Automated preservation workflows:
How not to think about preservation every day by thinking about it a lot on the first day

Thursday, 05th November 2021

UCT Libraries

Digital Library Services (DLS)

- Dr. Sanjin Muftic (Digital Scholarship Specialist)

Based on the work of Zimmer, N. (2019): **Lost & found: some issues that digital preservation seeks to address**. Presentation. 10.25375/uct.10253069

CC BY SA

- Preservation at UCTL
- Automation
- Automation meets Preservation?

DIGITAL LIBRARY SERVICES

Where is Preservation at UCTL?

Lost: fidelity (integrity, beauty, veracity, …)

issue: Data rot

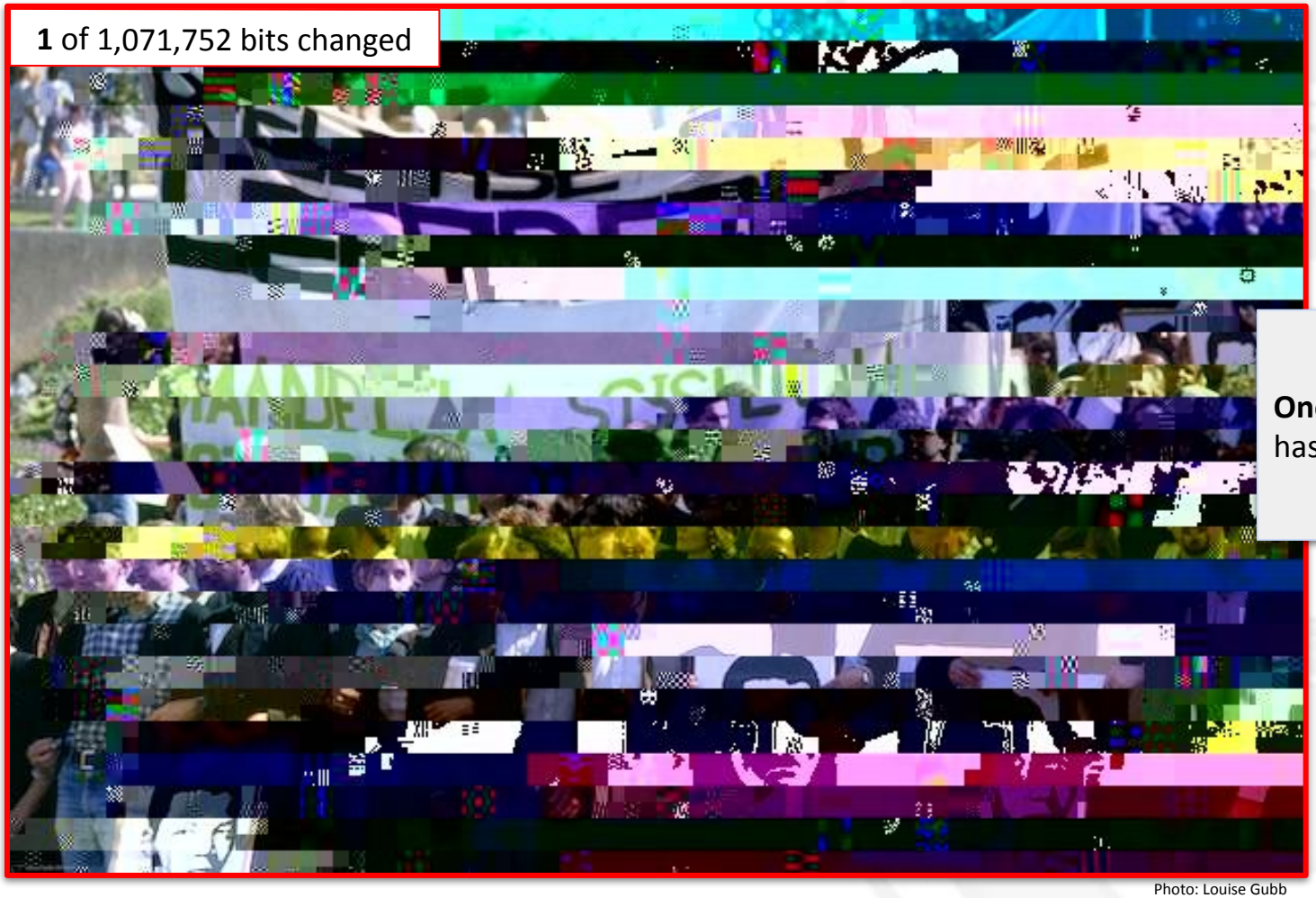
original file

Photo: Louise Gubb

What if “flip” one bit?

Lost: fidelity (integrity, beauty, veracity, …)

issue: Data rot

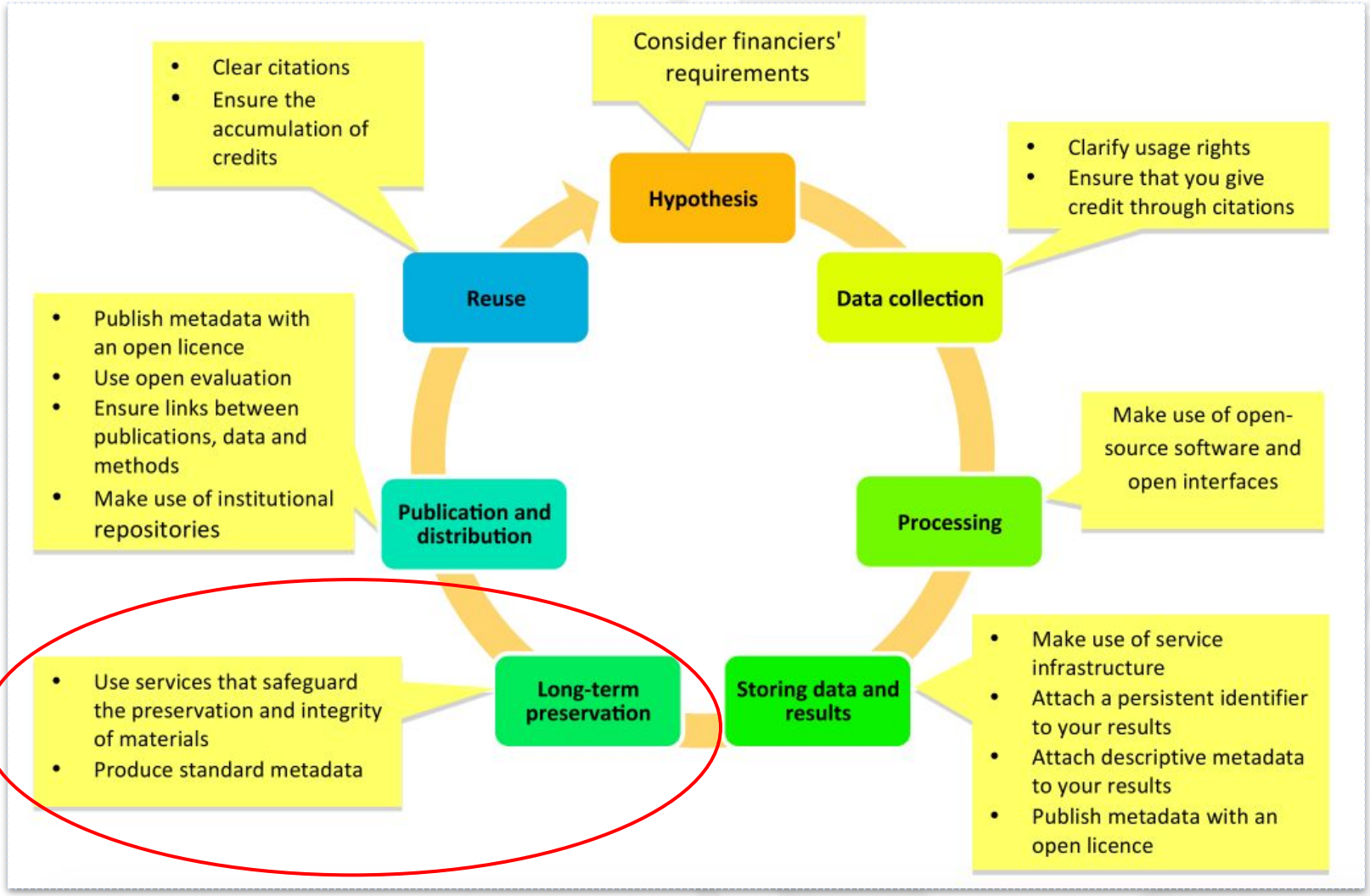


One of over a million bits has been ‘flipped’

Photo: Louise Gubb

See: Omar Shehata (May 1, 2019): Issue 01 Science + Society: **Unraveling the JPEG**. (Online). Available: https://parametric.press/issue-01/unraveling-the-jpeg/

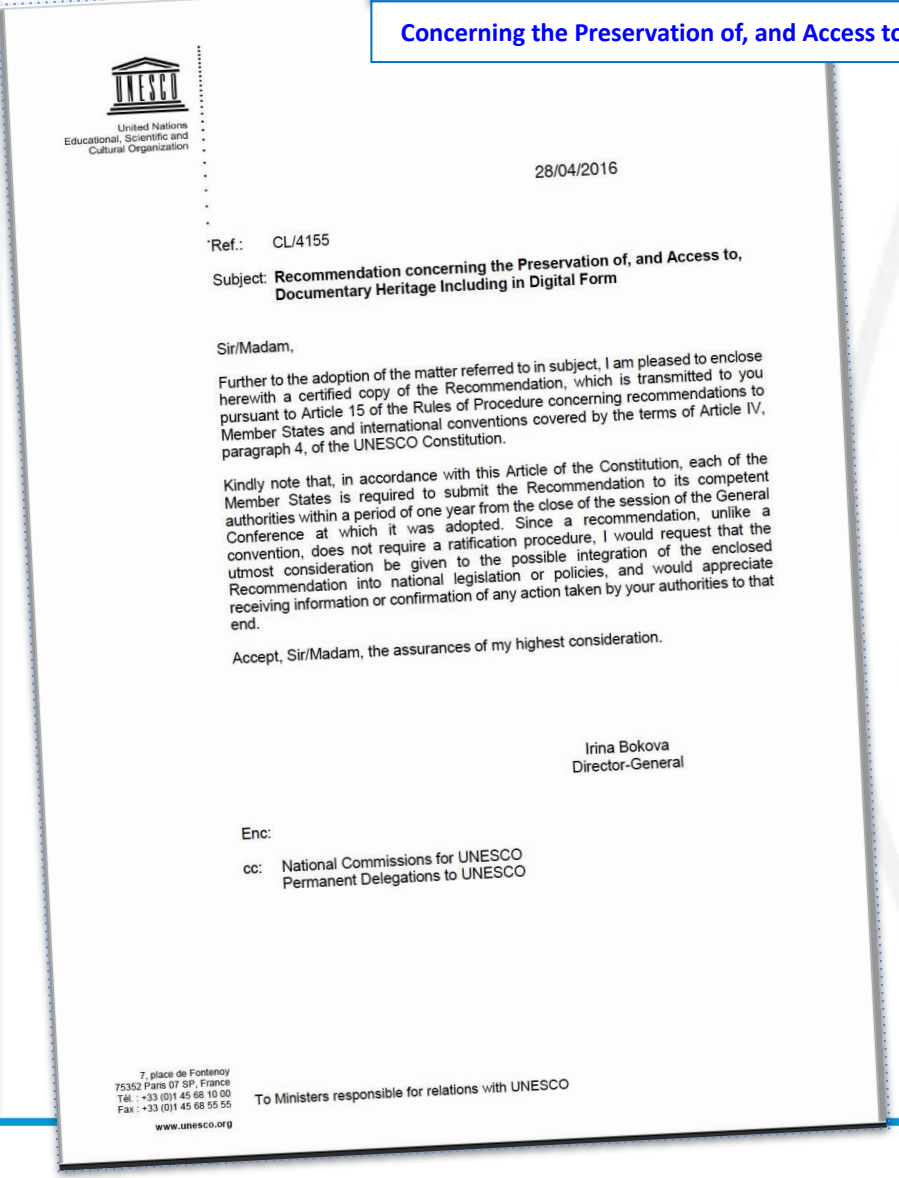
Preservation is foundational to **Open Science**



Source: Foster Open Science: **What is Open Science?** Figure 1. **Promoting openness at different stages of the research process**. https://www.fosteropenscience.eu/content/what-open-science-introduction

UNESCO

Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form (2016)



United Nations
Educational, Scientific and
Cultural Organization

28/04/2016

Ref.: CL/4155

Subject: **Recommendation concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form**

Sir/Madam,

Further to the adoption of the matter referred to in subject, I am pleased to enclose herewith a certified copy of the Recommendation, which is transmitted to you pursuant to Article 15 of the Rules of Procedure concerning recommendations to Member States and international conventions covered by the terms of Article IV, paragraph 4, of the UNESCO Constitution.

Kindly note that, in accordance with this Article of the Constitution, each of the Member States is required to submit the Recommendation to its competent authorities within a period of one year from the close of the session of the General Conference at which it was adopted. Since a recommendation, unlike a convention, does not require a ratification procedure, I would request that the utmost consideration be given to the possible integration of the enclosed Recommendation into national legislation or policies, and would appreciate receiving information or confirmation of any action taken by your authorities to that end.

Accept, Sir/Madam, the assurances of my highest consideration.

Irina Bokova
Director-General

Enc:

cc: National Commissions for UNESCO
Permanent Delegations to UNESCO

To Ministers responsible for relations with UNESCO

7, place de Fontenoy
75352 Paris 07 SP, France
Tél. : +33 (0)1 45 68 10 00
Fax : +33 (0)1 45 68 55 55
www.unesco.org

'The world's documentary heritage is of global importance and responsibility to all, and should be fully preserved and protected for all, with due respect to and recognition of cultural mores and practicalities. It should be permanently accessible and re-usable by all without hindrance. It provides the means for understanding social, political, collective as well as personal history. It can help to underpin good governance and sustainable development. For each State, its documentary heritage reflects its memory and identity, and thus contributes to determine its place in the global community.' (p.6)

Izolo powered by Arkivum Perpetua

- ***Izolo***
  - isiXhosa word for “yesterday”
- Institutional on-site preservation platform
- Powered by Arkivum Perpetua - every bit Archived
  - “a long- term data archiving service that is designed to be scalable, secure and trustworthy... the service will preserve your data at the bit-level over decade long timescales”
  - “fully supports and fulfils the functions of the Open Archival Information System (OAIS) reference model and allows for a range of digital preservation activities to be undertaken including: ***ingest, storage, management, administration, access, and preservation planning***.”

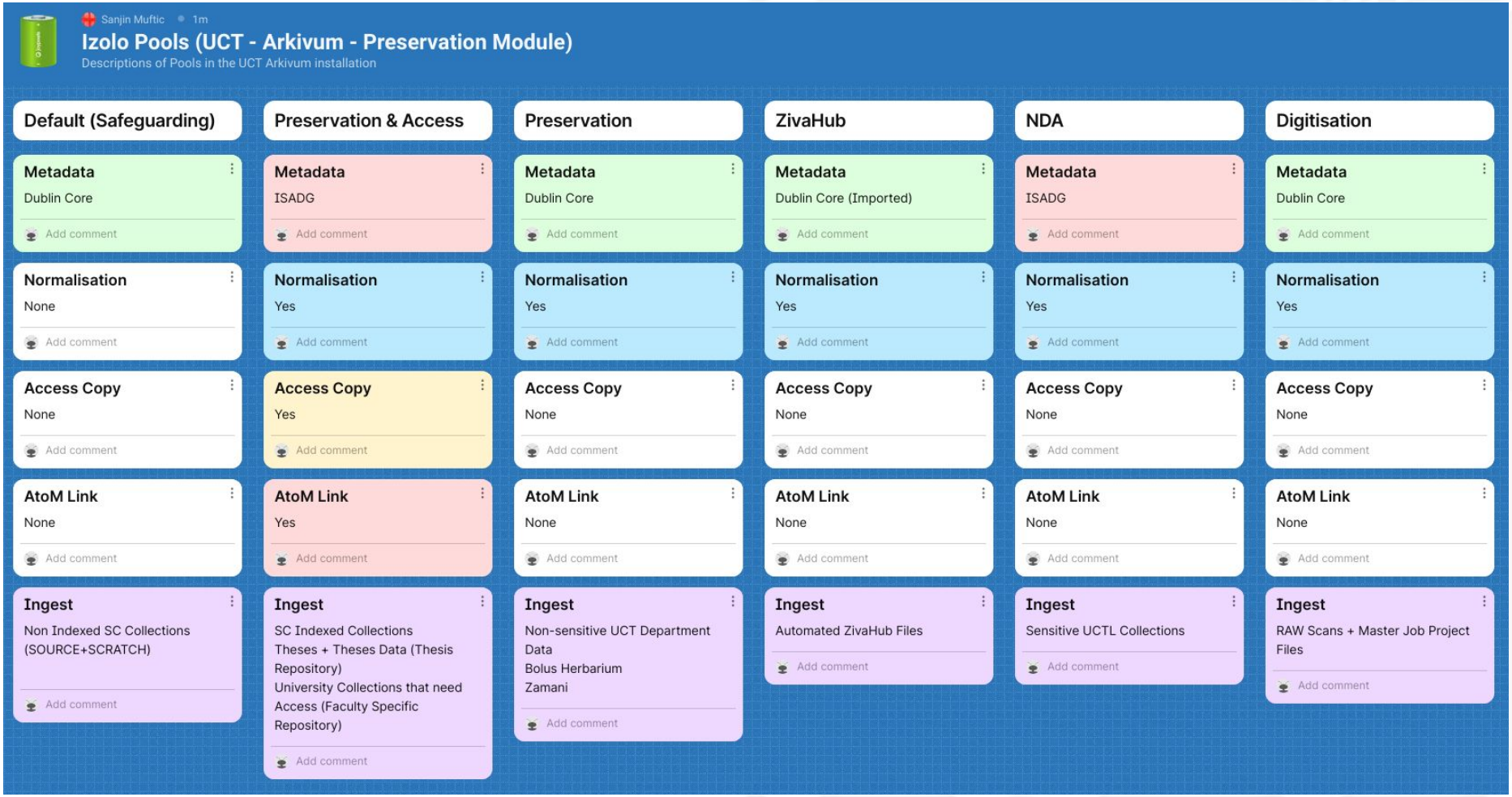
Source: Arkivum On-site brochure

Two parts of Izolo (powered by Arkivum Perpetua)

- Perpetua is architected using the industry leading **open source and open standard** technologies, **Archivematica** and **AtoM** from Artefactual.

archivematica	atom
Preservation Module *...to create standards-based, self-documenting and system-independent archival information packages (AIPs).*	**Accessibility Module** *...to provide web access and listing for standards-based archival description objects*
• Ingest of files • Processes that check integrity of each bit • Processes that convert the files ingested into formats that will remain accessible into the future • Access to original and normalized files	• Web-browser based system • Presentation of ingested files with human readable **metadata** • Provides multilingual access (if desired) • Organizes files into repositories and collections (“fonds”)

The Pools of Izolo



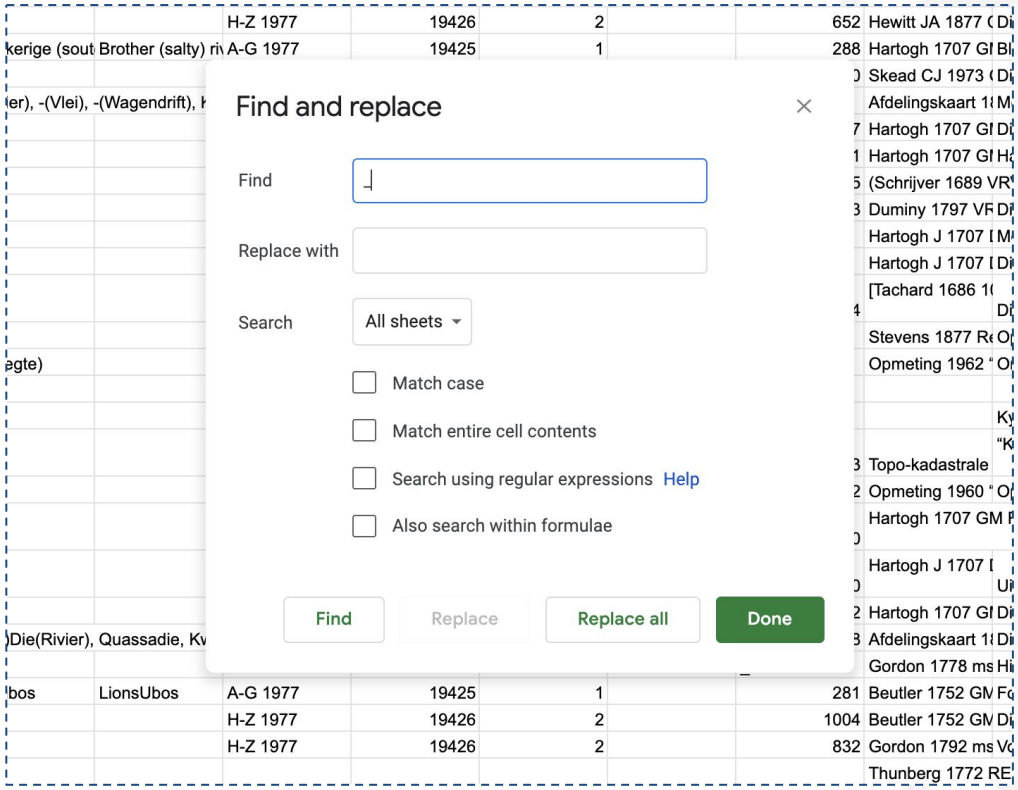
What is Automation?

Benefits of automation in digital preservation

Automation is about using computers or machines to do stuff that people would otherwise do manually.

Solving a problem (usually a very boring/monotonous one) through a number of finite but repeatable steps.

Example 1 - Find & Replace and Lookup

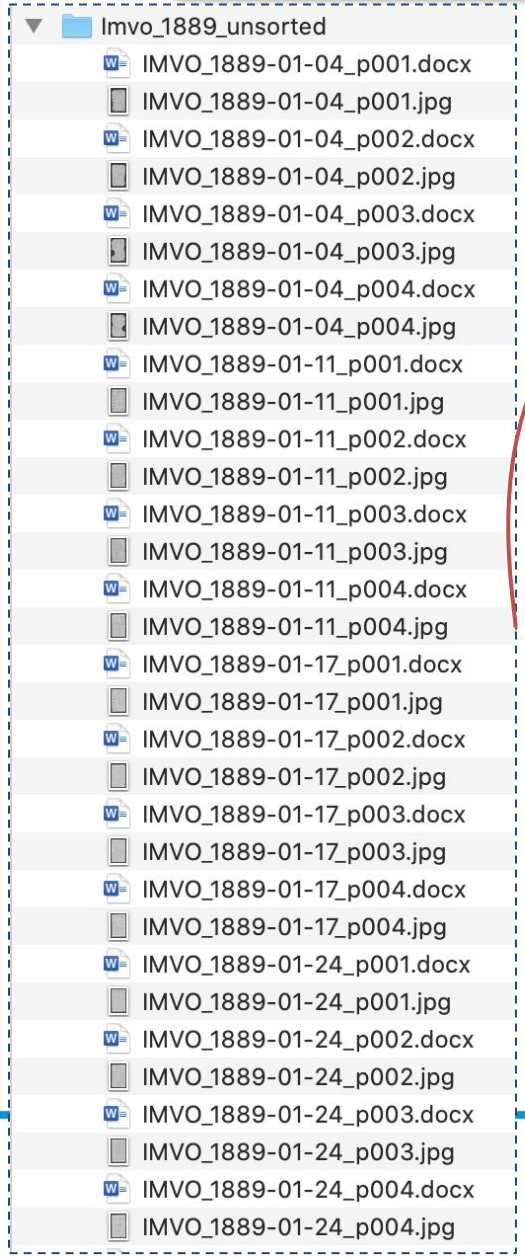


Jagger Store Crate Look Up:

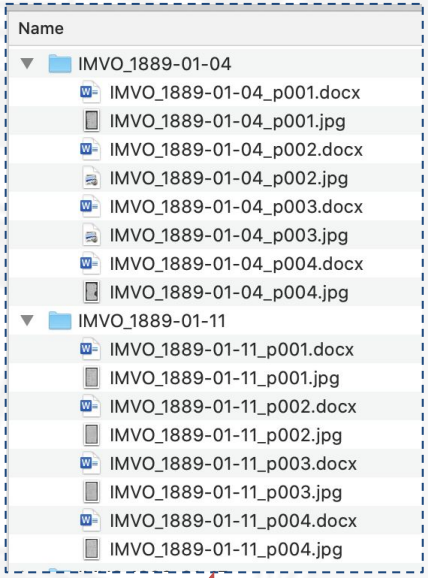
A	B	C	D
Jagger Store Crate Look Up:			
1st Number	2nd Number	Lookup code	Collection with Detail:
4	12	4.12	B 888.9 AEL to B 910 BUES
AV Store Crate Look Up:			
1st No.	2nd No.	Lookup code	Collection with Detail:
40	2	AV40.2	BVF14— Environmental

	A	B
1	Crate Number	Collection
2	AV37.2	???
3	AV43.1	ASL vhs
4	AV43.2	ASL vhs
5	AV44.1	ASL vhs
6	AV30.2	BC1310/ BVF31—Howes-Howell
7	AV30.1	BC718—Bertram Woods ; BC1233— Binckes(?) ; BC1310/BVF31— Howes-Howell
8	AV48.2	Binders
9	AV45.1	Boxes
10	AV49.1	Boxes
11	AV50.1	Boxes
12	AV50.2	Boxes
13	AV51.1	Boxes
14	AV51.2	Boxes
15	AV1.1	BVF01_0559 BVF01_1771
16	AV2.1	BVF01_1772 BVF01_2265
17	AV3.2	BVF01_1798 BVF01_3303
18	AV4.2	BVF01_1798 BVF01_3303
19	AV2.2	BVF01_2266 BVF01_2797
20	AV3.1	BVF01_3304 BVF01_3791
21	AV4.1	BVF01_3792 BVF01_4301
22	AV5.2	BVF01_4811 BVF01_5317
23	AV5.1	BVF01_5318 BVF01_5825
24	AV6.1	BVF01_5826 BVF01_6333

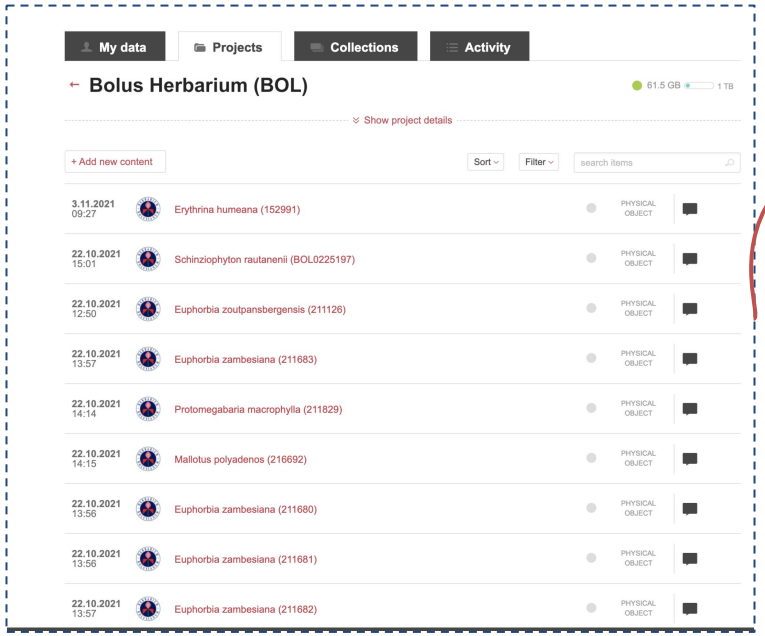
Example 2 - Group files with similar names into folders




```
In [7]: count=0;
for file in os.listdir(folder_path):
    filename, file_extension = os.path.splitext(file)
    newedition = file[0:4]
    year = file[5:9]
    print(year)
    month = file[10:12]
    print(month)
    day = file[13:15]
    print(day)
    pages = file[16:20]
    print(pages)
    print(file_extension)

    newfilename = newedition + '_'+ year + '-' + month + '-' + day + '_' + pages + file_extension
    newdir = "/Users/sanjinmuftic/Desktop/IXIT/Imvo_1889_sorted" + "/" + newedition + '_'+ year + '-' + month + '-' + d
    final_path = newdir + '/'+ newfilename
    print("NEW FILE")
    print("ORIGINAL FILE NAME: "+ file)
    print("NEW DIR: "+ newdir)
    print("NEW FILENAME: "+ newfilename)
    print("NEW FILE PATH: "+ final_path)
    #dir_path = dir + dir_name
    #print(f'dir_path: {dir_path}')

    #check if directory exists or not yet
    if not os.path.exists(newdir):
        os.makedirs(newdir)

    if os.path.exists(newdir):
        #file_path = dir + file
        #print(f'file_path: {file_path}')

        # move files into created directory
        #shutil.move(file_path, dir_path)
        copyfile(file, final_path)
        count += 1
        print(count)

p002
.tif
NEW FILE
ORIGINAL FILE NAME: IMVO_1889-09-05_p002.tif
NEW DIR: /Users/sanjinmuftic/Desktop/IXIT/Imvo_1889_IIIF_sorted/IMVO_1889-09-05
```



Example 3 - Capture details about a number of items




```
In [*]: #create a dataframe for the list of items with a DOI
#construct rows of the dataframe which wil have articles and their files together per row
#use ArticleID from rslt data frame to call file details and attach those details to article row
#THIS WILL NOT WORK FOR ITEMS WITH MULTIPLE FILES
filelist = pd.DataFrame(columns=['Id','DOI', 'ZivaHubURL','Thumb','Name', 'Downloadurl'])
filelistcount = 0
for entry in rslt: #for every entry in the project get files of article
    articleid = str(entry['id'])
    #using articleID call to call a HTTP request to get details of files
    url= "https://api.figshare.com/v2/articles/"+articleid+"/files"
    print(url)
    getfiles = raw_issue_request("GET", url, token)
    for thangs in getfiles:
        print('Name:', thangs['name'],'DownloadURL:',thangs['download_url'])
    #construct the row with all of the details - note that name of file without extension is the BOLUS CODE
    filelist.loc[filelistcount]=entry['id'], entry['doi'], entry['url_public_html'], entry['thumb'], thangs['name'], th
    print(filelist.loc[filelistcount]) #display entry
    filelistcount += 1 #move counter
```

```
https://api.figshare.com/v2/articles/14566014/files
Name: 47987.jpg DownloadURL: https://ndownloader.figshare.com/files/27950034
Id                                                  14566014
DOI                                          10.25375/uct.14566014
ZivaHubURL      https://zivahub.uct.ac.za/articles/physical_ob...
Thumb           https://s3-eu-west-1.amazonaws.com/ppreviews-u...
Name                                                47987.jpg
Downloadurl       https://ndownloader.figshare.com/files/27950034
Name: 0, dtype: object
https://api.figshare.com/v2/articles/14566026/files
Name: 47997.jpg DownloadURL: https://ndownloader.figshare.com/files/27950046
Id                                                  14566026
DOI                                          10.25375/uct.14566026
ZivaHubURL      https://zivahub.uct.ac.za/articles/physical_ob...
Thumb           https://s3-eu-west-1.amazonaws.com/ppreviews-u...
Name                                                47997.jpg
Downloadurl       https://ndownloader.figshare.com/files/27950046
Name: 1, dtype: object
https://api.figshare.com/v2/articles/14566074/files
```

	Id	DOI	ZivaHubURL	Thumb	Name	Downloadurl
0	14566074	10.25375/uct.14566074	https://zivahub.uct.ac.za/articles/physical_object/Diastella_proteoides_51633_/14566074	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950112/thumb.png	51633.jpg	https://ndownloader.figshare.com/files/27950112
1	14566119	10.25375/uct.14566119	https://zivahub.uct.ac.za/articles/physical_object/Buckinghamia_celsissima_60254_/14566119	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950157/thumb.png	60254.jpg	https://ndownloader.figshare.com/files/27950157
2	14566026	10.25375/uct.14566026	https://zivahub.uct.ac.za/articles/physical_object/Diastella_divaricata_subsp_divaricata_47997_/14566026	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950046/thumb.png	47997.jpg	https://ndownloader.figshare.com/files/27950046
3	14566014	10.25375/uct.14566014	https://zivahub.uct.ac.za/articles/physical_object/Brabejum_stellatifolium_47987_/14566014	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950034/thumb.png	47987.jpg	https://ndownloader.figshare.com/files/27950034
4	14566113	10.25375/uct.14566113	https://zivahub.uct.ac.za/articles/physical_object/Hakea_sericea_60252_/14566113	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950151/thumb.png	60252.jpg	https://ndownloader.figshare.com/files/27950151
5	14566167	10.25375/uct.14566167	https://zivahub.uct.ac.za/articles/physical_object/Euplassa_hoehnei_60734_/14566167	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950214/thumb.png	60734.jpg	https://ndownloader.figshare.com/files/27950214
6	14565999	10.25375/uct.14565999	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_album_44778_/14565999	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950019/thumb.png	44778.jpg	https://ndownloader.figshare.com/files/27950019
7	14566164	10.25375/uct.14566164	https://zivahub.uct.ac.za/articles/physical_object/Euplassa_inaequalis_60724_/14566164	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950211/thumb.png	60724.jpg	https://ndownloader.figshare.com/files/27950211
8	14566095	10.25375/uct.14566095	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_argenteum_59656_/14566095	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950133/thumb.png	59656.jpg	https://ndownloader.figshare.com/files/27950133
9	14566098	10.25375/uct.14566098	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_argenteum_59658_/14566098	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950136/thumb.png	59658.jpg	https://ndownloader.figshare.com/files/27950136
10	14566101	10.25375/uct.14566101	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_argenteum_59660_/14566101	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950139/thumb.png	59660.jpg	https://ndownloader.figshare.com/files/27950139
11	14566104	10.25375/uct.14566104	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_argenteum_59661_/14566104	https://ndownloader.figshare.com/files/27950142/preview/27950142/thumb.png	59661.jpg	https://ndownloader.figshare.com/files/27950142
12	14566107	10.25375/uct.14566107	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_argenteum_59662_/14566107	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950145/thumb.png	59662.jpg	https://ndownloader.figshare.com/files/27950145
13	14566065	10.25375/uct.14566065	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_sp_51626_/14566065	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950100/thumb.png	51626.jpg	https://ndownloader.figshare.com/files/27950100
14	14566086	10.25375/uct.14566086	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_levisanus_51640_/14566086	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950124/thumb.png	51640.jpg	https://ndownloader.figshare.com/files/27950124
15	14566071	10.25375/uct.14566071	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_concavum_51632_/14566071	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950109/thumb.png	51632.jpg	https://ndownloader.figshare.com/files/27950109
16	14566047	10.25375/uct.14566047	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_spissifolium_subsp_spissifolium_48735_/14566047	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950079/thumb.png	48735.jpg	https://ndownloader.figshare.com/files/27950079
17	14566050	10.25375/uct.14566050	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_spissifolium_subsp_spissifolium_48736_/14566050	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950082/thumb.png	48736.jpg	https://ndownloader.figshare.com/files/27950082
18	14566125	10.25375/uct.14566125	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_sp_60398_/14566125	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950163/thumb.png	60398.jpg	https://ndownloader.figshare.com/files/27950163
19	14566137	10.25375/uct.14566137	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_sp_60401_/14566137	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950169/thumb.png	60401.jpg	https://ndownloader.figshare.com/files/27950169
20	14566053	10.25375/uct.14566053	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_spissifolium_subsp_spissifolium_48737_/14566053	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950085/thumb.png	48737.jpg	https://ndownloader.figshare.com/files/27950085
21	14566110	10.25375/uct.14566110	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_xanthoconus_60251_/14566110	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950148/thumb.png	60251.jpg	https://ndownloader.figshare.com/files/27950148
22	14566029	10.25375/uct.14566029	https://zivahub.uct.ac.za/articles/physical_object/Leucadendron_xanthoconus_47998_/14566029	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950049/thumb.png	47998.jpg	https://ndownloader.figshare.com/files/27950049
23	14566032	10.25375/uct.14566032	https://zivahub.uct.ac.za/articles/physical_object/Leucospermum_conocarpodendron_subsp_conocarpodendron_47999_/14566032	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950052/thumb.png	47999.jpg	https://ndownloader.figshare.com/files/27950052
24	14566116	10.25375/uct.14566116	https://zivahub.uct.ac.za/articles/physical_object/Leucospermum_wittebergense_60253_/14566116	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950154/thumb.png	60253.jpg	https://ndownloader.figshare.com/files/27950154
25	14566017	10.25375/uct.14566017	https://zivahub.uct.ac.za/articles/physical_object/Paranomus_bracteolaris_47996_/14566017	https://s3-eu-west-1.amazonaws.com/ppreviews-uct-2876656789329/27950037/thumb.png	47996.jpg	https://ndownloader.figshare.com/files/27950037

DIGITAL LIBRARY SERVICES

Automation meets Preservation?

Benefits of automation in digital preservation

- **Reducing costs** -
  - ... getting more done with the same staff and the same budget.
- **Getting more done** -
  - ... many organisations struggle to keep up with the amount and complexity of the digital content they need to keep and provide access to.
- **Getting the right things done -**
  - Allows these staff to focus on interesting and challenging areas of digital preservation, for example selection, appraisal and curation rather than having to do more mundane stuff like checking fixity or analyzing file formats.
- **Reduce errors** -
  - People can make mistakes so automating tasks can help reduce the number of these – and in the digital preservation world this can mean lower long-term risks that content will be lost or won't be accessible or usable.

Source:Can digital preservation be automated in the real world?

Automation in the Preservation Module

- *A suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. From the Submission Information Package it generates:*
  - *Archival (AIP) | Dissemination (DIP)*

- *Archivematica implements default format policies based on an analysis of the significant characteristics of file formats.* […]
  - *Format identification, package extraction, transcription and* ***normalization*** *for preservation and access.*
    - *Normalization is the process of converting (transcoding) an ingested digital object to a preferred* ***preservation format****.*
    - *Access files may also be created by transcoding*
    - *Original File, Normalised File, Access File*

- *Archivematica generates* ***checksums*** *upon ingest of files into the system, and will verify those checksums before storing the AIP to insure* ***bit integrity****.*

Source: Archivematica website

All the information packages

NOT AUTOMATED	AUTOMATED	AUTOMATED
Submission Information Package (SIP)	**Archival Information Package (AIP)**	**Dissemination Information Package (DIP)**

• Content (digital object)
• Metadata
	• Content Information (CI)
	• Preservation Description Information (PDI)
	  ○ Reference Information
	  ○ Provenance Information
	  ○ Context Information
	  ○ Fixity Information
	• Packaging Information (PI)
	• Descriptive Information (DI)
		• Information Package derived from AIP

Source: McMaster University

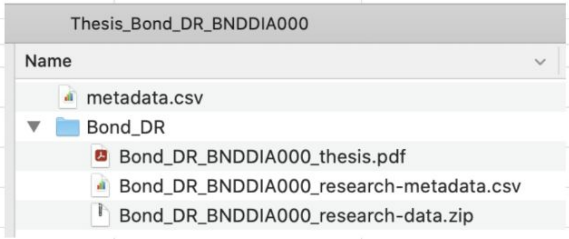
Things we want to Automate

- Creating the **Submission Information Package (SIP)**
  - **SIP** is required to be in a particular **format**
  - Packaging **the content and metadata** for Ingest
  - **Assemble** and **verify** prior to Ingest
- What we need to think about
  - Map through the whole process in detail
  - Reducing Outliers - there will always be special cases
  - Automation in Stages
  - Providing a solution for UCT departments needing a service

Import of Thesis Data

			ISADG Metadata Sheet Details			
Faculty	Faculty of Science	*required*	**slug**	uct-sci-1997		
Year	1997	*required*	**Name of Your Archive**	Thesis_Nash_K_NSHKAT000	*once you create the zip file, change its name to this before ingesting*	
First Name	Katherine Lindsay Gifford	*required*	Extent	1 thesis and 1 research data package with research metadata file		
Last Name	Nash	*required*				
Initials	K	*required*	Put the files in folder with this name and ZIP:	Thesis_Nash_K_NSHKAT000		
Student Number		*required*	**Structure of your Archive**	metadata.csv	File	
Thesis PDF?	☑	*required*		Nash_K_NSHKAT000	Folder	
Title of Thesis	Inclusion of Compounds of multipedal hosts	required		**CSV:**	Nash_K_NSHKAT000_research-metadata	research-metadata
				ZIP:	Nash_K_NSHKAT000_research-data	research-data
Suggested Student Number	NSHKAT000			**PDF:**	Nash_K_NSHKAT000_thesis	thesis

STEP 2: Follow these steps	
1. Use this name to create a folder, this is the name of your archive	Thesis_Nash_K_NSHKAT000
2. Create a sub-folder with this name:	Nash_K_NSHKAT000
3. Inside the subfolder place:	
3.1. A ZIP of all the thesis data with this name:	Nash_K_NSHKAT000_research-data
3.2. The metadata CSV sheet for the thesis data with this name:	Nash_K_NSHKAT000_research-metadata
3.3. If applicable a copy of the thesis PDF with this name:	Nash_K_NSHKAT000_thesis
4. Download the ingest metadata sheet.	
4.1 Click on the **metadata** sheet, choose **File** and **Download** a **CSV**	
4.2 Rename the downloaded file with this name	metadata
4.3 Copy the metadata.csv into the archive folder (Point 1 above)	
5. Compress the archive folder into a ZIP file and ingest into Izolo	



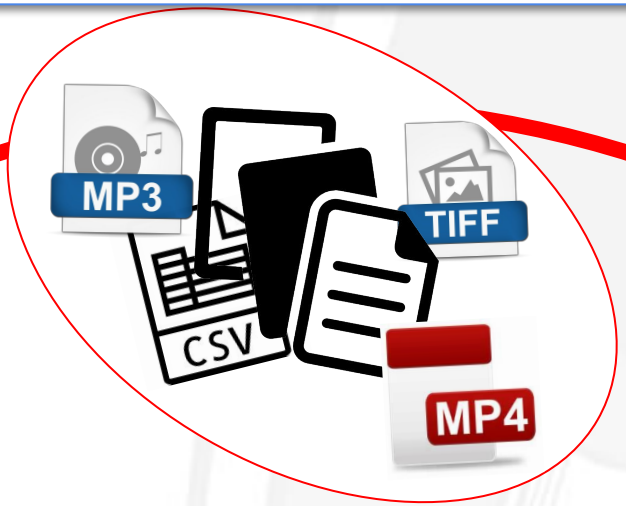
All the Submission Packages Needed

Ingest	Size	Files
Indexed Digital Objects for AtoM	10GB	5,000+
Safeguarding Collections (awaiting Indexing)	50TB	1,000,000+
Theses Data	20GB	10,000+
Digitisation High Scans	10TB	200,000+
UCT Departments	100TB	500,000+

How do others do it?

- Many institutions are looking for automation
- Usually a custom built solution that ties in with their respective systems, repositories, processes
- A lot of code is shared on online platforms such as GitHub/ GitLab
- An example from LSE: https://git.lse.ac.uk/bywell/lse-digital-toolkit-perl-version/-/wikis/LSE-Digital-Toolkit

All published ZivaHub items to be preserved on Izolo



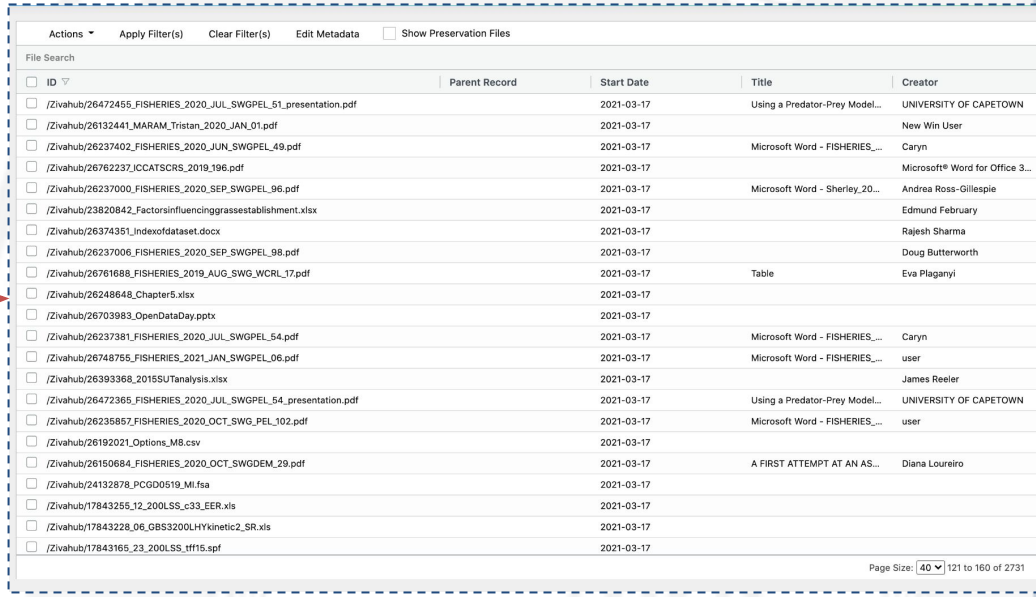
ZivaHub
Open Data UCT

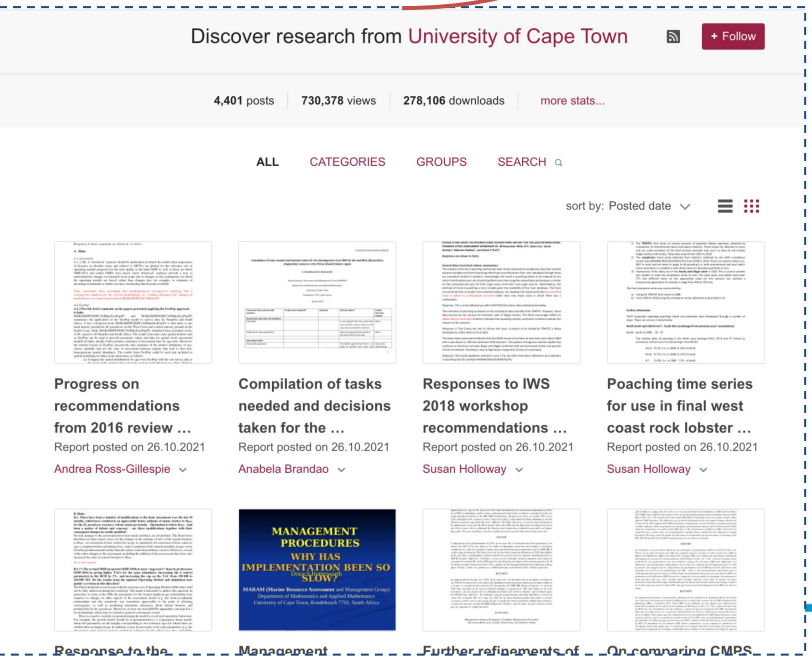
Izolo
Digital Preservation UCT

Application **P**rogramming **I**nterface

arkivum
Bringing archived data to life

All published ZivaHub items to be preserved on Izolo





UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Automation is GREAT … but ...

- Before it works you need to THINK about it a LOT
- Need to be able to set it up
  - Think through the process and articulate the steps
- Need the skills to code and infrastructure
- Will not solve all the problems
  - The outliers
  - The difficult decisions on WHAT to preserve

Source:Can digital preservation be automated in the real world?

Automation Advantages

- Reduce the human error and labour in monotonous processes
- Re-allocate the time for human staff to think through other issues
- Collaboration while building solutions
- Speed up “breaking-barriers” by providing access and enabling preservation of “our” resources

Let's change what we value in research.

Sign DORA

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
LIBRARIES

DIGITAL LIBRARY SERVICES

dls@uct.ac.za

@DigitalUct

uct-dls.slack.com

Data Stewards Community

DIGITAL LIBRARY SERVICES

http://www.digitalservices.lib.uct.ac.za/

UCT DMP

omeka s

ibal

iiif

arkivum
Bringing archived data to life

ZivaHub
Open Data UCT