OPEN SOURCE TOOLS FOR RECORDS MANAGEMENT

NARA/OMB M-12-18, Managing Government Records Directive
Reporting on Requirement A3.2

NATIONAL ARCHIVES *and* RECORDS ADMINISTRATION
AGENCY SERVICES
OFFICE OF THE CHIEF RECORDS OFFICER

INTRODUCTION

The Managing Government Records Directive, released in August of 2012 by the acting director of the Office of Management and Budget and the Archivist of the United States, sets two ambitious goals for Federal agencies. First, agencies are required to implement electronic recordkeeping to ensure transparency, efficiency, and accountability and to demonstrate compliance with Federal records management statutes and regulations. Second, by the end of 2019, agencies must, to the fullest extent possible, manage all permanent electronic records in an electronic format, and, by the end of 2016, agencies must manage all email records in an electronic format

The Directive encourages NARA, agencies, and stakeholders to investigate and stimulate applied research in automated technologies to reduce the burden of records management responsibilities in agencies. Item A3.2 specifically states that, "By December 31, 2014, the Federal Chief Information Officers Council, and the Federal Records Council, working with NARA, will obtain external involvement for the development of open source records management solutions."

The use of tools and technology could assist agencies in automating records management tasks. This will not only reduce the burden of records management responsibilities on individuals, but will make Federal government records and information easier to access because they are more consistently managed. The Directive promotes greater transparency, efficiency, accountability in Federal government and automating records management helps achieve that vision.

In particular, NARA is interested in exploring open source tools for automating records management. Open source tools have the potential for lower costs and could be reusable from one agency to another. Many of the open source tools available are robust and are driven by an active user and developer base. This means that the tools are constantly improving and could supply the Federal records management community with economically-viable tools to automate records management tasks. In addition, such active user development communities can often identify and provide remedies to security vulnerabilities that arise in faster time-frames than those provided by vendors of proprietary software products.

In this document, NARA identified open source tools that could be used for records management tasks. NARA recognizes the extensive work by a number of individuals and groups who have compiled lists of free, open source, and commercial tools for use in digital preservation and archival processing, such as the Community Owned digital Preservation Tool Registry (COPTR) and the Digital POWRR Tool Grid. However, these efforts have not analyzed the tools for how they could be used for records management tasks and have not been addressed by the records management community.

Therefore, NARA has developed this list as a means to introduce these efforts and tools to the Federal records management community and highlight how they could be used for records management tasks. Through the sharing of tools and experiences, we hope to identify opportunities for further tool development. We welcome ideas for working with external groups to incorporate records management functions into open source tools developed for other information management tasks.

SCOPE

This document only focuses on currently available open source tools. We did not include proprietary free software. NARA will not be developing any new records management solutions in the scope of this project. However, NARA is exploring how to build relationships with the open source community to identify gaps in open source records management tools and identify opportunities for external involvement to develop new records management solutions. The intended audience for this document is not only records management and IT staff in Federal agencies, but also developers in the open source community and any other interested parties.

WHY OPEN SOURCE

There are renewed efforts within the Federal Government to move towards open source solutions, such as the efforts of the recently formed 18F group and the US Digital Service. The Digital Services Playbook encourages agencies to "consider using open source, cloud based, and commodity solutions across the technology stack, as these solutions have seen widespread adoption and support by the most successful private-sector consumer and enterprise software technology companies." The TechFAR Handbook provides acquisition support to implementing the "plays" in the Digital Services Playbook. In addition, the Obama Administration's Digital Government Strategy calls for agencies to "participate in open-source communities."

Open source tools are generally free and available in a time of shrinking agency budgets. They often have very robust user and developer communities that are actively working to report bugs and improve the tools. Agencies are under significant constraints to minimize costs and operate more efficiently in all areas, including records management. NARA recognizes that some tools may not be immediately scalable in Federal agencies or may be in the early stages of development or even abandoned, but they have potential to meet agency needs in a cost-effective manner.

Open source software may be available under one of the various open source licenses that may ease agencies ability to acquire these tools. These licenses generally make the source code available with the proviso that any "local" developments, additions, or modifications to the code be likewise made openly available, in turn.

NARA recognizes that security is a concern with some implementations of open source tools in Federal agencies. Records management staff in agencies should work closely with information technology staff to test and download these applications within a proper test environment. Agencies should make use of the many resources available that address security concerns, including those listed at the end of this document.

NEXT STEPS

This list represents NARA's renewed efforts in the area of sharing open source tools for records management with Federal agencies. NARA recognizes that open source tools may require different degrees of customization and skills to deploy and may only address a piece of the lifecycle for managing records. While the tools themselves may be free, the expertise and time to customize must be considered.

NARA acknowledges that this list does not provide a roadmap for the deployment of a fully-operational open source electronic records management system for agencies. This is just the beginning of a process that we hope will lead to practical, affordable automated solutions for agencies. The first step is to release this compilation of existing open source tools so that we can begin to identify gaps in their ability to meet records management requirements. NARA wishes to encourage development of new solutions or the improvement of existing tools.

The next step will be to engage the Federal Records Council, the Federal Records Officer Network (FRON), and the ERM Automation Working Group to work through practical aspects of procurement and support of open source software in agencies. We will be moving this discussion to the ERM Automation Wiki hosted on Max.gov where agencies will be able to share their experiences and best practices for deployment and possibly improved versions of software that could be used by other agencies.

LIST OF TOOLS

The following table lists a sample of available tools and software available at the time of publication (October 2014) that could act as a "toolkit" to assist Federal agencies in automating and improving records management functions. Tools were included if they were described as accomplishing a function related to records management. They were neither tested nor are they endorsed by NARA. It remains the responsibility of agency records officers to evaluate software functionality and compliance with recordkeeping requirements and their agency needs. The list represents the range of services available together with software descriptions from their developers and available product reviews. The "Use Cases/Tags" column describes some of the possible uses for the tool when agencies manage their records. It may not be the original intended use of the tool by the developers. This list is not comprehensive and NARA appreciates additional suggestions for inclusion on the list.

We want to know:

- Have you used these tools at your agency or organization?
- Do you see a potential need for a tool?
- Are you looking to develop tools?

This document is intended to start the discussion on what tools are available and what the current and future needs are.

	Tool	Creator/ Developer	Developer's Tool Description	Tags for RM Functions	Notes from NARA
1	ACE (Audit Control Environment)	University of Maryland Institute for Advanced Computer Studies	"ACE (Auditing Control Environment) is a system that incorporates a new methodology to address the integrity of long term archives using rigorous cryptographic techniques. ACE continuously audits the contents of the various objects according to the policy set by the archive, and provides mechanisms for an independent third-party auditor to certify the integrity of any object. ACE consists of two components, the first an Audit Manager(AM) that checks files locally to ensure they have not been compromised. The second part, the Integrity Management Service (IMS), issues tokens that the AM can use to verify that its local store of file digests has not been tampered with."	Integrity checks Records audits	
2	Alfresco Community	Alfresco	"Alfresco Community Edition allows organizations to manage any type of content from simple office documents to scanned images, photographs, engineering drawings and large video files. It is commonly used as a: • Document management system • Content platform • CMIS-compliant repository There are also Add-ons that might do exactly what you're looking for."	Document management Content management Process management	

3	APACHE™ OODT	NASA's Jet Propulsion Laboratory	"It's metadata for middleware (and vice versa): Transparent access to distributed resources Data discovery and query optimization Distributed processing and virtual archives But it's not just for science! It's also a software architecture: Models for information representation Solutions to knowledge capture problems Unification of technology, data, and metadata"	Data grid framework Metadata management
4	AVI-MetaEdit	National Archives and Records Administration	"The software gives you ability to perform various metadata editing for AVI files. You can use the tool to embed, edit, import, and export metadata."	Digitization Metadata management
5	BagIt Library	Library of Congress	"The BAGIT LIBRARY is a software library intended to support the creation, manipulation, and validation of bags." "Bags are based on the concept of "bag it and tag it," where a digital collection is packed into a directory (the bag) along with a machine-readable manifest file (the tag) that lists the contents. Bags have a sparse structure that envelopes any institutional data architecture and format. It can hold documents, pictures, music, movies and even other folders. Anything digital can fit into a bag."	Transfer format Transferring records
6	BitCurator	School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and the Maryland Institute for Technology in the Humanities (MITH)	"The BitCurator project uses open source digital forensics tools to help collecting institutions manage born-digital materials. BitCurator packages forensics and data analysis software in an environment where users can create disk images, rapidly sort through files and file systems, extract and transform metadata, and identify and redact sensitive information."	Digital forensics Disk imaging File system analysis Metadata export PII detection

7	BWF MetaEdit	Federal Agencies Digitization Guidelines Initiative	"This tool permits embedding, editing, and exporting of metadata in Broadcast WAVE Format (BWF) files."	Metadata management Audiovisual formats Metadata export	
8	browser-shots	Internet Memory SCAPE Project	"The browser-shots tool is developed by Internet Memory in the context of SCAPE project, for the preservation and watch (PW) sub-project. The goal of this tool is to perform automatic visual comparisons, in order to detect rendering issues in the archived Web pages and report it to SCOUT via C3PO."	Appraisal analysis Preservation	
9	C3PO: Clever, Crafty Content Profiling of Objects	SCAPE Project	"C3PO – or 'Clever, Crafty, Content Profiling of Objects' is a software tool, which uses metadata extracted from files of a digital collection as input to generate a profile of the content set. The tool transforms the data for faster and scalable analysis and stores it, then post-processing solves issues like conflict resolution and provides a machine-readable overview, and a web application enables the user to filter and explore any part of the data further."	Metadata export Content profiling	Official release is coming soon.

10	CINCH	Elon University, Belk Library, NC LIVE (North Carolina Libraries for Virtual Education), North Carolina State Archives, State Library of North Carolina (lead), University of North Carolina at Charlotte, J. Murrey Atkins Library	"CINCH is a web-based, open source, lightweight tool that was designed to help libraries, archives, and agencies with similar mandates to collect and authenticate digital content that is freely available on the web."	Audit trail Checksum generator Checksum validation File renaming Metadata extraction Web capture
11	Cloud Deployment Toolkit	SCAPE Project	"Cloud Deployment Toolkit facilitates the deployment of various Scape software components on top of public or private (on-premises) clouds."	Cloud computing
12	CollectiveAccess	Collaboration between Whirl-i-Gig and partner institutions in North America and Europe with projects in 5 continents.	"CollectiveAccess is open-source collections management and presentation software designed for museums, archives, and special collections. As it is highly flexible and easily customized, it is also increasingly used by libraries, non-profits, private collectors, artist studios, performing arts organizations and other groups around the world. At its core, CollectiveAccess is a relational database that enables complex cataloging, powerful searching and browsing and nuanced web-based collection discovery."	Description Visualization

13	ContextMiner	Chirag Shah	"ContextMiner is a framework to collect, analyze, and present the contextual information along with the data. It is based on an idea that while describing or archiving an object, contextual information helps to make sense of that object or to preserve it better. This website provides tools to collect data, metadata, and contextual information off the Web by automated crawls. At present, ContextMiner supports automated crawls from blogs, YouTube, Flickr, Twitter, and open Web. It also collects inlinks information for YouTube videos from the Web. Additional sources will continue to be added."	Metadata management Social media capture Social media
14	Curator's Workbench	UNC University Libraries	"The Curator's Workbench is an extensible digital collection and appraisal tool for the desktop. It is designed to acquire and process batch data efficiently while giving the user control over work flow."	Workflow Appraisal
15	CSV Validator	The National Archives (United Kingdom)	"CSV Validator is a CSV validation and reporting tool which implements CSV Schema."	File validation CSV file validation
16	Data Accessioner	Duke University Libraries	"The DataAccessioner was built out of the need for a simple GUI interface to allow ... staff an easy way of migrating data off disks and onto a file server for basic preservation, further appraisal, arrangement, & description. It also provides a way to integrate common metadata tools at the time of migration rather than after the fact. With a simplified interface and being written in Java it is intended to be easily adopted by smaller institutions with little or no IT staff support."	Accessioning Checksum generator Digital Preservation Migration Preservation
17	DeDuplicator	National and University Library of Iceland	"The DeDuplicator is an add-on module for Heritrix to reduce the amount of duplicate data collected in a series of snapshot crawls."	Duplicate detection Web archiving

18	DELOLD	Alex Issakoo	"DELOLD is a command line tool used to delete old files with a create date older then a set amount of days. It can do verbose and rescursive with any path given. Perfect for Scheduled and/or batch file jobs."	File cleanup File management Shared drive cleanup	
19	Dependency Discovery Tool	School of Engineering and Computer Science at Victoria University, Wellington, Archives New Zealand	"The Dependency Discovery Tool searches through binary office files (.doc, .xls and .ppt) and tries to find any documents or files that are linked to the document."	Dependency detection	
20	DROID (Digital Record Object Identification)	The National Archives	"DROID stands for Digital Record Object IDentification. It's a free software tool developed by The National Archives of the United Kingdom that will help you to automatically profile a wide range of file formats. For example, it will tell you what versions you have, their age and size, and when they were last changed. It can also provide you with data to help you find duplicates. Profiling your file formats helps you to manage your information more effectively. It helps you to identify risks (and therefore plan mitigating actions)."	File format identification Duplicate detection	DROID can accurately identify more file formats than any other widely available tool.
21	Duplicate Files Finder	Matthias Böhm	"Duplicate Files Finder is an application which searches for duplicate files (files which have the same content, but not necessarily the same name) and lets the user remove duplicate files, either by deleting them or by creating links. The search is very fast compared to other similar programs which use hashing algorithm."	Duplicate detection	

22	DVA Profession	Österreichische Mediathek	"Professional digital video archiving system solution, developed and used by the Austrian national audio/video archive, designed to handle vast amounts of video content from ingest to long-term storage, including analysis, transcoding and metadata."	Audiovisual formats Workflow management Metadata production Digitization	
23	Email Parser	Rockefeller Archive Center and the Smithsonian Archives	"Downloads emails, scans the message body looking for "From:... To:..." blocks, reconstructs social relations and saves everything into a database. A navigator then allows you to browse friendship relations and perform searches through the contacts."	Email management Appraisal analysis mbox format	No longer being updated, but can be used.
24	ePADD (Email: Process, Appraise, Discover, Deliver Discovery) Module - BETA SITE	MUSE project at the Stanford University Computer Science Department	The Discovery Module displays summary metadata extracted from email collections held by Special Collections & University Archives. This metadata includes: collection extent, monthly summaries of frequently used terms, named entities, subjects, and correspondents. The module also features visualizations of entities and correspondents.	Email visualization Email processing Appraisal analysis Email analysis	As of October 2014, this tool is only available in beta version. The NHPRC grant runs until 2015.
25	EXIF to DC XML Normalizer	Open Planets Foundation	"The purpose of this tool is to extract EXIF data and normalise it to DC XML. This OS-independent tool enables you to read EXIF-data from an image file using ExifTool and normalise this to Dublin Core compatible XML adding specific metadata which is contained in an .ini file. The XML output templates can be tailored to one's needs. Tool is able to handle single files, or recurse trough a folder. Output is written to STDOUT, unless an output XML file is specified. Please note this tool is a prototype and adds very specific institutional metadata from The State University of New York at Binghamton, though the code can be changed very easy to your own needs."	Metadata normalization/ transformation	

26	Exiftool	Phil Harvey	"A command-line application that can read, write, and edit embedded metadata in files."	Metadata management	
27	Exsite9	Intersect Australia for the Australian National Data Service	"ExSite9 is a desktop application that was built to facilitate researchers easily and quickly tagging their data files with descriptive metadata and subsequently packaging their data files and associated metadata ready for submission to a repository. ExSite9 also allows for the structural organisation of said files within actually moving their physical location on your local file storage; allowing you to correctly organise your files and metadata ready for packaging."	Metadata management Metadata creation	
28	FFmpeg		"FFmpeg is a complete, cross-platform solution to record, convert and stream audio and video."	File format conversion Audiovisual formats	
29	FIDO (Format Identification for Digital Objects)	Open Planets Foundation	"Format Identification for Digital Objects (FIDO) is a Python command-line tool to identify the file formats of digital objects. It is designed for simple integration into automated workflows."	File Format Identification	This is a command-line python tool that uses parts of the PRONOM signature files. DROID is a GUI-based tool that uses all of the PRONOM signature and container files.
30	FITS (File Information Tool Set)	Open Planets Foundation	"FITS allows data curators to identify, validate, and extract technical metadata for the objects in their digital repository. It does this by incorporating a range of mostly third-party open source tools, normalizing and consolidating their output."	File format identification Metadata creation Metadata management	

31	Fixity v0.4	AV Preserve	"Fixity creates a manifest of files stored in directories identified by the user, documenting file names, locations, and checksums. The user can then schedule regular reviews of the directories to monitor for any changes to files that may point to data corruption or loss. Fixity is ideal for monitoring of files in long term storage, complimenting tools such as BagIt that check fixity at points of transition."	File validation Directory manifest creation Checksum generation Fixity monitor File integrity Shared drive management
32	GDuplicate Finder	Guillermo Campelo	"GDuplicateFinder is a FREE cross-platform application, with the ability to search among not just local files, but files on the network, such as a Windows or a Linux share using VFS library. Taking advantage of Groovy facilities and GPars power to process in parallel, GDuplicateFinder will help you get rid of those duplicates you always wanted to dispose in an easy and friendly way."	Duplicate detection
33	GNU Diffutils	Free Software Foundation	"GNU Diffutils is a package of several programs related to finding differences between files. Computer users often find occasion to ask how two files differ. Perhaps one file is a newer version of the other file. Or maybe the two files started out as identical copies but were changed by different people."	File differentiation
34	Hawarp	SCAPE Project	"Hawarp is a set of tools for processing web archive data by means of the Hadoop framework. The different tools are available as command line interface applications, each with it's own purpose, documentation, and usage modalities."	Web archiving Web content processing

35	Heritrix	Internet Archive	"Heritrix is an open-source web crawler, allowing users to target websites they wish to include in a collection and to harvest an instance of each site. The software is most often used as a powerful back-end tool incorporated into a web archiving workflow."	Web archiving Web crawling
36	ImageMagick	ImageMagick Studio LLC	"ImageMagick® is a software suite to create, edit, compose, or convert bitmap images. It can read and write images in a variety of formats (over 100) including DPX, EXR, GIF, JPEG, JPEG-2000, PDF, PhotoCD, PNG, Postscript, SVG, and TIFF. Use ImageMagick to resize, flip, mirror, rotate, distort, shear and transform images, adjust image colors, apply various special effects, or draw text, lines, polygons, ellipses and Bézier curves."	File format conversion Image processing
37	Interstitial	AV Preserve	"Interstitial is a tool designed to detect dropped samples in audio digitization processes. These dropped samples are caused by fleeting interruptions in the hardware/software pipeline on a digital audio workstation. The interstitial tool Follows up on our work with the Federal Agencies Digitization Guidelines Initiative (FADGI) to define and study the issue of Audio Interstitial Errors."	Audiovisual formats Digitization

38	iRODS	Supported and maintained by the iRODS Consortium at RENCI in partnership with the Data Intensive Cyber Environments (DICE) Center at UNC-CH	"This is iRODS, the integrated Rule-Oriented Data System, a distributed data-management system for creating data grids, digital libraries, persistent archives, and real-time data systems."	Data grids Data management Data systems Digital Libraries	NARA supported its development. It is used by a number of government agencies including, NASA, NIH, NOAA, Dept of Energy. Not for the faint of heart when it comes to deployment and maintenance. You need good IT support. This software is used to maintain and manage some of the largest data collections in the world. It can scale from a laptop to supercomputers.
39	iText		"iText is a PDF library that allows you to CREATE, ADAPT, INSPECT and MAINTAIN documents in the Portable Document Format (PDF)"	PDF processing	
40	Jpylyzer	SCAPE Project	Jpylyzer is a validator and feature extractor for JP2 images. JP2 is the still image format that is defined by Part 1 of the JPEG 2000 image compression standard (ISO/IEC 15444-1). Jpylyzer tells you if a JP2 image really conforms to the format's specifications (validation). It also reports the image's technical characteristics (feature extraction).	Metadata extraction File format validation	Section 1.2 of the User Manual - Validation: scope and restrictions provides useful information to determine the suitability of the software for your usage.

41	Kepler	UC Davis, UC Santa Barbara, and UC San Diego	"Kepler is designed to help scien-tists, analysts, and computer programmers create, execute, and share models and analyses across a broad range of scientific and engineering disciplines. Kepler can operate on data stored in a variety of formats, locally and over the internet, and is an effective environment for integrating disparate software components, such as merging "R" scripts with compiled "C" code, or facilitating remote, distributed execution of models."	Workflow Data reuse Data processing	
42	LogicalDOC Community Edition	Logical Objects	"LogicalDOC is a modern document management system with a nice interface, easy to use and very fast. It uses open source Java technologies such as GWT, Spring, Lucene in order to provide a flexible and scalable DMS solution."	Document management system	Comparison Matrix
43	Maarch	Maarch	"Maarch is a consistent set of tools and solutions to manage your document flows and their archiving. Maarch components are distributed under open source licenses, because archiving and conservation need open solutions! Maarch ecosystem offers document import and retrieval functionalities to make long term conservation and exploitation of digital resources possible, according to French and international standards for archiving"	Workflow Validation Physical archives management Authorization and clearance management File type management	
44	Mallet	Andrew McCallum	"MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text."	Natural language processing Document classification Clustering Topic modeling Machine learning	

45	Matchbox	SCAPE Project	"Matchbox is an open source tool which: • provides decision-making support for duplicate image detection in or across collections • identifies duplicate content, even where files are different (in format, size, rotation, cropping, colour-enhancement etc.), and if they have been scanned from different original copies of the same publication • applies state-of-the art image processing works where OCR will not, for example images of handwriting or music scores • is useful in assembling collections from multiple sources, and identifying missing files."	Duplicate detection Images
46	MDQC	AV Preserve	"MDQC reads the embedded metadata of a file or directory and compares it against a set of rules defined by the user, verifying that the technical and administrative specs of the files are correct. This automates and minimizes the time needed to QC large batches of digitized assets, increasing the efficiency of managing digitization projects. MDQC can be used on any file type supported by ExifTool and MediaInfo. Both ExifTool and MediaInfo will need to be installed on your system in order for MDQC to work."	Metadata management
47	Metadata Extraction Tool	National Library of New Zealand	"Developed by the National Library of New Zealand to programmatically extract preservation metadata for resource discovery. The Metadata Extraction Tool is based on a library of adapters. Open source software, yet with both a Microsoft Windows interface and a UNIX command line interface. The tool can be automated for batch processing or on demand. It is written in Java and XML and distributed under the Apache Public License (version 2)."	Metadata extraction

48	Muse	The Stanford Mobile and Social Computing Research Group	"Muse is a research tool from Stanford Computer Science for browsing large email archives. It was originally meant for people to browse their own long-term email archives. We have now started adapting it for journalists, archivists and researchers."	Email browsing Appraisal analysis
49	Nanite	SCAPE Project	"The Nanite project builds on DROID and Apache Tika to provide a rich format identification and characterization system. It aims to make it easier to run identification and characterisation at scale, and helps compare and combine the results of different tools."	Format identification Format characterization
50	NARA File Analyzer and Metadata Harvesting Tool	National Archives and Records Administration	The application allows a user to analyze the contents of a file system or external drive and generates statistics about the contents of the contained directories. The application can be used to generate checksum values to ensure the bit-level integrity of files after they have been copied to a new device. After a collection of files have been converted from one digital format to another, this application can verify that there is a one-to-one match of before and after files.	Checksum generator Directory creation

51	NetarchiveSuite	The Royal Library and The State and University Library	"The NetarchiveSuite is a complete web archiving software package developed from 2004 and onwards. The primary function of the NetarchiveSuite is to plan, schedule and run web harvests of parts of the Internet. It scales to a wide range of tasks, from small, thematic harvests (e.g. related to special events, or special domains) to harvesting and archiving the content of an entire national domain. The software has built-in bit preservation functionality. The systems architecture allows for the software to be distributed among several machines, possibly on more than one geographical location. The NetarchiveSuite is built around the Heritrix web crawler, which it uses to harvest the web."	Web archiving Web crawling
52	Omeka	Roy Rosenzweig Center for History and New Media	"Omeka is a free, flexible, and open source web-publishing platform for the display of library, museum, archives, and scholarly collections and exhibitions. Its "five-minute setup" makes launching an online exhibition as easy as launching a blog."	Web publishing Web content management Archival management
53	OSSIM	OSSIM	"OSSIM is a powerful suite of geospatial libraries and applications used to process imagery, maps, terrain, and vector data. The software has been under active development since 1996 and is deployed across a number of private, federal and civilian agencies."	Geospatial data
54	Pagelyzer	SCAPE Project	"Pagelyzer is a tool which compares two web pages versions and decides if they are similar or not. It is based on: a web page segmentation algorithm a combination of structural and visual comparison methods embedded in a statistical discriminative model a visual similarity measure designed for Web pages that improves change detection a supervised feature selection method adapted to Web archiving."	Web page comparison Appraisal analysis

55	ParaView	Sandia National Labs and CSimSoft	"ParaView is an open-source, multi-platform data analysis and visualization application. ParaView users can quickly build visualizations to analyze their data using qualitative and quantitative techniques. The data exploration can be done interactively in 3D or programmatically using ParaView's batch processing capabilities."	Data visualization
56	PDF Box	The Apache Software Foundation	"The Apache PDFBox™ library is an open source Java tool for working with PDF documents. This project allows creation of new PDF documents, manipulation of existing documents and the ability to extract content from documents. Apache PDFBox also includes several command line utilities."	PDF/A validation
57	Personal Renamer	Balisteor	"Renames files just about any way you like. Monitor folders for files and auto rename (Have program rename images when downloaded). Undo, Save settings, Imageview,Drag-drop, and more. Service program included (monitor and rename files while logged out)"	File renaming
58	Plato: The Preservation Planning Tool	SCAPE Project	"Plato is a decision support tool which guides you through the preservation planning workflow. To do this efficiently it integrates information from external sources like control policies, content-profiles, and component registries; it can run these components to automate tool evaluation, and connects to repositories via open interfaces. During this process Plato collects all the information to enable the planner to take an informed decision, and finally generates an evidence-based preservation plan which can be executed on suitable platforms."	Workflow planning Preservation planning

59	RTextTools	Timothy P. Jurka, Loren Collingwood, Professor Amber E. Boydstun, Professor Emiliano Grossman, Professor Wouter van Atteveldt	"RTextTools is a free, open source machine learning package for automatic text classification that makes it simple for both novice and advanced users to get started with supervised learning."	Autoclassification	Requires access to and some knowledge of the R language and development environment.
60	Spider	Cornell University	"Spider scans your hard drive, web site, or other collection of files to identify confidential data such as social security, credit card, or bank account and routing numbers. When the scan is complete, Spider produces a list of files that may potentially contain confidential data."	Personally Identifiable Information (PII) Search	
61	Taverna	Hosted at the School of Computer Science, University of Manchester, UK.	"Taverna is an open source and domain-independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation." "The Taverna tools include the Workbench (desktop client application), the Command Line Tool (for a quick execution of workflows from a terminal), the Server (for remote execution of workflows) and the Player (Web interface plugin for submitting workflows for remote execution). Taverna Online lets you create Taverna workflows from a Web browser."	Workflow management	

63	ToMaR	SCAPE Project	"When dealing with large volumes of files, e.g. in the context of file format migration or file characterisation tasks, a standalone server often cannot provide sufficient throughput to process the data in a feasible period of time. ToMaR provides a simple and flexible solution to run preservation tools on a Hadoop MapReduce cluster in a scalable fashion. ToMaR enables the use of existing command-line tools and Java applications in Hadoop's distributed environment in a similar way to a Desktop computer without needing to rewrite the tools to take advantage of the specialised environment. By utilizing SCAPE tool specification documents, ToMaR allows users to specify complex command-line patterns as simple keywords, which can be executed on a computer cluster or a single machine."	Digital preservation Workflow management Used to make other tools run faster	Assumes availability and familiarity with Hadoop and related tools.
64	vRenamer	Carlos Verdier	"vRenamer is an easy to use mass renamer with a lot of options. It's able to insert, remove, and replace strings, extract audio and images metadata, write audio metadata, change length, numbrering, undo/redo, and much more. Tested in Win and Linux."	Audiovisual formats File renaming	
65	WarcManager	University of Maryland Institute for Advanced Computer Studies	"The Warc Manager is a tool to help archives quickly browse, search, and analyze archives of web crawl data. The manager is lightweight database web application which indexes and provides a nice browsing interface to a collection of warc data."	Web archiving	

66	WCT (Web Curator Tool)	National Library of New Zealand and the British Library, initiated by the International Internet Preservation Consortium	"The Web Curator Tool (WCT) is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix web crawler and supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata."	Web archiving Workflow management
67	xcorrSound	SCAPE Project	"xcorrSound provides four tools: overlap-analysis to detect overlaps in two audio files; sound-match finds occurrences of smaller WAV within a larger WAV; waveform-compare splits two audio files into equal sized blocks and outputs the correlation for each block; and sound-index builds an index for sound-match to work within."	Audiovisual formats Duplicate detection
68	Xinco Document Management System	Originally developed by Alexander Manes at the University of Cooperative Education, Heidenheim, Germany. Today, xinco DMS™ is led by Javier Ortiz.	"xinco [eXtensibe INformation COre] is a powerful Web-Service based Information and Document Management System (DMS) for files, text, URLs and contacts, featuring ACLs, versioning, full text search, an FTP-like client (easy install, J2EE+MySQL/PostgreSQL)."	Document management Version control

Resources for Further Information:

- From Theory to Action: Good Enough Digital Preservation for Under-Resourced Cultural Heritage Institutions
- General Study 08 – Open-Source Records Management Software: Final Report
- open source as alternative

- Open Source for America
- Open Source Initiative - "Open Source Bibliography"

Lists of Tools:

- AVPreserve tools
- Community Owned digital Preservation Tool Registry (COPTR)
- Digital POWRR Tool Grid
- Free/Open Source Software for Libraries
- Library of Congress Tools Showcase
- Lifecycle Management Tools (IMLS-funded project)
- Scape Project tools

Open Source in Federal Government

- code.NASA
- Content Management Systems Used by Government Agencies
- DARPA Open Catalog
- Digital Services Playbook
- NARA's Records Management Services Program
- NASA Technology Transfer Program Software Catalog - see especially Chapters 2 and 14
- Project Open Data
- Sandia's Open Source Software Portal
- TechFAR Handbook
- The Contributor's Guide to 18F: Code for the Common Good
- The GovLoop Guide - Agency of the Future: Open Source