

Digital Preservation



**NANCY Y. MCGOVERN AND
KARI R. SMITH**

**NETSL ANNUAL SPRING CONFERENCE
APRIL 2013**

Overview



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

Sources

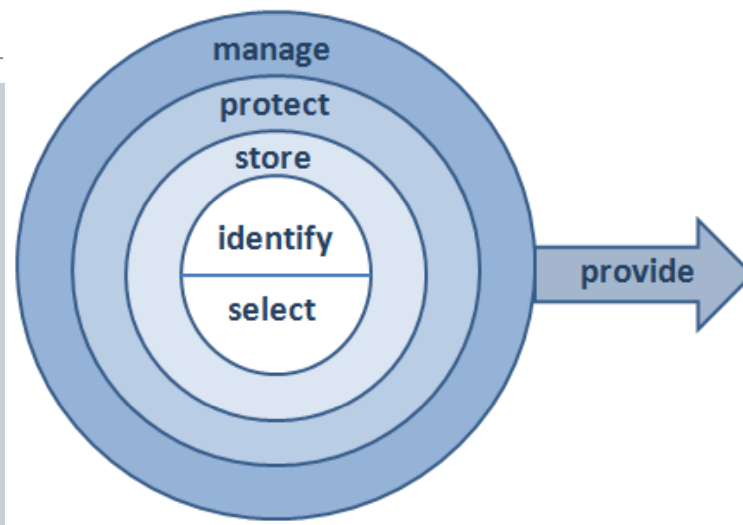


Sources:

- 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

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?

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?”

Content Categories



Inventories should include all relevant, e.g.:

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

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)

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?

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?

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

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

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)

Storage Considerations



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



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

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

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

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

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

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)

Planning



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

Also

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

Provide: Long-term Access



Preservation makes long-term access possible...

Preservation

proven

accumulate

access over time

future users

vs.

<- technologies ->

<- metadata ->

<- purpose ->

<- focus ->

Access

cutting edge

relevant now

access now

current users

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

Sustainable Access



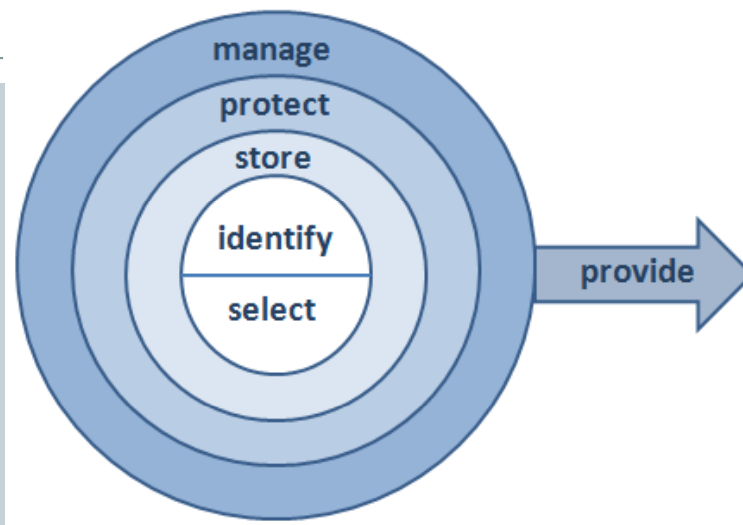
Effective and sustainable DP programs address:

- 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

DPOE Management Model



Identify - what digital content do you have?

Select - what portion of that content will be preserved?

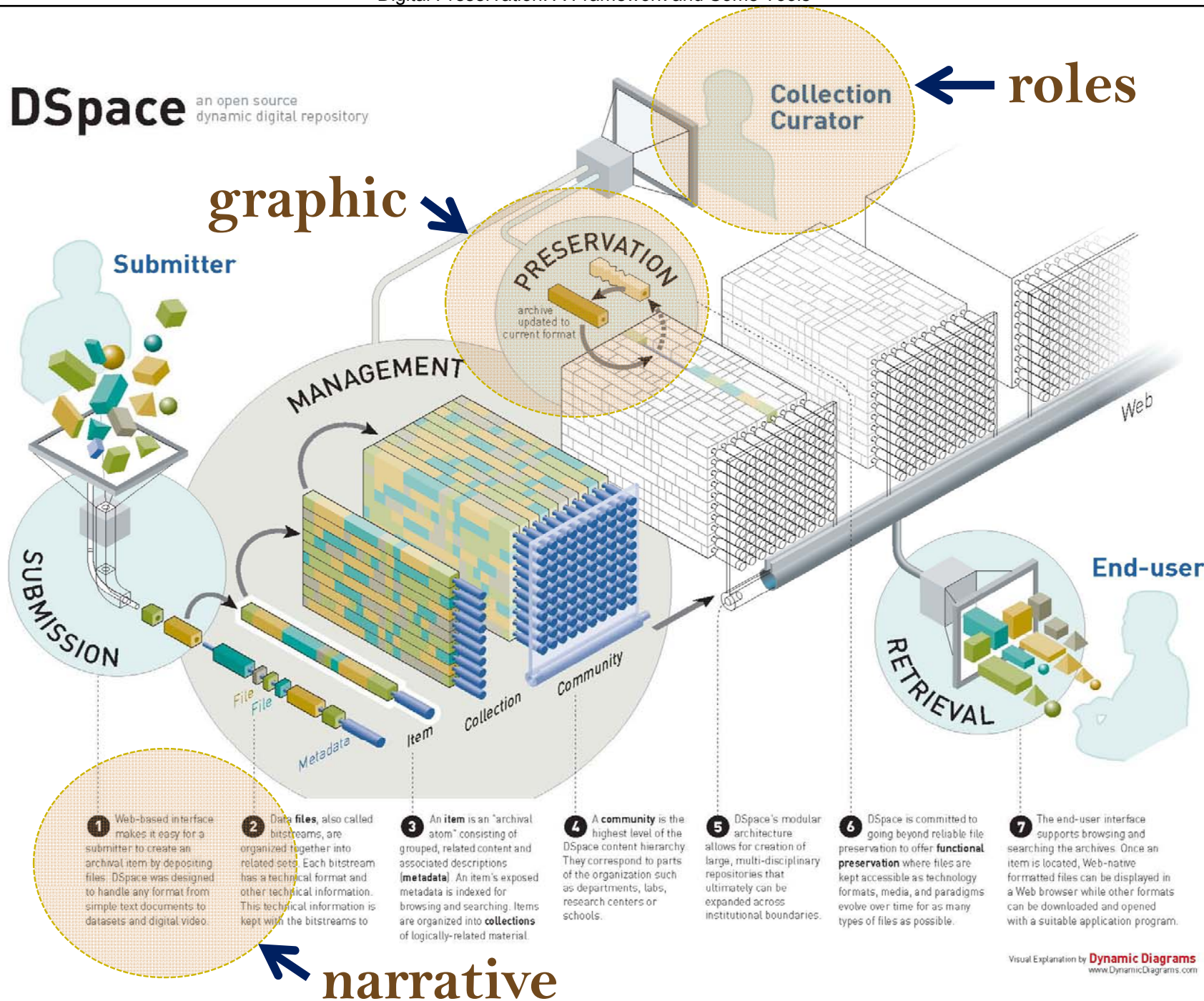
Store - what issues are there for long term storage?

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?

DSpace an open source dynamic digital repository



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

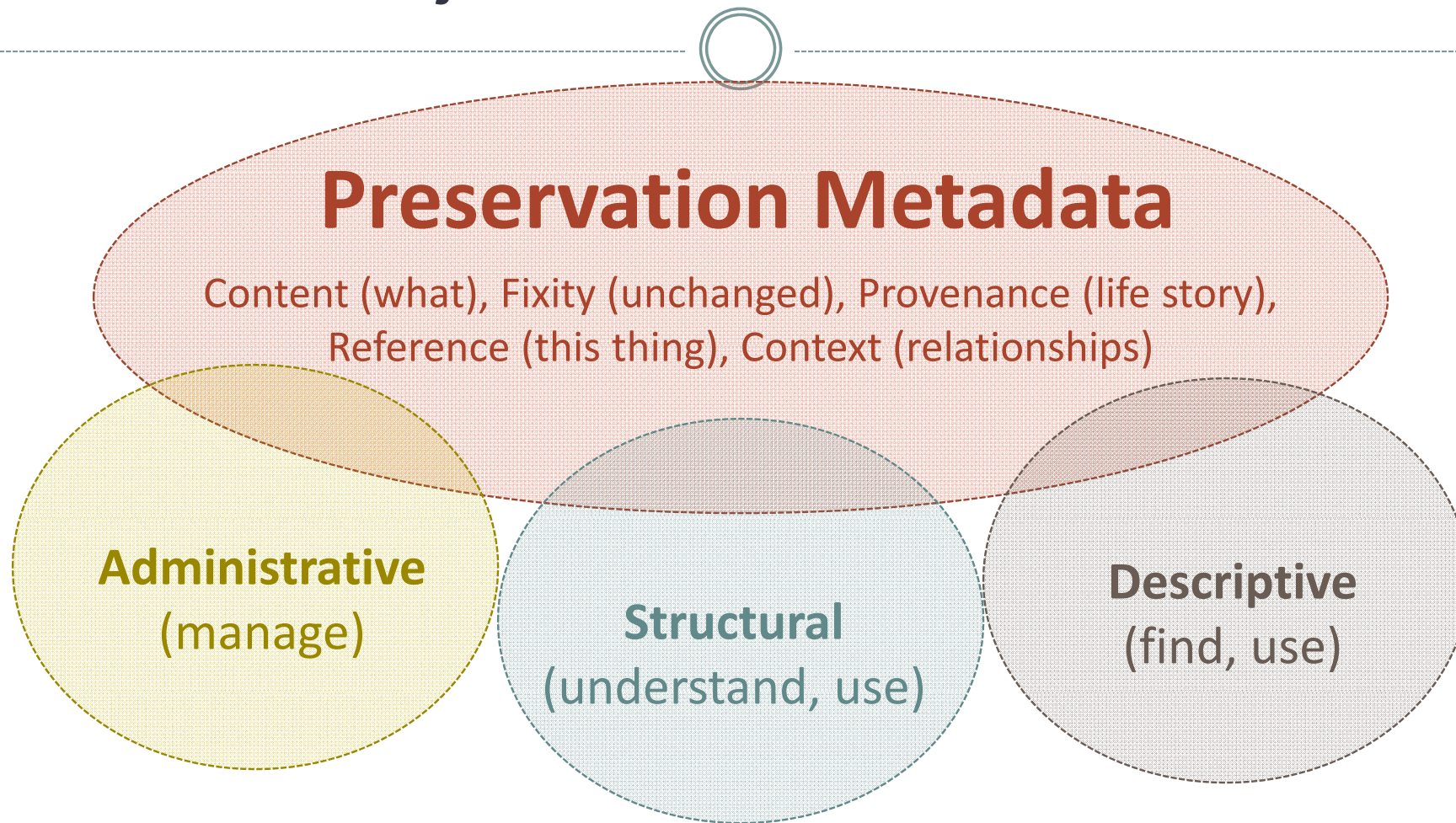
Importance of Metadata



- 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

Object-level Metadata



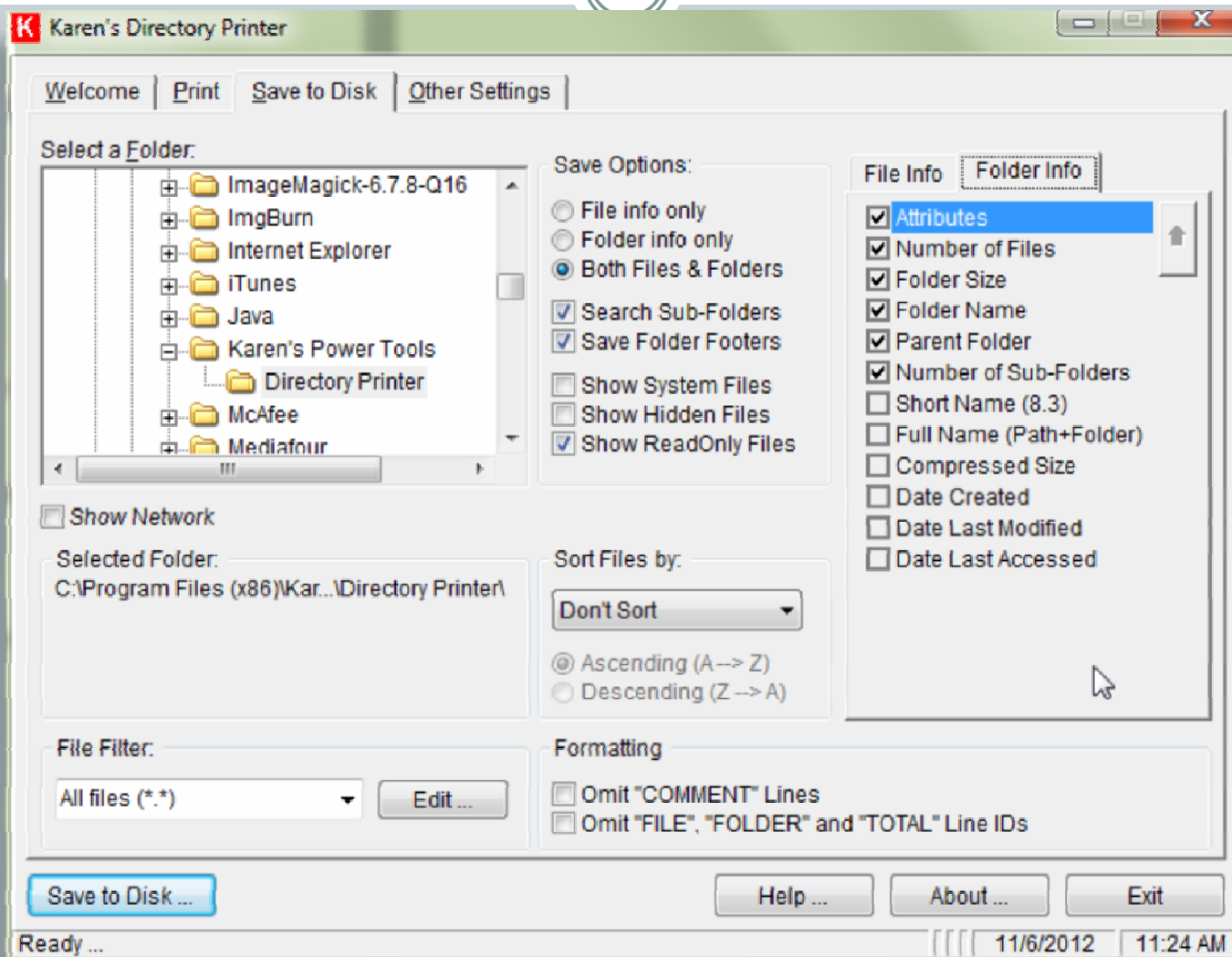
Source: DPM Workshops

Tasks during transfer/ingest



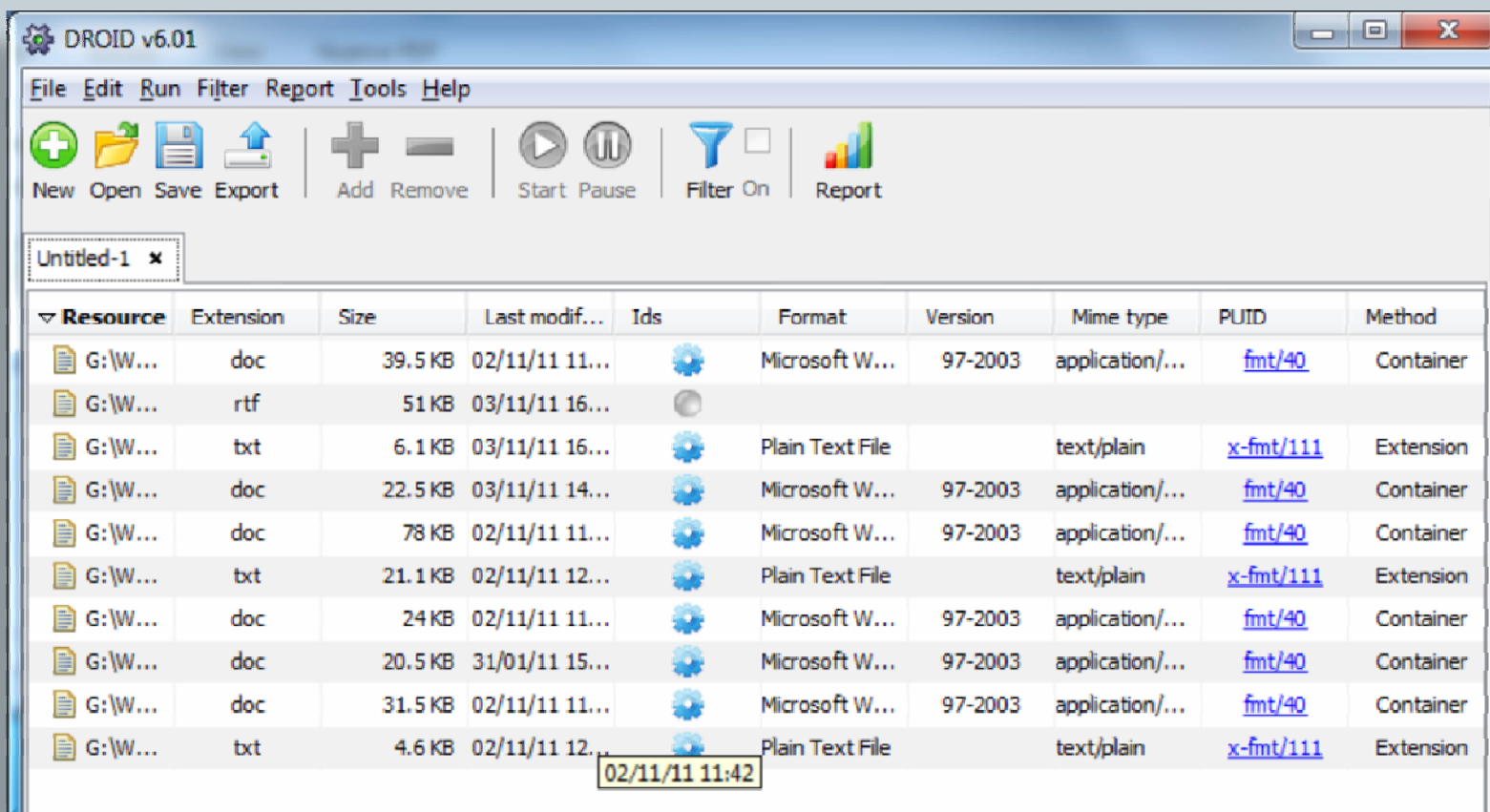
- Virus checking
- Inventory
- File type identification/verification/validation
- Metadata extraction (admin/technical)
- File normalization
- Fixity verification
- Persistent identifiers (assign and manage)
- Arrangement and description

Karen's Directory Printer



File Type identification / verification

DROID (digital record object identifier)



The screenshot shows the DROID v6.01 application window. The title bar reads 'DROID v6.01'. The menu bar includes 'File', 'Edit', 'Run', 'Filter', 'Report', 'Tools', and 'Help'. The toolbar contains icons for 'New', 'Open', 'Save', 'Export', 'Add', 'Remove', 'Start', 'Pause', 'Filter On', and 'Report'. Below the toolbar is a tab labeled 'Untitled-1'. The main area displays a table with the following columns: Resource, Extension, Size, Last modif..., Ids, Format, Version, Mime type, PUID, and Method. The table contains 10 rows of data, each representing a file with its path, extension, size, last modification date, and identified format and PUID.

Resource	Extension	Size	Last modif...	Ids	Format	Version	Mime type	PUID	Method
G:\W...	doc	39.5 KB	02/11/11 11...		Microsoft W...	97-2003	application/...	fmt/40	Container
G:\W...	rtf	51 KB	03/11/11 16...						
G:\W...	txt	6.1 KB	03/11/11 16...		Plain Text File		text/plain	x-fmt/111	Extension
G:\W...	doc	22.5 KB	03/11/11 14...		Microsoft W...	97-2003	application/...	fmt/40	Container
G:\W...	doc	78 KB	02/11/11 11...		Microsoft W...	97-2003	application/...	fmt/40	Container
G:\W...	txt	21.1 KB	02/11/11 12...		Plain Text File		text/plain	x-fmt/111	Extension
G:\W...	doc	24 KB	02/11/11 11...		Microsoft W...	97-2003	application/...	fmt/40	Container
G:\W...	doc	20.5 KB	31/01/11 15...		Microsoft W...	97-2003	application/...	fmt/40	Container
G:\W...	doc	31.5 KB	02/11/11 11...		Microsoft W...	97-2003	application/...	fmt/40	Container
G:\W...	txt	4.6 KB	02/11/11 12...		Plain Text File		text/plain	x-fmt/111	Extension

File Fixity



- Checksums (digital thumbprint, Hash)
- Create checksum on receipt
- Verify during processing
- Create new checksum for normalized files
- Verify during processing

Generated Checksums



Karen's Directory Printer

Clipboard	Font	Alignment	Number	Formatting as Table	Styles	Cells
A11	fx	Avengers.doc				
Filename	Date last accessed	Date created	File extension	File Size	Checksum	Filepath
1 Bookplate.pub	18/02/2009 20:27	08/01/2009 17:55	pub	26,112	B6881A1D799335CDA4989C9F4C878FE	D:\born digital\Stephen Gallagher\Bookplate.pub
2 CompSlip.pub	24/06/2008 23:42	08/01/2009 17:56	pub	20,992	868A497B18FAA988597B48300240D48	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	BBB8419E5BE327D4D7D515EC0986EBCD	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

Clipboard		Font		Alignment		Number		Styles		Cells		
A1		fx		MD5								
	A	B	C	D	E	F	G	H	I	J	K	L
1	MD5	SHA1	FileNames									
2	4845a682802c12564e	Annie\R\InstArch\Student & temps\Annie\Kari\DigitalCollectionWorkflow_Draftv.2.doc										
3	176e19046454ee4ce1	Annie\R\InstArch\Student & temps\Annie\Kari\draftPREMIS RIGHTS_09_fillInForm.pdf										
4	fef985e080f90a8dba	Annie\R\InstArch\Student & temps\Annie\Kari\draftPREMIS RIGHTS_09_fillInForm_KRS1.pdf										
5	6b1d278355dc201222	Annie\R\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInForm.doc										
6	cfa879dc9fed2a502c	Annie\R\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInForm.pdf										
7	e7ebf5595f60bf4390	Annie\R\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInFormv2.1.doc										
8	41039def6469ed9e5f	Annie\R\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInFormv2.1.pdf										
9	1709c2b1cf57a7282a	Annie\R\InstArch\Student & temps\Annie\Kari\PREMIS RIGHTS_09_fillInFormv2.doc										
10	ee2c40b6f4e71c7505	Annie\R\InstArch\Student & temps\Annie\Kari\Screenshots\capture_004_06112012_105431.png										
11	07cda40f3f337f1036	Annie\R\InstArch\Student & temps\Annie\Kari\Screenshots\capture_005_06112012_105450.png										
12	bb14acfbcbabae6a5a	Annie\R\InstArch\Student & temps\Annie\Kari\Screenshots\capture_006_06112012_105600.png										

FTK
Imager

Review files for Appraisal and Selection

QuickViewPlus

AccessData FTK Imager 3.1.0.1514

File View Mode Help

Evidence Tree

- Annie
 - R:\InstArch\Student & temps\Annie
 - Kari
 - Nora
 - Inauguration
 - International Students
 - Libraries
 - WWI

File List

Name	Size	Type	Date Modified
libraries.doc	25	Regular File	10/23/2012 3:5...
The Tech-Dec111916-...	269	Regular File	10/2/2012 7:41:...
TheTech-Jan1961.pdf	5,034	Regular File	10/25/2012 5:1...
Thumbs.db	27	Regular File	10/25/2012 6:4...

1 / 16 29.8%

Collaborate Sign Find

THE TECH
Established at MIT in 1981

Caryell To Address RADP On Proposed 'Walk To Washington'

Professor Charles Caryell will speak this evening concerning the proposed student-led walk in Washington demonstration scheduled for February 18 and 19 at a meeting of the MIT Student Organization for a National Approach to Disarmament.

The meeting will be held in the Hayden Library Lounge at 8:30 p.m. RADP will also show a video, entitled "March to Disarmament" and provide full information to all interested persons regarding the Walk to Washington.

Corvett, Professor of Chemistry, also spoke last week at a meeting of the MIT Student Chapter of the American Chemical Society on "Fuel, State, and Nuclear War." He is one of the 600 scientists who signed an open letter to the President Kennedy in November protesting a nuclear fallout shelter program.

Revised Student Union Plans To Be Presented At Incomm Session

Professor Edmund Condon of the Department of Architecture will present the revised plans for the Student Union at the next meeting of the Institute Committee. The new plans incorporate some of the latest suggestions which have been brought forward.

Also to be discussed at the next meeting are the questions of student life, and the present relationship between the student government and student business activities.

Unfair To Industry?

The awarding of a National Aeronautics and Space Administration contract to the MIT Instrumentation Laboratory has caused a furor in aviation circles, as reported in an article in the Week story of Jan. 4. The contract is for the development of the reconnaissance system of the Apollo spacecraft, and according to Laboratory Director Frank Brown, could eventually reach a total amount of \$30-40 million.

The controversy is a major dispute in MIT's growing desire in establishing its own private research industry and its own staff, as a private government contractor. In the past, certain MIT people have taken note of the reluctance of private industry to come to them with consulting questions when they may find themselves in direct competition with MIT laboratories for government work.

The controversy is a major dispute in MIT's growing desire in establishing its own private research industry and its own staff, as a private government contractor. In the past, certain MIT people have taken note of the reluctance of private industry to come to them with consulting questions when they may find themselves in direct competition with MIT laboratories for government work.

Australian Astronomer Speaks On Radio Telescopes

MIT Lincoln Laboratory to commemorate its tenth anniversary.

As early as 1953, Dr. Bowen was a member of the British team under Sir Robert Watson Watt that carried out the pioneering development of radar, the radio detection technique that contributed so much to winning World War II.

During the war, he was a member of the Tizard Mission to the United States, sharing British secrets with this country. He was also a staff member at the MIT Radiation Laboratory.

Ernst Creates 'Robof' Hand From Remote-Control Unit

Hand, Computer Provide Tactile Sense

A computer-controlled mechanical hand that, among other things, manipulates objects with very much in the manner of a human hand has been developed by a graduate student at MIT.

The hand, called the "Robof" hand, was developed by a graduate student at MIT.

The Tech Directors Choose New Board; Brydges Is Chairman

At a meeting last Saturday in the office of Vice Tech, the Board of Directors elected their members for the coming year. The Board is given the following appointments:

The new chairman, succeeding Charles Murray, will be Thomas Brydges of East Greenwich, Rhode Island. Mr. Brydges was Managing Editor of Volume Eighty-One.

The Managing Editor, succeeding Mr. Brydges, will be Joseph Haden of Boston Street and Lawrenceville, New Jersey. Mr. Haden has been Associate Managing Editor.

The Business Manager, succeeding Peter Shannon, will be Joseph Haden of Boston Street and Lawrenceville, New Jersey. Mr. Haden has been Associate Managing Editor.

The Editor, succeeding Carl Wastak, will be Allen Wastak of Boston Street and Lawrenceville, New Jersey. Mr. Wastak has been Associate Editor.

The Acting News Editor will be Roger Weinstein of FBI, Chicago, Illinois, and Stoughton, New York. Mr. Weinstein has been the position he has filled the past year.

The Sports Editor, succeeding Thomas Shannon, will be John W. Haden of FBI, Chicago, Illinois, and Stoughton, New York. Mr. Haden has been the position he has filled the past year.

The position of Features Editor, newly placed on the Board of Directors, will continue to be filled by Thomas Haden of Boston Street and Lawrenceville, New Jersey.

The Board of Directors of Volume Eighty-Two will be installed at the annual staff meeting this week.

For User Guide, press F1

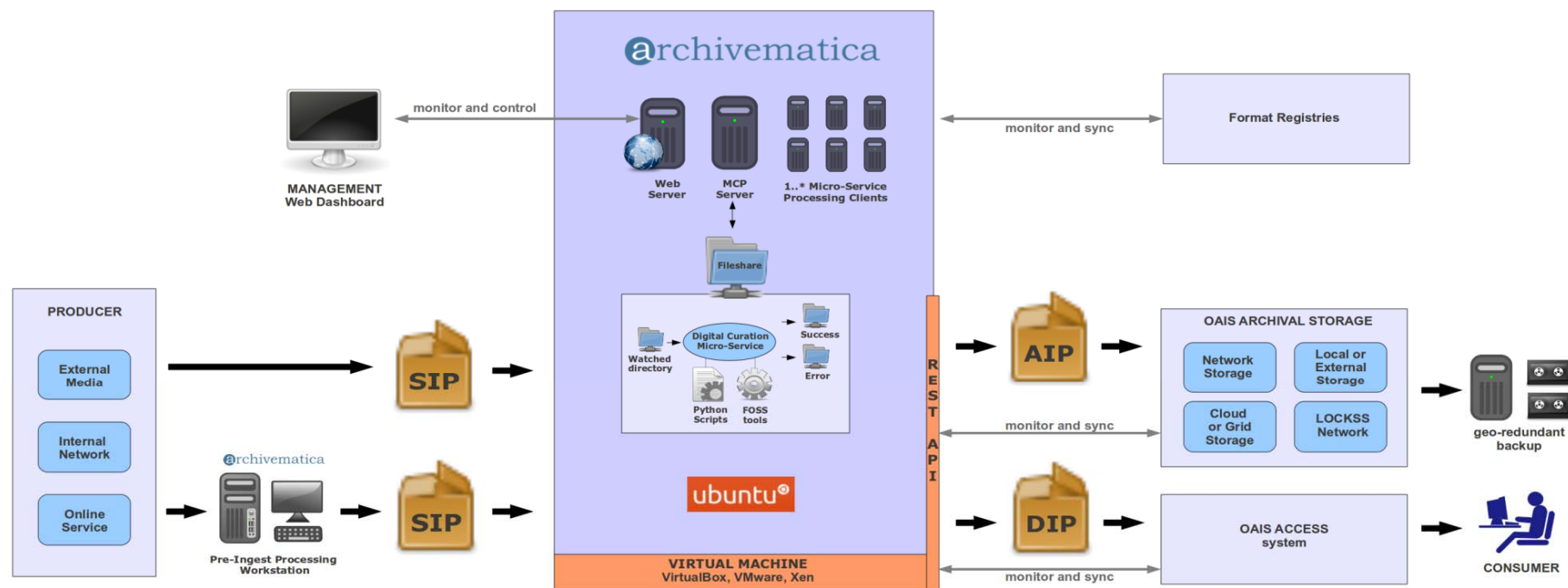
NUM

Tools: Preservation Planning



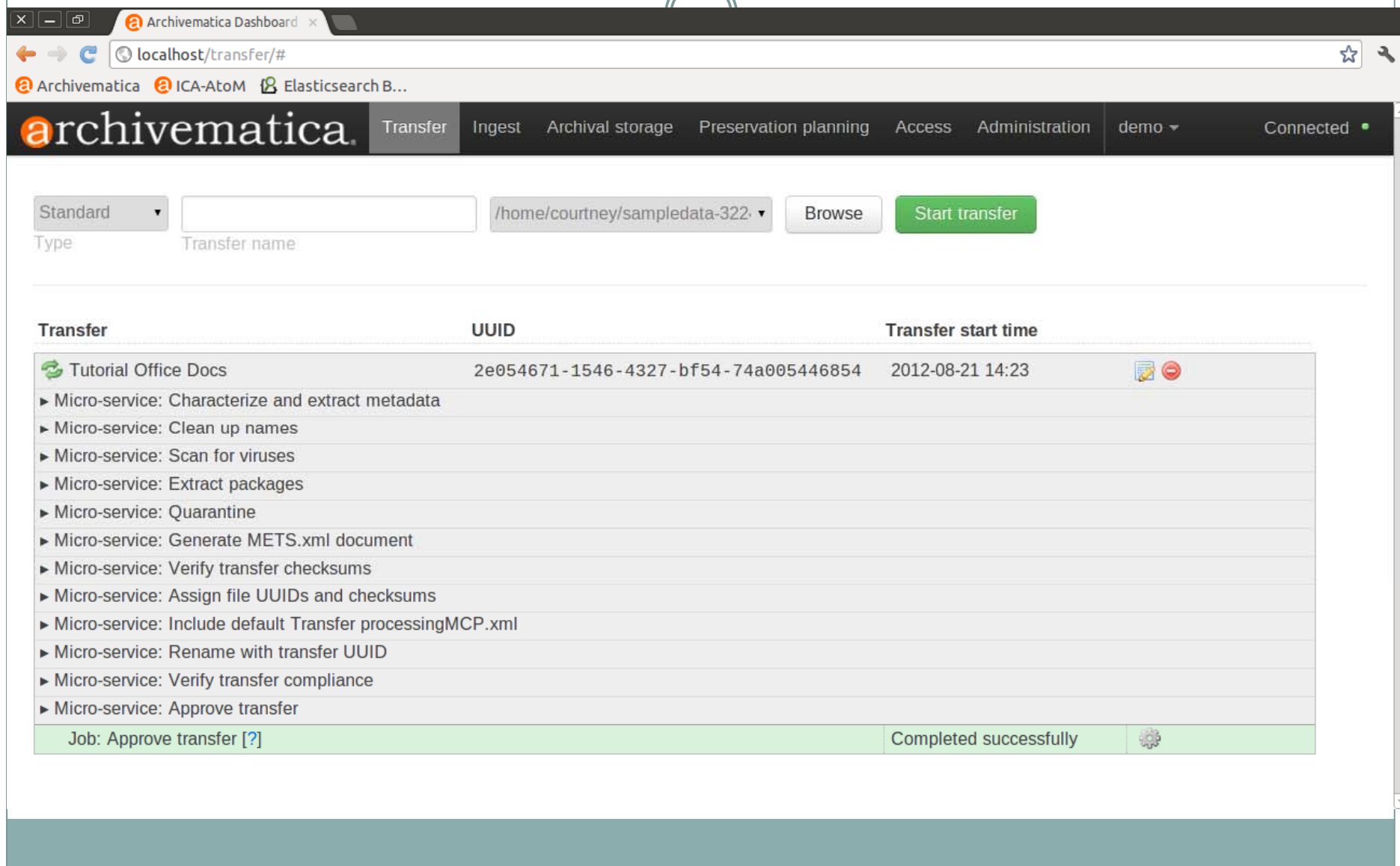
- Technology watch services
- Current awareness services
 - RSS feeds
 - listservs
- Information compendia
- Metadata frameworks
- User preference data
- Risk management procedures

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





alpha version 0.9 in 2013

workflow of tools (SIP to AIP, DIP)

The screenshot shows the Archivematica Dashboard in a web browser. The browser tab is labeled "Archivematica Dashboard" and the address bar shows "localhost/transfer/#". The dashboard has a navigation bar with links: Transfer, Ingest, Archival storage, Preservation planning, Access, Administration, demo, and Connected. Below the navigation bar, there is a form for starting a transfer. The form includes a "Type" dropdown set to "Standard", a "Transfer name" input field, a "Browse" button, and a "Start transfer" button. Below the form, there is a table showing the transfer workflow.

Transfer	UUID	Transfer start time
 Tutorial Office Docs ▶ Micro-service: Characterize and extract metadata ▶ Micro-service: Clean up names ▶ Micro-service: Scan for viruses ▶ Micro-service: Extract packages ▶ Micro-service: Quarantine ▶ Micro-service: Generate METS.xml document ▶ Micro-service: Verify transfer checksums ▶ Micro-service: Assign file UUIDs and checksums ▶ Micro-service: Include default Transfer processingMCP.xml ▶ Micro-service: Rename with transfer UUID ▶ Micro-service: Verify transfer compliance ▶ Micro-service: Approve transfer	2e054671-1546-4327-bf54-74a005446854	2012-08-21 14:23
Job: Approve transfer [?]	Completed successfully	



CDL Microservices

Curation Micro-services

[About CDL](#)
[Services and Projects](#)
[Information Gateways](#)
[Comm](#)
[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](#).

[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](#)

Microservices launched 2010 plus Camp Curate

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

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

Transfer and Ingest Package

- o 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](#))

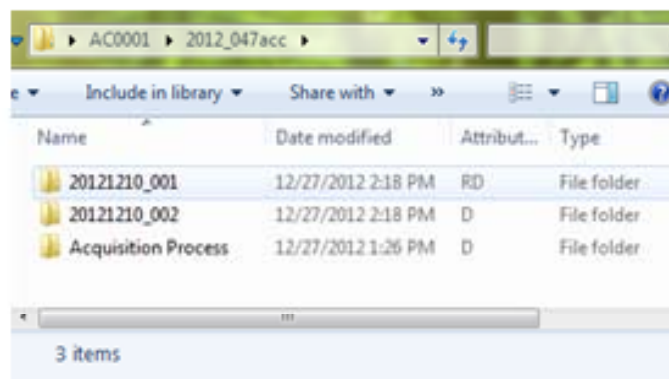


Figure 1 Ingest Package

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



Wrap Up

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

DP Program Balance

