Handling a Digital Backlog and Analyzing Content in Archivematica

August 7, 2012
Research Forum
Society of American Archivists Annual Conference, Beyond Borders
San Diego, California

Courtney C. Mumma, MAS/MLIS, Systems Analyst and Archivematica Community Manager

archivematica

Overview

- Action research in the Archivematica project
- Digital backlog
- Processing digital objects
- Managing transfer backlogs
- Creating SIPs in Archivematica

Action research

- To quote action research's instigator Kurt Lewin: "if you want truly to understand something, try to change it". This kind of work is not simply about changing, but also improving an environment. As John Elliott says, action research is “the study of a social situation with a view to improving the quality of action within it”

(Elliott, J. (1991). Action research for educational change. Buckingham: Open University Press, p. 69). http://en.wikiversity.org/wiki/Action_research

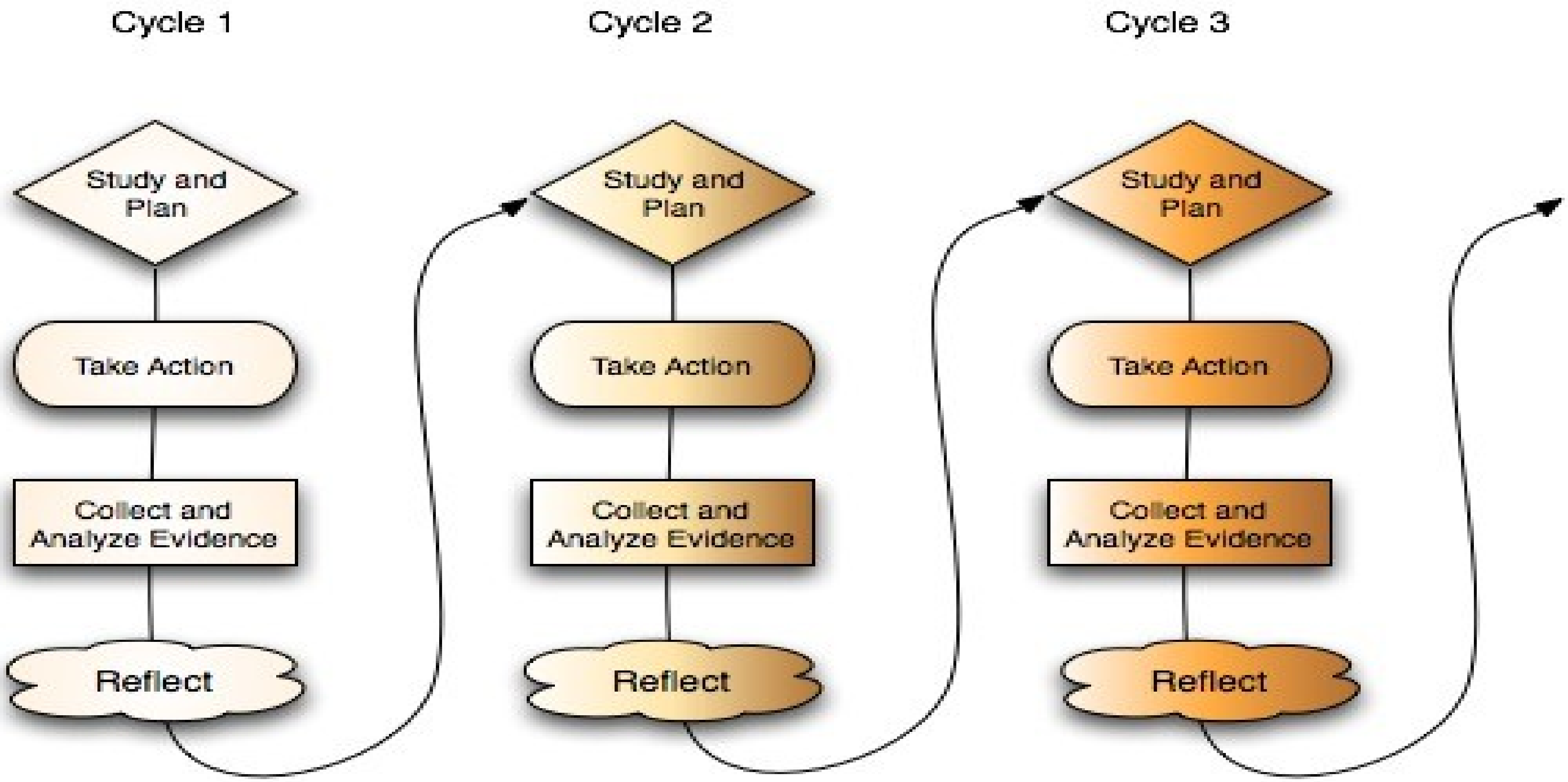
Goals of action research

- The improvement of professional practice through continual learning and progressive problem solving;

- A deep understanding of practice and the development of a well specified theory of action;

- An improvement in the community in which one's practice is embedded through participatory research.

Carr, W., & Kemmis, S. (1986) Becoming critical.
Lewes: Falmer Press



Progressive Problem Solving with Action Research

Carr, W., & Kemmis, S. (1986) Becoming critical.
Lewes: Falmer Press

Agile development

Individuals and interactions over processes and tools

Working software over comprehensive documentation

Customer collaboration over contract negotiation

Responding to change over following a plan

Beck, Kent; et al. (2001). "Manifesto for Agile Software Development". Agile Alliance. Retrieved 14 June 2010.

Six criteria for action research

Roles: Clarify the roles of researchers (Artefactual staff) and practioners (our clients)

Documentation: Explain the data collection approach and how data quality is managed (Observing and documenting local workflows and requirements)

Control: Explain the control measures (standards and best practices)

Usefulness: Establish the usefulness of the findings in the problem situation (assess community response)

Frameworks: Relate the action taken as well as the findings to frameworks to support the study (build up Archivematica tools and processes)

Transferability: Explicate conditions for transfer of findings to other situations (commonalities make up default workflows in Archivematica)

Peter Axel Nielsen. "IS Action Research and Its Criteria" in Information Systems Action Research: An Applied View of Emerging Concepts (2007), p. 366

Backlogs are inevitable

- Most of us have backlogs
- Most of us lack resources (tools, staff, funding) to prioritize and process those backlogs

Minimum data about the backlog

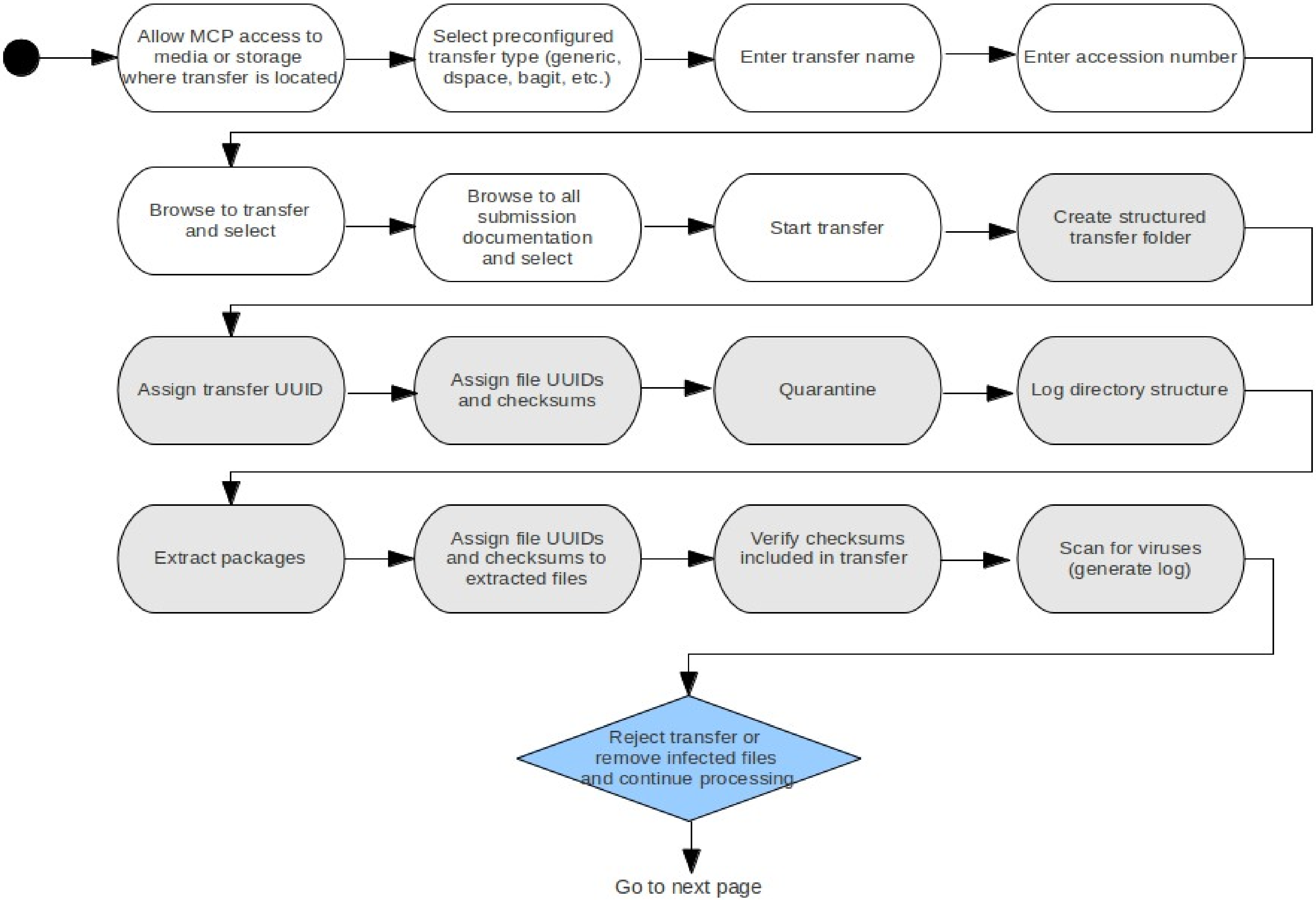
- Verify transfer compliance
- Create transfer UUID and assign file UUIDs to objects
- Verify metadata directory checksums (verifies any checksums included with the transfer)
- Assign checksums to objects
- Generate METS.xml document
- Extract packages
- Sanitize object's file and directory names
- Scan for viruses and malware
- Characterize and extract metadata
- 0.9 added transfer indexing, 1.0 accessioning as PREMIS event

Accession metadata

* PREMIS Event = Accession

```
•        <event>
•          <eventIdentifier>
•            <eventIdentifierType>UUID</eventIdentifierType>
•            <eventIdentifierValue>35cbe00d-d661-4174-b11a-e203f5608008</eventIdentifierValue>
•          </eventIdentifier>
•          <eventType>accession</eventType>
•          <eventDateTime>2012-03-14</eventDateTime>
•          <eventDetail>accession#2012-029</eventDetail>
•          <eventOutcomeInformation>
•            <eventOutcome></eventOutcome>
•            <eventOutcomeDetail>
•              <eventOutcomeDetailNote></eventOutcomeDetailNote>
•            </eventOutcomeDetail>
•          </eventOutcomeInformation>
•          <linkingAgentIdentifier>
•            <linkingAgentIdentifierType>archivist</linkingAgentIdentifierType>
•            <linkingAgentIdentifierValue>Courtney Mumma</linkingAgentIdentifierValue>
•          </linkingAgentIdentifier>
•         </event>
```

0.9 Transfer workflow



Allow MCP access to media or storage where transfer is located

Select preconfigured transfer type (generic, dspace, bagit, etc.)

Enter transfer name

Enter accession number

Browse to transfer and select

Browse to all submission documentation and select

Start transfer

Create structured transfer folder

Assign transfer UUID

Assign file UUIDs and checksums

Quarantine

Log directory structure

Extract packages

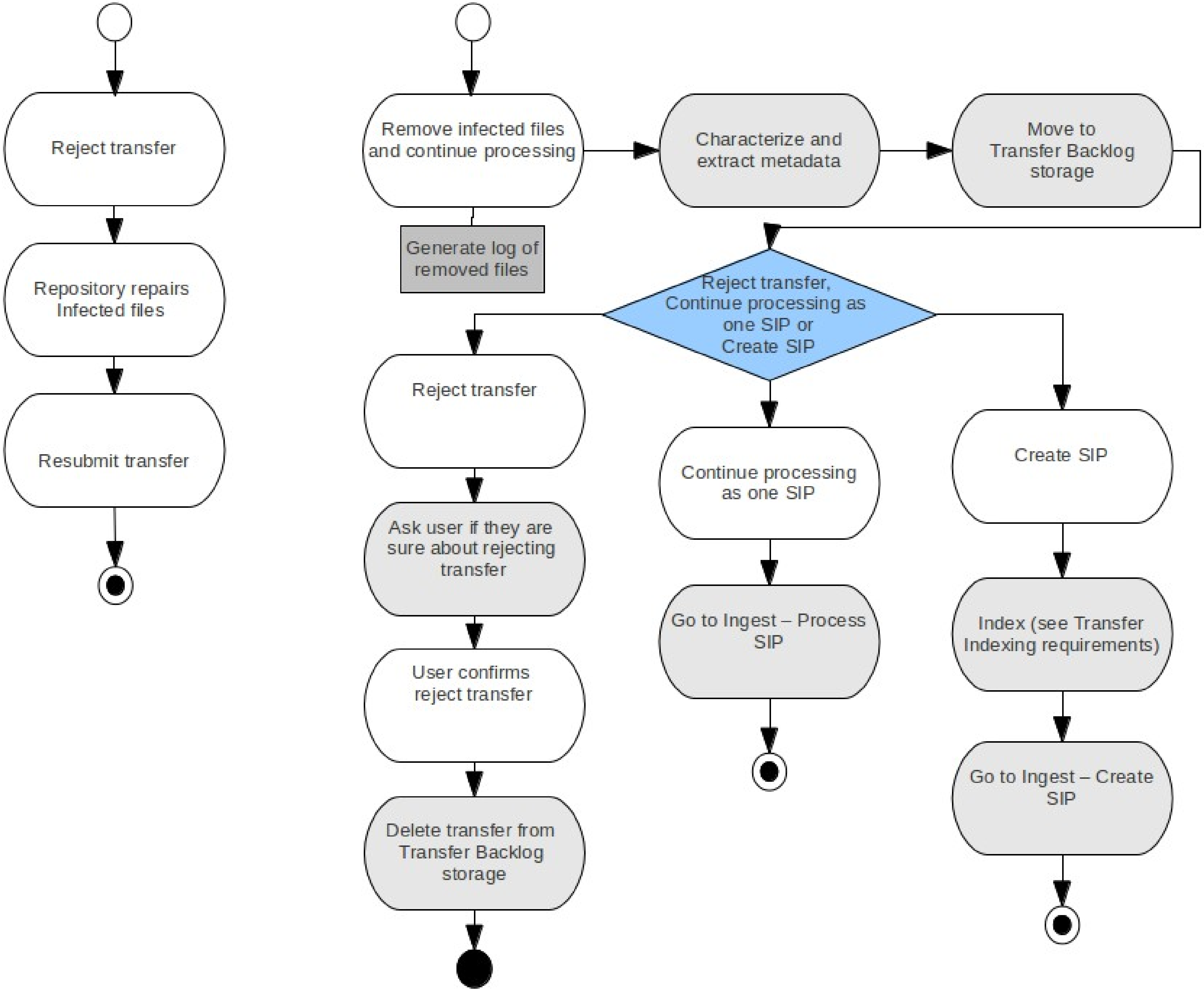
Assign file UUIDs and checksums to extracted files

Verify checksums included in transfer

Scan for viruses (generate log)

Reject transfer or remove infected files and continue processing

Go to next page


Reject transfer
Repository repairs Infected files
Resubmit transfer
Remove infected files and continue processing
Generate log of removed files
Characterize and extract metadata
Move to Transfer Backlog storage
Reject transfer, Continue processing as one SIP or Create SIP
Reject transfer
Ask user if they are sure about rejecting transfer
User confirms reject transfer
Delete transfer from Transfer Backlog storage
Continue processing as one SIP
Go to Ingest – Process SIP
Create SIP
Index (see Transfer Indexing requirements)
Go to Ingest – Create SIP

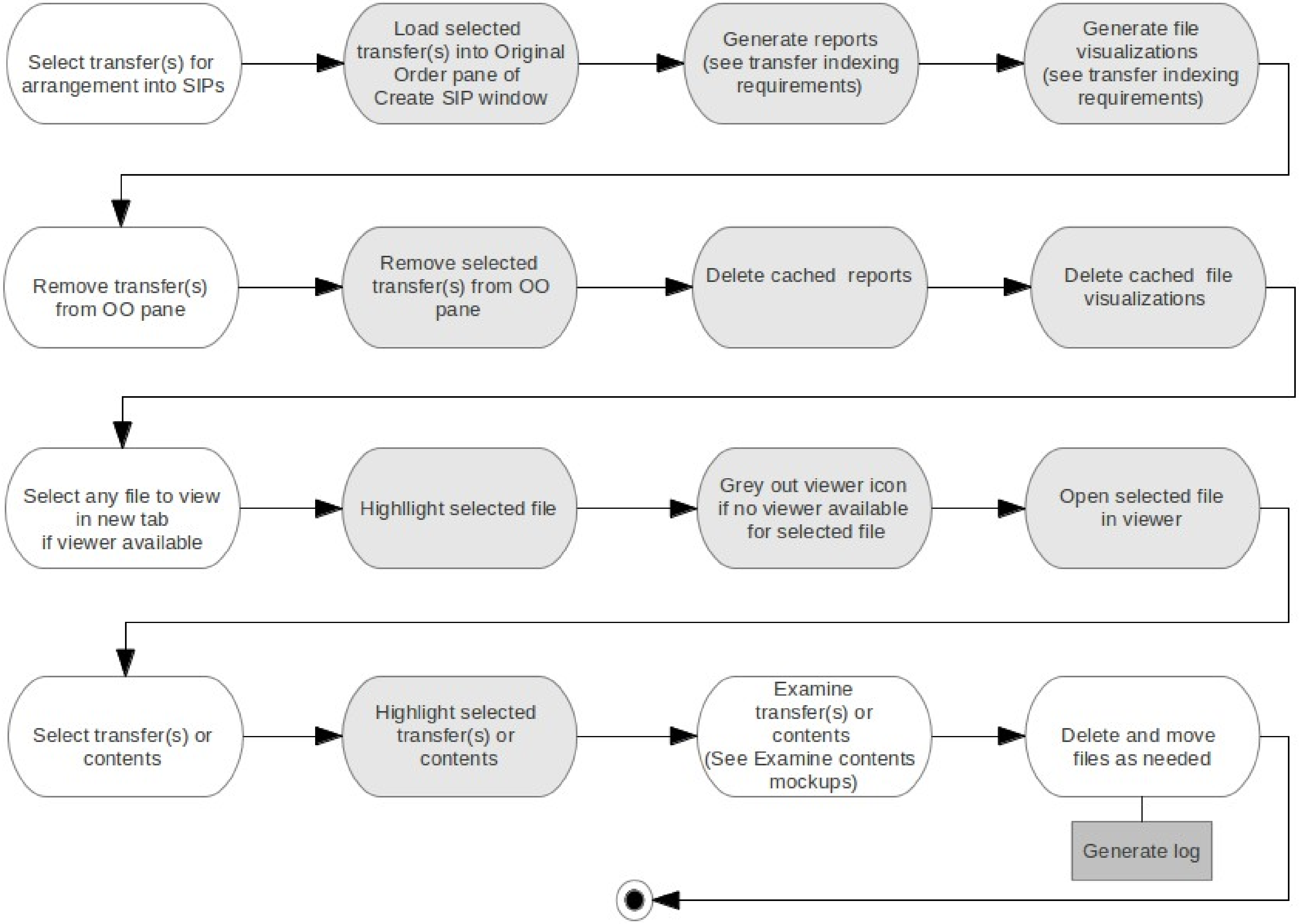
Transfer indexing

- Full text content
- File embedded metadata
- Formats - by folder, by transfer
- Keyword & pattern matching for privacy/security sensitive information (e.g. social insurance numbers/social security numbers, credit card numbers, email addresses security keywords like 'private', 'confidential' - find or generate domain-specific taxonomies)
- Possible reports: PDFs that have not been OCR'ed, password protected / encrypted files, duplicates with their file paths

What to do with that data

- Examine transfer(s)
- Use visualization tools and index to access transfer content
- Assign transfer(s) to appropriate intellectual levels of arrangement
- Create SIP(s) from transfer(s)

0.9 Ingest – Create SIP workflow



Select transfer(s) for arrangement into SIPs

Load selected transfer(s) into Original Order pane of Create SIP window

Generate reports (see transfer indexing requirements)

Generate file visualizations (see transfer indexing requirements)

Remove transfer(s) from OO pane

Remove selected transfer(s) from OO pane

Delete cached reports

Delete cached file visualizations

Select any file to view in new tab if viewer available

Highllight selected file

Grey out viewer icon if no viewer available for selected file

Open selected file in viewer

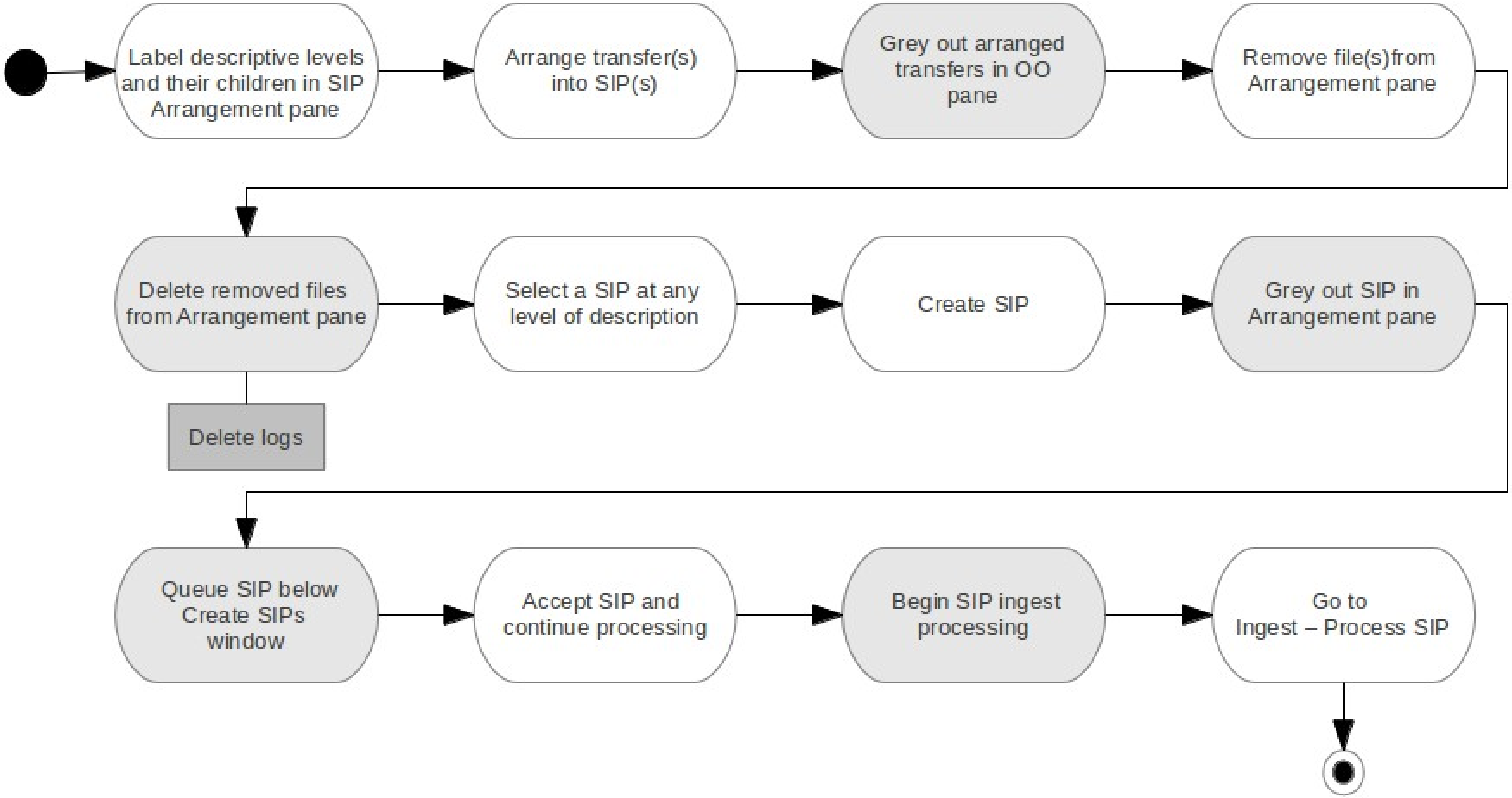
Select transfer(s) or contents

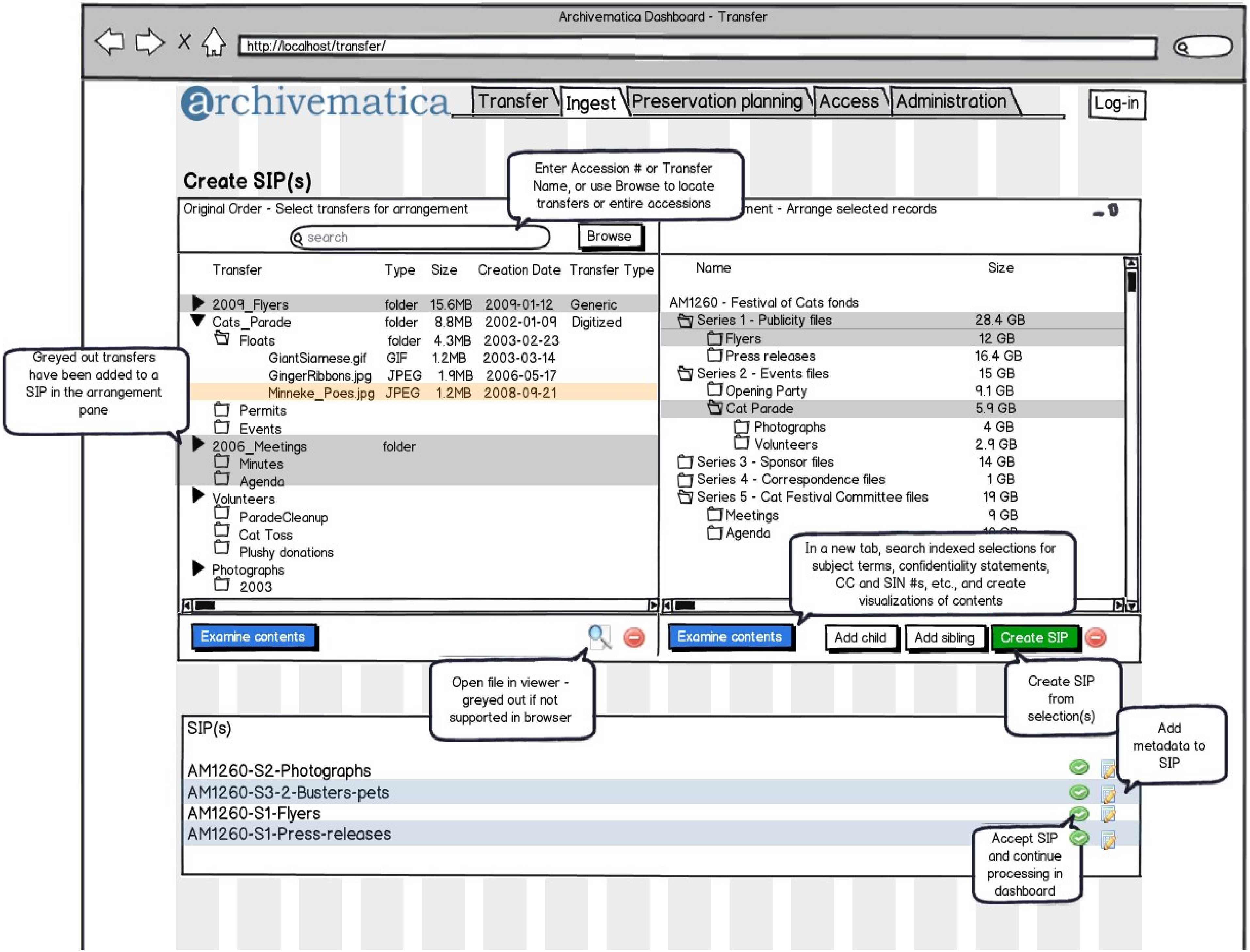
Highlight selected transfer(s) or contents

Examine transfer(s) or contents (See Examine contents mockups)

Delete and move files as needed

Generate log


Label descriptive levels and their children in SIP Arrangement pane
Arrange transfer(s) into SIP(s)
Grey out arranged transfers in OO pane
Remove file(s)from Arrangement pane
Delete removed files from Arrangement pane
Delete logs
Select a SIP at any level of description
Create SIP
Grey out SIP in Arrangement pane
Queue SIP below Create SIPs window
Accept SIP and continue processing
Begin SIP ingest processing
Go to Ingest – Process SIP



Create SIP(s)

Enter Accession # or Transfer Name, or use Browse to locate transfers or entire accessions

Original Order - Select transfers for arrangement

Transfer	Type	Size	Creation Date	Transfer Type
▶ 2009_Flyers	folder	15.6MB	2009-01-12	Generic
▼ Cats_Parade	folder	8.8MB	2002-01-09	Digitized
Floats	folder	4.3MB	2003-02-23	
GiantSiamese.gif	GIF	1.2MB	2003-03-14	
GingerRibbons.jpg	JPEG	1.9MB	2006-05-17	
Minneke_Poes.jpg	JPEG	1.2MB	2008-09-21	
Permits				
Events				
▶ 2006_Meetings	folder			
Minutes				
Agenda				
▶ Volunteers				
ParadeCleanup				
Cat Toss				
Plushy donations				
▶ Photographs				
2003				

Greyed out transfers have been added to a SIP in the arrangement pane

ment - Arrange selected records

Name	Size
AM1260 - Festival of Cats fonds	
Series 1 - Publicity files	28.4 GB
Flyers	12 GB
Press releases	16.4 GB
Series 2 - Events files	15 GB
Opening Party	9.1 GB
Cat Parade	5.9 GB
Photographs	4 GB
Volunteers	2.9 GB
Series 3 - Sponsor files	14 GB
Series 4 - Correspondence files	1 GB
Series 5 - Cat Festival Committee files	19 GB
Meetings	9 GB
Agenda	10 GB

In a new tab, search indexed selections for subject terms, confidentiality statements, CC and SIN #s, etc., and create visualizations of contents

Open file in viewer - greyed out if not supported in browser

Create SIP from selection(s)

Add metadata to SIP

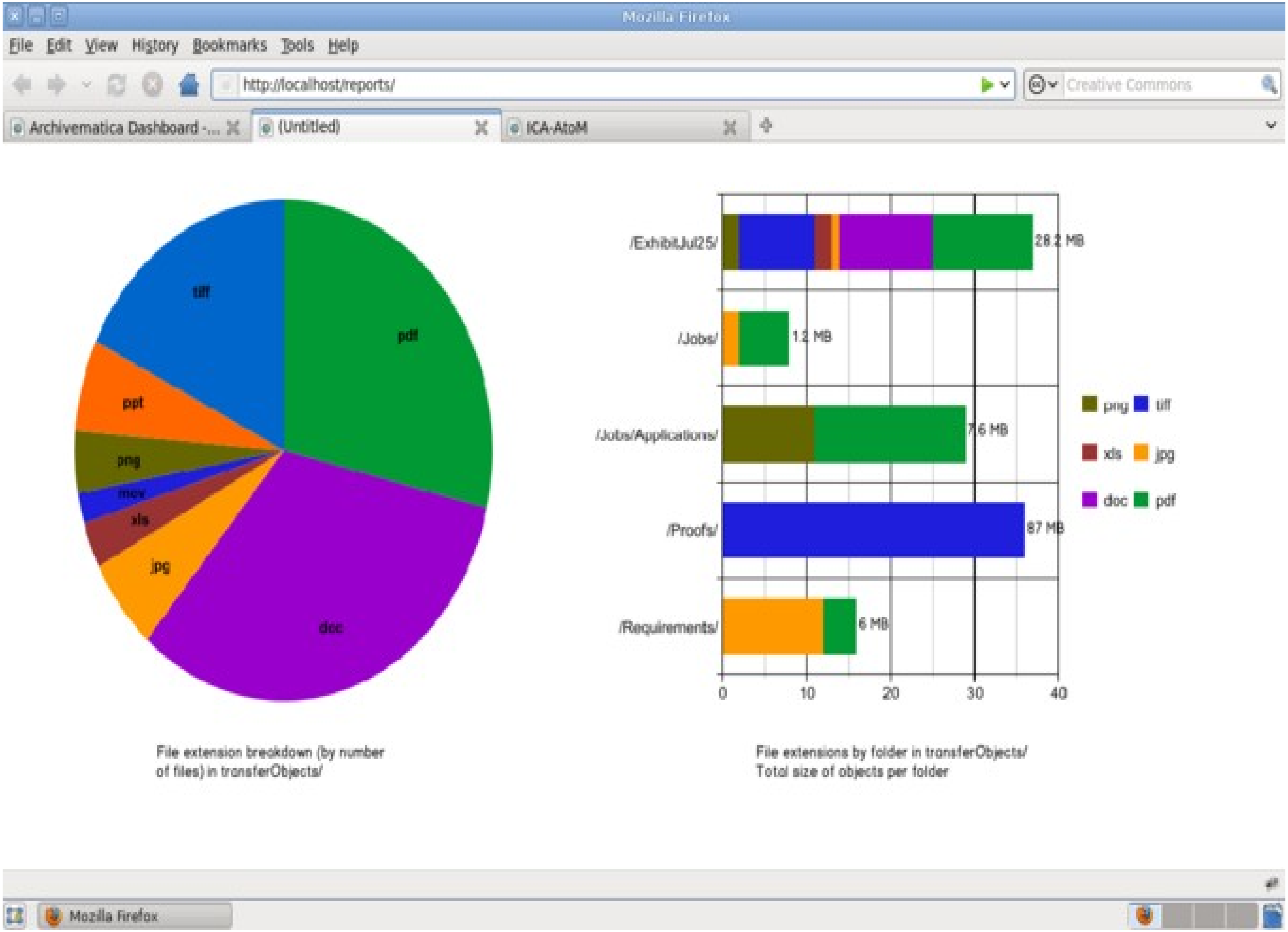
SIP(s)

AM1260-S2-Photographs
AM1260-S3-2-Busters-pets
AM1260-S1-Flyers
AM1260-S1-Press-releases

Accept SIP and continue processing in dashboard

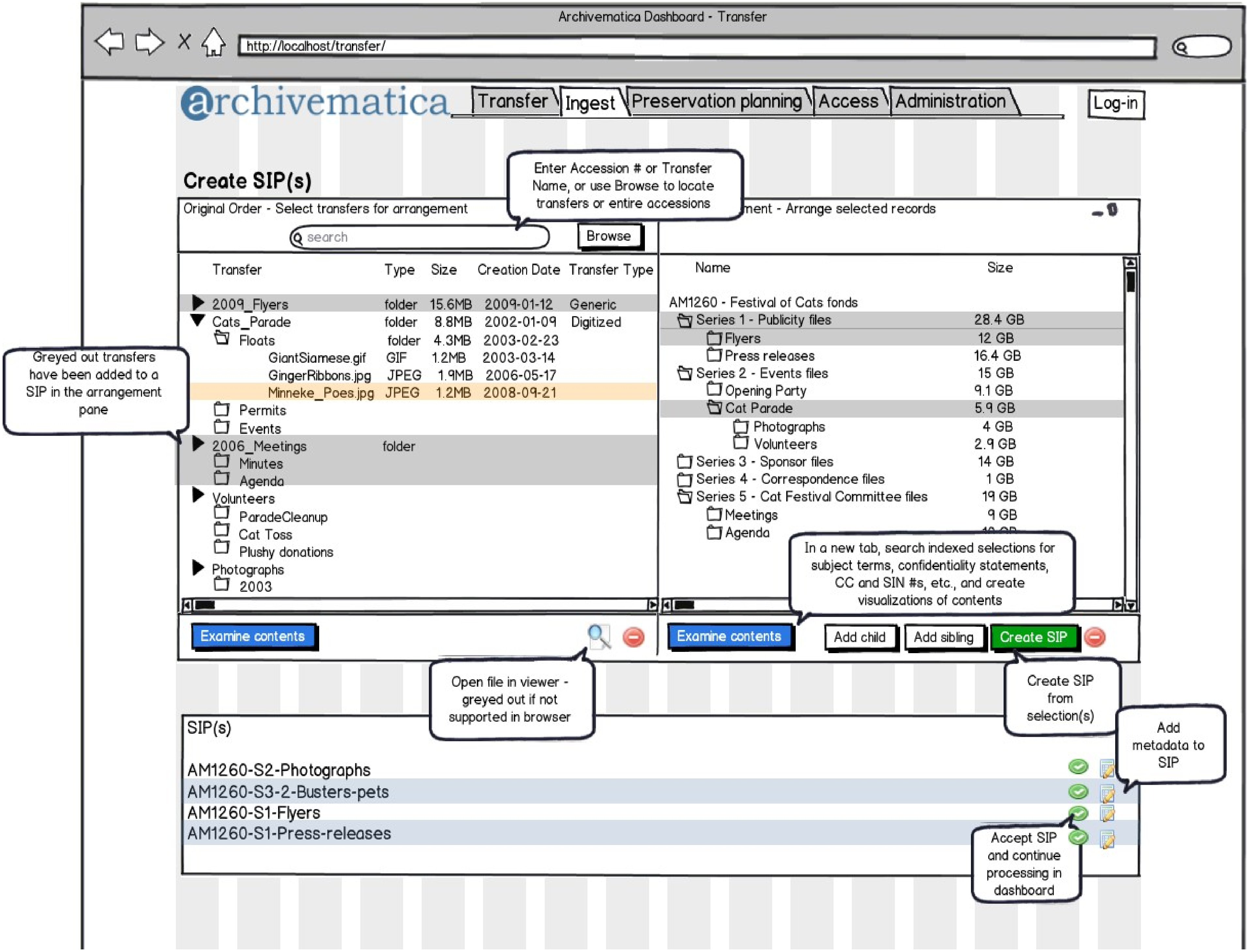
Examine contents splash page

- Information about the transfer or selected file group (number of files, size, name, uuid, accession #, and?)
- Pie graph showing file type distribution overall and bargraph showing file type by folder and ordered by size)
- Clickable links: file type opens into new tab with file browser interface of all of specified format), folders opens into new tab with file browser interface of entire folder in context of rest of transfer)
- Search box to search index (opens in new tab)
- Report options (each opens in new tab): duplicates with full path locations, security keywords, CC numbers, SIN/SS#s, email addresses (with distribution graph), see password protected files with distribution (graph?)



File extension breakdown (by number of files) in transferObjects/

File extensions by folder in transferObjects/
Total size of objects per folder



Create SIP(s)

Enter Accession # or Transfer Name, or use Browse to locate transfers or entire accessions

Original Order - Select transfers for arrangement

Transfer	Type	Size	Creation Date	Transfer Type
▶ 2009_Flyers	folder	15.6MB	2009-01-12	Generic
▼ Cats_Parade	folder	8.8MB	2002-01-09	Digitized
Floats	folder	4.3MB	2003-02-23	
GiantSiamese.gif	GIF	1.2MB	2003-03-14	
GingerRibbons.jpg	JPEG	1.9MB	2006-05-17	
Minneke_Poes.jpg	JPEG	1.2MB	2008-09-21	
Permits				
Events				
▶ 2006_Meetings	folder			
Minutes				
Agenda				
▶ Volunteers				
ParadeCleanup				
Cat Toss				
Plushy donations				
▶ Photographs				
2003				

Greyed out transfers have been added to a SIP in the arrangement pane

Open file in viewer - greyed out if not supported in browser

ment - Arrange selected records

Name	Size
AM1260 - Festival of Cats fonds	
Series 1 - Publicity files	28.4 GB
Flyers	12 GB
Press releases	16.4 GB
Series 2 - Events files	15 GB
Opening Party	9.1 GB
Cat Parade	5.9 GB
Photographs	4 GB
Volunteers	2.9 GB
Series 3 - Sponsor files	14 GB
Series 4 - Correspondence files	1 GB
Series 5 - Cat Festival Committee files	19 GB
Meetings	9 GB
Agenda	10 GB

In a new tab, search indexed selections for subject terms, confidentiality statements, CC and SIN #s, etc., and create visualizations of contents

Create SIP from selection(s)

SIP(s)

- AM1260-S2-Photographs
- AM1260-S3-2-Busters-pets
- AM1260-S1-Flyers
- AM1260-S1-Press-releases

Add metadata to SIP

Accept SIP and continue processing in dashboard

File viewers

- This will allow the user to see individual documents in the transfer to get a better idea of their contents and technical metadata before assigning them to SIPs.
- Viewers are browser-dependent; viewer option is greyed out if viewer is not supported in browser
- Examine Contents window allows for viewing technical MD and other metadata available after Transfer microservices as well as indexing MD

@archivematica (official Twitter)
@snarkivist (me)

archivematica.org (wiki with links to documentation, downloads, user group and issues list)

courtney@artefactual.com (email)

artefactual
systems inc.