is the DDI-DExT Tools application and the underlying SPSS Reader component. Both products have been released under the GNU Lesser General Public License[9] and donated to the DDI Foundation Tools Program[10] as a contribution from the UKDA and the Open Data Foundation. The DDI-DExT installer and SPSS Reader can be downloaded from the DDI Tools web site[11] and the source code is publicly available on the ODaF forge repository.

To advocate the availability of the tools, presentations and demos are planned at the ODaF Europe 2008 meeting and the IASSIST 2008 conference in Stanford, CA in May. Announcement will also be

8 http://sourceforge.net/projects/spss-writer/
9 http://www.gnu.org/copyleft/lesser.html
10 http://tools.ddialliance.org/?lvl1=ftp
11 Http://tools.ddialliance.org/?lvl1=product&lvl2=dext

made through relevant channels such as the DDI Users, IASSIST and Open Foundation mailing lists.

Next Steps

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

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

Potential future enhancements include:

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

Conclusions

The DDI-DExT project successfully demonstrated the feasibility of using open source solutions, ASCII text files and standards based metadata for the conversion, preservation and dissemination of microdata. While not a production grade product, the DDI-DExT application can already be used as a free stand alone import/export utility for SPSS files and the automated creation of DDI metadata.

The availability of an open source Java library for reading SPSS data is also an important contribution as it greatly facilitate access to data and metadata stored in the proprietary SPSS format by any applications and web based services. The library as already been integrated in the DANS MIXED project[12] and has drawn interest from several other individuals and open source product developers.

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

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

Overall, this initial effort was successful in more way than initially anticipated. We hope to have the

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

opportunity to further developed the package t turn it into a versatile set f utilities for the management of microdata.

Annex 1: Screenshots

Illustration 1: DExT Splash Screen

Illustration 2: DExT About Box

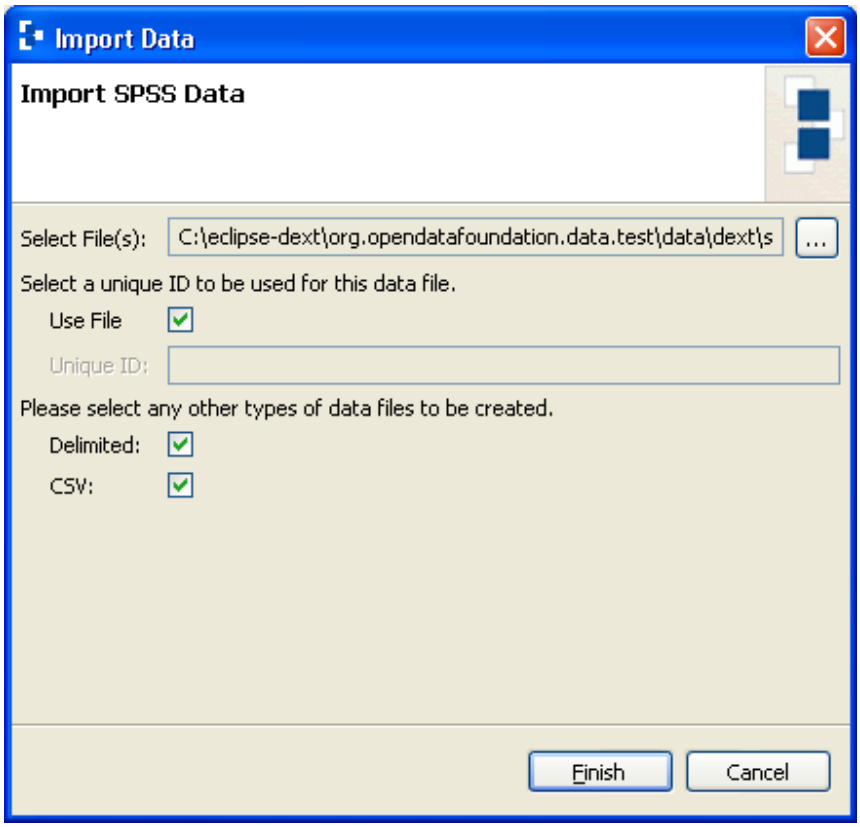

Illustration 3: DExT Import Data Wizard

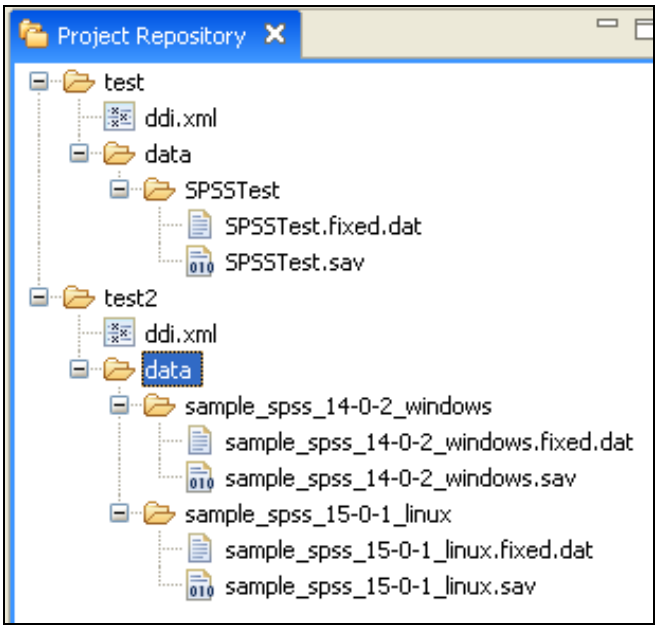

Illustration 4: Sample DExT project repository

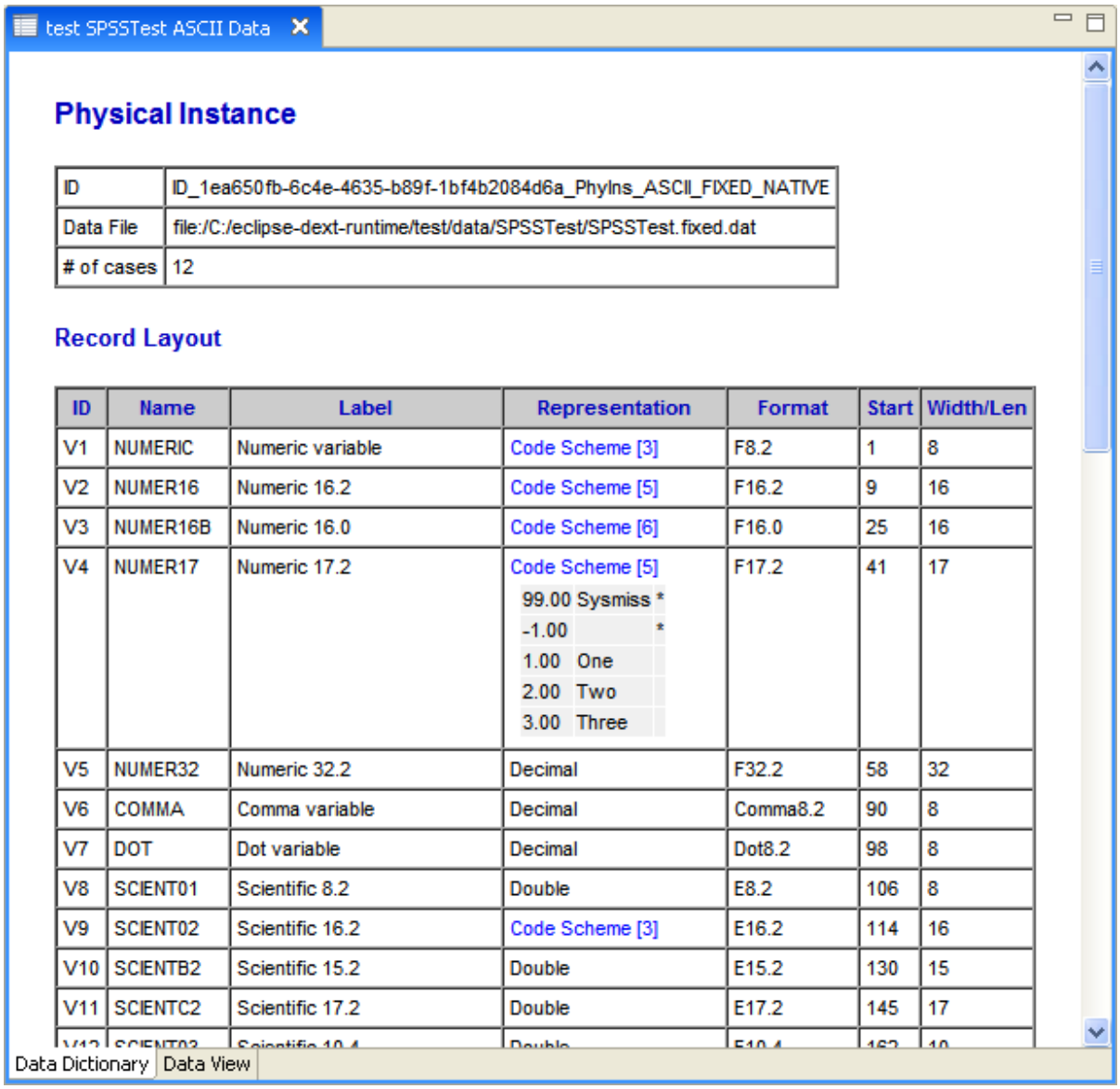

Illustration 5: DDI 3.0 Metadata for an converted Fixed ASCII file

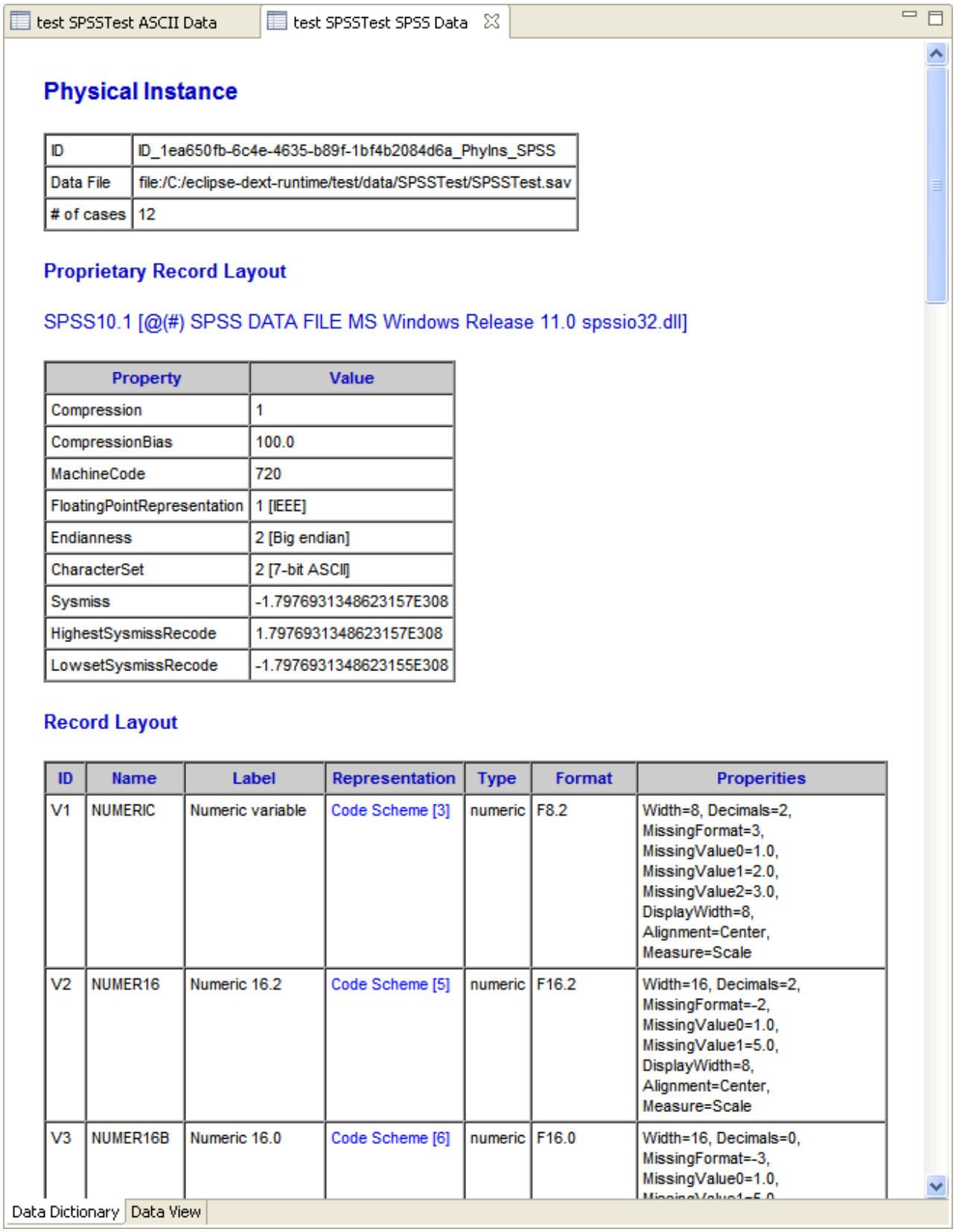

Illustration 6: DDI 3.0 Metadata for a native SPSS file (showing proprietary information)

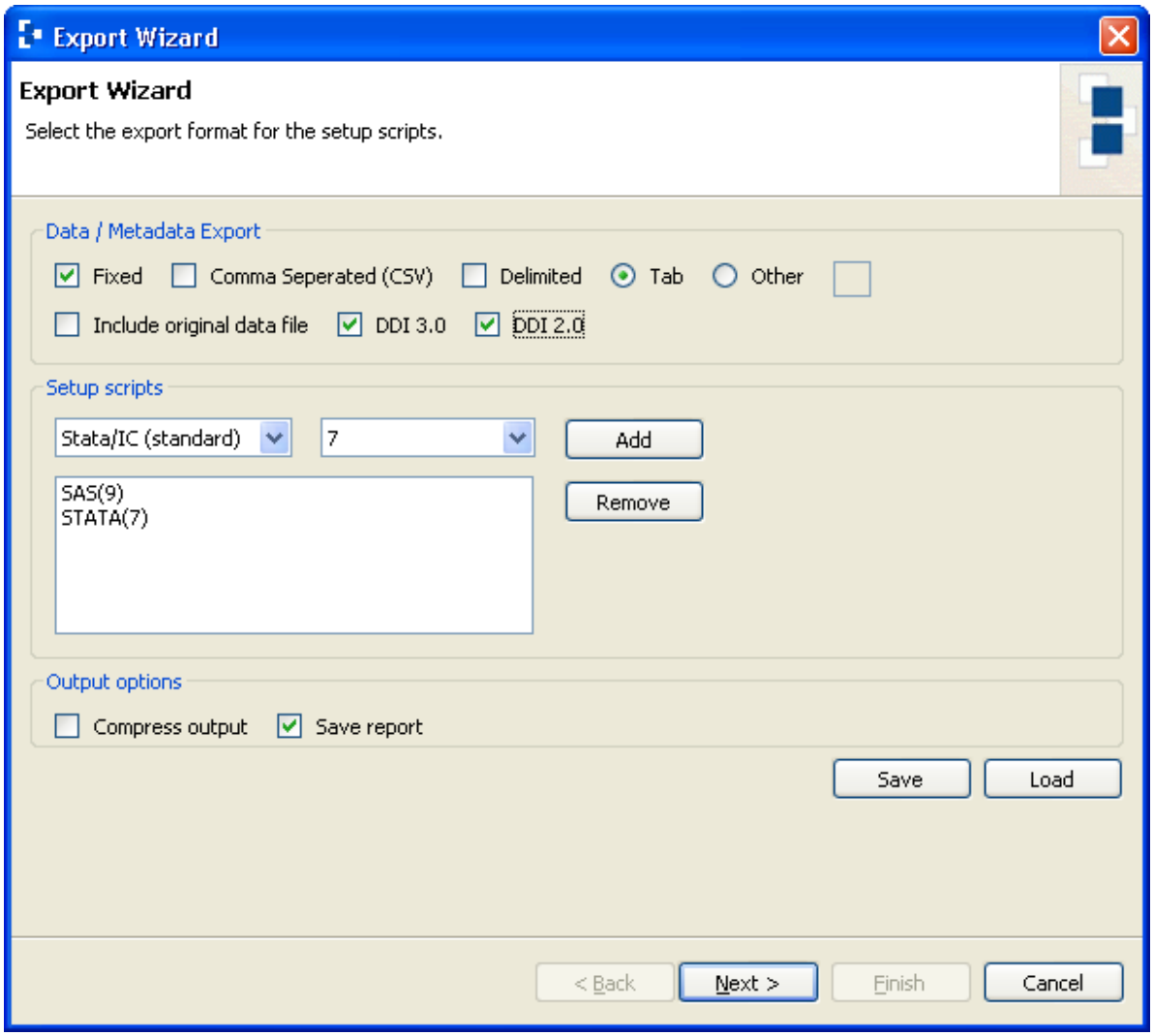

Illustration 7: DExT export wizard