Digital Preservation

NANCY Y. MCGOVERN AND
KARI R. SMITH

NETSL ANNUAL SPRING CONFERENCE
APRIL 2013

Overview

- Framework for practice
- Tools and workflow
- Wrap-up

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Sources

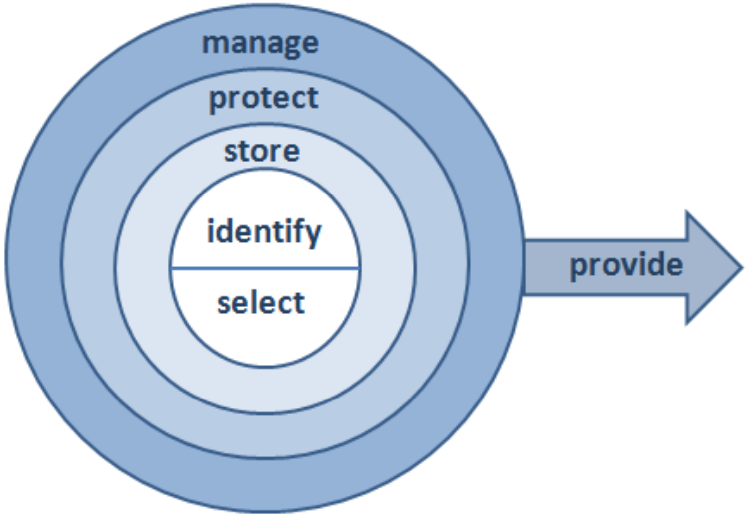
Sources:

- Community standards and practice (examples cited)
- Digital Preservation Outreach and Education (DPOE) http://www.digitalpreservation.gov/education/
- Digital Preservation Management (DPM) Workshops http://dpworkshop.org/
- Examples from MIT Libraries http://libguides.mit.edu/digitalarchivestools
- http://libraries.mit.edu/sites/digital-archives

NOTE: Tool developers cited individually

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

DPOE Management Model



Identify - what digital content do you have?

Select - what portion of that content will be preserved?

Store - what issues are there for long term storage?

Protect - what steps are needed to protect your digital content?

Manage - what provisions are needed for long-term management?

Provide - what considerations are there for long-term access?

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Identify: How will an inventory help?

Good preservation decisions are based on an *understanding* of the possible content to be preserved

The Identify stage addresses:

"what content do I (or will I) have?"

LIBRARY OF CONGRESS

DPOE Baseline Modules: Identify, version 2.0, Nov 2011

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Content Categories

Inventories should include all relevant, e.g.:

- Institutional records
- Special collections
- Scholarly content – licensed and open
- Research data
- Web content

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Select: Selection Criteria

- Acquisition or collection development policy
- Departmental criteria (priorities, precedents)
- Core record/content types (need no review)
- Research criteria (interests, significance)
- Uniqueness (only source)
- Value (historical, evidential, can't reproduce)
- Preserved elsewhere (avoid duplication)

LIBRARY OF CONGRESS

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Considerations during Review

Stop if or when the answer is 'no'…

1. Content
   - does the content have value?
   - does it fit your scope?
2. Technical
   - is it feasible for you to preserve the content?
3. Access
   - is it possible to make the content available?

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Documentation

Supplement inventory from Identify

- Descriptions – more granular
  - Not item level, but enough to specify categories
- Extent
  - How much content is there/will there be?
- Use
  - When will content no longer be active?
- Rights
  - Who owns rights to preserve and disseminate?

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Store: What are storage needs?

Archival Storage manages content as objects

Digital content (files + metadata = object)

- May include any type of content
  - e.g., images, text, sound, video, maps
- Requires some identification and description
  - Captured as metadata
- Needs at least two copies at least two places

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Number of Copies

How many copies are enough for you?

Minimum: two (2) copies in two location

Optimum: six (6) copies

Examples of storage factors:

- Video files are too large to store 6 copies
- Possible legal restrictions (e.g., storage locations)
- Types of media used for storing the content

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Storage Media Options

- Content (objects) are kept on storage media
- Options include: online, near-line, offline
- Factors for choosing options include
  - Cost (available resources for preservation)
  - Quantity (size and number of files)
  - Expertise (skills required to manage)
  - Partners (achieving geographic distribution)
  - Services (outsourcing)

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Storage Considerations

- Multiple, geographically distributed copies
- Storage Partners
- Hosted services, e.g.

DURACLOUD™

This is a service to make it easy for organizations to use cloud services to manage content over time

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Protect: From what?

- Change and loss – accidental and intentional
- Obsolescence – as technology evolves
- Inappropriate access – e.g., confidential data
- Non-compliance – standards and requirements
- Disasters – emergencies of all kinds

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Everyday Protection

- Know where your content is located
  - Onsite and offsite; online and offline
- Know who can have access to it
  - DP staff, IT staff, others?
- Manage authentication information
  - For staff, depositors, users
- Track and review usage then adjust practices
  - Web use, internal use and activities, maintenance

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Emergency Protection

- Engage in ongoing disaster planning
  - Establish committee and share information
  - Develop and maintain documents
- Identify possible outcomes and prepare
  - e.g., server goes down, media is damaged

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Manage: Achieve Balance

An effective approach will address:

- Organizational – requirements and objectives
- Technological – opportunities and change
- Resources – funding, staff, equipment, etc.

A sustainable program will:

- Align with community standards and practice

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Trusted Digital Repository

A TDR should have these characteristics:

- **community standards** (OAIS Compliance)
- **commitment** (Administrative Responsibility)
- **management** (Organizational Viability)
- **resources** (Financial Sustainability)
- **infrastructure** (Technological ... Suitability)
- **protection and control** (System Security)
- **documentation** (Procedural Accountability)

LIBRARY OF CONGRESS

DPOE Baseline Modules: Identify, version 2.0, Nov 2011

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Planning

- Preservation Planning (ongoing)
- Self-assessment (internal process)
- Audit (external review by peers)

Also

- Business Continuity (Protect Module)
- Disaster Planning (Protect Module)

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Provide: Long-term Access

Preservation makes long-term access possible…

Preservation	**vs.**	**Access**
proven	<- technologies ->	cutting edge
accumulate	<- metadata ->	relevant now
access over time	<- purpose ->	access now
future users	<- focus ->	current users

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Requirements for providing content

Content should be delivered to users over time:

- **Easily** – using current and known technologies
- **Coherently** – well-documented and presented
- **Completely** – intact and well-formed
- **Correctly** – accurately representing deposits
- **Reliably** – using well-managed technologies
- **Consistently** – in accordance with policies
- **Fairly** – with equity and precedent

LIBRARY OF CONGRESS

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Sustainable Access

Effective and sustainable DP programs address:

- Value – understand and stress content value
- Roles – identify stakeholders and involve them
- Incentives – identify “carrots” for preserving

Identify and address costs across life cycle

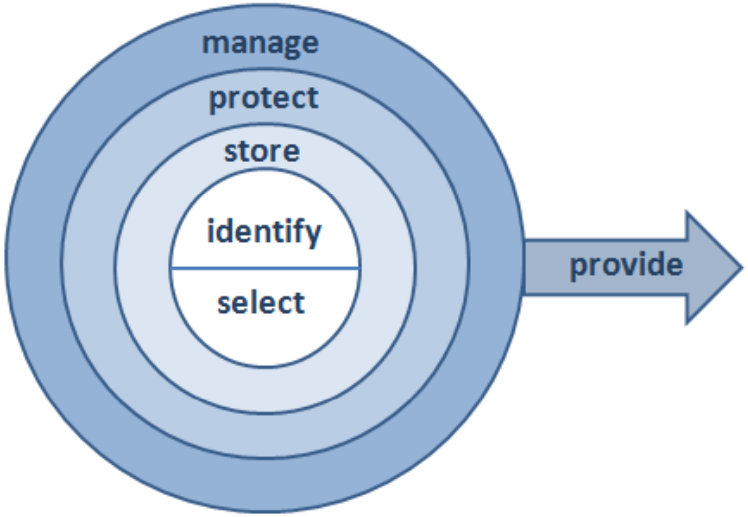
See: Blue Ribbon Task Force Report on Sustainable Preservation and Access Report

LIBRARY OF CONGRESS

DPOE Baseline Modules: Identify, version 2.0, Nov 2011

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

DPOE Management Model



Identify - what digital content do you have?

Select - what portion of that content will be preserved?

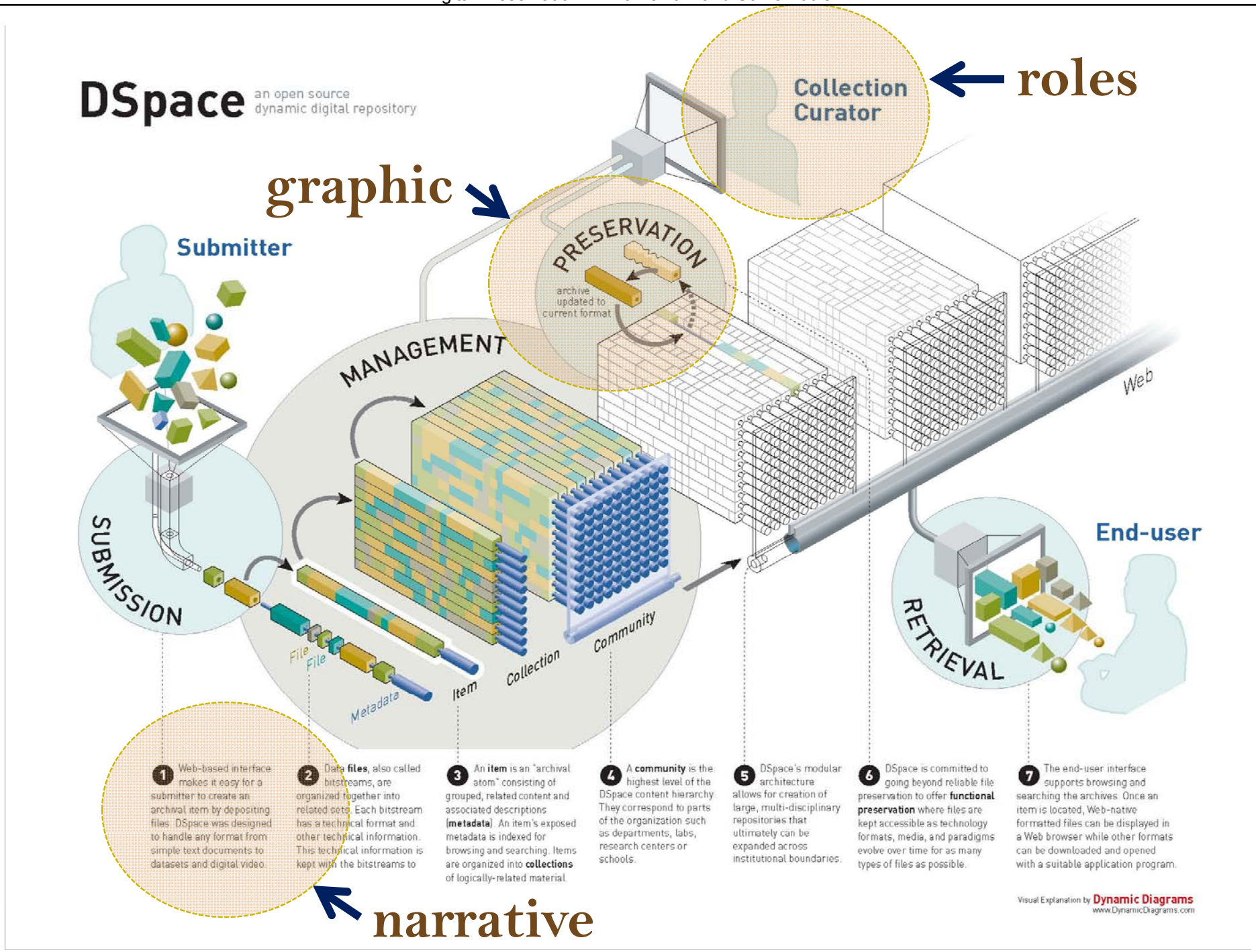
Store - what issues are there for long term storage?

Protect - what steps are needed to protect your digital content?

Manage - what provisions are needed for long-term management?

Provide - what considerations are there for long-term access?

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu



Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Tools Overview

- Often initiated to solve a specific problem
- Tools plug into workflows and repositories
- Digital Preservation tool developments began with Ingest
- Increasing numbers of open source tools
- Generic tools suffice in some cases
- Status: emerging and evolving rapidly

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Importance of Metadata

- How do you know what an object is?
  - Metadata uniquely identifies digital objects
- How do you use content in the future?
  - Metadata makes digital objects understandable
- How do you know an object is authentic?
  - Metadata allows objects to be traced over time

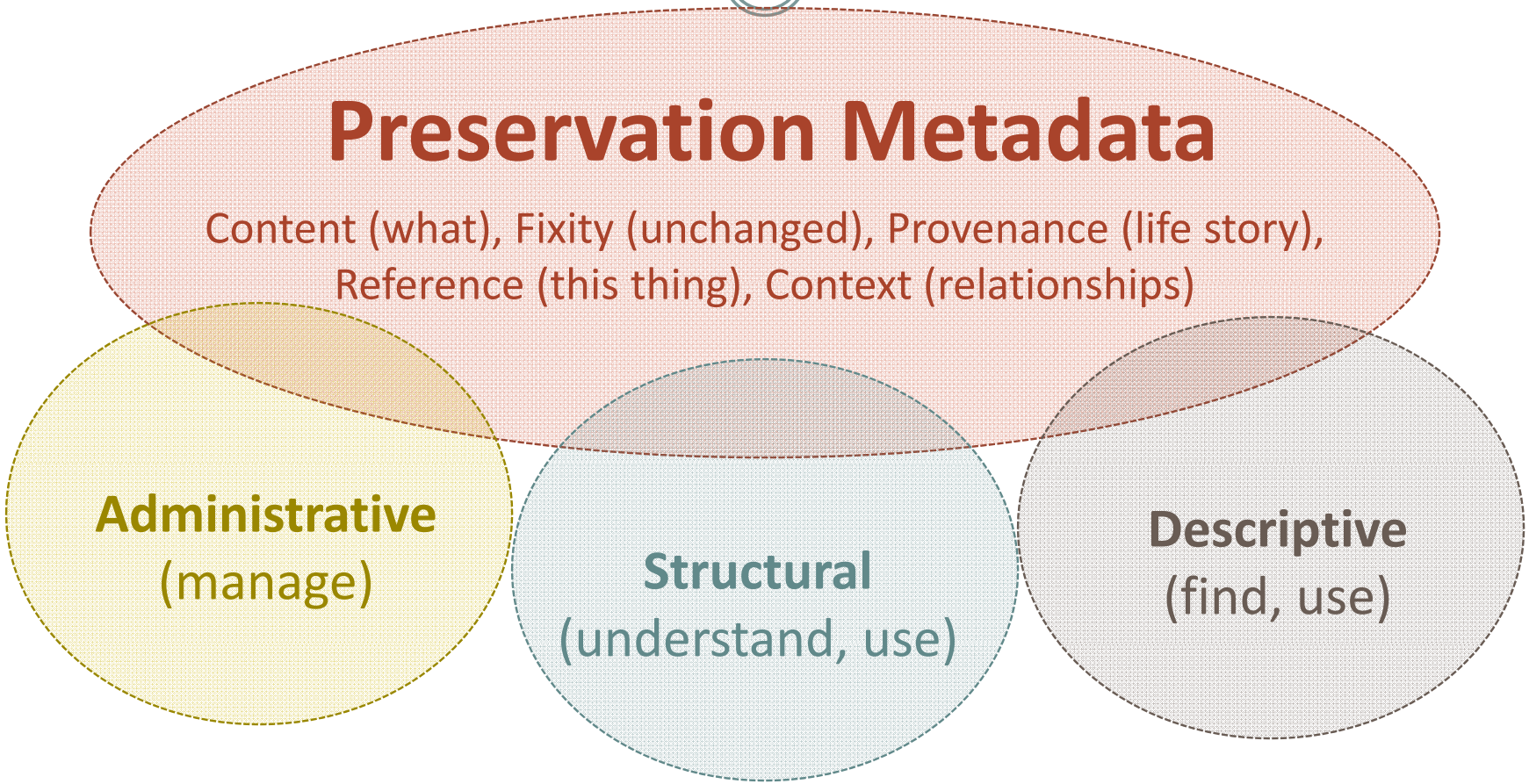
Metadata enables long-term preservation

LIBRARY OF CONGRESS

DPOE Baseline Modules: Identify, version 2.0, Nov 2011

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Object-level Metadata



Source: DPM Workshops

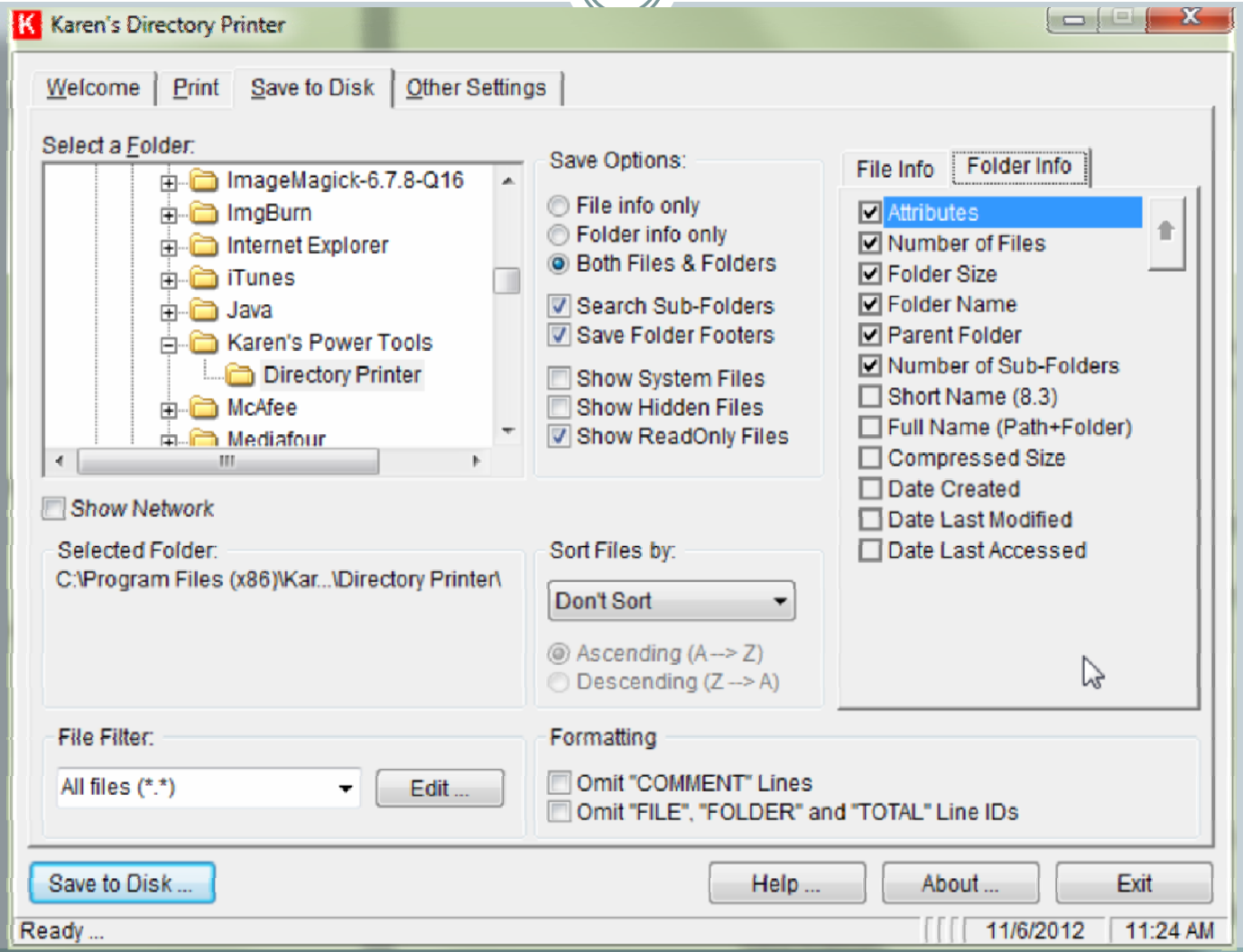
Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Tasks during transfer/ingest

- Virus checking
- Inventory
- File type identification/verification/validation
- Metadata extraction (admin/technical)
- File normalization
- Fixity verification
- Persistent identifiers (assign and manage)
- Arrangement and description

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

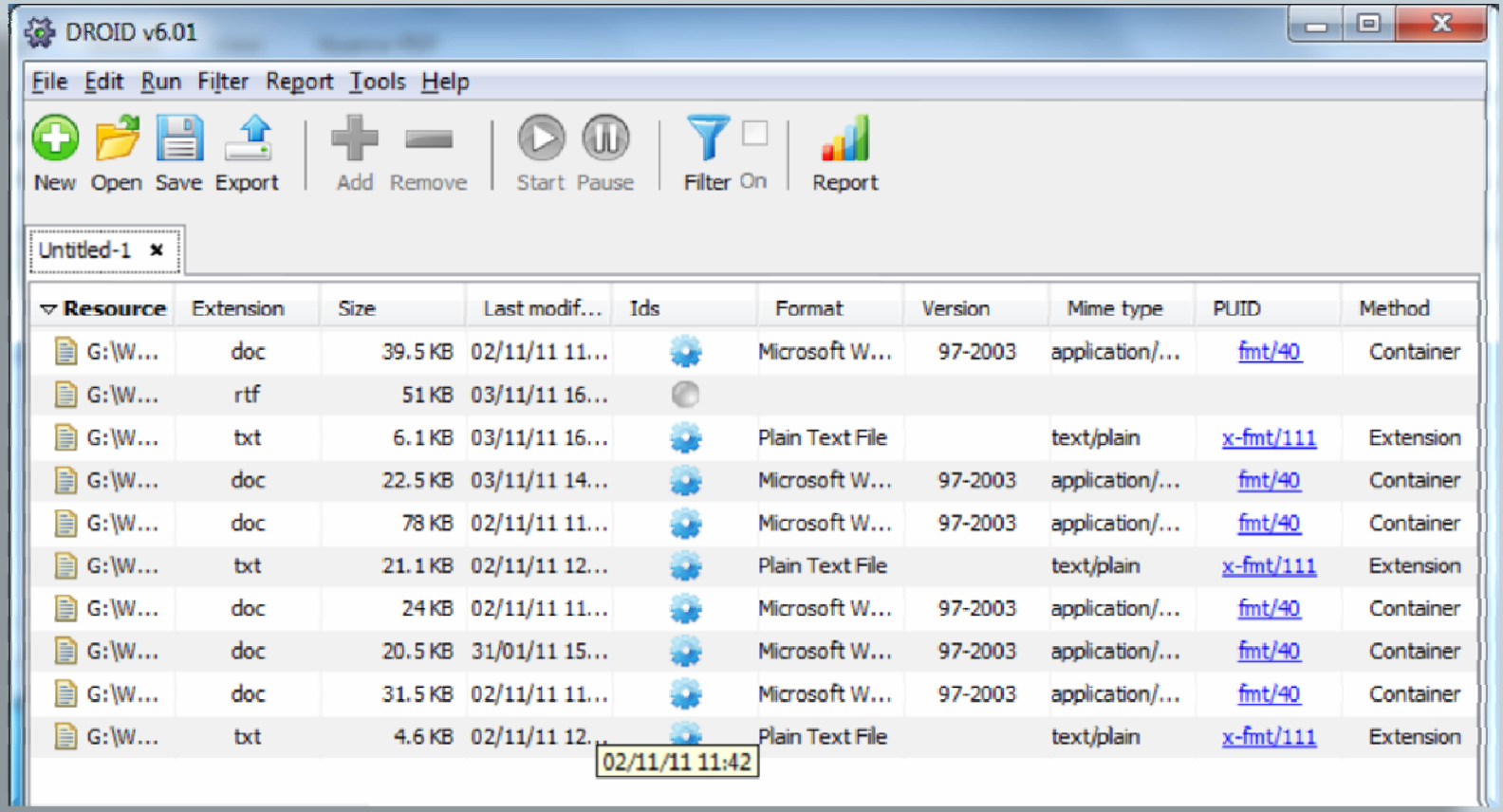
Karen's Directory Printer



Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

File Type identification / verification

DROID (digital record object identifier)



Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

File Fixity

- Checksums (digital thumbprint, Hash)

- Create checksum on receipt
- Verify during processing
- Create new checksum for normalized files
- Verify during processing

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Generated Checksums

Karen's Directory Printer

	Filename	Date last accessed	Date created	File extension	File Size	Checksum	Filepath
2	Bookplate.pub	18/02/2009 20:27	08/01/2009 17:55	pub	26,112	B6B81A1D799335CDAA4989C9F4CB78FE	D:\born digital\Stephen Gallagher\Bookplate.pub
3	CompSlip.pub	24/06/2008 23:42	08/01/2009 17:56	pub	20,992	868A497B18FAAA988597B48300240D48	D:\born digital\Stephen Gallagher\CompSlip.pub
4	Letterhead template.doc	24/06/2008 23:42	06/09/2007 19:14	doc	21,504	40265233F0064BF21F0386D99A4774BF	D:\born digital\Stephen Gallagher\Letterhead template.doc
5	A Year in Oktober.doc	24/06/2008 23:42	16/10/1998 17:32	doc	55,296	AD0E4497263F85CE3C15694C35C61FF0	D:\born digital\Stephen Gallagher\Articles\A Year in Oktober.doc
6	Abner Stein Books CV April 2006.doc	24/06/2008 23:42	17/04/2006 15:14	doc	32,768	B95AC91D8B84EB21145E9C53BA6032AA	D:\born digital\Stephen Gallagher\Articles\Abner Stein Books CV April 2006.doc
7	Agency CV November 2005.doc	24/06/2008 23:42	22/11/2005 01:28	doc	40,960	CB63602C9473330A4C39B8585411015A	D:\born digital\Stephen Gallagher\Articles\Agency CV November 2005.doc
8	American notes 2007.doc	24/06/2008 23:42	13/07/2007 12:15	doc	72,704	F5C662A91F11EA9FC65DF8AA31575EEC	D:\born digital\Stephen Gallagher\Articles\American notes 2007.doc
9	ARCHIVET.doc	24/06/2008 23:42	16/09/2001 21:17	doc	34,304	849B48A6EF4756C65E9F232B123DCE3E	D:\born digital\Stephen Gallagher\Articles\ARCHIVET.doc
10	AUTHOR QUESTIONNAIRE.doc	24/06/2008 23:42	23/10/2006 20:14	doc	45,056	3A08299E4D4121D934A1A21AB29247A6	D:\born digital\Stephen Gallagher\Articles\AUTHOR QUESTIONNAIRE.doc
11	Avengers.doc	24/06/2008 23:42	09/04/2002 00:33	doc	60,416	BBBB419E5BE327D4D7D515EC0986EBCD	D:\born digital\Stephen Gallagher\Articles\Avengers.doc
12	BLOG IDEAS FROM FM.doc	24/06/2008 23:42	26/05/2008 14:14	doc	32,768	F68BC97A9705EACBF824924959B1288E	D:\born digital\Stephen Gallagher\Articles\BLOG IDEAS FROM FM.doc
13	blog-06-17-2009.xml	14/07/2010 12:03	17/06/2009 23:00	xml	2,105,247	71B7471E12429EA2BDCB80130CEDF4B9	D:\born digital\Stephen Gallagher\Articles\blog-06-17-2009.xml
14	BOOKFILM.doc	24/06/2008 23:42	28/06/2002 17:47	doc	37,376	E4430D087DF3BEDC8CF2DF70069BDDA4	D:\born digital\Stephen Gallagher\Articles\BOOKFILM.doc
15	BRIANCLEMENS.DOC	14/07/2010 12:03	08/08/2009 21:56	DOC	26,624	1E7582AB2E385C896AB2DC94FA893AC9	D:\born digital\Stephen Gallagher\Articles\BRIANCLEMENS.DOC

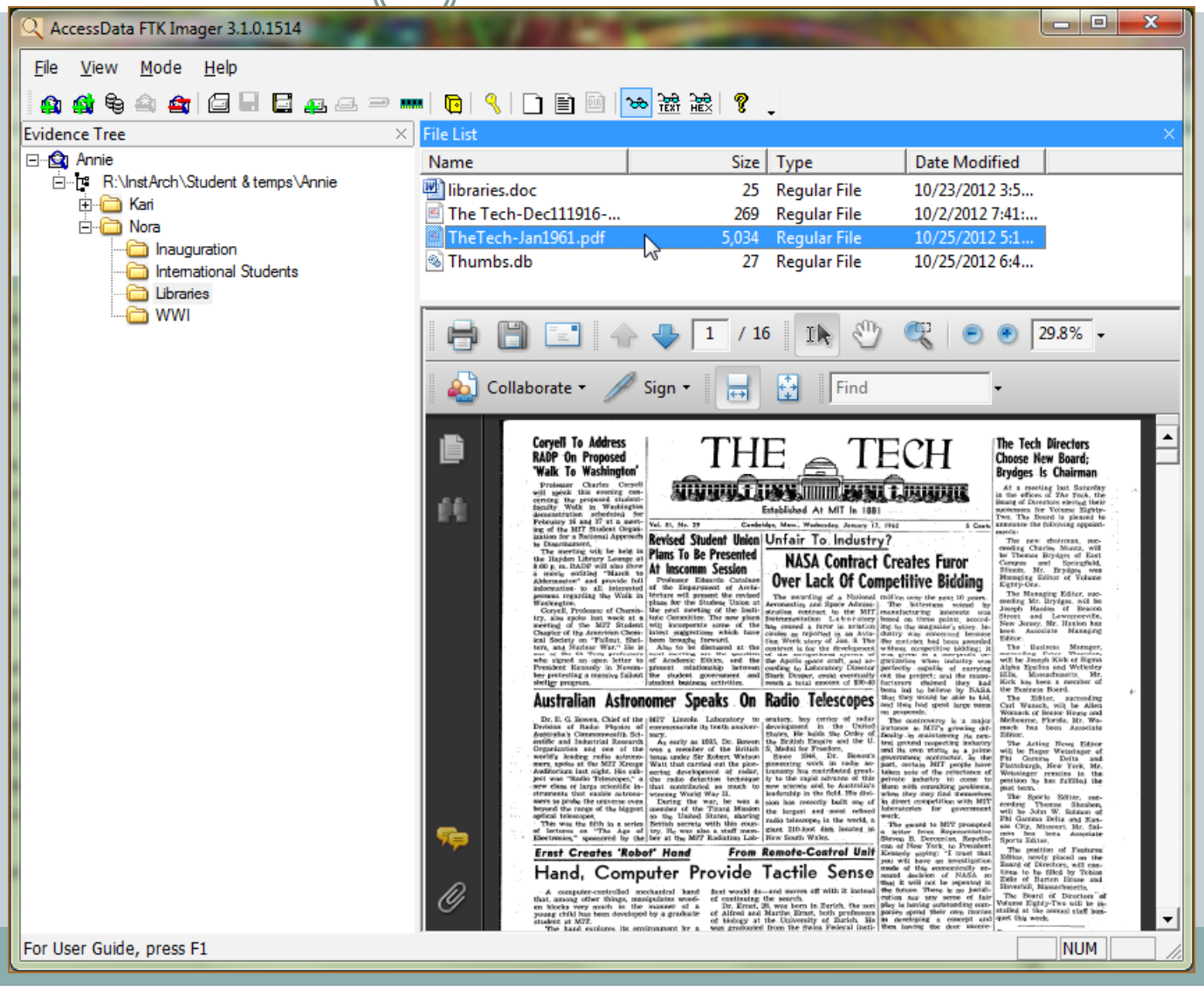
	MD5	SHA1	FileNames
2	4845a6828	02c12564e	Annie\R:\InstArch\Student & temps\Annie\Kari\DigitalCollectionWorkflow_Draftv.2.doc
3	176e1904(	454ee4cel	Annie\R:\InstArch\Student & temps\Annie\Kari\draftPREMIS RIGHTS_09_fillInForm.pdf
4	fef985e08	0f90a8dba	Annie\R:\InstArch\Student & temps\Annie\Kari\draftPREMIS RIGHTS_09_fillInForm_KRS1.pdf
5	6b1d2783!	5dc201222	Annie\R:\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInForm.doc
6	cfa879dc9	fed2a502(	Annie\R:\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInForm.pdf
7	e7ebf559!	f60bf4390	Annie\R:\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInFormv2.1.doc
8	41039defe	469ed9e5(	Annie\R:\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInFormv2.1.pdf
9	1709c2b1c	f57a7282a	Annie\R:\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInFormv2.doc
10	ee2c40b6l	4e71c750!	Annie\R:\InstArch\Student & temps\Annie\Kari\Screenshots\capture_004_06112012_105431.png
11	07cda40f3	f337f1036(	Annie\R:\InstArch\Student & temps\Annie\Kari\Screenshots\capture_005_06112012_105450.png
12	bb14acfbc	abae6a5a(	Annie\R:\InstArch\Student & temps\Annie\Kari\Screenshots\capture_006_06112012_105600.png

FTK Imager

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Review files for Appraisal and Selection

QuickViewPlus



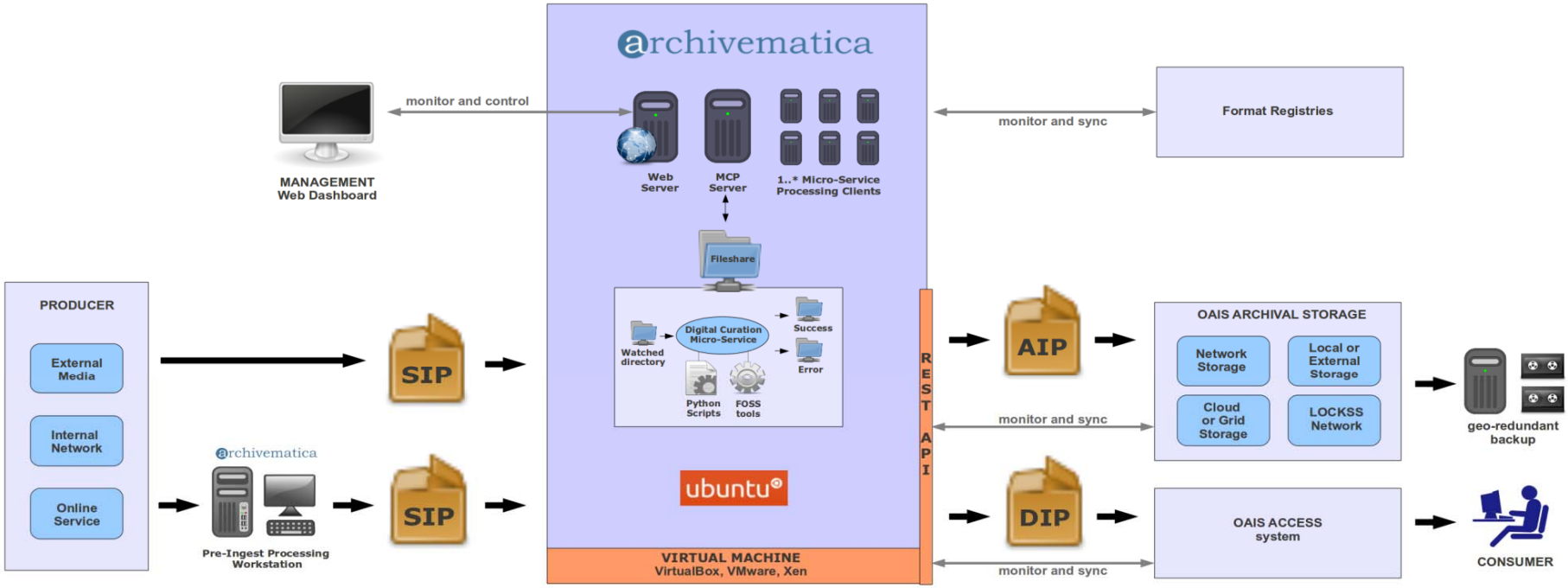
Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Tools: Preservation Planning

- Technology watch services
- Current awareness services
  - RSS feeds
  - listservs
- Information compendia
- Metadata frameworks
- User preference data
- Risk management procedures

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Archivematica is a series of tools in a workflow to produce Archival and Dissemination content packages

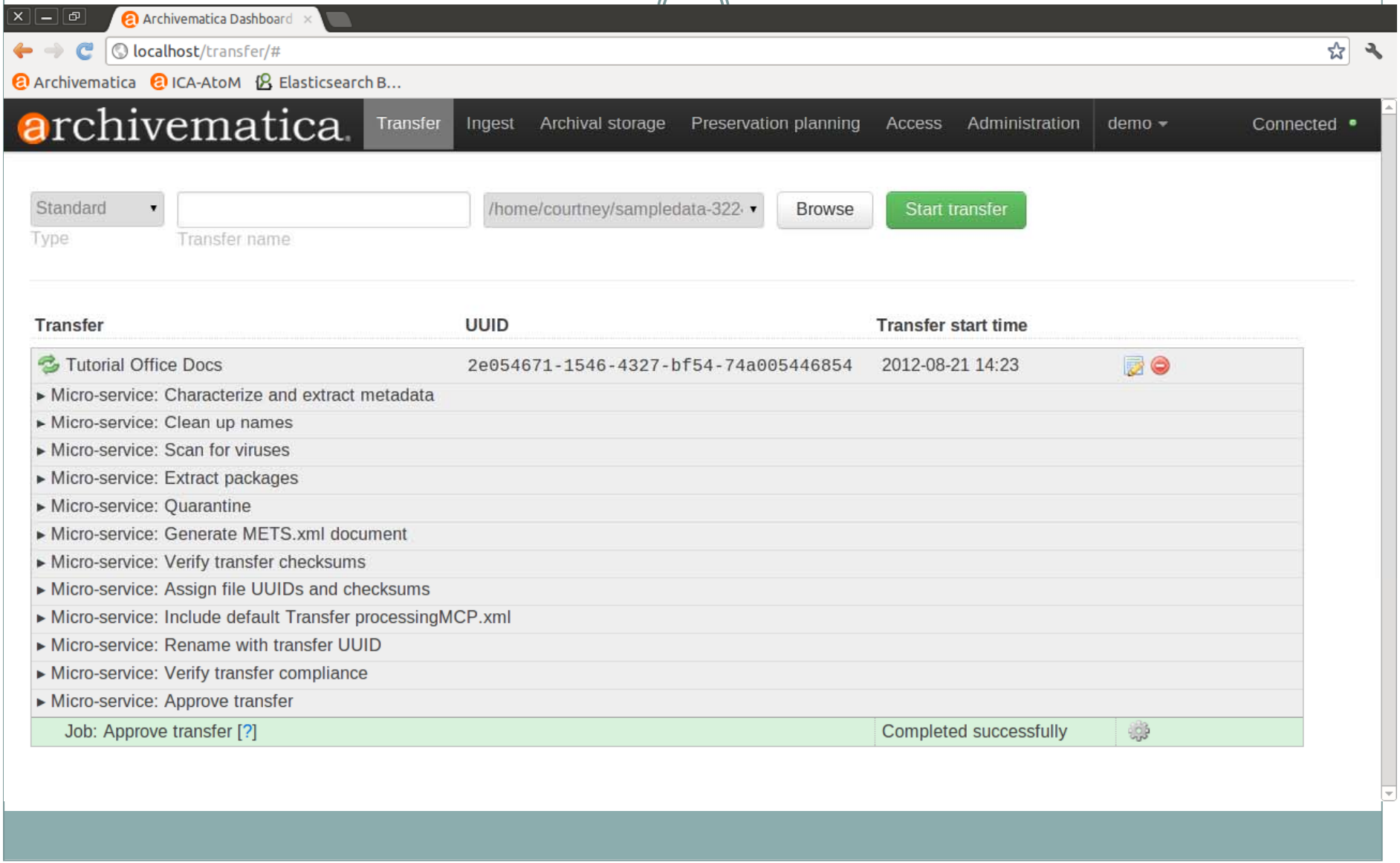


alpha version 0.9 in 2013

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

workflow of tools (SIP to AIP, DIP)

archivematica.



Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

University of California
CDL
California Digital Library

CDL Microservices

Curation Micro-services

- Identity Service
- Storage Service
- Fixity Service
- Replication Service
- Inventory Service
- Characterization Service
- Ingest Service
- Index Service
- Search Service
- Transformation Service
- Notification Service
- Annotation Service
- Common Services

About CDL | Services and Projects | Information Gateways | Commi

CDL Home > Services and Projects > UC3 > Curation

Curation Micro-Services

Micro-services are an approach to digital curation based on devolving curation function into a set of independent, but interoperable, services that embody curation values and strategies. Since each of the services is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance. Equally as important, they are more easily replaced when they have outlived their usefulness. Although the individual services are narrowly scoped, the complex function needed for effective curation emerges from the strategic combination of individual services.

Micro-services provide a curation environment that is comprehensive in scope, yet flexible with regard to local policies and practices and the inevitability of disruptive technological change. Micro-services can be deployed in environments in which it makes most sense, both technically and administratively. UC3 will use micro-services as the basis for its centrally-managed curation activities (for example, the Digital Preservation Repository); micro-services can also be operated in local campus environments either individually or in strategic combinations.

The initial set of micro-services can be grouped into four categories that provide incrementally increasing levels of preservation assurance and curation value. For more information and documentation, see the UC3 Curation wiki.

Microservices launched 2010 plus Camp Curate

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

3. Ingest

Brings the files and associated processing documentation into the physical and intellectual custody of the archives. These steps are to be completed after returning to the archives.

File Verification & Summarization

- Create a README file for the acquisition.
- Make all folder(s) in the acquisition read-only.

Transfer and Ingest Package

- Review the package you have created.

README.txt
Use the template provided (*README_template.txt*). Name the file "README_[transfer no].txt" and save in the Acquisition Process folder. Do not save over the template. Transfer the information from the On-Site Data Transfer Documentation Form or the Materials Receipt Form. (See: Appendix B:2)

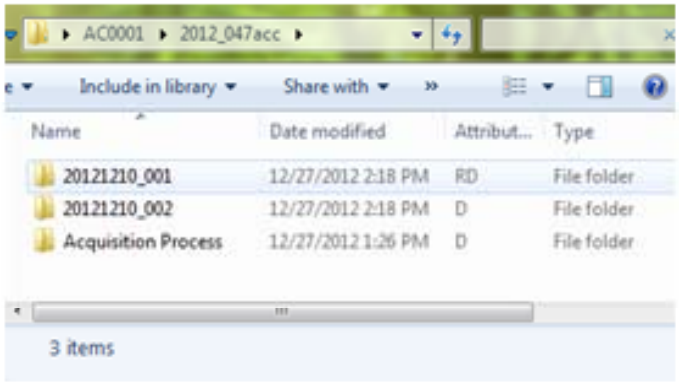


Figure 1 Ingest Package

- Within the Acquisition Process folder there should be contained at least the following processing documentation:
  - README .txt file
  - Directory Printer.txt file
  - File Hash List .txt (from FTK Imager)
  - .xml file (from Data Accessioner) **OR** DROID .txt file

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Wrap Up

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

Well-managed Collections

Well-managed status makes preservation easier

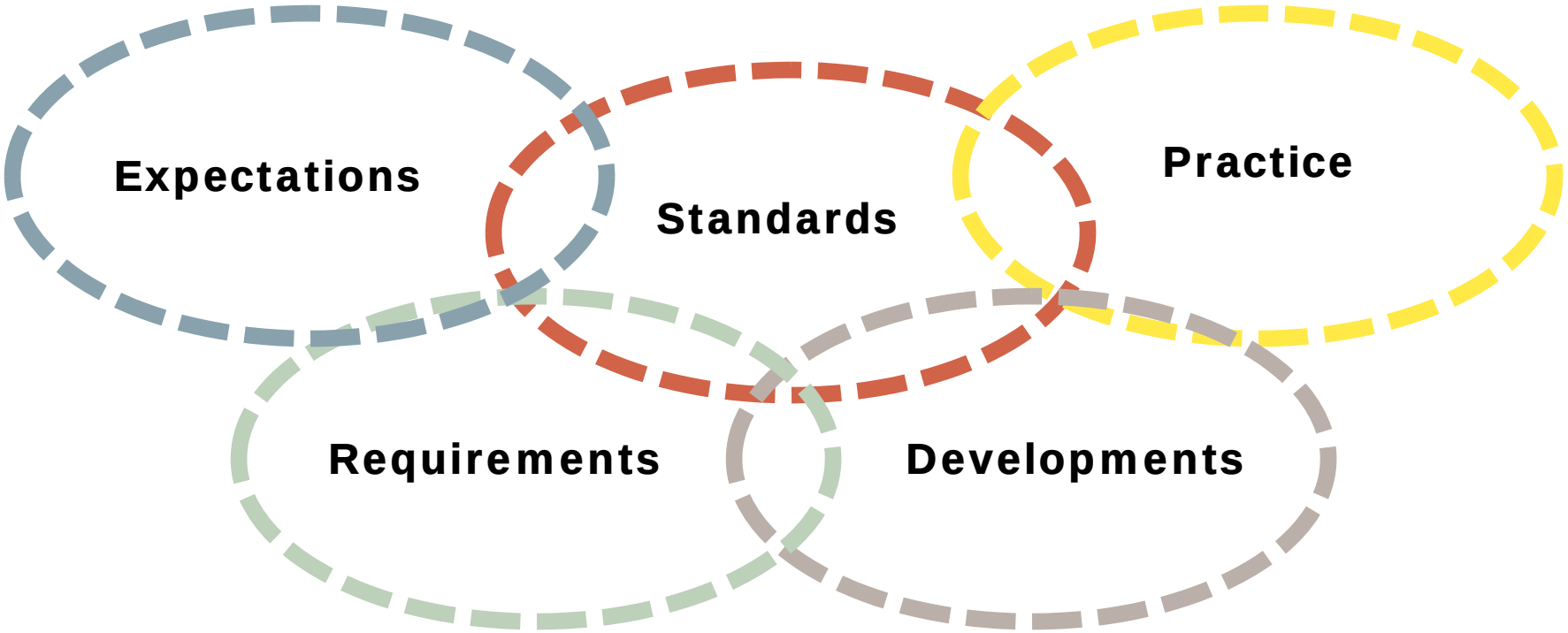
Sample characteristics of well-managed:

- Basic information about each deposit
- Minimal metadata for objects (you define)
- Common (or *normalized*) file formats
- Controlled and known storage of content
- Multiple copies in at least 2 locations

Source: DPM Workshops

Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu

DP Program Balance



Nancy Y. McGovern and Kari R. Smith
nancymcg@mit.edu | smithkr@mit.edu