The Conversion Software Registry

Michael Ondrejcek, Kenton McHenry, and Peter Bajcsy
National Center for Supercomputing Applications, University of Illinois, Urbana, IL 61801
ondrejce@illinois.edu, kmchenry@ncsa.uiuc.edu, pbajcsy@ncsa.uiuc.edu

We have designed web based Conversion Software Registry (CSR) for collecting information about software that are capable of file format conversions. The work is motivated by a community need for finding file format conversions inaccessible via current search engines and by the specific need to support systems that could actually perform conversions, such as the NCSA Polyglot[2]. In addition, the value of CSR is in complementing the existing file format registries such as the Unified Digital Formats Registry (UDFR before GDFR) and PRONOM, and introducing software quality information obtained by content-based comparisons of files before and after conversions. The contribution of this work is in the CSR data model design that includes file format extension based conversion, as well as software scripts, software quality measures and test file specific information for evaluating software quality. We have populated the CSR with the help of the National Archives and Records Administration (NARA) staff. The CSR has been deployed at http://isda.ncsa.uiuc.edu/NARA/CSR *and provides multiple search services. As of May 28th, 2010, CSR has been populated with 183,142 conversions, 544 software packages, 1316 file format extensions associated with 273 MIME types and 154 PRONOM identifications.*

1. Introduction

With an increasing number of software packages and file formats used by the US federal agencies preservation of electronic records has become one of the major challenges for the National Archives and Records Administration (NARA) [1]. Over the past years it has become apparent that file format migrations will become one part of preservation as the information technology is changing very rapidly. Furthermore, with the plethora of software on the market, the problem arises of converting files from one file format to another. The desired functionalities are distributed across multiple software packages as each supports only a subset of file formats. There is currently no system providing answers to questions such as “what file format conversion paths exist from file format A to B?” and “what would be the best file format conversion in terms of information preservation?” This work aims at filling the gap from the end user perspective and at providing services to make the file format consolidation more efficient. Our approach to the problem described above is to design a Conversion Software Registry (CSR) and provide services on top of the CSR repository. Furthermore, the CSR system serves as the source of information and a test bed for the system that can execute the conversions automatically by using the third party software, for example, NCSA Polyglot [2]. The CSR system is a database with a web-based interface that provides services related to a)

finding a conversion path between formats b) uploading information about the 3rd party software packages and file extensions, c) uploading files for testing, and finally d) uploading scripts in operating system (OS) specific scripting languages (Windows AutoHotKey, AppleScript and Perl) for automated conversions according to the idea of imposed code reuse used by NCSA Polyglot [2].

2. CSR design

In order to provide file format conversion services, we have included the following components into CSR related to software capable of conversions: input and output file formats (extensions), scripts operating on the software, validated files to be used for information loss measurements, as well as quantitative measures of the information loss for conversions. These components define the data entities of the CSR database as illustrated in Figure 1. There exist several file format services for collecting information about digital formats [3]. In contrast to the existing systems, the CSR focuses on software and finding the format conversion paths described by a number of software packages and unique input and output formats. The formats themselves are represented by extensions. While not always unique, extensions are often the only accessible information when the 3rd party software is installed (often listed under the File/Open menu in most packages). There are two other file format identifiers we have considered for encoding format conversions in CSR. First, it is the MIME, an internet standard for message content types which is not unique either. Second, it is the PRONOM id (PUID) [4] which is promising but does not provide PUIDs for all file formats we have encountered. We record and use PUID where different versions of the same format are represented by the numerical portion of the PUID. For example, a tiff extension (MIME 'image/tiff') is represented as PUID 'fmt/10' for the version 6.0, 'fmt/155' for GeoTiff. The problem of multiple extensions for the same file format has been tackled by allowing alternative extension names (e.g., jpg and jpeg). and treating them as equivalent through an additional table in the database.

A test file in the CSR is defined as any file which can be used for conversion accuracy and software validation. The files are uploaded and verified through the UNIX File command and against the file extension entry in the CSR database. Additional file validation has been performed semi-automatically by NARA using GTRI (Georgia Tech Research Institute) File Type Identifier [5].

The CSR also contains information about the software, operating system, software interface and scripts to execute the software. The scripts are important for the automating conversions with the 3rd party software and can be implemented using AutoHotkey scripts (Windows), AppleScript (Mac) or one of a variety of scripting languages for

Unix. Each software package requires more than one script depending on the complexity of accessing 3rd part software functionality. In general, the following script types are present in the CSR: 'Convert' for a full conversion from file A to the final product B, 'Monitor' for monitoring software behavior and 'Kill' for terminating the software in unexpected circumstances. Script headers are standardized with up to four lines with Software name and version, software domain (image, 3d, document, etc.), and input/output formats. The order of execution of the scripts is managed by the conversion service (i.e. Polyglot). The information loss due to file format conversions is measured externally by different techniques within the NCSA file-to-file comparison framework (called Versus). The comparison is relevant to the software domain, for example for 3D applications surface area or spin images are used and the loss (0-100 range with 100 representing no loss) for a particular software-conversion pair is stored in the database. The information loss also represents edge weights to Input/Output (I/O) Graph [2], a simple workflow used for finding the shortest conversion path.

The CSR is written as a web service. It consists of three main components: Query, Add, and Edit. In the Query mode users can a) view list of all software packages with their conversion options, b) select subsets of software in the I/O-Graph, c) search the database by conversions (see Figure 2 left), software, extensions, MIME and PUID. The I/O-Graph contains all information about installed applications and the conversions they allow. The JAVA applet front end is part of the CSR web visualization interface. Section Add allows users to add new software packages with their conversion capabilities and upload the software scripts to automate them (Figure 2 right). The last section, Edit is designed for adding detailed information about the software, extensions and for uploading the test files. CSR requires users to login for adding and editing. The web fields are searchable 'as one types' with options being presented to the user.

3. Summary

The CSR is part of a broader initiative to understand and solve the file format conversions problem. We designed a system for documenting all software packages with such functionalities and adding services for finding conversion paths under different constraints (i.e., the shortest path, the path with minimum information loss) as well as a system for making the automated file format conversion easier. The main purpose of the CSR is to provide an information repository for a new generation of scalable, multi-OS Polyglot conversion service framework.

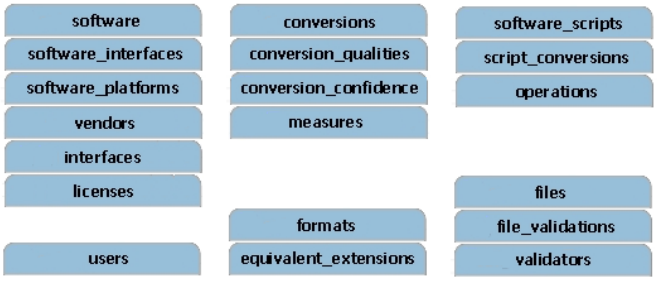

Figure 1. *The CSR pseudo-tables block design. The CSR includes information about software, file formats, software scripts and quantitative conversion measures, as well as the information about test files.*

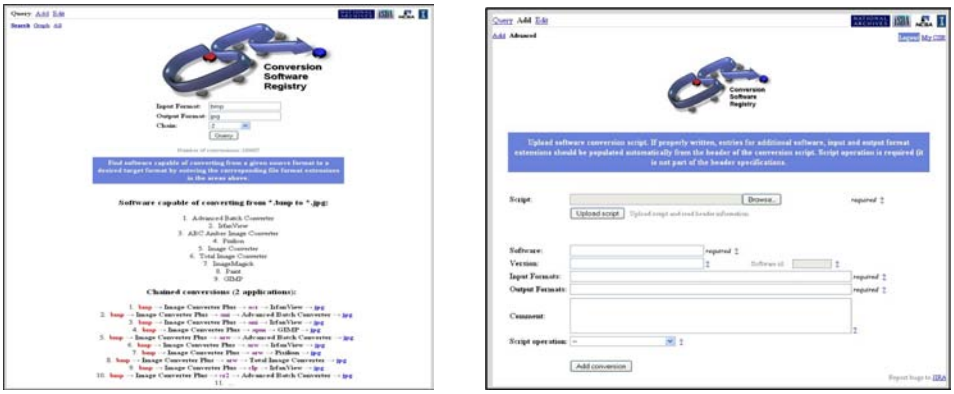

Figure 2. ***Left****: The CSR web interface showing a conversion query to find the shortest conversion path between the two formats entered. The single and multiple path conversions are listed alphabetically.* ***Right****: A form for inserting software information.*

Acknowledgement

This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019.

References

[1] The Strategic Plan of the National Archives and Records Administration (NARA) 2006-2016, *Preserving Past to Protect Future* 2006, URL http://www.archives.gov/about/plans-reports/strategic-plan/

[2] K. McHenry and P. Bajcsy, "Key Aspects in 3D File Format Conversions," Meeting of the Society of American Archivists and the Council of State Archivists, 2009, August 11, Austin, TX

[3] Global Digital Format Registry Data Model, 2009, http://www.gdfr.info/docs/GDFR-data-model-5_0_8.pdf; Wotsit.org, the file and data format resource, 2010, URL: http://www.wotsit.org/

[4] PRONOM is a resource registry (information) about the file formats, software products and other technical components, 2006, URL http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm

[5] W. Underwood, "Extensions of the UNIX file command and magic file for file type identification", Technical report ITTL/CSITD 09-02, Georgia Tech Institute, 2009, URL: http://perpos.gtri.gatech.edu/publications/index.htm