

Preserving digital content through improved EPrints repository integration with Archivematica

George Macgregor ^{a, b}

^a Scholarly Publications & Research Data (SPRD), IS Information Management, University of Strathclyde

^b iSchool, Department of Computer & Information Sciences, University of Strathclyde

 [0000-0002-8482-3973](https://orcid.org/0000-0002-8482-3973) \ \  [@g3om4c](https://twitter.com/g3om4c)

Tomasz Neugebauer ^c

^c Digital Projects & Systems Development Librarian, Concordia University

 [0000-0002-9743-5910](https://orcid.org/0000-0002-9743-5910) \ \  [@photomediathink](https://twitter.com/photomediathink)

Background



- Context:
 - [EPrints](#), open source repository platform
 - [Strathprints](#) – implementation at University of Strathclyde
 - Different flavours = OA publication, data, OER
 - Plugin to enhance functionality ([Bazaar](#))
 - Heterogeneous digital objects, e.g. text (PDF), video, etc.
- Demands of the scholarly record: persistence, citability, accessibility, maintenance
 - Need for better digital preservation capacity among repositories needed through [Archivematica](#) interoperation
 - [Concordia University](#) spearheaded conceptual & development work, see [\[1\]](#) – support from [EPrints Services](#) & [Artefactual](#) & community
- **EPrints - Archivematica Integration @ GitHub:**
<https://github.com/eprintsug/EPrintsArchivematica>

Digital preservation export: overview

@ the EPrints side...

- [Digital preservation export plugin](#)
- Config. files
- Batch scripts (scheduled crons)
- Identifies new/updated items worthy of preservation
 - **process_transfers**
 - 'Triggers'
- Generates Archivemata optimized export of EPrints content using directory structure
- Export includes rich repository metadata (XML, JSON, md5), digital objects (documents & derivatives)

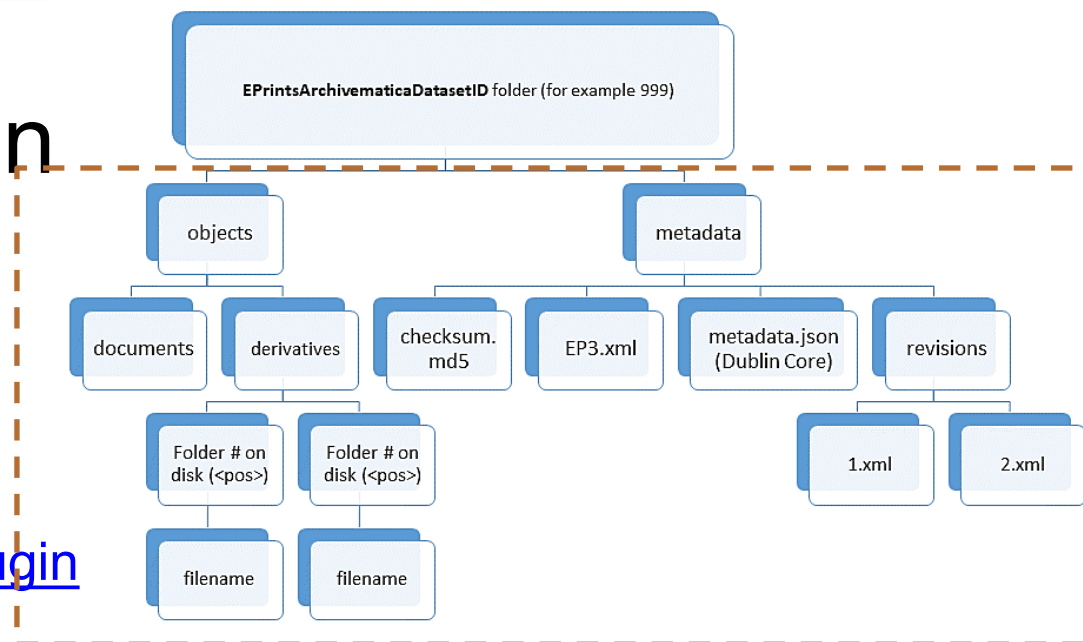


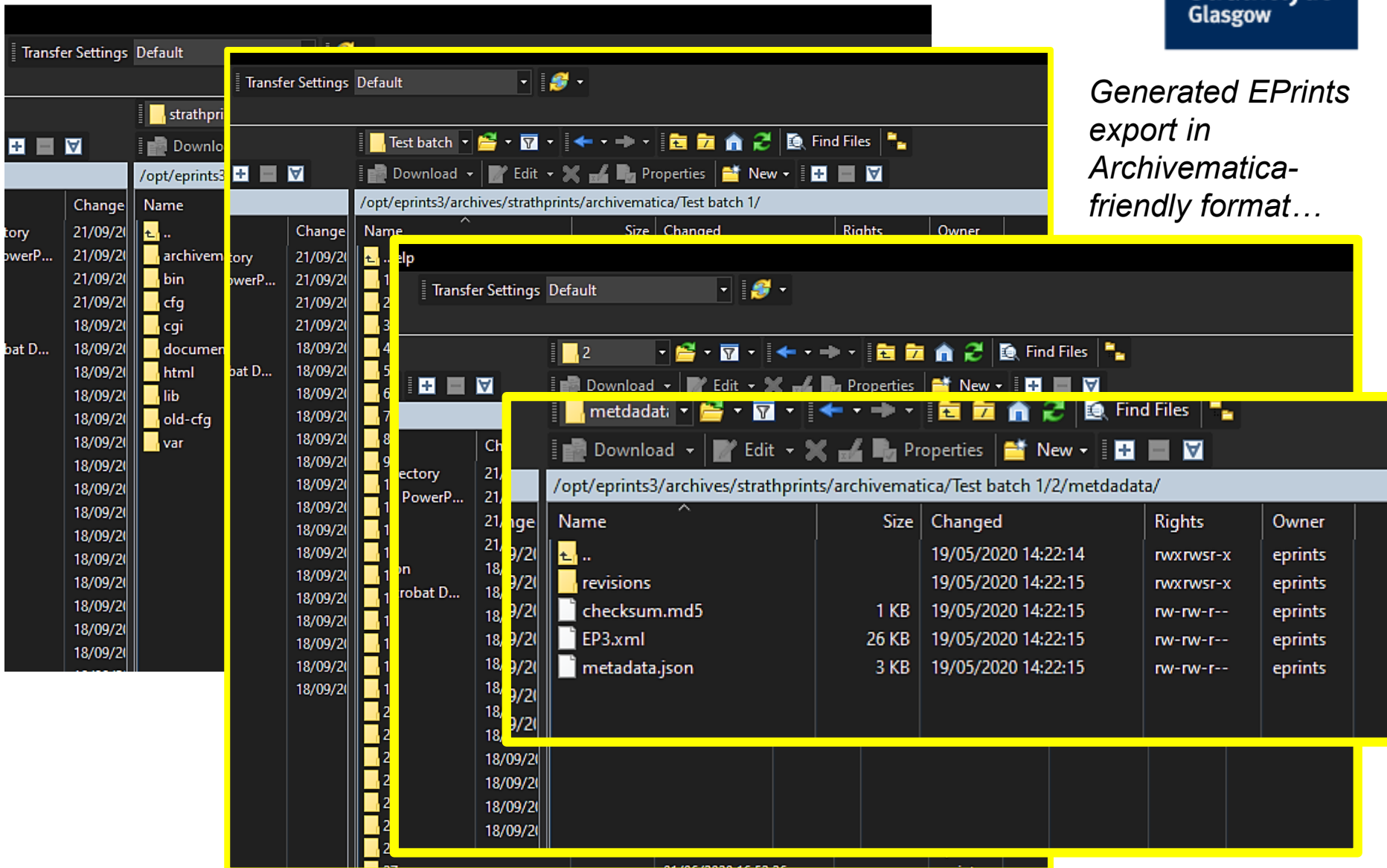
Diagram: T. Neugebauer, EPrints - Archivemata Integration
@ GitHub: <https://github.com/eprintsug/EPrintsArchivemata>

@ the Archivemata side...

- Exported content written to shared location
- [Archivemata automation tools](#) used to pick up export at location & begin processing transfers within the Archivemata pipeline

Export: abstraction to reality...

*Generated EPrints
export in
Archivematica-
friendly format...*








The image shows a series of overlapping screenshots of a file manager window, illustrating the export process. The top screenshot shows the root directory of the export: `/opt/eprints3/archives/strathprints/archivematica/Test batch 1/`. The middle screenshot shows a subdirectory named `metdatadi`. The bottom screenshot shows the contents of the `metdatadi` directory, which includes a `revisions` folder and several files: `checksum.md5`, `EP3.xml`, and `metadata.json`.

Name	Size	Changed	Rights	Owner
..		19/05/2020 14:22:14	rw-rwsr-x	eprints
revisions		19/05/2020 14:22:15	rw-rwsr-x	eprints
checksum.md5	1 KB	19/05/2020 14:22:15	rw-rw-r--	eprints
EP3.xml	26 KB	19/05/2020 14:22:15	rw-rw-r--	eprints
metadata.json	3 KB	19/05/2020 14:22:15	rw-rw-r--	eprints

Anatomy of EPrints export ingest... (1)

Use case is for automation via export cron jobs @ Eprints side
With automation @ the Archivematica side
Useful to examine manually though...

Submission Information Package	UUID	Ingest start time	
✔ Testprints2	f7ee1176-1280-4f31-9153-02473d5e2ab3	2020-09-21 13:29	
▶ Microservice: Store AIP			
▶ Microservice: Prepare AIP			
▶ Microservice: Add README file			
▶ Microservice: Generate AIP METS			
▶ Microservice: Bind PIDs			
Job: Bind PIDs?		Completed successfully	
▶ Microservice: Process metadata directory			
▶ Microservice: Process submission documentation			
▶ Microservice: Transcribe SIP contents			
Job: Transcribe SIP contents?		Completed successfully	
▶ Microservice: Add final metadata			
Job: Reminder: add metadata if desired		Completed successfully	
Job: Move to metadata reminder		Completed successfully	
▶ Microservice: Policy checks for derivatives			
▶ Microservice: Process manually normalized files			
▶ Microservice: Normalize			
▶ Microservice: Change SIP filenames			
▶ Microservice: Remove cache files			
▶ Microservice: Include default SIP processingMCP.xml			
▶ Microservice: Rename SIP directory with SIP UUID			
▶ Microservice: Verify SIP compliance			

Anatomy of EPrints export ingest... (AIP)

Archival storage / Testprints2

Testprints2 Archival Information Package

UUID	f7ee1176-1280-4f31-9153-02473d5e2ab3
Size	5.88 MB
Date stored	2020-09-21 13:31
Status	Stored
Encrypted	False
Location	Download C:\Users\georg\Desktop\Testprints2-f7ee1176-1280-4f31-9153-02473d5e2ab3.7z

C:\Users\georg\Desktop\Testprints2-f7ee1176-1280-4f31-9153-02473d5e2ab3.7z\Testprints2-f7ee1176-1280-4f31-9153-02473d5e2ab3\

File Edit View Favorites Tools Help

Add Extract Test Copy Move Delete Info

Name	Size	Packed Size	Modified	Created	Accessed	Attributes	CRC	Encrypted	Method	Block
data	12 173 630	0	2020-09-21 13:31	2020-09-21 13:31	2020-09-21 13:31	D drwxr-xr-x	D412DF4D	-		
bag-info.txt	198	6 162 676	2020-09-21 13:31	2020-09-21 13:31	2020-09-21 13:31	A -rw-r--r--	A0B5E611	-	BZip2	0
bagit.txt	55		2020-09-21 13:31	2020-09-21 13:31	2020-09-21 13:31	A -rw-r--r--	CB58FA90	-	BZip2	0
manifest-sha256.txt	118 258		2020-09-21 13:31	2020-09-21 13:31	2020-09-21 13:31	A -rw-r--r--	33352B57	-	BZip2	0
tagmanifest-sha256.txt	238		2020-09-21 13:31	2020-09-21 13:31	2020-09-21 13:31	A -rw-r--r--	4D0432F2	-	BZip2	0

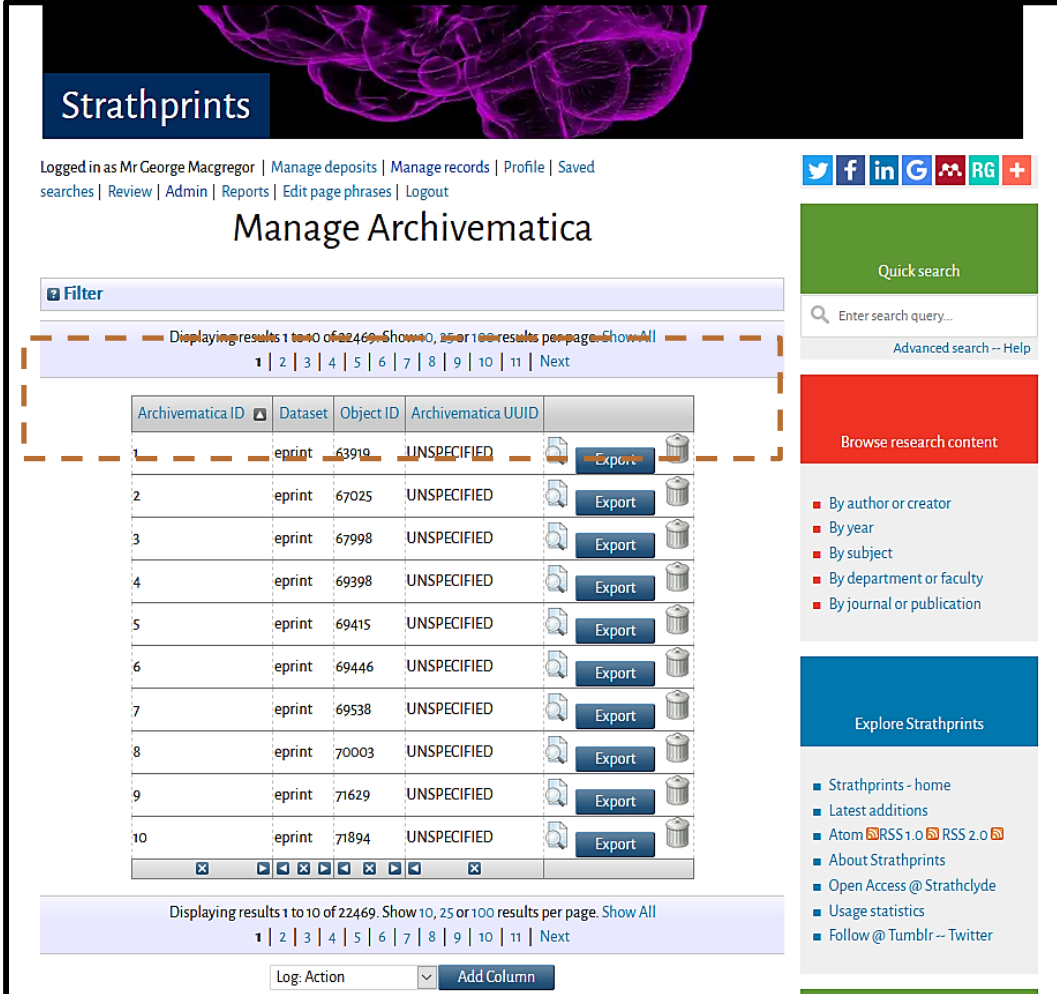
Repository dashboard management...

Preservation exports can be monitored and managed via EPrints dashboard, e.g.

- Archivemata ID
- Local ID

Archivemata UUID(?!)

- Unspecified! ;-)
- Work to be done:
 - RESTful endpoint supported by Eprints
 - API call back from Archivemata with UUID – coming soon!



Strathprints





















Logged in as Mr George Macgregor | [Manage deposits](#) | [Manage records](#) | [Profile](#) | [Saved searches](#) | [Review](#) | [Admin](#) | [Reports](#) | [Edit page phrases](#) | [Logout](#)

Manage Archivemata

Filter

Displaying results 1 to 10 of 22469. Show 10, 25 or 100 results per page. Show All

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Next

Archivemata ID	Dataset	Object ID	Archivemata UUID	
1	eprint	63910	UNSPECIFIED	 Export 
2	eprint	67025	UNSPECIFIED	 Export 
3	eprint	67998	UNSPECIFIED	 Export 
4	eprint	69398	UNSPECIFIED	 Export 
5	eprint	69415	UNSPECIFIED	 Export 
6	eprint	69446	UNSPECIFIED	 Export 
7	eprint	69538	UNSPECIFIED	 Export 
8	eprint	70003	UNSPECIFIED	 Export 
9	eprint	71629	UNSPECIFIED	 Export 
10	eprint	71894	UNSPECIFIED	 Export 

Displaying results 1 to 10 of 22469. Show 10, 25 or 100 results per page. Show All

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Next

Log: Action

Quick search

Enter search query...

Advanced search -- Help

Browse research content

- By author or creator
- By year
- By subject
- By department or faculty
- By journal or publication

Explore Strathprints

- [Strathprints - home](#)
- [Latest additions](#)
- [Atom](#) [RSS 1.0](#) [RSS 2.0](#)
- [About Strathprints](#)
- [Open Access @ Strathclyde](#)
- [Usage statistics](#)
- [Follow @ Tumblr -- Twitter](#)

Experimentation: results

- *Testing, testing, testing!*
- Currently experimenting in our test environment (circa 6 months)
 - Lots of KE with Concordia
 - Experimentation positive; minor niggles
 - Repositories interesting use case, involving large volume of content
- Rolling out to production = autumn 2020!
 - Probable rollout without API call back
 - Repository ecosystem at Strathclyde: repository & Current Research Information System (CRIS) running in parallel
 - Compromise in data that can be written to Strathprints [\[2\]](#)
 - Possible hacks to make call back via write protection of EPrints DB fields (John Salter @ [WRRO](#))

Other work to be done, or already done...

- Call back... *in hand*
- Improved metadata modelling in the JSON export
- Additional export job options
 - **touch_transfers**: Updates all Archivemata records which are not pending a transfer
 - **create_transfers**: Create Archivemata records which do not already exist
- Packaging as EPrints plugin & depositing in the EPrints Bazaar

References

- [1] Neugebauer, Tomasz , Simpson, Justin and Bradley, Justin (2018) Digital Preservation through EPrints- Archivematica Integration. In: *13th International Conference on Open Repositories (OR2018)*, June 3-7, 2018, Bozeman, Montana, USA. Available: <https://spectrum.library.concordia.ca/983933/>
- [2] Macgregor, George (2019) Repository and CRIS interoperability issues within a 'connector lite' environment. In: *14th International Conference on Open Repositories (OR2019)*, June 10-13, 2019, Universität Hamburg, Germany. Available: <https://strathprints.strath.ac.uk/68240/>