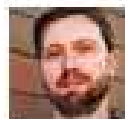




Tackling File Characterization & Analysis with Archivematica

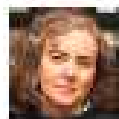
Courtney C. Mumma
CurateGear 2013
Wednesday, January 9, 2013

digital preservation consulting
open-source software for archives and libraries



Peter Van Garderen

President



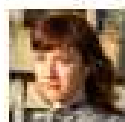
Evelyn McLellan

Director, Consulting Services



David Juhasz

Director, Technical Services



Courtney C. Mumma

Archivematica Product Manager



Jessica Bushey

AtoM Product Manager



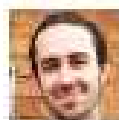
Jesús García Crespo

Software Developer



Joseph Perry

Software Developer



Austin Trask

Systems Technician



Mike Cantelon

Software Developer



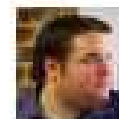
Dan Gilleen

Systems Analyst



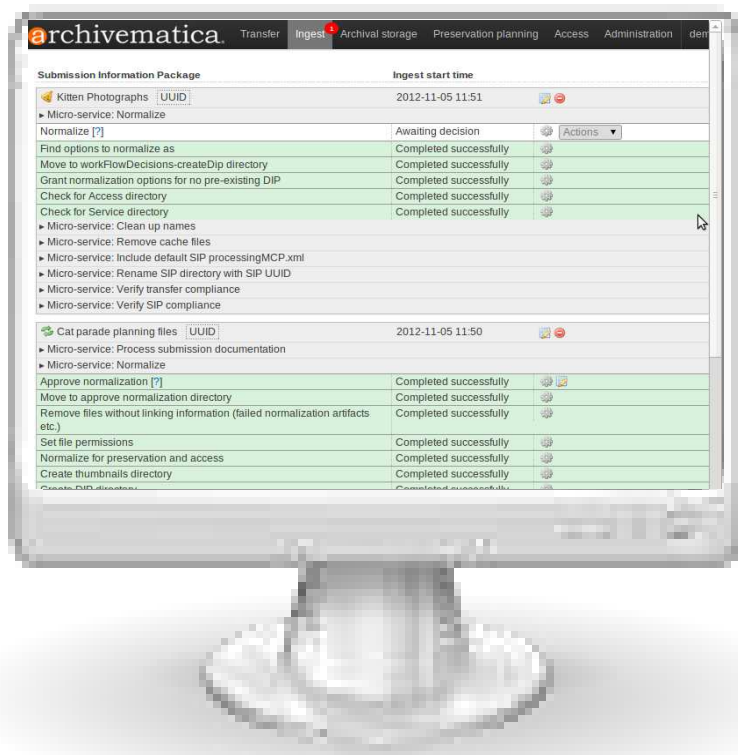
Justin Simpson

Software Developer

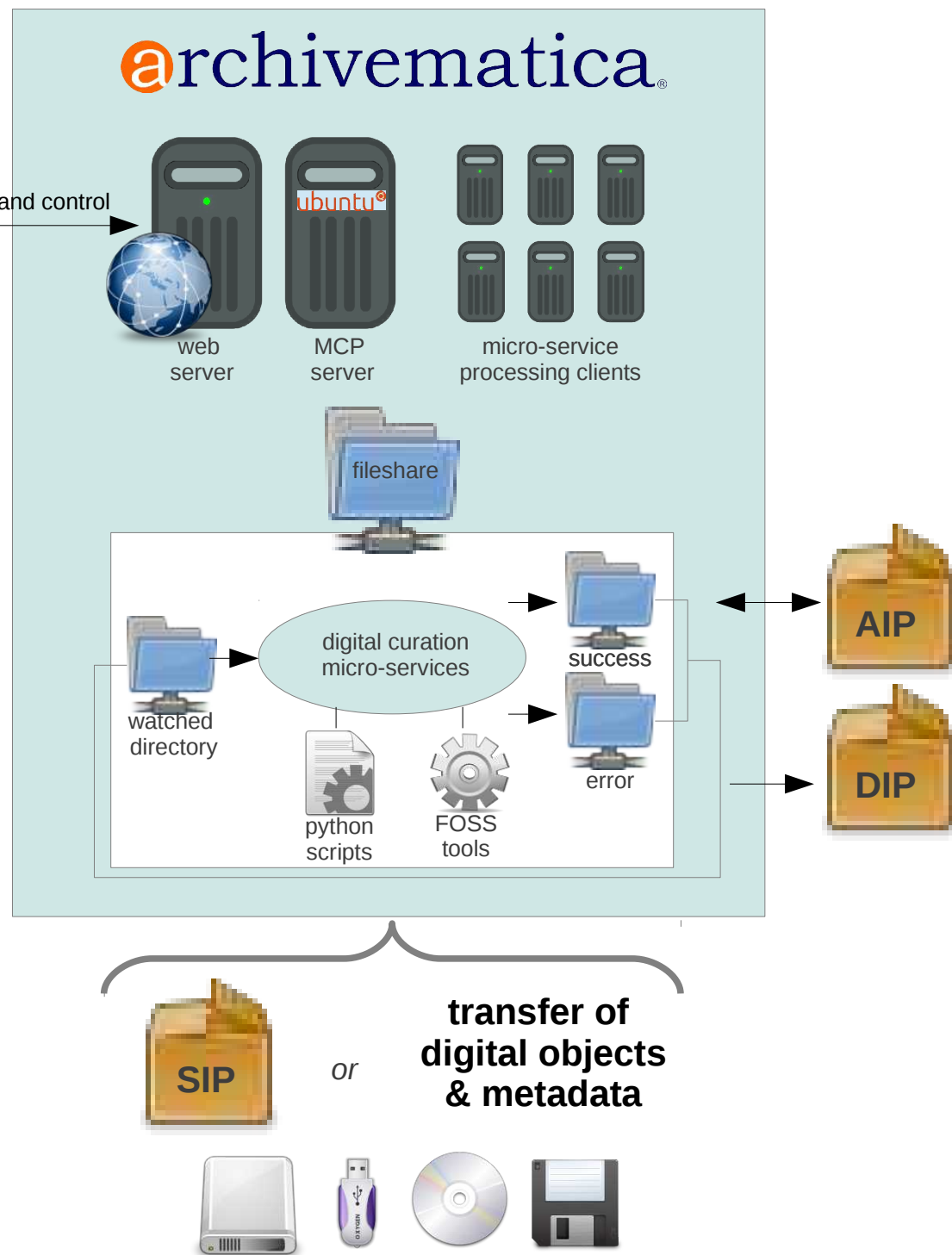


Mike Gale

Software Developer



web-based dashboard



Preservation planning

- A two-pronged approach:
 - Normalization on ingest
 - Preservation of the original file to support future strategies such as emulation
- Normalization relies on *format policies* based on an analysis of the significant characteristics of file formats
 - A format policy indicates the actions, tools and settings to apply to a file of a particular file format (e.g. normalization to preservation and/or access format)

Archivematica format policies

- Criteria for selecting default formats:
 - Non-proprietary
 - Freely available specifications
 - Widely used/endorsed by major repositories
 - No compression/lossless video compression
 - Tools available to write and render the format
- Format policies will change as community standards, practices and tools evolve.

Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	PDF for PPTX	OpenOffice
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	OpenOffice
Raster Images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera files/Digital Negative format**	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Original format	JPEG	ImageMagick/UFraw
Spreadsheets	XLS	Original format	Original format	None
Vector Images	AI, EPS, SVG	SVG	PDF	Inkscape
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM in MKV	MPEG-1	FFmpeg
Word processing files	DOC, WPD, RTF	<ul style="list-style-type: none"> • ODF (WPD and RTF) • Original format (DOC) 	PDF	OpenOffice

https://www.archivematica.org/wiki/Format_policies

Format confessional

iPres 2012 Toronto CurateCamp





**JUST SOLVE THE
PROBLEM**

24-hour CurateCamp/OPF worldwide file id hackathon

- participants from every time zone
- rapid, iterative testing
- OpenFITS (github fork)
- Open Planets Format Corpus (github)












Choose normalization tool

Archivematica Dashboard x ICA-AtoM x

localhost/ingest/

Archivematica ICA-AtoM Elasticsearch B... AM FAQ - Archi...

archivematica Transfer Ingest ¹ Archival storage Preservation planning Access Administration demo ▾

Submission Information Package	UUID	Ingest start time	
 Felis catus photographs	4ababa45-5c81-44f3-b152-d78fd983dc	2013-01-04 14:29	 
► Micro-service: Normalize			
Job: Select normalization file identification tool		Awaiting decision	 Actions
Job: Set resume link after tool selected.		Completed successfully	 Actions
Job: Find options to normalize as		Completed successfully	 - Tika
Job: Move to workFlowDecisions-createDip directory		Completed successfully	 - FITS
Job: Grant normalization options for no pre-existing DIP		Completed successfully	 - file utility
Job: Set remove preservation and access normalized files to renormalize link.		Completed successfully	 - fident
Job: Check for Access directory		Completed successfully	 - file extension
Job: Check for Service directory		Completed successfully	 - FIDO
► Micro-service: Clean up names			
► Micro-service: Remove cache files			
► Micro-service: Include default SIP processingMCP.xml			
► Micro-service: Rename SIP directory with SIP UUID			
► Micro-service: Verify transfer compliance			

- JHOVE
- DROID
- mediainfo

Select normalization type

Archivematica Dashboard x ICA-AtoM

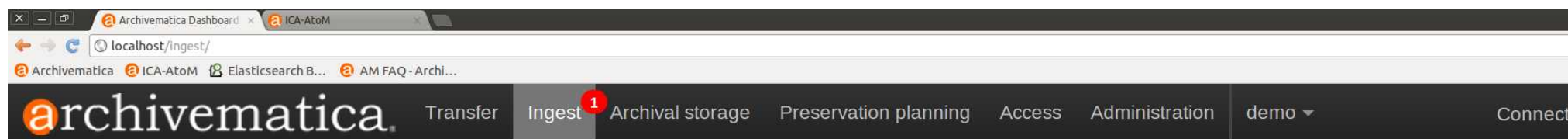
localhost/ingest/


Archivematica ICA-AtoM Elasticsearch B... AM FAQ - Archi...

archivematica Transfer **Ingest** 1 Archival storage Preservation planning Access Administration demo ▾

Submission Information Package	UUID	Ingest start time	
Felis catus photographs	4ababa45-5c81-44f3-b152-d78dfd983dc	2013-01-04 14:29	
► Micro-service: Normalize			
Job: Normalize [?]		Awaiting decision	Actions
Job: Resume after normalization file identification tool selected.	Completed		Actions <ul style="list-style-type: none"> - Normalize for preservation and access - Normalize for preservation - Reject SIP - Normalize service files for access - Do not normalize - Normalize for access
Job: Set SIP to normalize with file extension file identification.	Completed		
Job: Select normalization file identification tool	Completed		
Job: Set resume link after tool selected.	Completed		
Job: Find options to normalize as	Completed		
Job: Move to workFlowDecisions-createDip directory	Completed successfully		
Job: Grant normalization options for no pre-existing DIP	Completed successfully		
Job: Set remove preservation and access normalized files to renormalize link.	Completed successfully		
Job: Check for Access directory	Completed successfully		
Job: Check for Service directory	Completed successfully		
► Micro-service: Clean up names			
► Micro-service: Remove cache files			
► Micro-service: Include default SIP processingMCP.xml			
► Micro-service: Rename SIP directory with SIP UUID			
► Micro-service: Verify transfer compliance			

Review normalization results



Submission Information Package	UUID	Ingest start time	
 Felis catus photographs	4ababa45-5c81-44f3-b152-d78fd983dc	2013-01-04 14:29	 
► Micro-service: Normalize			
Job: Approve normalization (review) [?]		Awaiting decision	  Actions
Job: Move to approve normalization directory		Completed successfully	 Actions
Job: Remove files without linking information (failed normalization artifacts etc.)		Completed successfully	 - Reject
Job: Set file permissions		Completed successfully	 - Approve
Job: Normalize for preservation and access		Completed successfully	 - Redo normalization
Job: Create thumbnails directory		Completed successfully	
Job: Create DIP directory		Completed successfully	
Job: Move to processing directory		Completed successfully	
Job: Normalize [?]		Completed successfully	
Job: Resume after normalization file identification tool selected.		Completed successfully	
Job: Set SIP to normalize with file extension file identification.		Completed successfully	
Job: Select normalization file identification tool		Completed successfully	
Job: Set resume link after tool selected.		Completed successfully	
Job: Find options to normalize as		Completed successfully	
Job: Move to workFlowDecisions-createDip directory		Completed successfully	
Job: Grant normalization options for no pre-existing DIP		Completed successfully	
Job: Set remove preservation and access normalized files to renormalize link.		Completed successfully	
Job: Check for Access directory		Completed successfully	

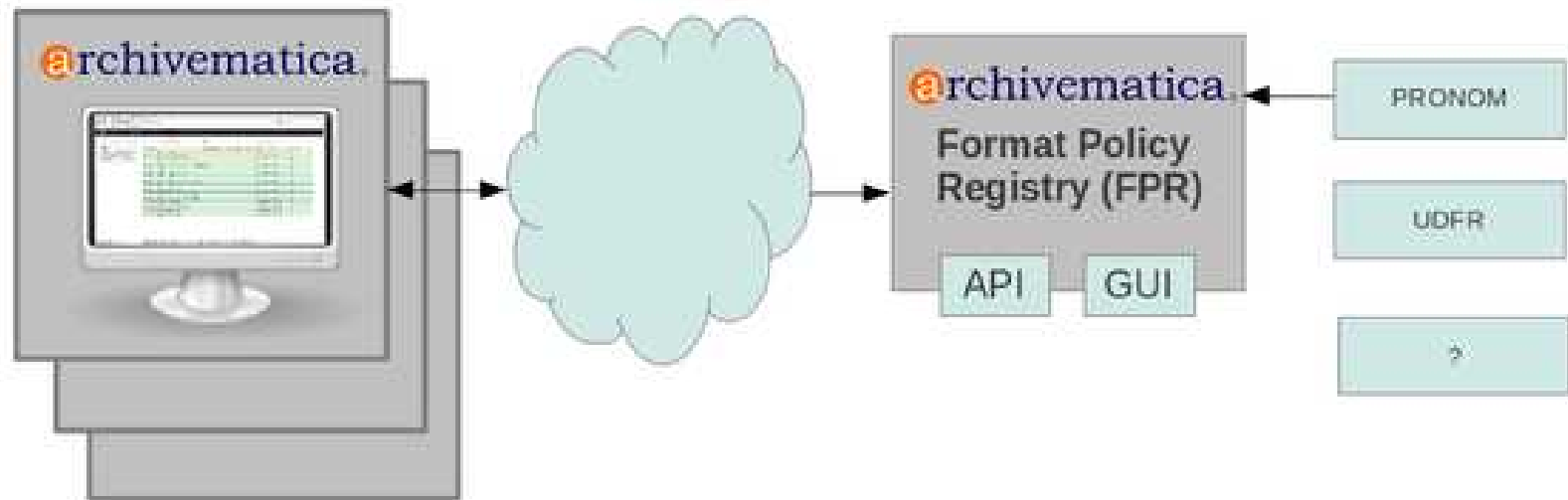
Format Policy Registry - FPR

- To share and test format choices and tool commands for normalization
 - ie “format policies”
- To support the community-wide evolution of best practices
- Hosted at archivematica.org/FPR with local dashboard customization and updates

Format Policy Registry - FPR

- Allow users to choose:
 - normalization tool
 - preservation and/or access format
 - file properties (eg resolution, bitrate)
 - to use local tool/process outside of the system
- Allow users to add new format policies

Format Policy Registry - FPR



Home Transfer Import Actual storage Preservation planning Access Administration				
Media type	Show advanced details			
Audio	Extension	Normalization description	Command	Purpose
	w3	Transcoding to mp3 with ffmpeg	Show	access
	w3	Transcoding to wav with ffmpeg	Show	preservation
	all	Transcoding to mp3 with ffmpeg	Show	access
	all	Transcoding to wav with ffmpeg	Show	preservation
ffmpeg -i "%(input_filename)s" -acodec libmp3lame -acodec_params 128 -vn -y "%(output_filename)s" && prep "%(input_filename)s" && prep "%(output_filename)s"				
	mp3	Transcoding to mp3 with ffmpeg	Show	access

