

DIGITAL PRESERVATION FOR THE MASSES:

**Using Archivematica and DSpace as
Solutions for Small-sized Institutions**

(and other options)

Digital Commonwealth Annual Conference 2012

Joseph Fisher

Database Management Librarian @ UMass Lowell

- Electronic Resources
- Digitization Projects
 - MBLC ILS grant to digitize the Paul E. Tsongas Congressional Papers
 - Additionally included Lowell Historical Building Surveys
 - Current proposal to digitize Tewksbury Almshouse records
- Digital Commons repository
- Digital Scholarly Services – NSF data management planning

Vice President Digital Commonwealth



AGENDA

- Why Digital Preservation
- For whom
- What it is
- How to approach it
- OAIS and TRAC
- Basic requirements
- Solutions
 - DuraCloud
 - LOCKSS
 - DSpace
 - Archivematica



WHERE THIS INFORMATION ORIGINATES

Graduate (2011) University of Arizona SIRLS

Graduate Certificate Program in Digital Information Management (DigIn) digin.arizona.edu

Digital Preservation Management Workshop:

Implementing Short-term Strategies for Long-term Problems
(attended 2004 (Cornell) and 2010 (ICPSR) @ MIT)

SAA Digital Archives Specialist (DAS) program

- Nine workshops and exams required for DAS Certificate
- 24 workshops currently in four sections with 8 online



WHY IS DIGITAL PRESERVATION IMPORTANT??

Obsolescence!! Bit Rot!!



NOT JUST FOR LIBRARIES & ARCHIVES ANYMORE

- **Researchers** – coming soon to a government grant near you – Data Management Planning
- **Record Managers** – born digital tsunami
- **People** – personal archiving

“Indeed, we are now all our own librarians.”

Ellysa Stern Cahoy, Penn State University Libraries

The Signal: Digital Preservation, Library of Congress blog, 4/9/2012

<http://blogs.loc.gov/digitalpreservation/2012/04/the-challenge-of-teaching-personal-archiving/>



DIGITAL PRESERVATION: WHAT IS IT?

- “The series of managed activities to ensure continued access to digital materials for as long as necessary.” *DCP Handbook*. Digital Preservation Coalition (2008)
- **Managed activities:** “defined very broadly...refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change.”
- **Access:** “continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy, and functionality deemed to be essential for the purposes the digital material was created and/or acquired for.” [see “significant properties”]
- **Authenticity:** “the trustworthiness of the electronic record as a record.... that whatever is being cited is the same as it was when it was cited unless the accompanying metadata indicates any changes.”

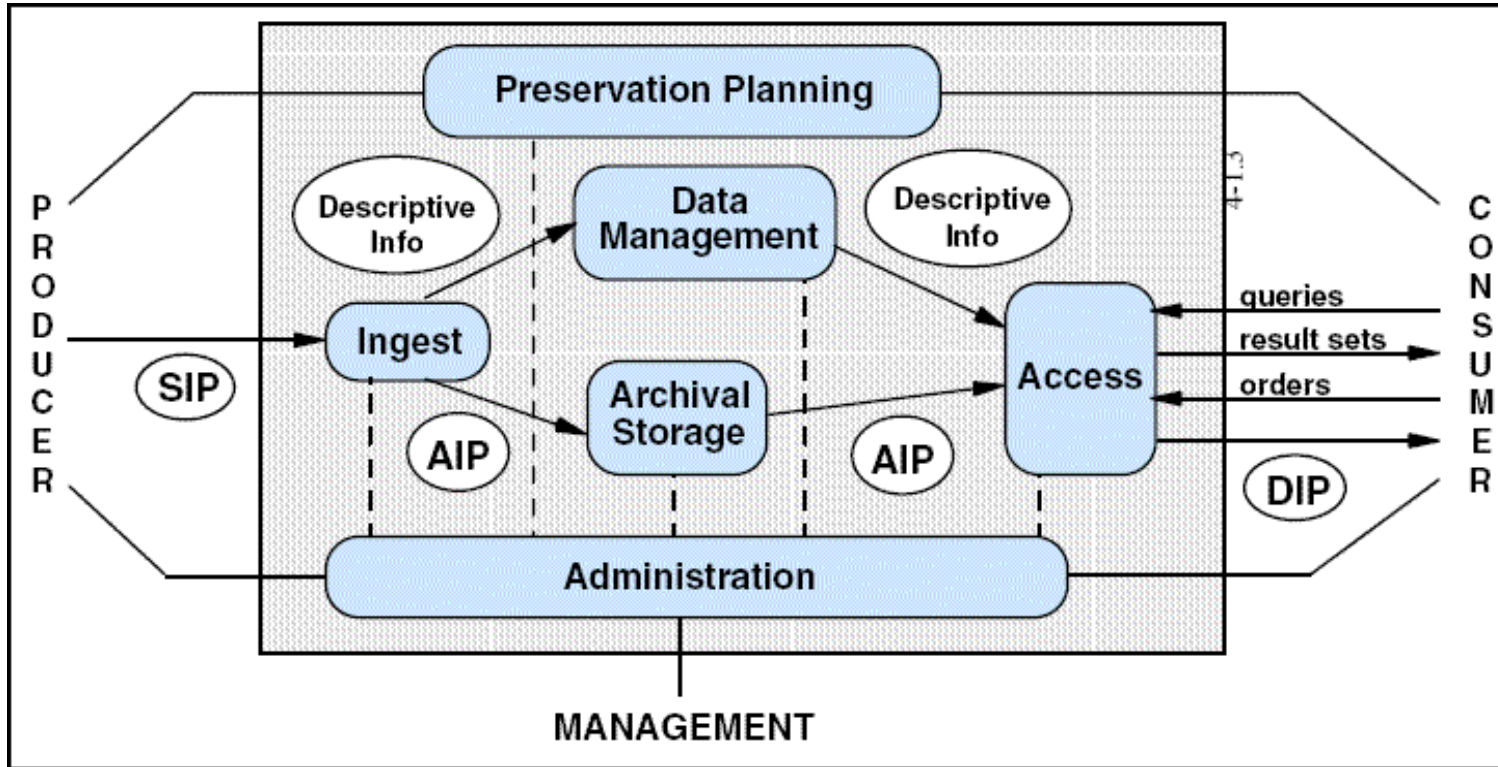


FIVE ORGANIZATIONAL STAGES

1. *Acknowledge*: Understanding that digital preservation is a local concern
2. *Act*: Initiating digital preservation projects
3. *Consolidate*: Segueing from projects to programs
4. *Institutionalize*: Incorporating the larger environment and rationalizing programs
5. *Externalize*: Embracing inter-institutional collaboration and dependency.

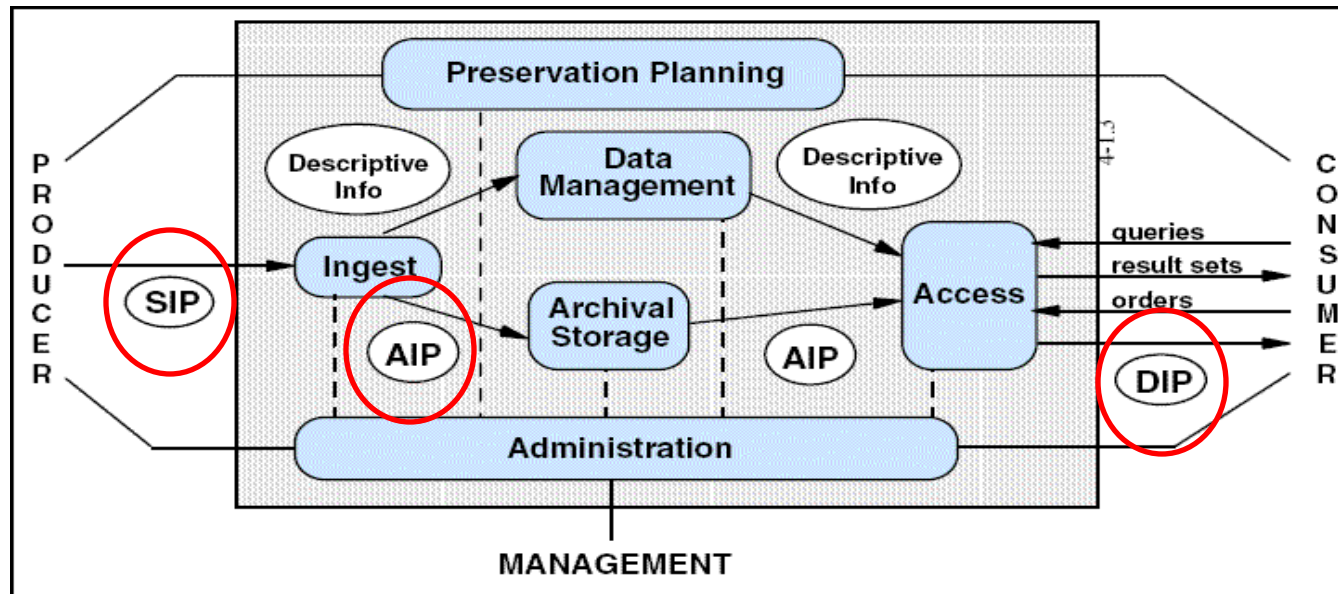


OAIS REFERENCE MODEL (OPEN ARCHIVAL INFORMATION SYSTEM)



The Consultative Committee for Space Data Systems
(CCSDS) released in 1999





SIP – Submission Information Package (Producer)

- Appraisal & Accession – Validate & Verify
- Virus protection & Checksum
- file normalization (PDF/A)
- metadata – description, preservation, structural

AIP – Archival Information Package (Management)

- Store digital object(s) and associated metadata
- Dublin Core, MODS, PREMIS, METS package
- Refresh, migrate, error-check, replace

DIP – Dissemination Information Package (Consumer)

- Retrieval, delivery, and security
- Monitor Designated Community for changing needs



WHAT IS THE OPEN ARCHIVAL INFORMATION SYSTEM?

- It's "Open" in the flexible sense of an outline, framework, or blueprint.
- And an "Information System" in the sense of a comprehensive, integrated, and complex conceptual construct.
- ISO 14721:2003
- a collection of six high-level services, or functional components, that, taken together, fulfill the OAIS's dual role of preserving and providing access to the information in its custody.



SIX CORE OAIS REQUIREMENTS

1. Negotiate and accept appropriate information from Information Producers
2. Obtain sufficient intellectual control of the information to ensure Long-term preservation
3. Determine the scope of the Designated Community
4. Ensure the information is understandable by the Designated Community without the assistance of the information producers
5. Follow clearly documented policies & procedures to ensure the information is preserved against all reasonable contingencies
6. Make the information available to Designated Community



TDR AND TRAC

TRUSTWORTHY REPOSITORIES AUDIT & CERTIFICATION

Categories:

A. Organizational Infrastructure

- Governance, organizational structure, staffing & viability
- Procedural accountability & policy framework
- Financial sustainability, contracts, licenses, & liabilities

B. Digital Object Management

- Ingest -- preservation strategies & processing procedures
- Workflows, documentation, records, & audit procedures
- Unique identifiers, metadata, & verification testing
- preservation planning & strategies
- Access policies & designated community interaction

C. Technologies, Technical Infrastructure, & Security

- Software, updates, security
- Checksum error-checking
- Backups & disaster recovery



ISO 16363

- The standard is titled the *Trusted Digital Repository (TDR) Checklist*
- Based upon the *Trusted Digital Repositories and Audit Checklist (TRAC)*
- CCSDS publication (Magenta Book) Sep. 2011
(The Consultative Committee for Space Data Systems)
- ISO approved standard for publication in Mar. 2012
- working group also wrote and submitted ISO 16919, entitled, *Requirements for Bodies providing Audit and Certification*



BASIC REQUIREMENTS OF DIGITAL PRESERVATION

- The more copies the safer
 - Replicate data on multiple storage systems
- The more independent the copies the safer
 - Save in different geographical locations
 - Save on different technology system types
- The more frequently the copies are audited by checksum error checking the safer
 - Audit or scrub the replicas to detect damage, and repair by overwriting the bad copy with a good copy

David S. H. Rosenthal

“Bit Preservation: A Solved Problem.” *International Journal of Digital Curation*. 1.5 (2010)



SIP TO AIP

- Save and maintain at least one copy of file kept exactly as is in it's original file format
- Convert copy for public use to PDF or JPEG
- Plan to migrate use copy as format changes
- Normalize copy to preservation format if necessary
 - Word doc to PDF/A1b
- Possibly migrate copy of Word doc as format changes
- Dublin Core descriptive record and maybe a MODS record also in XML
- PREMIS record in XML – preservation metadata
- METS record in XML – structural metadata
























SO WHAT ARE SOME OPTIONS?

- DuraCloud
- LOCKSS
- Dspace
- Archivematica



DuraCloud Subscription Plans - [Get Plan Information Now!](#)

	DuraCloud Preservation Basic i	DuraCloud Preservation Plus i	DuraCloud Enterprise i	
			Standard	Premium
Price (all bandwidth and compute charges included) this price includes	\$1,500/year for first TB \$1,300/year for additional TBs	\$3,000/year for first TB \$2,600/year for additional TBs	\$5,900/year for first TB \$1,300/year for additional TBs	\$7,200/year for first TB \$2,600/year for additional TBs
Number of redundant copies i	2	4	2	4
Number of cloud data centers storing content i	1	2	1	2
Online backup i				
Web-based administrative dashboard i				
Automatic content health checks and reports i				
Dynamic usage reports i				
Online sharing i				
Automatic synchronization and file recovery across cloud data centers i				
Shibboleth Authentication -- available to Internet2 and InCommon members i	coming soon	coming soon	coming soon	
Media serving i	optional	optional		
Sub-account creation i				
Account management and access controls per sub-account i				
Access to additional cloud services, as they become available i				
Users i	2	2	unlimited	

- *Pricing quoted above is for storing up to 10TB. If interested in storing more than 10TB, please contact us for a custom quote as pricing decreases for storage above 10TB.*





- Began development 1991 (beta release 2001)
- Still managed out of Stanford
- Global LOCKSS hosted at Stanford
- **Private LOCKSS Networks (PLN)** to preserve manuscript and image collections, data sets, etc.
- Example is MetaArchive Cooperative
 - First year server purchase \$4,600
 - \$1 /GB/year + \$5,500 or \$3,00 annual membership
 - 1 TB = \$24,100 for 3 years for sustaining member
 - Good example of a TRAC audit report (PDF available)
- At least 6 nodes (so 6 copies)
- Maintain storage server



DSPACE

- HP-MIT Libraries Alliance (2002)
- DuraSpace (2009)
- Current version 1.8.2 (24 Feb. 2012)
- Linux / Windows (Java)
- “DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets.”
- Beginning with 1.7 (Dec. 2010) began adding significant digital curation functionalities



DSPACE DEVELOPMENT

- 1.7.0 released 17 Dec. 2010
 - Discovery – enables faceted searching
 - AIP backup and restore – Duracloud integration
 - Export/import entire hierarchy, community, or collection
 - Curation System (CS)
 - Profile collection based on format type
 - Check that required metadata fields are present
 - Enhance/replace/normalize an item's metadata or content
 - Checksum checker
- 1.8.0 released 4 Nov 2011
 - Bulk metadata editing
 - SWORD client – push content to other SWORD repositories
 - Rewrite Creative Commons license
 - Virus checking during submission
- 3.0 projected Oct/Nov 2012
 - Version number scheme changing to 2 digits
 - Major release increments 1st digit & bug fixes 2nd digit
 - Item-level versioning – features from Dryad Project



DSPACE INSTALLATION

Prerequisite Software :

- Linux or Windows
- Oracle Java JDK
- Maven (Java build tool for stage 1)
- Ant (Java build tool for stage 2)
- PostgreSQL or Oracle
- Tomcat
- Perl



Paul E. Tsongas Digital Congressional Collection

Search within this community and its collections:

[Advanced Search](#)

Collections in this community

- [Paul E. Tsongas Congressional Papers](#)

Recent Submissions

[Letter from Jack E. Stark to Paul E. Tsongas](#)

Stark, Jack E. (*University of Massachusetts Lowell Libraries*, 1976-12-27)

[Fifth District Report](#)

Tsongas, Paul E. (*University of Massachusetts Lowell Libraries*, 1975-06-23)

[Tsongas finds economy and energy primary district concerns](#)

Tsongas, Paul E. (*University of Massachusetts Lowell Libraries*, 1974)

[Letter from Ronald Bryan Ginn to Paul Tsongas](#)

Ginn, Ronald Bryan (*University of Massachusetts Lowell Libraries*, 2012-05-06)

Search DSpace

Search DSpace

This Collection

[Advanced Search](#)

Browse

- All of DSpace
 - ◇ [Communities & Collections](#)
 - ◇ [By Issue Date](#)
 - ◇ [Authors](#)
 - ◇ [Titles](#)
 - ◇ [Subjects](#)
- This Community
 - ◇ [By Issue Date](#)
 - ◇ [Authors](#)
 - ◇ [Titles](#)
 - ◇ [Subjects](#)

My Account

- [Logout](#)
- [Profile](#)
- [Submissions](#)

Context

- [Edit Community](#)
- [Export Community](#)
- [Export Metadata](#)

Edit Item

[Item Status](#) [Item Bitstreams](#) [Item Metadata](#) [View Item](#) [Curate](#)

Welcome to the item management page. From here you can withdraw, reinstate, move or delete the item. You may also update or add new metadata / bitstreams on the other tabs.

Item Internal ID:	4
Handle:	123456789/8
Last Modified:	2012-05-06 11:30:34.071
Item Page:	http://localhost:8080/xmlui/handle/123456789/8
Edit item's authorization policies:	Authorizations...
Withdraw item from the repository:	Withdraw...
Move item to another collection:	Move...
Completely expunge item:	Permanently delete

[Return](#)

Search DSpace

[Advanced Search](#)

Browse

- [All of DSpace](#)
 - [Communities & Collections](#)
 - [By Issue Date](#)
 - [Authors](#)
 - [Titles](#)
 - [Subjects](#)

My Account

- [Logout](#)
- [Profile](#)
- [Submissions](#)

Administrative

- [Access Control](#)
 - [People](#)
 - [Groups](#)
 - [Authorizations](#)
- [Registries](#)
 - [Metadata](#)
 - [Format](#)



Edit Item

[Item Status](#) [Item Bitstreams](#) [Item Metadata](#) [View Item](#) [Curate](#)

Bitstreams

Name	Description	Format	View	Order	
Bundle: ORIGINAL					
<input type="checkbox"/> pt10189.pdf	Public access PDF	Adobe PDF	[view]	1 (Previous:1)	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> pt10189_1.tif	Preservation TIFF page 1	TIFF	[view]	2 (Previous:2)	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> pt10189_2.tif	Preservation TIFF page 2	TIFF	[view]	3 (Previous:3)	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> pt10189_dublin-core.xml	Dublin Core xml	XML	[view]	4 (Previous:4)	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> pt10189_pdf-a.pdf	Preservation PDF/A	Adobe PDF	[view]	5 (Previous:5)	<input type="checkbox"/> <input type="checkbox"/>
Bundle: LICENSE					
<input type="checkbox"/> license.txt		License	[view]	1 (Previous:1)	<input type="checkbox"/> <input type="checkbox"/>

[Upload a new bitstream](#)

Update bitstream order

Delete bitstreams

Return

Browse

- All of DSpace
 - [Communities & Collections](#)
 - [By Issue Date](#)
 - [Authors](#)
 - [Titles](#)
 - [Subjects](#)

My Account

- [Logout](#)
- [Profile](#)
- [Submissions](#)

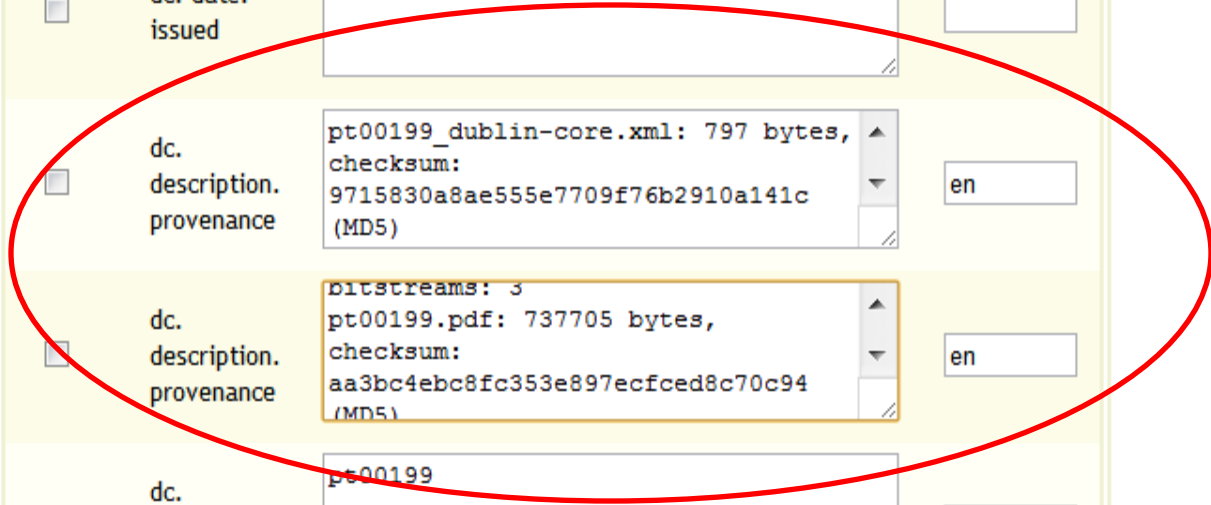
Administrative

- Access Control
 - [People](#)
 - [Groups](#)
 - [Authorizations](#)
- Registries
 - [Metadata](#)
 - [Format](#)
- [Items](#)
- [Withdrawn Items](#)
- [Control Panel](#)
- [Statistics](#)
- [Import Metadata](#)
- [Curation Tasks](#)



<input type="checkbox"/>	dc. date. accessioned	2012-05-06T15:29:19Z	<input type="text"/>
<input type="checkbox"/>	dc. date. available	2012-05-06T15:29:19Z	<input type="text"/>
<input type="checkbox"/>	dc. date. issued	1976-12-27	<input type="text"/>
<input type="checkbox"/>	dc. description. provenance	pt00199_dublin-core.xml: 797 bytes, checksum: 9715830a8ae555e7709f76b2910a141c (MD5)	<input type="text" value="en"/>
<input type="checkbox"/>	dc. description. provenance	bitstreams: 3 pt00199.pdf: 737705 bytes, checksum: aa3bc4ebc8fc353e897ecfced8c70c94 (MD5)	<input type="text" value="en"/>
<input type="checkbox"/>	dc. identifier. other	pt00199	<input type="text"/>
<input type="checkbox"/>	dc. identifier. uri	http://hdl.handle.net/123456789/8	<input type="text"/>

- ◊ [Format](#)
- [Items](#)
- [Withdrawn Items](#)
- [Control Panel](#)
- [Statistics](#)
- [Import Metadata](#)
- [Curation Tasks](#)



System Curation Tasks

Handle of DSpace
Object:

Hint: Enter [your-handle-prefix]/0 to run a task across entire site (not all tasks may support this capability)

Task:

- Profile Bitstream Formats
- Profile Bitstream Formats
- Check for Required Metadata
- Check Links in Metadata

Perform

Queue

Search DSpace

Go

[Advanced Search](#)

Browse

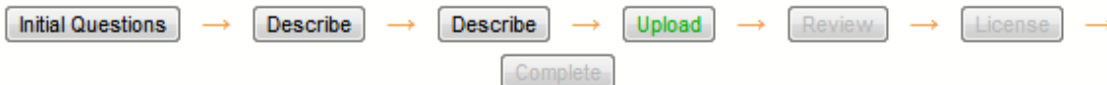
- [All of DSpace](#)
 - [Communities & Collections](#)
 - [By Issue Date](#)
 - [Authors](#)
 - [Titles](#)
 - [Subjects](#)

My Account

- [Logout](#)
- [Profile](#)
- [Submissions](#)



Item submission



Search DSpace

- Search DSpace
- This Collection

[Advanced Search](#)

Upload File(s)

File:

No file chosen

Please enter the full path of the file on your computer corresponding to your item. If you click "Browse...", a new window will allow you to select the file from your computer.

File Description:

Optionally, provide a brief description of the file, for example "Main article", or "Experiment data readings".

Browse

- All of DSpace
 - [Communities & Collections](#)
 - [By Issue Date](#)
 - [Authors](#)
 - [Titles](#)
 - [Subjects](#)
- This Collection
 - [By Issue Date](#)
 - [Authors](#)
 - [Titles](#)
 - [Subjects](#)

Files Uploaded

Primary	File	Size	Description	Format	
<input type="radio"/>	<input type="checkbox"/> pt00442.pdf	1759910 bytes	Public access PDF	application/pdf (Supported)	<input type="button" value="Edit"/>
File checksum: MD5:a92c30f503f03262aec5843ae897bed					
<input type="button" value="Remove selected files"/>					
<input type="button" value=" < Previous"/> <input type="button" value=" Save & Exit"/> <input type="button" value=" Next >"/>					

My Account

- [Logout](#)
- [Profile](#)
- [Submissions](#)

Context

- [Edit Collection](#)
- [Item Mapper](#)
- [Export Collection](#)
- [Export Metadata](#)

Administrative



ARCHIVEMATICA

- A free and open-source digital preservation system.
- Uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model.
- Managed by Artefactual Systems (Toronto) in collaboration with the UNESCO Memory of the World's Subcommittee on Technology, the City of Vancouver Archives, the University of British Columbia Library, the Rockefeller Archive Center, Simon Fraser University Archives and Records Management, and a number of other collaborators.



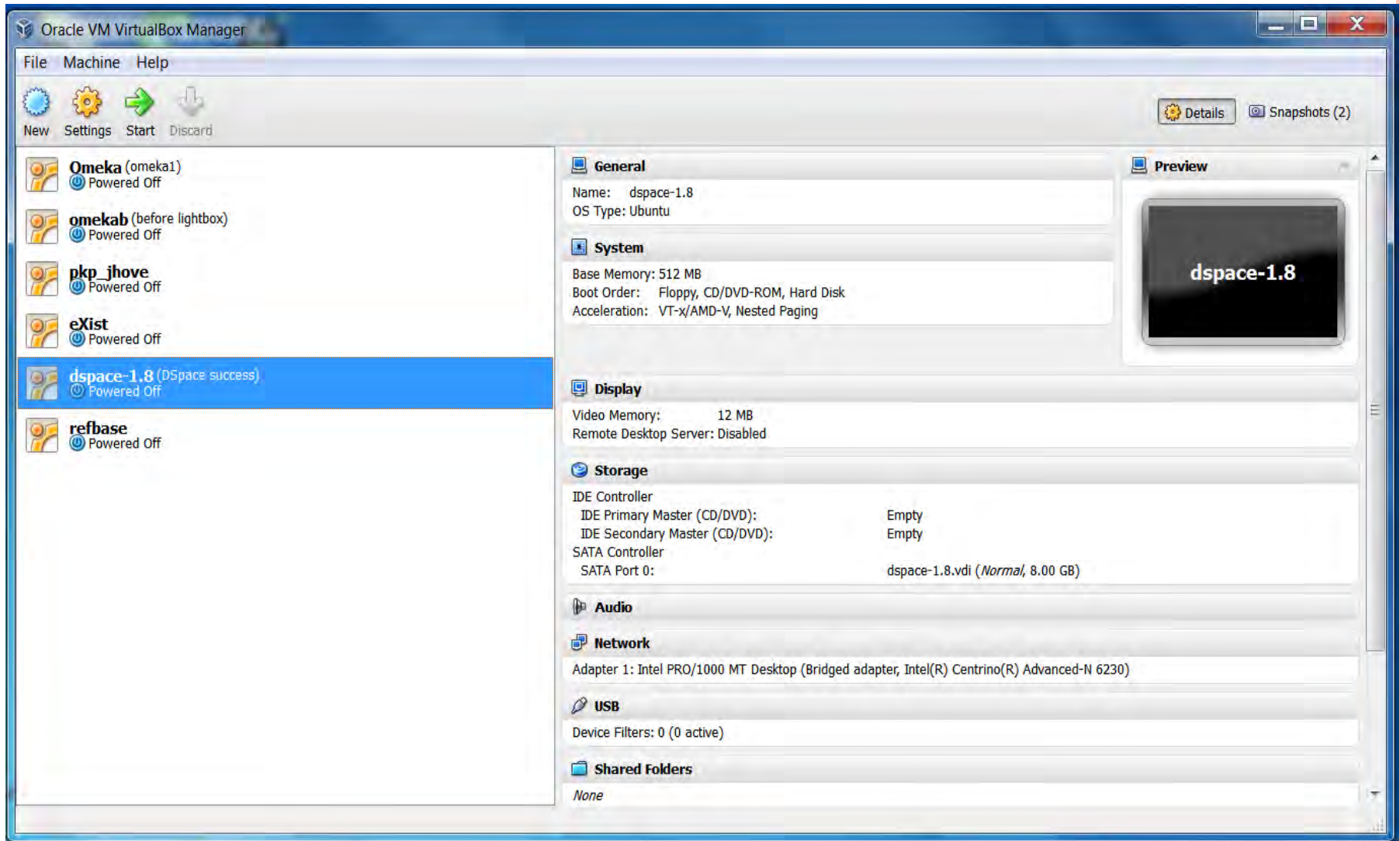
ARCHIVEMATICA DEVELOPMENT

- 0.6 alpha release 19 May 2010
- 0.7 alpha release 18 Feb. 2011
- 0.8 alpha release 3 Feb 2012
 - Complete standards-compliant PREMIS in METS implementation
 - Multiple normalization options
 - Ability to ingest DSpace exports



Archivematica Appliance Installation in Oracle VM VirtualBox

1. Install Open Source VirtualBox

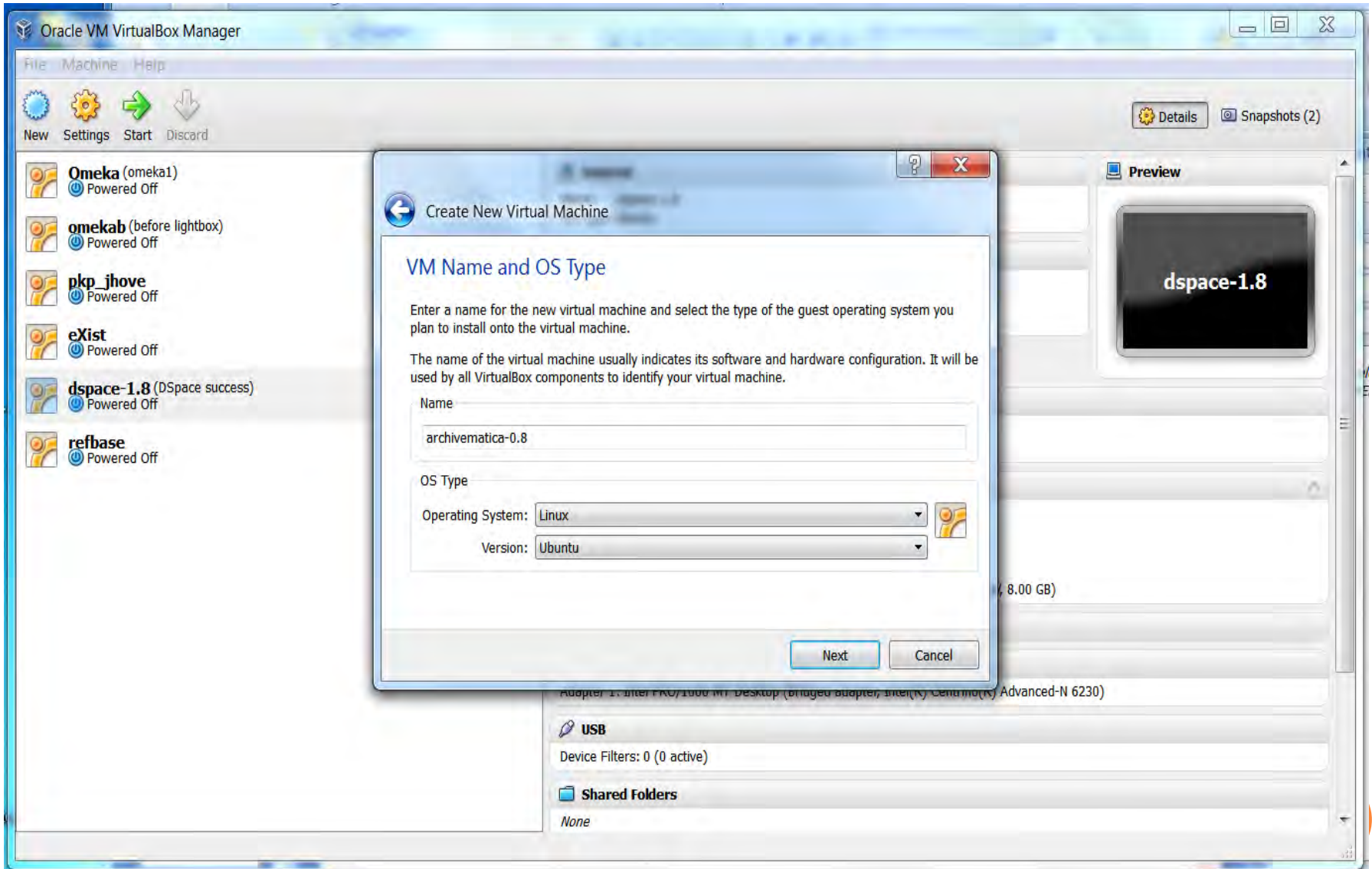


DOWNLOAD ARCHIVEMATICA APPLIANCE FILE

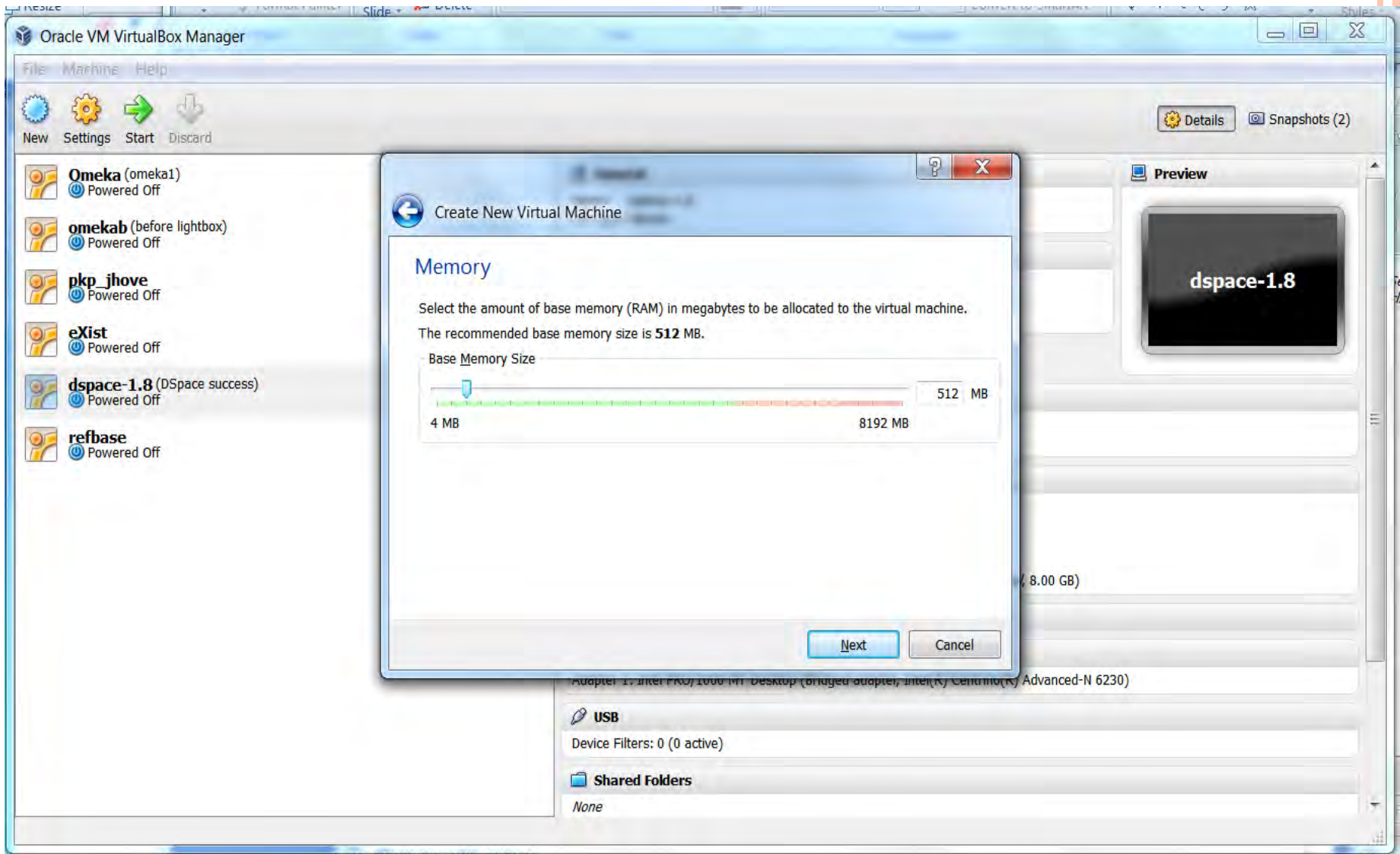
1. <http://archivematica.org/downloads/archivematica-0.8-alpha-vmrk.tbz>
2. **Requires something like 7Zip to unpack to this tar file:**
`archivematica-0.8-alpha-vmrk2.tar`
3. **Which you then unpack yet again to the appliance installation file:**
`archivematica-0.8-alpha.vmrk`



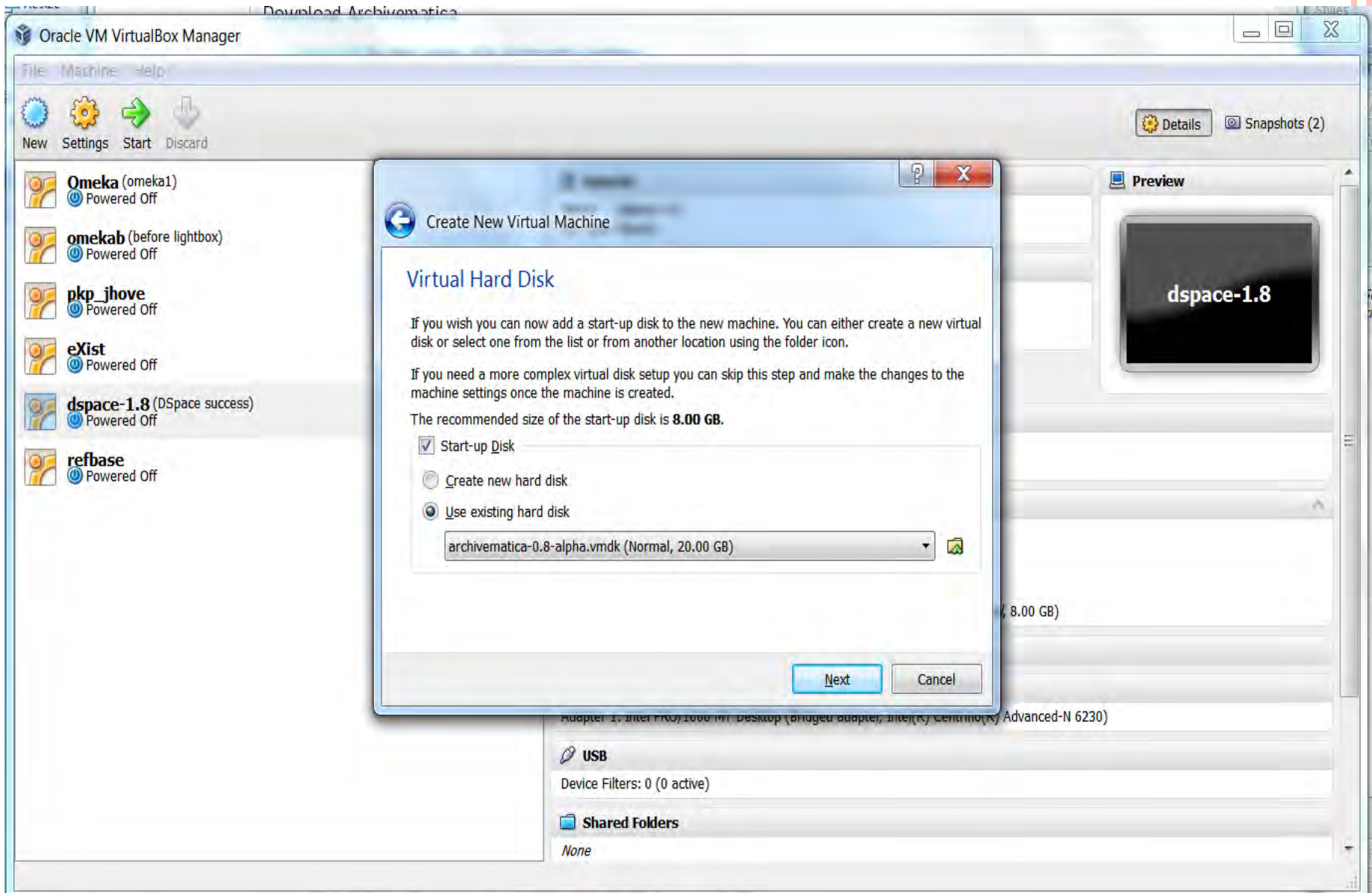
Create New VM and Assign OS to Linux/Ubuntu



Accept default Memory allocation



Point to the Archivematica vmdk appliance file



Additional recommended configurations outlined on Archivemata site

Requires some knowledge of Linux command line

- type `ifconfig` in terminal you should see a IP address like '192.168.56.101' (likely eth1 interface)

```
$ ifconfig
eth1      Link encap:Ethernet  HWaddr fe:54:00:9d:92:64
          inet addr:192.168.56.101  Bcast:192.168.56.255  Mask:255.255.255.0
          inet6 addr: fe80::1c6b:7bff:fe07:d6b6/64  Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:24  errors:0  dropped:0  overruns:0  frame:0
          TX packets:45  errors:0  dropped:0  overruns:0  carrier:0
          collisions:0 txqueuelen:0
          RX bytes:1400 (1.4 KB)  TX bytes:5815 (5.8 KB)
```

- From here your machine should be connectable via SFTP. Download a SFTP client, a popular opensource option is FileZilla, which works on Linux and Windows. If using OSX cyber duck is reported to be a decent opensource SFTP client.
- the connection information should be as follows

```
username: demo
password: demo
IP/Hostname: 192.168.56.101 < results of ifconfig likely '192.168.56.*'
port: 22
destination folder: /home/demo/ < if this is not set you will have to navigate to /home/demo directory
```

Using Virtual Box Guest Additions

Instructions contributed by R. Pearce-Moses at http://groups.google.com/group/archivemata/browse_frm/thread/6ee0ef53832706e1?hl=en

1. update

```
sudo aptitude update
```

```
sudo aptitude upgrade
```

2. Determine which linux headers to get using `uname -a`

```
uname -a
```

3. Install the appropriate headers headers

```
$apt-get install linux-headers-2.6.32-38-generic
```

- Replace version based on info from `uname` command as necessary

4. install gcc

```
sudo apt-get install gcc
```

5. Restart to complete installation of the update.

6. From the VBox Device menu click Install Guest Additions

List of MicroServices and Tools used by Archivematica

- Receive SIP
 - verifyChecksum
 - Review SIP
 - extractPackage
 - assignIdentifier
 - parseManifest
 - clean Filename
 - Quarantine SIP
 - lockAccess
 - virusCheck
 - Appraise SIP
 - identifyFormat
 - validateFormat
 - extractMetadata
 - decidePreservationAction
 - Prepare AIP
 - gatherMetadata
 - normalizeFiles
 - createPackage
 - Review AIP
 - decideStorageAction
 - Store AIP
 - writePackage
 - replicatePackage
 - auditfixity
 - readPackage
 - updatePackage
 - Provide DIP
 - uploadPackage
 - updateMetadata
 - Monitor Preservation
 - checkFormatRegistry
 - migrateFormat
 - synchronizeAIPsandDIPs
- EXT3, Thunar, incron, flock
- UUID, Detox, Easy Extract, ClamAV
- FITS, JHove, DROID, NLNZ Extractor
- FFident, Unoconv, Ffmpeg, OpenOffice
- ImageMagick, Inkscape, Xena
- Bagit, SAMBA, NFS-common, Poster
- ICA-AtoM, DCB Dashboard



Live demo of Exercise One in this Archivematica Tutorial:

<https://www.archivematica.org/mediawiki/images/0/05/Tutorial-08.pdf>

Another good introductory tutorial is a YouTube video available on the home page of the Archivematica Wiki:

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



RECOMMENDATIONS:



Library of Congress Digital Preservation Outreach & Education (DPOE)

<http://www.digitalpreservation.gov/education/courses/index.html>

DPOE Webinars: Intro to Digital Preservation 1-3 by Jody DeRidder

<http://www.aserl.org/archive/>

DCC Curation Lifecycle Model: How to use the Curation Lifecycle Model

<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

