The OpenChoice Project Server: Decommissioning and Archiving Report

David Bliss, Andrew Childress, Nicolette Khan, Amanda Winograd
Digital Archiving and Preservation
Dr. Pat Galloway
4 May 2016

Introduction

OpenChoice was an IMLS-grant funded project that began in 2007 at the University of Texas at Austin iSchool. The two co-principal investigators were Drs. Don Turnbull and Miles Efron. Dr. Loriene Roy was also involved with the project, along with Dr. Arro Smith (who was, at the time, a postdoctoral student). OpenChoice was intended to be an open source tool for library communities to control their own web-filtering. The project was never completed. However, the documents, programs, and environment that the project was created with are preserved on the server detailed in this report.

The OpenChoice project server is a decommissioned RAID server that houses material from the iSchool project of the same name. The machine is a Dell PowerEdge 2500 with a Pentium III Processor, 1400 MHz. It houses an SCSI (Small Computer System Interface) controller and 4096 MB synchronous DRAM. It runs on a Linux operating system and contains GNU GRUB version 0.97, a Multiboot boot loader. The server consists of 5 drives and runs on Ubuntu 6.06 "Dapper Drake" operating system.

We were assigned the responsibility of imaging the contents of the server and archiving the disk image on DSpace so that the server could be officially decommissioned. We also wanted to provide a context for the disk image and its contents within the DSpace community, which can be found under the name "School of Information OpenChoice Project Server."

Research Correspondence

Our team corresponded with several members of the OpenChoice team to gain information about the project and collect documentation. This correspondence was added to DSpace under "Correspondence".

Andrew corresponded with former postdoctoral student, Dr. Arro Smith beginning on February 18th. In response to an email with questions about the OpenChoice machine, Arro provided background on the project and some explanation of how the project ended. OpenChoice was originally to have been Arro's dissertation research project, but the team was not able to complete the project in a way that let him use it for that purpose.

Andrew also corresponded with former co-principal investigator, Dr. Miles Efron, beginning on February 18th. In response to the same questions about the OpenChoice machine that were sent to Arro Smith, Miles wrote a short email in which he stated that he thought the machine was a VM (it was not), and that they had used it to "develop and test a highly configurable web proxy based on the Apache Squid proxy".

David corresponded with former co-principal investigator, Dr. Don Turnbull, beginning on March 7th. Dr. Turnbull provided preliminary information on the project and a number of publications and presentations on the development of OpenChoice, which were added to DSpace under "Publications". He also provided information on some of the tools used in development, and helped to identify contributors whose names we found on the server.

Methodology

Our methodology for this project consisted of four main components: disk imaging, web archiving, metadata, and ingest.

1. Disk imaging

Before we examined the server, our initial speculation was that a disk image could be created using either a Forensic Recovery of Evidence Device (FRED), or by booting from a live CD. After talking to Sam Burns, Senior IT Manager of the iSchool and major contributor to this project, it was determined that a live CD would be the best choice, and that an image could be

produced of the entire RAID. Our initial planning for the disk image process also involved reaching out to Don Turnbull, who supervised the OpenChoice project itself. We prepared to provide a large blank external hard drive to transfer the image. We hit an early roadblock during our first meeting with Sam when he powered the server on and found an unexpected root password prompt. Sam needed some weeks to locate the root password.

Once Sam had found the root password for the server, we met once again. Sam booted the server into read-only mode and found that it was running Ubuntu 6.06 "Dapper Drake". A live CD with the matching version of Ubuntu was written, and the server was successfully booted into rescue mode. Sam took the opportunity to explore the file structure for an initial probe into the server's contents. The next week, we reconvened with an external hard drive with the intent of copying the disk image onto it, but found that NTFS drives and partitions larger than 300GB could not be mounted onto the server. We returned the following week with a blank VFAT drive, but once again found that we could not mount the image: because of the system's software RAID configuration, the server could only be accessed in rescue mode, which mounts the hard drives and prevents the *dd* disk imaging command from being run on them.

In place of a disk image, Sam captured the contents of the server directory-by-directory onto the hard drive using the *cp* command. This process took several hours and did not provide us with the single disk image file we had hoped to be left with. Occasionally, the transfer process would return an error indicating that a file could not be copied because there were "too many symbolic links", but Sam explained that this simply indicated a loop of files referring to one another which the system knew to break. Sam did ultimately succeed in backing up the server onto the drive, which due to its formatting is readable only on a Linux system. We explored the disk's contents in BitCurator, which prevented any writing to the directories. Access to the disk gave us the names of two additional OpenChoice team members, as well as information on

some of the tools used on the project. The transferred directories include system files as well as

project documentation, and total 27.8 GB.

2. Web archiving

Since the original website for the OpenChoice project was still up on an iSchool server in

March 2016, we were able to archive the website and include it as part of the documentation of

the OpenChoice project. The website includes the following pages:

- An introduction to the project (index.html)
- A summary of the project (project.html). This is essentially the same as the paper "OpenChoice: A Platform for Web Content Classification & Filtering", by Don Turnbull & Miles Efron, which was sent to us by Don Turnbull in the early stages of our project.
- A list of people involved with the project (people.html)
- Links to related projects (related.html)
- An FAQ (faq.html)
- A newsletter sign-up form (newsletter.html)

The main navigational menu on the website also included a link titled "Blog", but this link

was broken. We assume that this is because it linked to another iSchool server, Sentra, which

is no longer active. (Note that when exploring the contents of the hard drive from the

OpenChoice machine we did find files that appear to have been elements of this blog. Further

investigation may prove that the blog is still stored on the archived hard drive image.) The

source HTML for the newsletter sign-up form also includes a reference to the Sentra server: the

opening form tag reads

```
<form action="http://sentra.ischool.utexas.edu/~choice/process.php"
method="post" style="text-align: right">.
```

The flat, non-dynamic structure of the website made it simple to create an accurate and

mostly complete record of the website. Text at the bottom of each page indicates that every

page was last updated either February 24, 2007, September 4, 2006, or September 1, 2006.

To create an archive of the website, we used the Terminal running Bash on a MacBook Pro with the following command:

```
wget --mirror https://ischool.utexas.edu/~choice/
```

This created a folder titled www.ischool.utexas.edu that contained a single folder titled ~choice, inside which were all the HTML files listed above and a folder titled web_files that contained two image files and a CSS file.

To create a tar file of the archived website, we navigated in the Terminal to the parent folder of the folder www.ischool.utexas.edu and ran

```
tar -cvzf choicewebarchive.tar.gz www.ischool.utexas.edu
```

Finally, we uploaded the tar file into the OpenChoice Project Server Documentation subcommunity in DSpace.

After this process, we were also able to upload the individual files into DSpace so that they a user can navigate the webpage inside DSpace. There was some confusion about how to do this, as the original folder structure resulting from wget had a nested folder, which is not supported in DSpace, but it turns out that HTML files in DSpace reference CSS and image files in a flat structure, that is, in a single DSpace object. When we created a DSpace item with bitstreams for the HTML files, the CSS file, and the image file, we were able to recreate the appearance of the website as it appears on the internet.

3. Ingest

As noted, the "School of Information OpenChoice Project Server" community consists of two collections: "OpenChoice Project Server Contents" and "OpenChoice Project Server Documentation." Ingest of the copies of each file directory for the servers will be initiated in DSpace under the "Server Contents" collection, to be completed over the Summer 2016

semester or the Fall 2016 semester, when Ford is available. The directory copies will be uploaded as a single item - a packaged folder labeled "open_choice." All supporting documentation will be submitted to the "Server Documentation" collection.

4. DSpace metadata creation/management

Despite our difficulties creating one whole disc image of the server, we decided to approach the description of OpenChoice as if it were one item. The first set of fields provided in the default registry for the OpenChoice collections on the iSchool's DSpace repository included "Authors," "Title," "Other Titles," "Date of Issue," "Publisher," "Citation," "Series/Report No.," "Identifiers," "Type," and "Language." We provided the following values for the OpenChoice collection OpenChoice Project Server Contents:

- Authors - "Turnbull, Don. Efron, Miles. Smith, Arro. Roy, Loriene. Jones, Clinton Duchicela, Dezbah. Ganesan, Manoj. Riddle, Prentiss. Sharma, Amit. Veeravagu, Laksman."

- Title - "OpenChoice Project Server Contents"

- Other Titles - N/A

- Date of Issue - Will be filled in when items can be ingested

- Publisher - N/A

- Citation - N/A

- Series/Report No. - N/A

- Identifiers - N/A

- Type - "Other"

- Language - "English (United States)"

The second set of fields provided in the default registry included "Subject keywords," "Abstract," "Sponsors," and "Description." We included the following values:

- Subject keywords - "OpenChoice, RAID, array, server, decommissioned"

- Abstract - "OpenChoice is a decommissioned RAID server that houses an IMLS grant-funded iSchool project of the same name. The project began in 2007 and was intended to produce an open source tool for communities to control their own web-filtering. The project was never completed. However, the documents, programs, and environment that the project was created with are preserved on this server. The server consists of 5 drives and runs on Ubuntu 6.06 "Dapper Drake" operating system. Documentation for the project can be found in the OpenChoice to the OpenChoice Project Server Documentation collection."

- Sponsors - N/A

- Description - N/A

For documentation about the project OpenChoice, we have provided our research and writings about the process of creating a disc image of OpenChoice to the OpenChoice Project Server Documentation collection. Due to the volume of documents, we have organized them by type and uploaded them in separate submissions. The types of documents include:

- Correspondence

- Images

- Publications

- Spring 2016 Report

- Reference Documents

- The Machine

The metadata fields for all of the OpenChoice Project Server Documentation collection are the same as the fields for the OpenChoice Project Server Contents collection. We have provided the following values for the first set of fields in the collection OpenChoice Project Server Documentation:

	Correspondence	Images	Publications	Spring 2016 Report	Reference Documents	The Machine
Authors	Bliss, David. Childress, Andrew. Efron, Miles. Smith, Arro. Turnbull, Don.	Bliss, David. Childress, Andrew.	Turnbull, Don.	Bliss, David. Childress, Andrew. Khan, Nicolette. Winograd, Amanda.	Ballou, Jullianne. Feldman, Nicole. Gasull, Christina. Gaylord, Lauren. Gilbreath, Richard. Hendricks, Toby. McCarther, Amye. Moloney, Justina. Moello, Raye. Rizkalla, Jim. Wallaca, Andrew.	Childress, Andrew.
Title	"Correspondence"	"Images"	"Publications"	"Spring 2016 Report"	"Reference Documents"	"Info About the Machine"
Other Titles	N/A	N/A	"HIDE & SEEK: The Information Architecture & Design for the OpenChoice filtering project," "Rating, Voting & Ranking: Designing for Collaboration & Consensus," and "OpenChoice: A Platform for Web Content Classification & Filtering"	"INF392K Final Report"	"From the Server to the Virtual Machine: Archiving the AERI 2013 Website for Preservation in a Trusted Digital Repository," "Business Case Plan: Vintage Computers in the DAL," "Preserving Old Servers: A Business Proposal"	N/A
Date of Issue	TBD	TBD	TBD	TBD	TBD	TBD

Publisher	N/A	N/A	N/A	N/A	N/A	N/A
Citation	N/A	N/A	N/A	N/A	N/A	N/A
Series/Report No.	N/A	N/A	N/A	N/A	N/A	N/A
Identifiers	N/A	N/A	N/A	N/A	N/A	N/A
Type	"Other"	"Image"	"Article"	"Working Paper"	"Working Paper"	"Other"
Language	English (United States)	N/A	English (United States)	English (United States)	English (United States)	English (United States)

The second set of fields provided in the default registry included "Subject keywords," "Abstract,"

"Sponsors," and "Description." We included the following values:

	Correspondence	Images	Publications	Spring 2016 Report	Reference Documents	The Machine
Subject keywords	Correspondence, OpenChoice	Images, OpenChoice	Publications, OpenChoice	Spring 2016 Report, OpenChoice	Reference Documents, OpenChoice	The Machine, OpenChoice
Abstract	Documentation for OpenChoice server project	Documentation for OpenChoice server project	Documentation for OpenChoice server project	Documentation for OpenChoice server project	Documentation for OpenChoice server project	Documentation for OpenChoice server project
Sponsors	N/A	N/A	IMLS	N/A	N/A	N/A
Description	N/A	N/A	N/A	N/A	N/A	N/A

Conclusions

Sam believes important work and data may already have been migrated elsewhere as

part of the IMLS grant compliance process, and as a result of the difficulties encountered by the

project team. The most significant contents of the server include project folders from teaching assistants that contained logos and design features for the OpenChoice blog, as well as downloads of Weka and Apache Nutch, which were presumably two base components of the OpenChoice content filter. Weka is a free suite of machine learning software written in Java and first developed at the University of Waikato, New Zealand in 1993. Apache Nutch is an open source web crawler developed by Doug Cutting and Mike Cafarella, first released in 2010. Given the dates of the OpenChoice project, we assume pilot or beta iterations of Nutch were downloaded.

The difficulties encountered during the disk imaging process delayed our project substantially. We first arrived in the Digital Archaeology Lab on April 5th with a hard drive which could not be mounted to have a disk image copied to it. The following week, with a blank and properly-formatted hard drive, Sam once again found he could not mount the disk to create an image, forcing him to use the *cp* command to copy each directory individually. This process took longer than disk imaging, since it required Sam to repeatedly step in and copy the next directory. It also required us to monitor the transfer, at least at the beginning, to make sure there were no problems copying the files. Once this process was completed, we were able to view the contents of the server in BitCurator, but we had less time to explore the contents and consider further action than if there had been no difficulty mounting the drives and creating an image.

Sam feels that, with more time, a greater effort might have been made to create an image. The effort was constrained by time, by the technical problems with mounting large drives, and by a shortage of PS/2 keyboards to use with the server. OpenChoice was backed up at the same time as the Herbie server, and with only one compatible keyboard between the two of them, inputting commands on both required repeatedly (delicately, given the age of the hardware) unplugging and plugging in the keyboard, a slight frustration and roadblock.

As stated before, we have had to postpone ingest of the directory copies until we have sufficient storage space. For the file upload, we may have to disable file size limit; however, the DSpace manual emphasizes that this does not guarantee successful upload of the file. We may encounter other major problems here depending on how slowly the file is uploaded and if we encounter further issues with Ford.

For the most part, we feel like we were successful in copying the server contents and related project information to contextualize the OpenChoice project. However, we also feel that our results were not adequate to serve as an example for future work, since there were several ways in which we were not able to follow the best practices we had identified at the beginning of the project. On a practical level, we think this may have been unavoidable in our case (and may prove to be so in others) — our problems with passwords, file systems, hardware compatibility, and time constraints meant that if we wanted to get results we had to violate some forensic guidelines for maintaining complete integrity of the data.

Luckily, our group had no ethics or privacy issues since almost all the information available on the server, including the names of teaching assistants in association with the project, was already public-facing. We also made every attempt to contact all individuals involved with the project to address any privacy concerns and collect further information about their contributions to the OpenChoice project. For our documentation, we have chosen the Creative Commons License "Attribution-NonCommercial CC BY-NC," which allows others to remix, tweak, and build upon our work non-commercially, and although their new works must also acknowledge us and be non-commercial, they don't have to license their derivative works on the same terms.

We would suggest that future work for Digital Archiving students could be to write a

specific workflow procedure for server decommissioning based on the various procedures, successes, and failures used by the three server groups. The assumed preservation standard for server decommissioning is a disk image of the entire server. However, as with our case, this is not always feasible. In some cases, a copy of the file directories will have to suffice. Some things to consider developing are a server decommissioning worksheet and specific DSpace registries for preservation metadata (or a preservation metadata form that would be uploaded with the disk image) to keep procedures and documentation thorough and consistent. We also need to designate ample dark storage space for these materials.

Appendices

Appendix A - "An aspirational schedule of work and proposed workflow, due by February 17"

Appendix B - "A formal report on management policy for the materials, due March 9 as part of your work on modeling stakeholders and workflow in DSpace (note this date has **not** been changed)."

Appendix C - "A formal report on workflow steps being achieved and how the workflow may have been modified when confronted with reality, due April 13."

Appendix A - "An aspirational schedule of work and proposed workflow, due by February 17"

Proposed Workflow for OpenChoice
1. Disk Imaging - David
    a. David will meet with Sam Burns to discuss disk imaging the server. We will need to determine the following:
        i.   What we are responsible for in terms of disk imaging: will we be doing this ourselves, or helping Sam do it?
        ii. Whether the server is a RAID, and whether it is a software or hardware RAID configuration.
        iii. What is the actual process of imaging? Andrew suggests that the -dd command would be the most natural choice for imaging the disk. Andrew suspects booting to a live disk will allow us to do this.
        iv. Another option may be to use a FRED to analyze and image the server. This may be more complicated and unnecessary, however, if the disk can be imaged more easily via a live disk.
    b. We should be able to analyze the disk image using software like BitCurator. David will investigate whether the DAL has a BitCurator machine.
    c. This work will rely on the availability of Sam Burns. David will reach out to Sam Burns and hopefully begin the imaging process in the next few weeks.
2. Web archiving - Andrew
    a. Andrew has experience with web archiving and will investigate the OpenChoice site structure to learn how to capture its contents, which we would like to preserve.
    b. This is not dependent on other work steps and can begin immediately.
3. Metadata - Amanda
    a. Once we have an idea of what is on the server, Amanda will devise a metadata scheme for the individual files we would like to ingest into DSpace.
    b. This cannot be done until the disk imaging process is completed.
4. Ingest - Nicolette
    a. Nicolette will handle the ingest of materials into DSpace. Nicolette will also identify which files should be preserved and uploaded individually.
    b. This cannot be done until the disk imaging process is completed, and metadata has been decided upon.
5. Background investigation - Shared
    a. We will correspond with OpenChoice team members and search for publications on the project to gather background documentation which we can archive alongside the server contents.

Appendix B - "A formal report on management policy for the materials, due March 9 as part of your work on modeling stakeholders and workflow in DSpace (note this date has **not** been changed)."

OpenChoice Server Project: Management Report

9 March 2016
INF 392K
David Bliss
Andrew Childress
Nicolette Khan
Amanda Winograd

"Each team will describe briefly and present schematically its proposed management policy for its designated community as a part of the topic discussion this week. Teams will have consulted the initial proposed policy document on pacer on overall policy for the iSchool repository as a context for their policy development. Use the document as a list of things that your handling of your particular materials might need to consider, as (for example) any privacy considerations or renegotiation if necessary of terms for materials for a project that continues a previous one. Especially, **work out how you see the structure around communities, subcommunities, collections, items, and bitstreams; as well as what materials you plan to ingest and how you propose to ingest them**--if you will be doing manual ingest, how you propose to set up any workflow you wish to use. The purpose here is to prepare for setting up the structure to receive your archival collections in DSpace and to have a plan for their maintenance."

Workflow / Roles

Create disk image: David
DSpace metadata creation/management: Amanda
Web archiving (https://www.ischool.utexas.edu/~choice/index.html): Andrew
Ingest into DSpace: Nicolette

Communities

Description of Community:

Documentation materials will fall into the Digital Archives Projects community. This community is described as "class projects undertaken by students at the School of Information" and includes "documentation of student projects whose archived products reside elsewhere."

Open Choice materials (disk image and website) might be better suited to the School of Information Faculty community, but it is unclear if we have the authority to create an Open Choice subcommunity here.

Description of Subcommunity: School of Information Technology Services
https://ford.ischool.utexas.edu/handle/2081/334
The subcommunity will be Open Choice Project. (?)

Community Expectations:

Website Archiving - Andrew

Goal: create an archival copy of https://www.ischool.utexas.edu/~choice/.

At this point, not having researched previous used practices for web archiving in the iSchool repository, I think the way to do this will be to upload to the repository a folder titled "~choice" (or a tar of said folder), obtained using wget with the --mirror command. The ~choice website is functionally separate from the rest of the iSchool website in that it does not share resources like external stylesheets or databases with any other pages, so everything needed to accurately render the page is in the ~choice folder as downloaded with wget. (This assumption is based on the fact that the pages downloaded with wget looked identical to those online when opened locally.)

One potential concern is that this method might not produce a perfectly accurate record of the website as it is on the server. Darnall et al seem to have addressed this concern by using HTTrack instead of wget. Further investigation is needed to determine whether HTTrack gives significant advantages. Another concern is metadata: how should we handle metadata for a small archived website? Following the report by Darnall et al, we should tar the folder and ingest it as single file with manually entered metadata. For this we would follow the steps under '"Describe" (item-level metadata)' in the Ingest section of this report. (And, of course, the rest of the Ingest process would also be followed.)

See also: https://ford.ischool.utexas.edu/handle/2081/30674?mode=full. DSpace does have HTML support. The DSpace documentation in section 4.3.10 notes that "DSpace can store and provide on-line browsing capability for self-contained, non-dynamic HTML documents". As the ~choice website is a collection of such documents, there should be no problems.

Subcommunity: Open Choice Project.

Metadata - Amanda

Although we do not know at this point what all of the qualifiers will be for our collection, we have looked at recent submissions to the Digital Archives Projects community as well as DSpace documentation and can speculate about certain DublinCore elements that will likely be necessary:

1. dc.contributor.author
    a. Efron, Miles
    b. Roy, Loriene
    c. Smith, Arro
    d. Turnbull, Don
2. dc.contributor.other
    a. Bliss, David
    b. Childress, Andrew
    c. Khan, Nicolette
    d. Winograd, Amanda
3. dc.date.accessioned
4. dc.date.available
5. dc.date.copyright
6. dc.date.created
7. dc.date.issued
8. dc.identifier.uri
9. dc.description.abstract
    a. This file is a disc image of Open Choice, a RAID array server.
10. dc.format.extent
11. dc.format.mimetype
12. dc.language.iso
    a. en_US
13. dc.subject
    a. server?
14. dc.title
15. dc.type

       a. disc image

Disk Imaging - David

Without further research, it is too early to lay out a specific plan for imaging the server disk. Two possible approaches are using a FRED, or booting the server from a live CD. The first step will be to talk to Sam Burns about these and other options, and to ascertain what imaging options are available for RAID arrays like OpenChoice.

It should be possible to boot from a live CD and use dd to copy the disk image onto an external hard drive. If this is possible, it would probably be the easiest route. We have begun a conversation with Don Turnbull, who supervised the OpenChoice project, and he may be able to provide some information on the server itself if Sam Burns is unable to answer all our questions.

Resources (possibly) needed:
- Large blank external hard drive
- Linux live boot CD
  - May need to create one for Ubuntu ~12.x

Ingest - Nicolette

The default item submission process for DSpace involves six steps:

1. "Select Collection"
2. "Describe" (item-level metadata)
3. "Upload" (described below)
4. "Review"
5. "License" (agreeing to repository distribution license)
6. "Complete" (item will either be available immediately or it may go through a "workflow approval process" - we will have to look into our collection policy)

We may also include the additional CC License step for materials generated by our group.

DSpace provides a submission configuration file in xml which can be used to define or edit submission steps. For the disk image upload, we may have to disable file size limit; however, the DSpace manual emphasizes that this does not guarantee successful upload of the file. We may encounter some major problems here since we have had major issues with the Ford server lately.

To ingest packages, we can use submit/ingest mode and use the command provided in the DSpace manual. We do not plan on ingesting multiple packages at once.

Appendix C - "A formal report on workflow steps being achieved and how the workflow may have been modified when confronted with reality, due April 13."

Open Choice Server Revised Workflow

Disk Imaging
David completed the server contents copying process on 8 April 2016. The contents are currently stored on a drive borrowed from Sam Burns.

Exploring Disk Image
We are hoping to meet in the DAL on 17 April 2016 to explore the server disk image on the BitCurator station.

Ingest
TBA - When the new server is up.

Metadata Description:
Amanda is meeting with the other metadata point people after class on 13 April 2016 to discuss consistent metadata description for server data.

Nicolette is drafting Scope and Contents and contributor bios.

Final Report
Amanda has created an outline of the project overview.

Completion
All tasks completed by 27 April 2016.