Using Scalable and Secure Web Technologies to Design a Global Digital Format Registry Prototype: Architecture, Implementation, and Testing

Muluwork Geremew, Sangchul Song, and Joseph JaJa; Institute for Advanced Computer Studies, Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, USA

Abstract

One of the most challenging problems confronting the long term preservation of digital information is how to handle format obsolescence. Several methodologies have been suggested in the literature, which include migration, emulation, and standardization to a few common formats. Clearly, a combination of these methodologies will be in use for the foreseeable future. Given this fact, some form of a community-based registry of digital representation formats and their supporting tools will be critical to managing format obsolescence. In this paper, we present the architecture, design, and testing of a Global Digital Format Registry (GDFR) based on scalable, extensible, and secure web technologies. Our architecture will easily be able to incorporate advances achieved through various methodologies and will easily adapt with emerging technologies as it is based on platform independent technologies.

Introduction

Digital file formats encode information into a form, which can only be accessed and processed comprehensibly by specific combinations of application and systems software. As today's technology evolves, it is uncertain which, if any, of the current formats will still be in use in the short or long term. Such uncertainty will leave most digital information at a high risk unless effective strategies are developed to manage the underlying technology evolution. There have been a number of ongoing international efforts to develop such strategies, which can be grouped roughly into three main approaches. The first consists of bringing various communities to agree on standardizing their digital content to a few common formats and structures and develop related open specifications. For example, JPEG [1], OMF [2] and PDF/A [3] are among the few selected open standard formats for long-term preservation. The advantage of adopting these open standard formats comes at the expense of some loss in processability, or loss at in structure and functionality as it is the case for PDF and ASCII respectively [1]. Moreover, it is not all clear whether these standard formats can address future needs or make use of future technologies.

The second approach relies on techniques to migrate from obsolete to current or emerging applications. A related strategy involves migration on batch basis (e.g. DAITSS [4], MyMorph [5]), on ingest (e.g. ANA, the Australian National Archives [6]) and on access (e.g. LOCKSS [7]). In practice, this approach probably has been the most widely used strategy. However migration will likely introduce some errors, which will be compounded after a series of migration. Care should be taken to ensure the integrity of the essential characteristics of the initial format. The third is based on emulation strategies of obsolete hardware and software to current systems. In principle, an emulation system enables the original bit-streams to be executed on the original hardware and systems software. By preserving not only the digital content but also the software on which it was written and originally intended to be run on, the digital resource will not undergo any changes and its preservation and authenticity can be assured. Different approaches for specifying an emulator has been suggested. R. Lorie [8] proposed a Universal Virtual Computer (UVC) as an emulation platform. Implementing such a strategy is extremely challenging and will in addition face difficulties similar to those that arise in the migration case. In particular, in essence emulation involves a transformation that is likely to introduce errors as in the migration case.

It seems clear that no strategy for handling format obsolescence will dominate in the near future and hence it is critical that we find a secure way to preserve reliable information about the various strategies being used to preserve formats. A promising such approach revolves around an international effort to establish a format registry, called the Global Digital Format Registry (GDFR) [9]. Such a registry is supposed to provide detailed authoritative representation information and possible transformations about formats. The concept of global digital format registry is not new. Attempts to make a universal digital format registry are being made through initiatives such as FRED [11]. However, it is not clear how FRED will interface with existing systems and current initiatives. In addition, we are not aware of any other working models of well–integrated global format registries or frameworks such as the one we have developed and describe here.

Our work builds on this concept by developing an efficient, scalable, and secure prototype format registry that captures all the essential features of GDFR, and attempts to offer a flexible framework for incorporating advances achieved through the approaches mentioned above. Our prototype, called FOCUS (FOrmat CUration Service), is platform independent, and is built on top of proven web technologies. Compared with other related systems such as PRONOM [10], FRED [11], and JHOVE [12], our architecture is more comprehensive and more importantly, it is more scalable and secure.

We have developed a framework within which the required representation information of digital formats and tools for accessing and processing such formats can be stored. In particular, FOCUS provides support for a wide variety of digital preservation functions such as digital object identification, validation, transformation and rendering. It contains general descriptive and

representation information on formats and how they can be preserved such as their specifications, and a list of available certified tools for transformation (conversion and emulation software), validation, rendering and processing. FOCUS can also incorporate anticipated expiration dates to monitor formats and technologies upon which they depend. Entries in FOCUS include input and output format types as well as information for processing and accessing each format, such as the URI to a web-service or download of stand-alone software.

Functions Supported by our Registry

Our prototype registry has been developed to support a range of preservation functions suggested in [9]. These include:

1. Automatic identification of file formats – given a digital object; what format is it?
2. Verification of digital objects compliance to a relevant file format specification – given an object that is supposed to be of format F; is it?
3. Delivery - given an object of format F; how can it be rendered?
4. Transformation – given an object of format F, to what formats can it be converted to?
5. Risk assessment – given an object of format type F; is it at risk of obsolescence?
6. Characterization – Given a format F; what are the representation specifications of F?

While our prototype currently provides limited options for each of these functions, the framework is general and scalable to accommodate the requirements for a community-based GDFR.

Prototype Architecture

Figure 1 provides a high-level overview of the architecture of FOCUS. It consists of three major components:

- A registry implemented through the LDAP (Light Directory Access Protocol) technology.
- Web-service Agent (WSA) that handles interactions between users and the registry, and includes a format identification component.
- Supplementary components used in validation, rendering, and conversion.

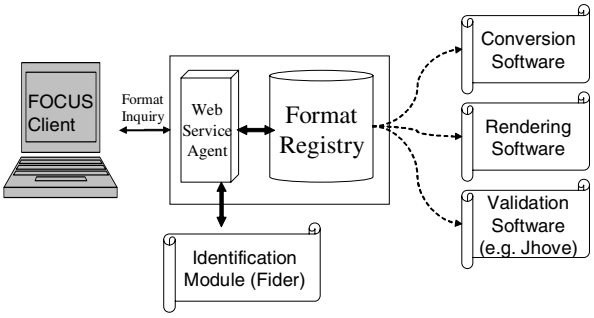

Figure 1. *The Overall Architecture of the FOCUS System*

Following, we describe the structure and the functionality of each software component in detail.

Registry Design

The design of FOCUS must ensure that our file format registry itself can survive various technology changes. Therefore, we have made important design considerations to address this. Our file format registry is intended to store frequently read information that needs to be distributed in order to support any of the functionalities described previously. Since directories are optimized for reading rather than writing, the directory becomes an ideal physical structure to store the file format information which does not change frequently. In addition, directories incorporate standard schemas that can be extended, replicated, and distributed.

The format registry in FOCUS makes use of the LDAP [16] directory technology, as it provides a mature and widely used directory service with proven security support. We have organized our registry around the LDAP technology using a hierarchical structure with two main subtrees. The left subtree corresponds to the information stored about all the applications software associated with any of the formats in the registry, while the right subtree contains detailed information about each format. The overall structure of our registry is shown in Figure 2.

Each leaf node in the applications subtree contains detailed information about the corresponding application including description, formats supported, download site and/or service location. Figure 3 shows an output to a request for information on a given application. Each leaf node in the formats subtree contains description, external signature information such as file extension, specification, owner, and available software tools related to the corresponding format. A sample of such format information is shown in Figure 4.

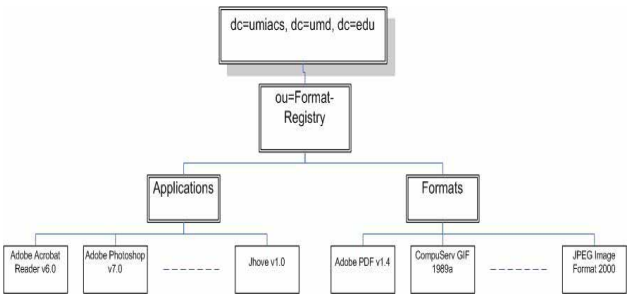

Figure 2. *A tree-like structure of our registry*

We have studied the physical structure of most popular digital format registries that currently exist. At this time, LCDF [13] stores the format information internally as MS word files and PRONOM [10] is a database system. The physical structure of GDFR [9] is not known as of today since it is only an idea rather than a reality. The NGDA [14] has developed hierarchical directory based local format registry. However, it is not clear to us what the underlying directory technology used for their registry. Hence, as far as we are aware, there is no directory based digital format registry that uses the LDAP technology.

Web-service Agent (WSA)

WSA is a web-service that acts as a mediator between the end user and the LDAP format registry. It provides two interfaces. The

first one is an outgoing interface to the digital format registry, which is used to query for information on a given format. This LDAP interface to the registry can also be used to determine potential format obsolescence. The second one is a SOAP-based [17] incoming web-interface which allows the end users and other external components to indirectly access our registry.

***Figure 3**. Application information screen*

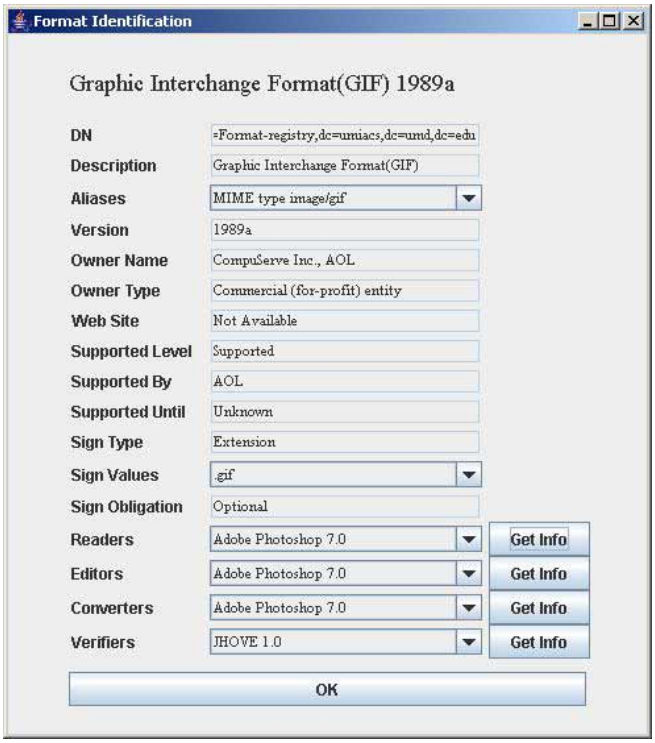

***Figure 4.** Format information screen*

Separate service modules can be independently developed and plugged into WSA to perform more complicated tasks. WSA currently contains an internal format identification module called *Fider* that identifies a file format using its internal signature (magic numbers). Fider is a Java module for identifying the file format of any file. This method, when applicable, offers fast and reasonably accurate output. However, in cases where the format specification does not include magic numbers, the only way to identify these formats will be to parse the entire file. In our current implementation, parsing is only done for the identification of US-ASCII format type. This provides the most accurate identification result, but the operation is very expensive.

Fider uses a sequential algorithm; it checks for magic number in order until the match is found. This order is defined from less to more generic format types. Fider has an extensible architecture; it can be configured at the time of invocation to add modules to identify more file format types. At present, Fider identifies most widely used formats including PDF, JPEG2000, GIF, TIFF, WAVE, and MP3.

The main advantage of WSA is that it can provide custom-tailored services to meet the end user's complicated needs. For example, a client can contact WSA through its SOAP-based web interface requesting to identify and obtain information on any supported format. With this identification module, the client can simply submit the digital file to the WSA. Upon receiving the file, WSA will identify the file format using this format identification module and query the registry for information on the identified format type behind the scene. Moreover, WSA makes the FOCUS system more flexible. From the system's perspective, if there are new operational requirements in the future, the new service modules can be implemented independently and easily plugged into WSA.

Another significant advantage of WSA is, like all other web applications, the ease with which it can be updated. Modifying existing services, or creating new services, can be done without affecting any client component, eliminating the need of redistributing new client-side components upon each update. Therefore, WSA offers advantages to both system administrators and clients.

Supplementary Software Components

Once the format is tentatively identified, WSA connects to FOCUS via LDAP to gather information on available validation software for the specific format. If the registered software is available as a web-service, the client contacts the available validation web-service of this file format to verify that the format is indeed the correct one. We currently make available a validation service through JHOVE, developed by JSTOR and Harvard University Library [12]. We have implemented JHOVE as a web-service locally and included the service location as well as the JHOVE module documentation into our format registry as shown in Figure 3. Afterwards, the registry can be consulted for available (and reliable) conversion, rendering, or emulation services for that particular format. Any of these services can be invoked as necessary.

Example Scenario

For the purposes of demonstrating and testing the FOCUS prototype, we have implemented client software that interacts with WSA. A user specifies that he/she requires a rendering application or service for a given digital content of unknown format using the FOCUS system. Client's request is then processed as follows. First, WSA invokes the internal identification module, Fider, and identifies the file format. Once the file format is identified successfully, WSA contacts our format registry and forwards user's request. When the search on the specific format is complete, WSA displays all the information such as all available editors, renderers and verification services back to the end user.

The user can then choose from the list of rendering applications returned by WSA. If the chosen application is stand-alone, the client software has the capability to check if the application is installed at the user's station locally. If so, it automatically launches and renders the digital content of interest. Otherwise, if the application exists as form of web-service, WSA can be tailored to interface to the web-service so that the user does not have to manually launch the application to view the file content. Currently, all the registered software tools are stand-alone except JHOVE which we have converted to a web-service locally to demonstrate how any web-service can be interfaced through WSA.

Conclusion

In this paper we have described the prototype FOCUS that is developed based on the concept of global digital format registry. It leverages current techniques on tackling format obsolescence for long-term preservation and accessibility of digital resources by integrating them and making them available through SOAP-based web interface.

Although our current prototype only supports a few widely used formats, the system can easily grow to accommodate many more file formats. The design of FOCUS offers maximum extensibility and scalability. Therefore, it has the potential to support all digital file formats. In summary, we believe that the FOCUS prototype represents a significant advance towards the development of global digital format registry, which is based on proven and platform-independent technologies.

References

[1] Selecting file formats for long-term preservation. http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_file_formats.pdf

[2] D. MacCarn, "Toward a Universal Data Format for the Preservation of Media," SMPTE, 106, pg. 477. (1997).

[3] L. William, "PDF/A: Developing a File Format for Long-Term Preservation," RLG DigiNews. (December 15 2003).

[4] DAITSS Overview. <http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf>, 2004.

[5] F. Walker and G. Thomas, "A Web-Based Paradigm for File Migration," Proc.Archiving, IS&T, pg. 93. (2004).

[6] H. Heslop, S. Davis and A. Wilson, "An Approach to the Preservation of Digital Records," (2002). http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf

[7] D. Rosenthal, T. Lipkis, T. Robertson and S. Morabito, "Transparent format migration of preserved web content". D-Lib Magazine, 11, 1 (January 2005).

[8] R. Lorie, "The UVC: A Method for Preserving Digital Documents - proof of concept," (2002). http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf

[9] S. Abrams and D. Seaman, "Towards a Global Format Registry," IFLA (2003) http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

[10] PRONOM, UK National Archives http://www.records.pro.gov.uk/pronom/

[11] Global Digital Format Registry: FRED http://tom.library.upenn.edu/fred/

[12] JHOVE, JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/

[13] Digital Formats for Library of Congress Collections http://www.digitalpreservation.gov/formats/index.shtml

[14] National Geospatial National Archive-local format registry http://www.ngda.org/reports/REPORT%7E1.pdf

[15] R. Lorie, "Preserving Digital Documents for the Long-Term," Proc.Archiving, IS&T, pg. 88. (2004).

[16] M. Wahl, T. Howes and S. Kille, "Lightweight Directory Access Protocol (v3)", RFC2251, (1997)

[17] M. Gudgin, M. Hadley, N. Mendelsohn, J. Moreau and H. F. Nielsen, "SOAP Version 1.2 Part 1: Messaging Framework ", W3C Proposed Recommendation, (2003)

Author Biography

Muluwork Geremew received the Bachelor of Arts degree in Computer Science and Mathematics in 2003 from Mount Holyoke College, MA, USA. Currently, she is pursuing her M.S. degree in Electrical and Computer Engineering at the University of Maryland, College Park, MD, USA. Her current research interests are in the fields of digital preservation and network security. Since joining Maryland, she has been actively involved in the long term digital preservation group led by Prof. Joseph JaJa.