NSRL and Its Potential Role in Digital Curation

CurateGear
January 6, 2011

Disclaimer

Trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Statement of Disclosure

This research was funded by the National Institute of Standards and Technology Office of Law Enforcement Standards, the Department of Justice National Institute of Justice, the Federal Bureau of Investigation and the National Archives and Records Administration.

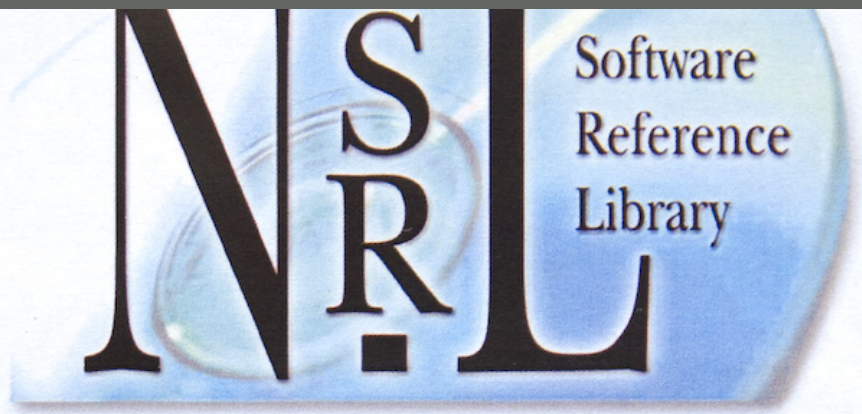
What is the NSRL?

A Library of Software

A Database of Metadata

NIST Special Database #28

A NIST Publication



Reference Data Set

Version 2.19 12/1/2007

NIST

A Research Environment

What is the NSRL data?

NSRL collects metadata that describes every file on all media in the physical collection.

Media tracking
Manufacturer information
Operating system requirements
Product description
File metadata includes:

- Directory path, File name, Bytes, Digital signature (hash), etc.

1,526 Manufacturers

most represented:
Adobe, Apple, Dell, HP, Intuit,
Microsoft, Oracle, Sun, Symantec

552 Operating Systems

most represented:
Windows (95,98,NT,2000,XP,Vista),
Linux, Mac OSX, Macintosh, Solaris, DOS

12,242 Products

most represented types:
Operating systems, games, office suite, database,
antivirus, financial, graphic/photo editor

101,451,035 Files **17,736 Media Images**

NSRL Environment

NSRL work occurs on an air-gapped (isolated) network.

Major process components are modularized.

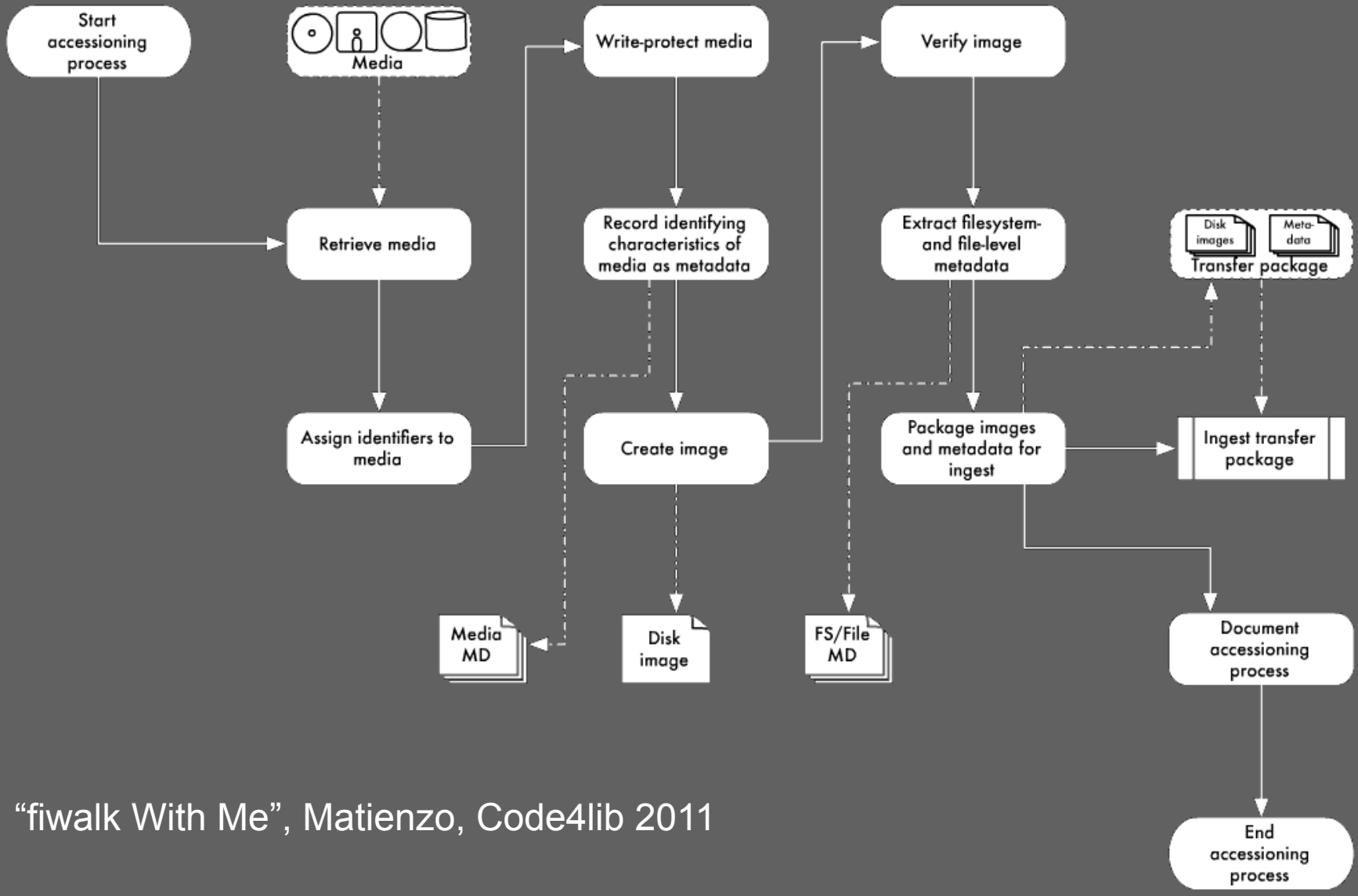
User interfaces are browser-based.

Infrastructure uses FOSS as much as possible.

Implementation philosophy: leverage existing knowledge and reuse tools as much as possible.

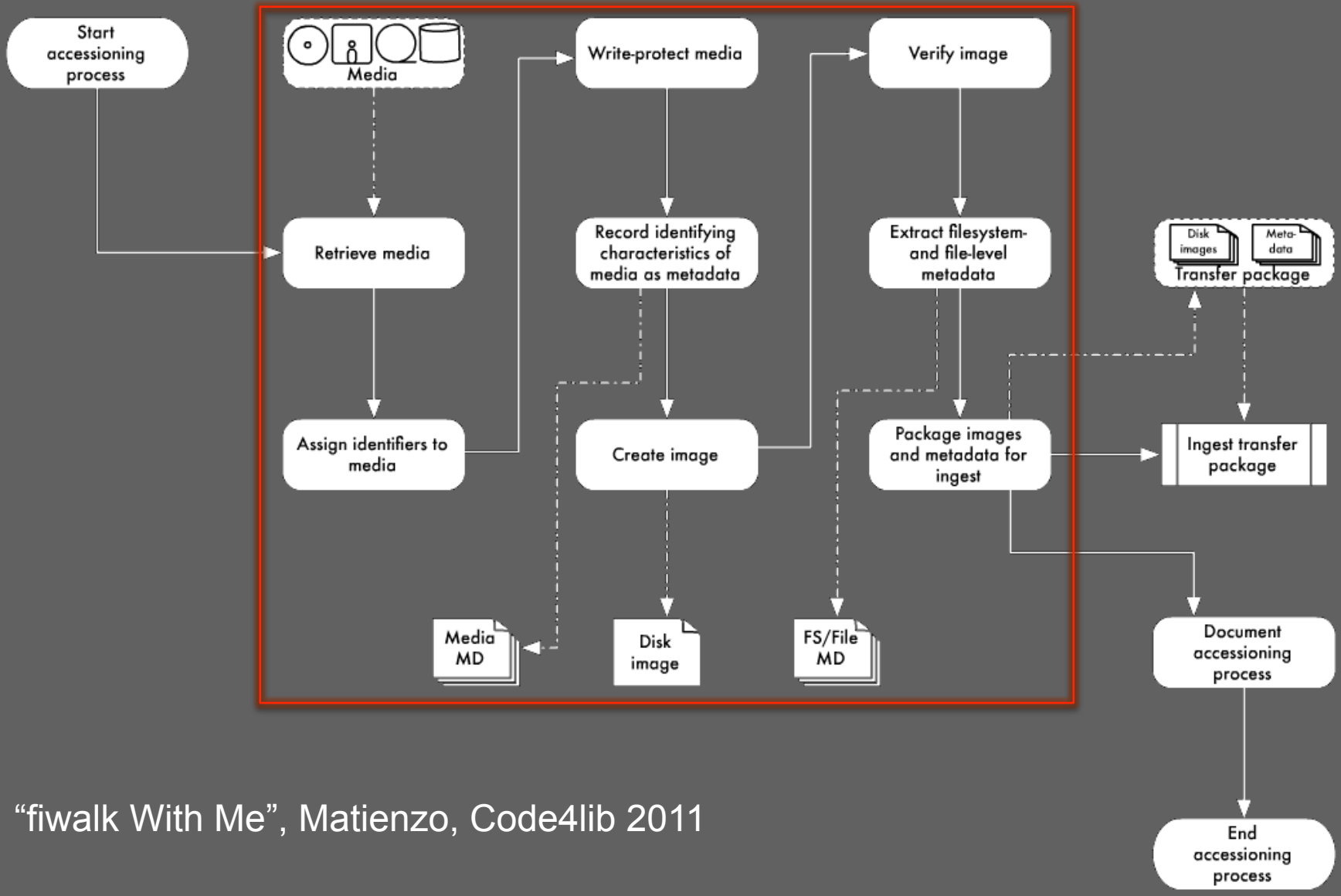
Role in Digital Curation

Accessioning Workflow



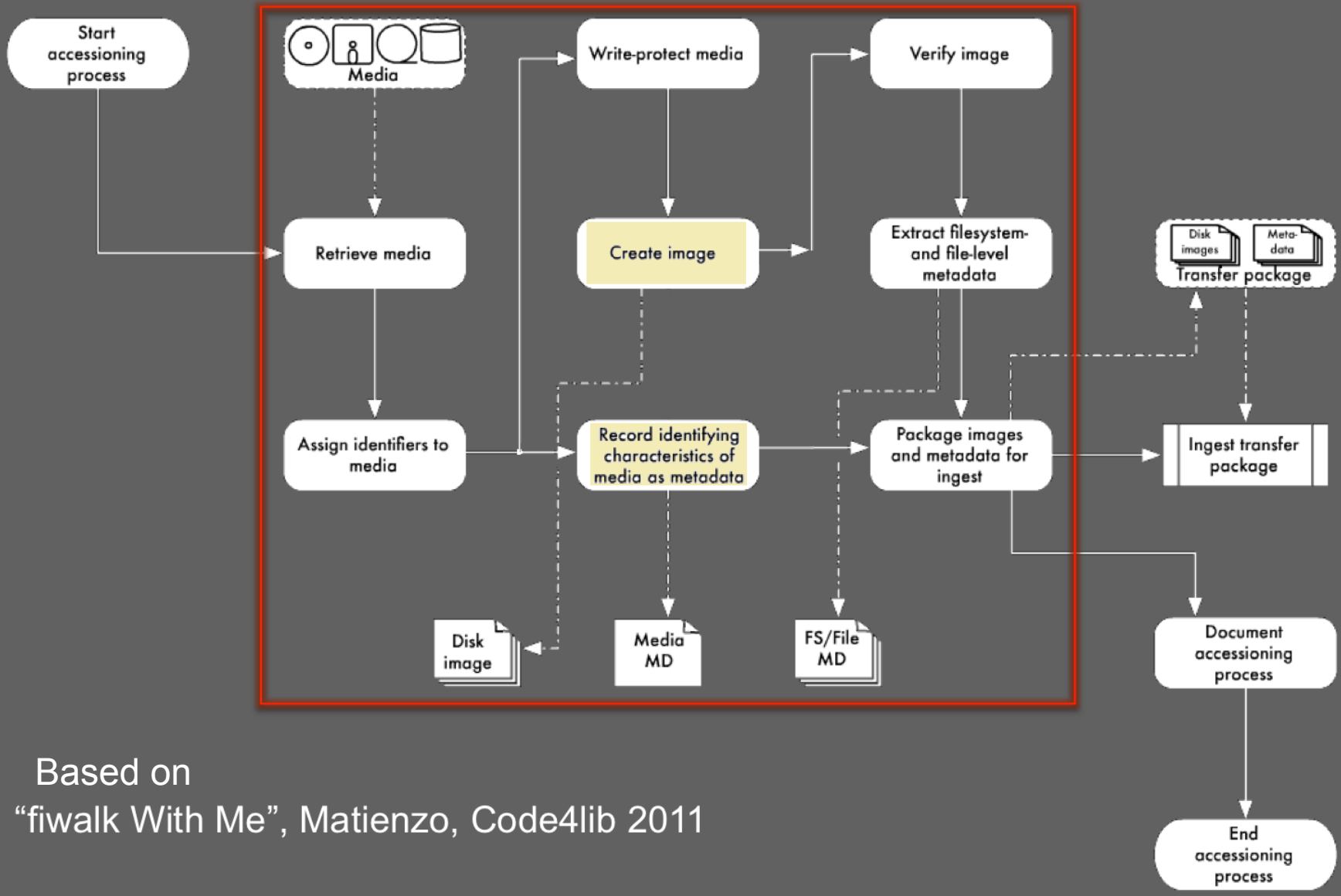
"fiwalk With Me", Matienzo, Code4lib 2011

Accessioning Workflow



"fiwalk With Me", Matienzo, Code4lib 2011

Accessioning Workflow



Based on
"fiwalk With Me", Matienzo, Code4lib 2011

Beyond Accession

Software Corpus

Researchers may run algorithms against the collection of 23,809,431 unique files.
(duplicates culled from 101M)

NSRL has collaborated on topics such as

- Identification of RAM objects
- Block hashes
- Non-cryptographic “fuzzy” hashes
- Diskprints

Hash Examples

Filename	Bytes	SHA-1
NT4\ALPHA\notepad.exe	68368	F1F284D5D757039DEC1C44A05AC148B9D204E467
NT4\I386\notepad.exe	45328	3C4E15A29014358C61548A981A4AC8573167BE37
NT4\MIPS\notepad.exe	66832	33309956E4DBBA665E86962308FE5E1378998E69
NT4\PPC\notepad.exe	68880	47BB7AF0E4DD565ED75DEB492D8C17B1BFD3FB23

WINNT31.WKS\I386\notepad.exe 57252 2E0849CF327709FC46B705EEAB5E57380F5B1F67

WINNT31.SRV\I386\notepad.exe 57252 2E0849CF327709FC46B705EEAB5E57380F5B1F67

Similarity Metrics

fourscore.doc
fd375c1f4fe60fb4fbcef5b3f1bb035042e34fdd
0e0a6f71bc90534877f6018b50b94c2e97cab8f7
fivescore.doc

fourscore.doc
96:f+pIKe/OQxx1av5BVh:QSKcRuEuGtXo2s2Rhf6Pe2Qx+fV
96:BmpIKe/OQxx1av5BVK:vcRuEuOw5us2GNGrBSPe2Qx+fV
fivescore.doc

Windows Registry

NSRL has published a strawman data set enumerating Windows registry keys and values.

NSRL collaborates with other researchers involved in registry forensics.

Security Content Automation Protocol (SCAP)

NSRL is collaborating with National Vulnerability Database (NVD) researchers to implement the Common Platform Enumeration (CPE) scheme.

This allows NSRL data to interoperate with other NIST data and a wider range of commercial products.

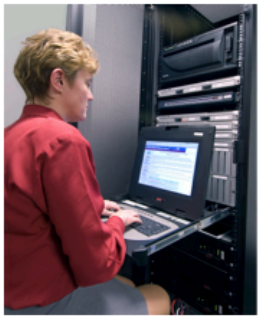
Mobile Devices

Mobile devices - smart phones, e-readers, etc. – run operating systems which are complicated enough to warrant applying NSRL reduction processes.

NSRL is cooperating with CFTT to identify file system based objects.

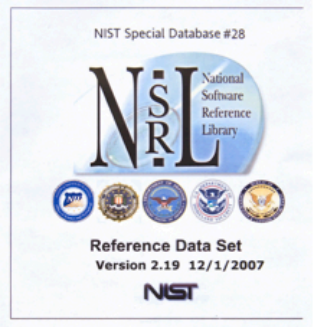
Physical software collection

Database of file metadata



File name, size, path, dates,
SHA-1, MD5, etc. are recorded

Reference data for investigative use



Data can be imported into many commercial digital forensics tools

Virtual software collection

Virtual machine installation

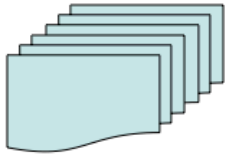
Reference data from installation

Metadata collection from installation, execution, patches, etc.

Registry data from installation

Metadata collection of registry key and value status after install, execution and removal

23,000,000+ unique
Software application files



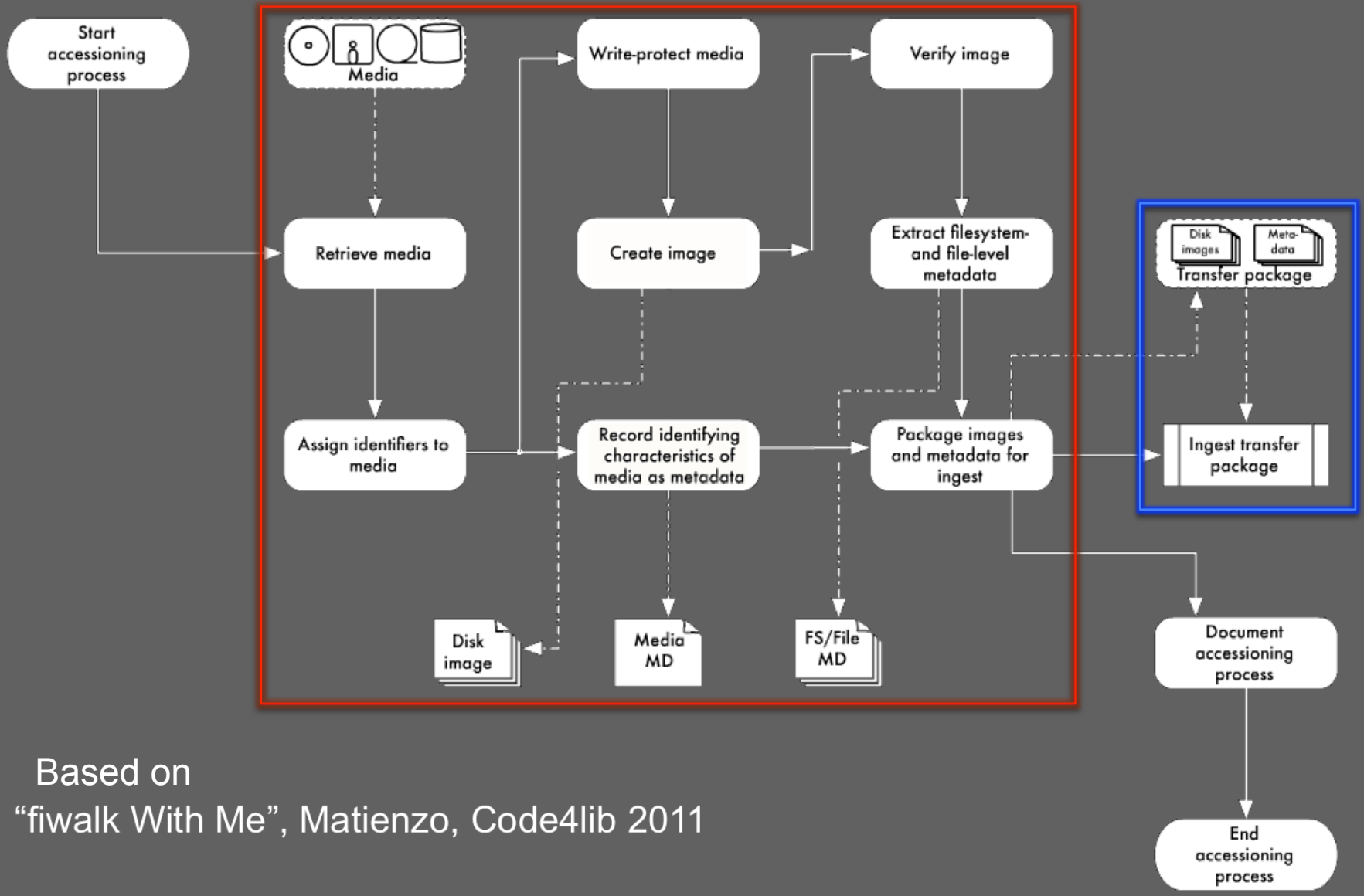
File corpus is available to researchers

RAM-based identifiers
Alternate cryptographic hashes
Smart unpacking
Common Platform Enumeration (CPE)
Security Content Automation Protocol (SCAP)

NSRL National Software Reference Library

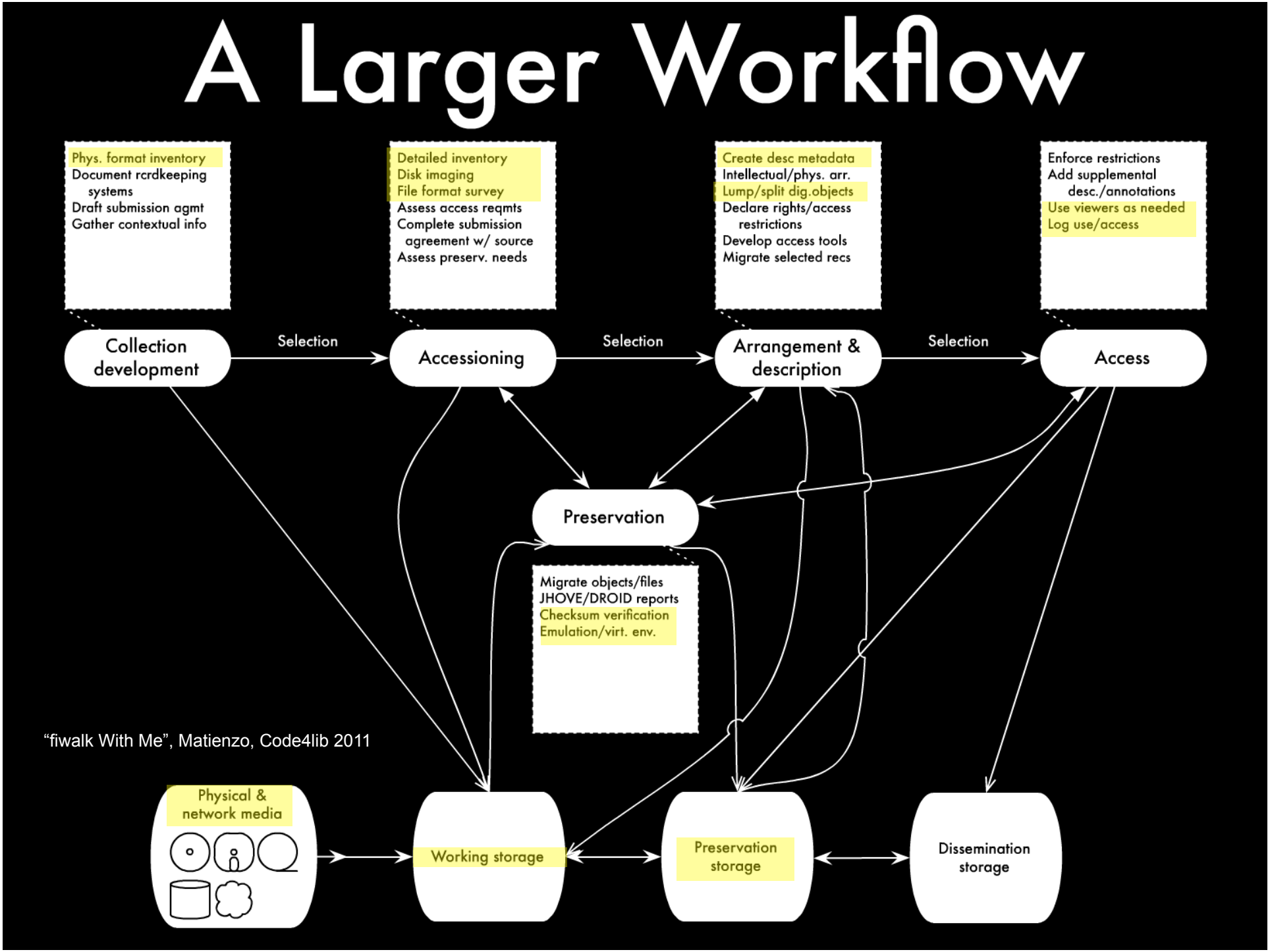
www.nsrl.nist.gov
nsrl@nist.gov

Accessioning Workflow



Based on
"fiwalk With Me", Matienzo, Code4lib 2011

A Larger Workflow



"fiwalk With Me", Matienzo, Code4lib 2011

Practical Use

NARA Bush(I) Research

93 subject computers containing 51,146 files, 2.3GB
11,118 unique files (78% duplicate)
8,077 files identified by SHA-1 (72%)

469 identical temporary installation files
161 zero-byte empty files
130 identical WordPerfect icon files
Several files had 90+ instances

Generated a “baseline” computer system

Obtained pedigree of operating system upgrades

Stanford - Cabrinety

Stanford University Libraries holds the
Stephen M. Cabrinety Collection in the History of Microcomputing.
This is a collection of 5,000 software titles spanning 1975-1995.
The collection is not easily accessible for research.

15 titles from the collection are on loan to NIST for processing.
NIST images media, collects metadata, photographs containers.
The original media, media images, metadata, and photographs are returned to Stanford.

Document Format ID

As NSRL builds a library of working operating systems on virtual machines, we have been requested to install various document creation software and create instances of common content.

This work may lead to a reference set of metadata more applicable to content-based archival.

nsrl@nist.gov

www.nsrl.nist.gov