

INF392K: Problems in the Permanent Retention of Electronic Records

Schmandt-Besserat (Denise) Papers

Digital Archiving Project Report

Javier Ruedas, Mark Firmin, Meredith Bush
Spring 2011

Contents

Contents.....	1
Project Overview (Mark Firmin).....	2
Biographical Sketch.....	2
Scope and Content.....	3
Inventory of Digital Collection	3
Nine-track magnetic tapes: A non-invasive investigation of their contents (Javier Ruedas)	4
DSpace File Structure and Mapping (Meredith Bush).....	9
Pacer Structure	10
UTDR Structure	11
Processing 3.5 inch diskettes for batch ingest (Javier Ruedas)	13
Nonobtrusive investigation	13
Disk Imaging (Mark Firmin)	13
Disk Image Analysis and File Extraction	14
Decision to manually construct file inventory	15
The batch ingest structure	15
Constructing xml metadata files	16
Contents files.....	17
Batch ingest.....	17
Observations on identifying dates of creation and file formats.....	18
Analysis and Recommendations (Meredith Bush, Mark Firmin, & Javier Ruedas)	28

Project Overview (Mark Firmin)

The Denise Schmandt-Besserat papers are a collection of both print and digital material held at the Dolph Briscoe Center for American History (CAH) at the University of Texas at Austin. The digital contents of this archive were identified for processing by the CAH digital archivist, Zachary Vowell, and a University of Texas School of Information archives professor, Patricia Galloway. Efforts to archive the collection, for access and preservation purposes, were a product of a class project for INF392K: Problems in the Permanent Retention of Electronic Records. The digital component of the archive is held in the University of Texas School of Information DSpace repository and will subsequently be transferred by CAH staff to the University of Texas Digital Repository (UTDR).

This report describes the workflow, decisions, and recommendations for future work.

Biographical Sketch

Denise Schmandt-Besserat is a professor emerita of Art and Middle Eastern Studies at the University of Texas at Austin. Dr. Schmandt-Besserat studies the art and archaeology of the ancient Near East. Her research has focused on the origin of writing and counting, and her published works include: *Before Writing* (1992), *How Writing Came About* (1996), *The History of Counting* (1999), and *When Writing Met Art* (2007). Dr. Schmandt-Besserat trained at the Ecole du Louvre in Paris, France. She received an honorary doctorate from Kenyon College in 2008.

Dr. Schmandt-Besserat began her academic career at Harvard's Peabody Museum in 1965. She received a Radcliffe fellowship in 1969, and her first grant to travel to the Middle East in 1971. Schmandt-Besserat intended to study clay pottery from the Near East created before 6000 BC (Neolithic Era). During her trip, Schmandt-Besserat encountered the clay tokens for the first time. Primarily kept in cigarette and cigar boxes, these small artifacts had largely been ignored by the archaeological community. The discovery of tokens inside a sealed clay envelope dating from circa 3000 BC by one of Schmandt-Besserat's professors from the Ecole de Louvre, led her to develop the theory that the tokens were part of a recording system that stretched back to the Neolithic Era. Schmandt-Besserat posited that this recording system spanned several thousand years and across the ancient Near East before rapidly evolving into complex tokens contained in sealed envelopes around 3500 BC, and eventually into writing. Put forward in *Before Writing*, Schmandt-Besserat's hypothesis pushed back humanity's understanding of the development of writing by 5000 years. Schmandt-Besserat refined her theory and published her second work, *How Writing Came About*. Her efforts resulted in *How Writing Came About* being named by the *American Scientist* as one of a hundred texts to shape science in the twentieth century. Based on her work at the Ain Ghazal excavation site in Jordan, Schmandt-Besserat has written a number of articles and delivered several presentations concerning her study of statuary, figurines, art, and tokens.

Scope and Content

Research materials, manuscript drafts, page proofs, and editorial materials document the creation of Denise Schmandt-Besserat's published works, including *Before Writing* (1992), *How Writing Came About* (1996), and scholarly articles. Also included in the collection are teaching materials, and research materials that document Schmandt-Besserat's scholarly activities, research, teaching, lectures, and travel. Of particular importance, are hundreds of pages of computer printouts from five, nine-track magnetic tapes containing the data pertaining to tokens and envelopes that is central to Schmandt-Besserat's ground-breaking study, *Before Writing*. The data from these printouts were the result of student assistants crafting FORTRAN statements derived from a coding scheme developed and refined by Schmandt-Besserat.

FORTRAN statements were keyed into a CDC 6600 mainframe computer housed at the UT Computation Center. The data was processed using SPSS (Statistical Package for the Social Sciences). The results were printed out and the five magnetic tapes were housed at the climate-controlled UT Computation Center till 1991, when they were transferred to Schmandt-Besserat's custody. Schmandt-Besserat stored the magnetic tapes in a shop with no climate control before they were transferred to the Briscoe Center for American History (CAH) in July 1992.

A collection of seventy-six, 3.5-inch floppy diskettes contain more than five hundred digital files. The diskettes contain drafts of *Before Writing*, articles, presentations, lectures, course materials, personal materials, spreadsheets, and materials pertaining to Schmandt-Besserat's work at the Ain Ghazal excavation site in Jordan. The bulk of the files are created in the Microsoft Word .doc format. Other file extensions, include DAT, BAK, DBF, DBH, EXE, and xls. The files were created and/or modified between 1989 and 2000. Much of the work to trace the provenance of this collection is a result of an oral history interview with Schmandt-Besserat, research into Schmandt-Besserat's numerous grant applications, and through perusing the control file to the Denise Schmandt-Besserat papers housed at the CAH.

Inventory of Digital Collection

Two boxes containing the seventy-six diskettes and five mainframe tapes (and some additional paper material) were obtained from the CAH in February 2011 and relocated to the UT iSchool Digital Archaeology Laboratory. An initial inventory of the physical media was conducted in the preservation lab. Each disk and tape was digitally photographed and all visible information (label text, color, brand, etc.) was recorded in a spreadsheet (DSB Inventory in the DSpace repository).

Nine-track magnetic tapes: A non-invasive investigation of their contents (Javier Ruedas)

The Schmandt-Besserat papers at the Briscoe Center for American History include five nine-track magnetic tapes, donated by Schmandt-Besserat in 1992 (figure 1). Although we were not able to recover the data on these tapes during the time allotted for our project (January to May, 2011), we carried out archival research aimed at elucidating their contents. We also interviewed the donor, Denise Schmandt-Besserat, and exchanged electronic correspondence with Solveig Turpin, one of the computer operators employed by Schmandt-Besserat. Based on our research, we can safely assert that these tapes contain Schmandt-Besserat's coded descriptions of approximately ten thousand small Middle Eastern ceramic objects called tokens, part of the research forming the basis for her two-volume opus, *Before Writing* (1992).

Denise Schmandt-Besserat developed a descriptive schema for ancient Middle Eastern tokens, which is presented in volumes 1 and 2 of *Before Writing*. The heart of her schema consisted of Type and Subtype attributes. Types included shapes such as cones, disks, and spheres. Each type was divided into multiple subtypes; for example, spheres could be large, incised, punched, or pinched, among other subtypes. Each type and subtype was assigned a number. In addition, she described tokens according to the country and site where they were discovered, the millennium they were dated to, their size, any markings found on them, and the material they are made out of (usually clay, but not always). All values for these attributes were assigned a number; for example, Iraq was country 6, Iran was country 7, and so forth, with each site within each country also assigned a number (figure 2).

Schmandt-Besserat visited over forty museums, in the United States, Europe, the Middle East, and Japan, to personally inspect and describe the tokens. Other tokens were described based on published reports. She described approximately ten thousand tokens during her research project, of which some seven thousand are catalogued in volume 2 of *Before Writing*. For each token, she also recorded, where possible, its museum location and number, or the site report and item number. Museum locations and bibliographical references were also assigned numbers.

The Schmandt-Besserat papers include archaeological data sheets representing descriptions of the tokens. These data sheets represent the first step in producing the data that eventually ended up on the computer tapes and, as printouts, in volume 2 of *Before Writing*. Each sheet describes one token in terms of Schmandt-Besserat's selected descriptive attributes. As described above, every value for each descriptive attribute was given a numerical code. Once each token was described and each value was assigned a number, the token descriptions were entered into Fortran coding forms (figure 3). After coding for computer entry, Schmandt-Besserat employed graduate students to enter the data on the University of Texas's CDC 6600 mainframe computers. The data were, in part, analyzed with SPSS (Statistical Package for the

Social Sciences). To produce volume 2 of *Before Writing*, Schmandt-Besserat requested that the tokens be sorted by site. The results of this analysis were printed out (figure 4). The printouts, stored in their entirety along with the rest of Schmandt-Besserat's papers at the Briscoe Center for American History, were published as the companion volume of *Before Writing*, a catalog of tokens arranged by country and site.

Schmandt-Besserat left behind enough information in her papers to decipher much of the code. A partial reading of a coded sheet is offered in figure 2. Further research in her papers and publications would likely yield enough clues to understand the majority of the coded sheets. It is likely that the tapes contain the computer-coded data on tokens that Schmandt-Besserat printed out as the basis for her catalog.

Figure 1



Above: One of the computer tapes in the Schmandt-Besserat Papers.

Below: Tape label showing the date and the name of the computer operator, and confirming that the tape contains data on tokens.

UNIVERSITY OF TEXAS AT AUSTIN COMPUTATION CENTER		
User Number	MYAE 623	Track
User Name	SOLVEIG TURPIN	9
Date	7-16-79	7
ID Number	F030	BPI
Reel Number	X 7457	800
	TOKENS	556
		200
		6250
		1600

Figure 2

Sample of Schmandt-Besserat's descriptive schema for tokens, with numeric codes assigned for each value. Source: Schmandt-Besserat Papers, Briscoe Center for American History.

IRAQ 06	IRAN 07
001 Tello	001 Susa
002 Tepe Gawra	002 Anau
003 Uruk	003 Choga Mish
004 Eridu	004 Tall-i-Bakun
005 Fara	005 Choga Mami
006 Larsa	006 Chaga Sefid
007 Arpachiyah	007 Djaffarabad
008 Babylon	008 Tepe Asiab
009 Ubaid	009 R.37 - Tepe Gaz Tavila
010 Ur	010 Sorkh-i-Dom
011 Billa	011 Tulai
012 Jarmo	012 Zagheh
013 Tell Asmar	013 Sarafabad
014 Khafaje	014 Tepe Sarab
015 Tell Agrab	015 Seh Gabi
016 Nineveh	016 KS.76
017 Nippur	017 KS.34
018 Uqair	018 KS.54 Abu Fandaweh
019 Hassuna	019 Ganj-Dareh
020 Karim Shahr	020 Yahya

Figure 3

Tepe Gawra - Baghdad Museum
FORTRAN CODING FORM *discs.*

PROGRAM _____ DATE *Dec 5. 1984* PUNCHING INSTRUCTIONS _____

PROGRAMMER _____

COMP	STATEMENT NUMBER	CONT	FORTRAN STATEMENT
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29	30	31	32
33	34	35	36
37	38	39	40
41	42	43	44
45	46	47	48
49	50	51	52
53	54	55	56
57	58	59	60
61	62	63	64
65	66	67	68
69	70	71	72
73	74	75	76
77	78	79	80
81	82	83	84
85	86	87	88
89	90	91	92
93	94	95	96
97	98	99	100

0600202001 -- 2604 . 1
0600202030 -- 1604 . 1
0600202030 -- 2816 . 1
0600202030 -- 2607 . 1
0600202030 -- 1611 . 1
0600202030 -- 1416 . 1
0600202001 -- 1510 . 1
0600202001 -- 1708 . 1
0600202001 -- 1207 . 1
0600202030 -- 2708 . 1
0600202030 -- 1008 . 1

Above: A Fortran coding form containing a description of tokens, to be entered into the University of Texas mainframe computers.

Below: the key to the code. The tokens described are from country 6, site 2 (Tepe Gawra, Iraq); they are of type 2 and subtypes 1 and 30. Their diameter ranges from 10 to 28 mm and their thickness from 4 to 16 mm. They are all plain.

ORIGIN	SHAPE	SIZE	SIGNS	MATERIAL
country	site	type	subtype	
		length/height		
		diam/width		
		thickness		
		notched		
		plain		
		1 punch		
		multiple punch		
		1 incision		
		multiple incisions		
		incised/punched		
		pointed		
		aplique		
		pinched		
		pattern		
		microrough		
		perforated		
		sealings		
		slanted		
		broken		
		stone		
		bitumen		
		ochre		
		plaster		
		ivory/bone		
		calciferous		
		unspecified		

Figure 4

Printout of data on tokens.

TYPE 1 [CONE]		SUBTYPE 11 [CARINATED]			
TOKEN NUMBER	SIZE IN MM L. W. TH.	MAT.	PERF.	MILL. LEVEL	MUS. FIELD/MUSEUM NUMBER
IRAN JEITUN					
1	20 X 24			6	HE
2				6	HE
3				6	HE
4				6	HE
5				6	HE
6				6	HE
7				6	HE
8	29 X 25			6	HE 2203-163
9	30 X 25			6	HE 2670-15
10	18 X 30			6	HE 2670-17
11	20 X 22			6	HE 2670-178
12	36 X 18			6	HE 2670-31
13	17 X 27			6	HE 2670-85
14	17 X 22			6	HE 2670-87
15	23 X 20			6	HE 2670-88
16	15 X 25			6	HE 2670-96
17	17 X 25			6	HE 2703-103
18	20 X 30			6	HE 2703-141
19	22 X 25			6	HE 2703-143
20	12 X 16			6	HE 2703-154
21	17 X 20			6	HE 2703-154
22	20 X 28			6	HE 2703-163
23	15 X			6	HE 2703-163
24	23 X 28			6	HE 2703-163
25	22 X			6	HE 2703-167
26	18 X 21			6	HE 2703-61
27	20 X 20			6	HE 2703-93
28	24 X 28			6	HE 2703-94
TOTALS: COUNTRY = IRAN					
				28	STONE 0 PERFORATED 0
TOTALS: SUBTYPE = 11					
				28	STONE 0 PERFORATED 0

In the oral history interview, Schmandt-Besserat stated that she had to submit requests for analyses and printouts of data to the computing center. An employee or a graduate student operated the computer. To fund these analyses, Schmandt-Besserat wrote many grant proposals, now on file at the Briscoe Center for American History. After requesting her data analyses and summaries, Schmandt-Besserat waited several days for notice until her request could be carried out. She then visited the computing center in person to pick up the printouts, now also on file at the Briscoe Center. For her catalog of tokens, Schmandt-Besserat organized the tokens by country and site.

DSpace File Structure and Mapping (Meredith Bush)

An archive, regardless of the state of its contents, should be designed according to the accepted practices. In the contemporary archival community, this means systems are designed to maintain original order and preserve provenance. The physical contents of the Schmandt-Besserat papers (including the physical media storing the digital content addressed in this project) have already been accessioned and archived by the Briscoe Center for American History. Thus the digital equivalent was designed to parallel that structure. Original order was

inferred from the order in the boxes as received. Information regarding provenance was derived from (in order of preference) the main archive, the metadata extracted from the digital files, the labels on the physical media, and the text contained within the digital files.

The structure was created in several stages. First we discussed the theoretical implications of the structure and created a schematic diagram of our ideal structure, which we presented to the class. Next, we began to structure the communities, subcommunities and collections within DSpace. As we processed (imaged, scanned, etc.) the digital files, we then had to translate these structures into a computer file/directory system. This file system went through several metamorphoses before it ended up on the Vauxhall (vauxhall.ischool.utexas.edu) for the batch ingest. The final stage was to merge the individual bitstreams (and corresponding metadata) from Vauxhall into the DSpace structure that had previously been created.

Pacer Structure

The two types of media (diskettes and mainframe tapes) were found in completely separate boxes in the original collection, so we knew that this separation must be maintained within the DSpace repository. Within DSpace, these vastly different physical media were represented as two top-level subcommunities in DSpace, "Schmandt-Besserat Diskettes" and "Schmandt-Besserat Mainframe Tapes".

The "Schmandt-Besserat Mainframe Tapes" subcommunity was created but not arranged any further. The next group that works with the tapes will have a subcommunity to structure according to that project. We had no way to speculate about the contents of the tapes and possible associated material.

The "Schmandt-Besserat Diskettes" subcommunity required several additional levels of structure. All the diskettes were stored in a single archival box, but within the box were 5 disk containers as well as a suite of disks not stored in any container. In order to retain the original order of the archival box, this level was designated by group number. This was largely due to the constraints of DSpace, which would have structured the subcommunities in alphabetical order (green box, red box, yellow box) followed by the unboxed disks. Instead, the structure contains subcommunities: "Group 1: Green Disk Container," "Group 2: Red Disk Container," "Group 3: Yellow Disk Container," "Group 4: Clear Disk Container," "Group 5: No Container," "Group 6: 3M Imation Box," "Group 7: No Container."

During the inventory of the physical media, new identification numbers were assigned to the objects that reflected the original order in the archival boxes (i.e. SB001 was immediately followed by SB002, etc.). These identification numbers were maintained throughout the project, and are used for naming of DSpace collections within Pacer. These collections were manually created by the DSpace subcommunity administrator. All smaller (item and bitstream level) structures were created within the Vauxhall server and uploaded to the corresponding DSpace

collection through the batch ingest process.

Structure of iSchool DSpace Repository

Community: Briscoe Center

Subcommunity: Schmandt-Besserat (Denise) Papers

Subcommunity: Schmandt-Besserat Diskettes

Subcommunity: Group 1

Collection: Disk SB001

Item: Disk Image

Bitstreams: Image, Checksums, Logs

Item: Extracted File 1

Bitstream: File

Collection: Disk SB002

Item: Disk Image

Bitstreams: Image, Checksums, Logs

Item: Extracted File 1 (etc.)

Bitstreams: File

UTDR Structure

The final destination of this collection is in the University of Texas Digital Repository (UTDR), where all digital components of the CAH collections are currently managed. This is another DSpace system, but is controlled through a different management team and has different standards for structure and metadata than the iSchool DSpace repository.

The UTDR structure, while not implemented in this project, is compatible with the existing structure. All necessary metadata is already created. Some metadata in the existing system will be lost in the transfer to the UTDR.

Structure of UTDR DSpace Repository

Community: Schmandt-Besserat (Denise) Papers

Collection: Disk 1

Item: Contents of Disk 1

Bitstreams: Image, Checksums, Logs, Extracted Files

Collection: Disk 2

Item: Contents of Disk 2

Bitstreams: Image, Checksums, Logs, Extracted Files

Processing 3.5 inch diskettes for batch ingest (Javier Ruedas)

The Schmandt-Besserat Papers at the Briscoe Center for American History included a box containing seventy-six, 3.5-inch disks, donated by Schmandt-Besserat in 2003. These disks contained much of her professional work—scholarly research and writing as well as teaching—from the time she obtained her first personal computer in the late 1980s until her retirement in 2001. Among our group’s objectives was to safely extract the contents of the disks and to ingest these into the iSchool’s DSpace repository.

Nonobtrusive investigation

We first gathered evidence nonobtrusively, by taking photographs of the disks and the disk boxes. Almost all the disks had labels. At the start of our work, the labels meant little to us, but after interviewing Denise Schmandt-Besserat and familiarizing ourselves with her work, they began to make more sense. Until 1992, Schmandt-Besserat was focused on the publication of her multi-volume opus, *Before Writing*. Afterwards, she concentrated on analyzing smaller sets of archaeological materials from the site of Ain Ghazal, a Neolithic site in Jordan. During the entire time span represented by the diskettes, she also taught Art History classes. The disks document these three main lines of effort, though they also contain many smaller projects such as book reviews and encyclopedia articles.

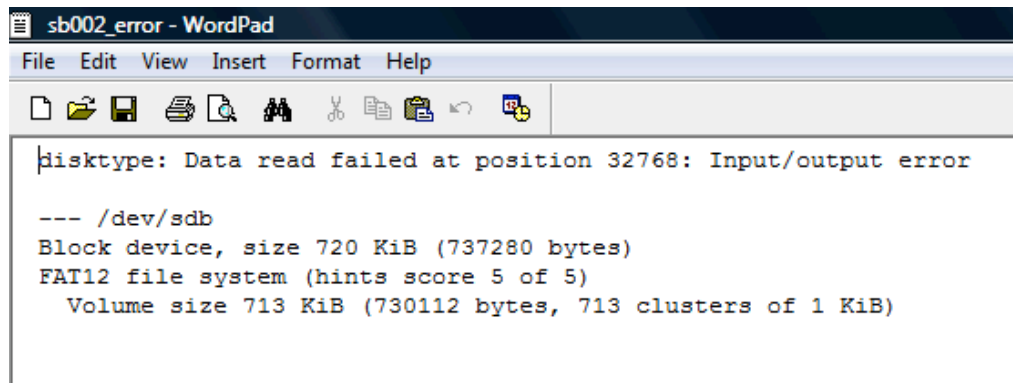
We took photographs of all disks and disk boxes (see figure 5). We then used a batch procedure in Adobe Photoshop to resize all photographs to a width of 440 pixels. Once the subcommunity and collection structure was ready in DSpace, we uploaded the photos as logos for the collections and, in the case of the disk box photos, for the subcommunities.

Disk Imaging (Mark Firmin)

To create images of the seventy-six diskettes, we opted to use Dracula and FRED, the 3.5-inch USB peripheral drive. Each diskette was physically write-protected before being inserted into FRED. To familiarize ourselves with the command-line procedures established by the Frankenstein II group, each one of us, using diskettes provided by Javier, conducted a trial run. The disk imaging process encountered errors with diskettes SB002, SB004, SB047, and SB048. In the latter two instances, FRED began making strange noises immediately after these diskettes were inserted. Upon entering the `sudo disktype` command, a “disk read failure” message appeared for both diskettes. A reference of the inventory revealed that these were the only 3M, 2 MB diskettes in the collection, which led us to deduce that the error may simply have been caused due to a defect with this particular type of diskette. Disk SB047 contained a draft of volume one of *Before Writing*, while SB048 contained a draft of volume two. Both diskettes were labeled March 1, 1990. Since *Before Writing* was published in 1992, it would seem logical to conclude that these two diskettes represent the draft Schmandt-Besserat submitted to the University of Texas Press and prior to the beginning of the copy-editing process. Thus, these

two diskettes may represent Schmandt-Besserat's final, unedited manuscript of *Before Writing*. Error reports for the two diskettes were filed in the directory. Some diskettes contained corrupted files. Consequently, complete images of a few diskettes were not obtained. These instances were noted on the checksum sheet we created and in text-files placed in the directory. Overall, the imaging process proceeded smoothly and with few surprises.

Example
of a Disk
Error Log
Report



```
sb002_error - WordPad
File Edit View Insert Format Help

Disktype: Data read failed at position 32768: Input/output error

--- /dev/sdb
Block device, size 720 KiB (737280 bytes)
FAT12 file system (hints score 5 of 5)
Volume size 713 KiB (730112 bytes, 713 clusters of 1 KiB)
```

Disk Image Analysis and File Extraction

Our second task was to extract the digital data from the disks. We used equipment in the iSchool Digital Archaeology Lab (DAL) and procedures developed by the student teams responsible for managing the DAL equipment. We first used procedures aimed at obtaining an image of the disk contents, or disk image, without in any way altering the data that were physically inscribed on the disk media. These procedures are described in more detail in the report by the Spring 2011 DAL team, located in the iSchool Institutional Repository Documentation DSpace community.

We connected a 3.5-inch disk drive to the DAL computer nicknamed "Dracula." The disks all had a write-protect tab, all of which were switched "off" to allow writing to the disk. In order to avoid altering the data on the disks, we manually switched the tabs to write-protect, then switched them back to their original condition after imaging. We then ran a **disktype** command through the command line to determine the file system type. Each disktype command resulted in a log file containing the command's results (see figure 6). After this, we created an image of the original disk by using the disk dump (**dcfldd**) command. We

simultaneously calculated the hash value, or checksum, of the original physical disk and output the result as a text file (figure 6). We then produced a checksum for the disk image that we had made on the DAL computer, and compared the two to verify that the data on the disk image were identical to the original disk.

After making the disk images and the checksums, we made a working copy of the disk images. This allowed us to leave the original images, with their authenticity confirmed by the checksums, untouched. We would use the working images to extract individual files, inspect their contents, and produce our file inventory. In order to verify the authenticity of the working copy disk images, we ran **hash** commands on a random sample of these images. Every checksum in our random sample of working images was identical to the checksums of the original disk images, which in turn were identical to the original disks. This allowed us to state with confidence that the files we extracted were authentic copies of the files on the original disks.

We then checked the disk images for viruses using ClamAV for Linux. In the collection of seventy-six disks, we found two infected disks. In these disks, all files were infected. The infected disk images were not ingested into the iSchool repository.

Finally, we ran an **fls** command on the original disk images (not the working copies) to list the files within the disk images, along with any information on date of creation and last modification that might have been preserved (figure 7). We would later use the file names and dates produced by the **fls** command as the basis for our xml metadata files.

Decision to manually construct file inventory

Once we had authenticated working copies of the disk contents, we carried out a file inventory. The complete inventory accompanies this report as part of our project documentation. We explored techniques for importing tabular text files—such as those produced by the **fls** command—into a spreadsheet program. However, the resulting spreadsheets required some manual corrections and the procedure took more time for each disk image than would be spent entering the information manually. We constructed our inventory manually by mounting the working disk images on Apple computers, visually inspecting the file names and dates, and typing these into a Google Docs spreadsheet. We then visually inspected the contents of the files in order to produce sentence summaries of the individual files, which became scope and content notes for the extracted files in DSpace.

The batch ingest structure

To prepare the disk images and extracted files for batch ingest into the iSchool institutional repository, we constructed a directory structure that matched the structure we had previously created in DSpace. In DSpace, each disk was represented as a “collection.” Each disk

image and extracted file was represented as an item within the collection. For each disk/collection, we created a directory. Within each disk/collection folder, we created as many subdirectories as there would be items within the collection. Each item subdirectory was labeled item_001, item_002, and so on sequentially. We created our directory in the iSchool's "Vauxhall" server, which is used for export to the institutional repository's server, "Pacer."

After creating the batch ingest directory structure, we populated it with the appropriate files. For each disk/collection, there were two types of items: the disk images and the files that were extracted from the disk images. Disk images were always the first item within each collection, and hence were always placed in item_001 subdirectories for each disk/collection. In addition, all the files describing the disk image file type, checksums, virus scans, and file listings were placed in the item_001 subdirectories, accompanying the disk images. Extracted files were placed in the sequentially numbered subdirectories (figure 8).

Constructing xml metadata files

Each item ingested into DSpace must be accompanied by an xml metadata file that describes the item. For batch ingest, each item-level directory had to contain an xml file named, always, dublin_core.xml. To construct our xml files, we consulted existing course resources and prior student teams' reports, as well as consulting with a current iSchool doctoral student (Sarah Kim), the Briscoe Center for American History digital archivist (Zach Vowell), and Professor Patricia Galloway. Our solutions were a best effort to describe the items accurately, to satisfy the needs of the destination repositories, and to select a strategy that best matched our specific project and our team's computer skills.

Our team first examined the available metadata harvesters, such as JHOVE, Droid, the New Zealand Metadata Extractor, and the meta-metadata extractor, FITS. Of all these, FITS was clearly the most useful since it incorporates all the others and generates a metadata file that includes results produced by several harvesters, each of which has its strengths and weaknesses. Sarah Kim had previously developed a script that automatically ran FITS on each item and output the results as an xml file. However, when we ran FITS on our files, we found a large quantity of spurious information, or information that we judged useless.

Figure 9 shows an example of metadata produced by applying the FITS metadata extractor to one of the files extracted from our disks. The portion shown was produced by the Droid metadata extractor, which is supposed to identify the file format. This would have been interesting because the earlier extracted files with .doc extensions do not open correctly in versions of Microsoft Word that are currently available at the time of this writing (May 2011). It would have been useful to know precisely what version of Word produced these earlier .doc files (this problem does not occur in .doc files from the late 1990s onwards). But Droid could only offer a sequence of tentative identifications, which appear to be a list of practically every word processor in mainstream use for the past twenty-five years. This is not useful information, and we judged that it should not be ingested into the institutional repository since it did not

describe the extracted files. If FITS could not identify the file type correctly, we also had our doubts as to the validity of the dates of creation and last modification that it proposed for the disk images and extracted files. Because of these doubts, we decided to use the file names and dates produced by the **fls** command (figure 7) to fill the required metadata fields for title and date.

Our xml templates incorporated required elements for both the iSchool repository and the University of Texas Digital Repository (UTDR). Since most of the files on the disks did not have any associated dates of creation, we used “unknown” in the date.created field. We added a date.modified field, where we placed the date of last modification if this was known. In most cases, files did have a date of last modification listed through the **fls** command. In some cases, however, the dates of last modification listed through the **fls** command were clearly spurious; for example, some dates were in 1983, when we know through our interviews that Denise Schmandt-Besserat did not own a computer and hence could not have produced the files in question. If we were absolutely certain that these dates were impossible, we entered “unknown” for date of last modification as well as date of creation.

After consulting with project stakeholders, we settled on two metadata templates, one for disk images and one for extracted files (figure 10). Since the xml files in item_001 subdirectories had to describe the entire item—including the imaging and virus scan logs, the checksums, and the file listing—we added a description.abstract field in their xml files to describe these extra contents. The template for xml files to describe extracted files was simpler because it did not require an abstract. After creating the template, we used our **fls** text files to copy and paste the titles and dates into the appropriate fields, then uploaded each file to the correct folder in our ingest structure in the Vauxhall server.

Contents files

In addition to an xml file, every item subdirectory for DSpace ingest must contain a contents file. This is a simple text file, which must be saved without a file extension. Different repositories have different policies for constructing contents files. For the iSchool institutional repository, the contents files should only list the files within each item, with each listing separated by a line break. UTDR contents files are more complicated because each file must be placed in a bundle, with the file name and its corresponding bundle separated by a tab, and each file-bundle combination separated by a tab or a line break. We followed iSchool repository procedures and produced contents files that were simple listings of contents separated by line breaks, with no bundles (figure 11).

Batch ingest

Once our batch ingest structure was complete and populated with the required files, we made our final preparations for batch ingest. We developed a workflow for verifying the structure. We divided the structure amongst the team, assigning a set of directories to each team member. Each team member opened every collection directory, verifying that all the item

files were correctly named and sequenced. Next, we opened each item folder, verifying that all the required files were in place. After that, we opened each xml and contents file to verify that these files were correctly placed, with titles matching the files in the item subdirectories. This process helped us to identify and correct a number of errors that if left unchecked would have resulted in a failed attempt at batch ingest. Once we were satisfied that our structure was correct, we contacted Sam Burns, the iSchool systems administrator, to request his assistance for batch ingest.

Sam Burns ingested the first collections one at a time until discovering an error in collection/disk number five. This error was an underscore in the wrong place. He then asked us to meet him at the iSchool computer lab to finish the ingest process so that we could be present to correct any further errors that might arise. Burns sent us a template for the ingest commands, which we modified to create a sequence of commands for batch ingest. Meredith Bush applied these commands and the remaining collections and items were ingested into the iSchool repository; there were no other errors.

Observations on identifying dates of creation and file formats

Among the problems we encountered during our preparations for batch ingest was the inability of available metadata extractors to identify dates of creation and file formats for our extracted files. Our solution for the dates of creation was to mark these as unknown, while our solution for the file format was to exclude the file format identification from our metadata template. However, some hints of possible ways to identify these missing metadata fields emerged during our research. We include these hints here in case they should be of use to future digital archivists.

When one of the older .doc files in the Schmandt-Besserat disk collection is opened on an Apple computer, where the default application for opening a .doc is Microsoft Word for Windows 2004, a dialog box opens (figure 12). When the files are opened using the “Recover Text From Any File” option, hidden headers and footers are revealed. The header provides information on the application and operating system, while the footer sometimes presented two dates. One of these dates always matched the date of last modification, while one was always prior to the date of last modification. This other date is not revealed by any other procedure, including the `fls` command in UNIX and the application of currently available metadata extractors. It is possible that this extra date is the date of the file’s creation. Future research might focus on recovering and authenticating these dates.

The header information generated by the aforementioned procedure describes the application as Word 3.1 for an IBM system. Version 3 of Word, according to Wikipedia articles, was produced only for MS-DOS and Mac. Since Denise Schmandt-Besserat never used a Mac, it is reasonable to conclude that she was working in Word for MS-DOS. As shown in figure 9, the

Droid metadata extractor did not make this identification clear. However, a close inspection of figure 9 reveals a hint concerning the possible identification of the file type. Throughout the sequence of tentative file format identifications, the MimeType elements are empty. However, when Droid tentatively identifies Microsoft Word for MS-DOS Version 3, the MimeType element is not empty, but rather contains the identification “application/msword.” Since this is the only tentative identification to provide a MimeType, and since independent evidence suggests that this is a correct file format identification, it is possible that we may take Droid’s tentative identification as a reliable identification in cases where the MimeType element is not empty. However, this is a hypothesis that would require further investigation prior to usage for constructing archival metadata files.

Figure 5



Above: the green diskette box in the Schmandt-Besserat papers. Contained drafts of *Before Writing*. Below: the first disk in the collection, inside the green box.

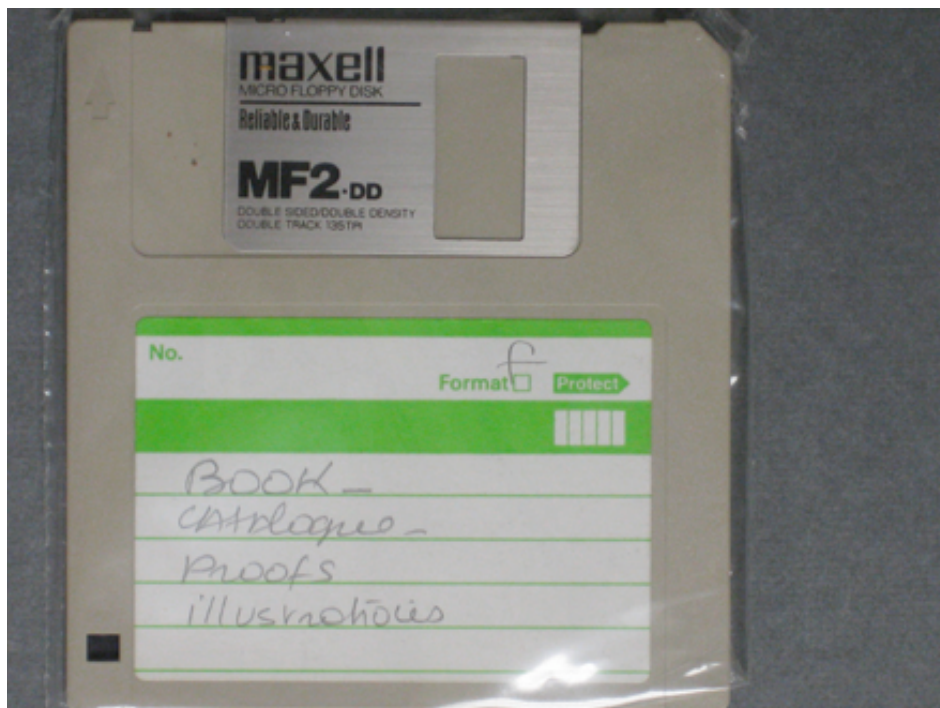
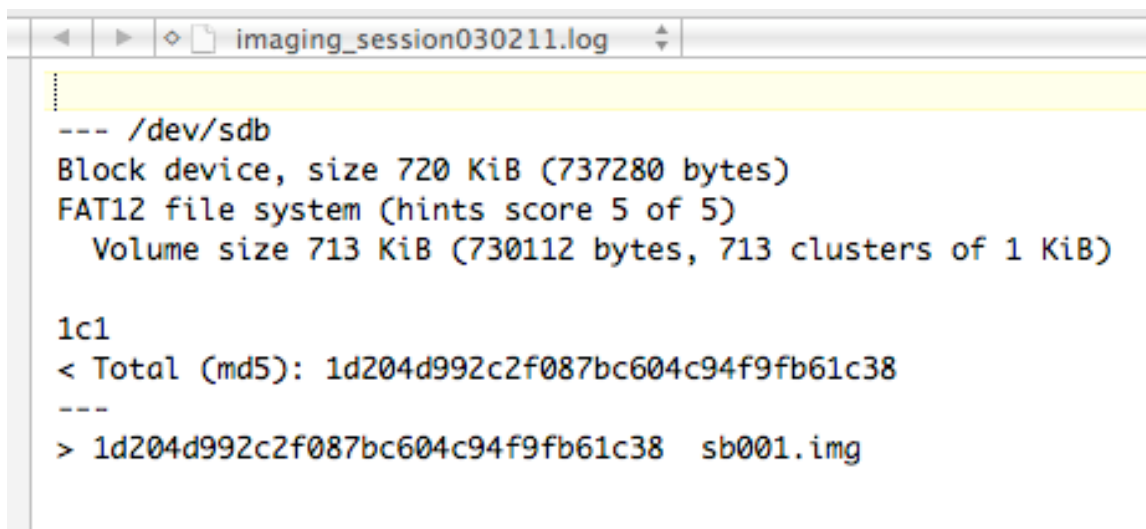


Figure 6



```
imaging_session030211.log
--- /dev/sdb
Block device, size 720 KiB (737280 bytes)
FAT12 file system (hints score 5 of 5)
  Volume size 713 KiB (730112 bytes, 713 clusters of 1 KiB)

1c1
< Total (md5): 1d204d992c2f087bc604c94f9fb61c38
---
> 1d204d992c2f087bc604c94f9fb61c38  sb001.img
```

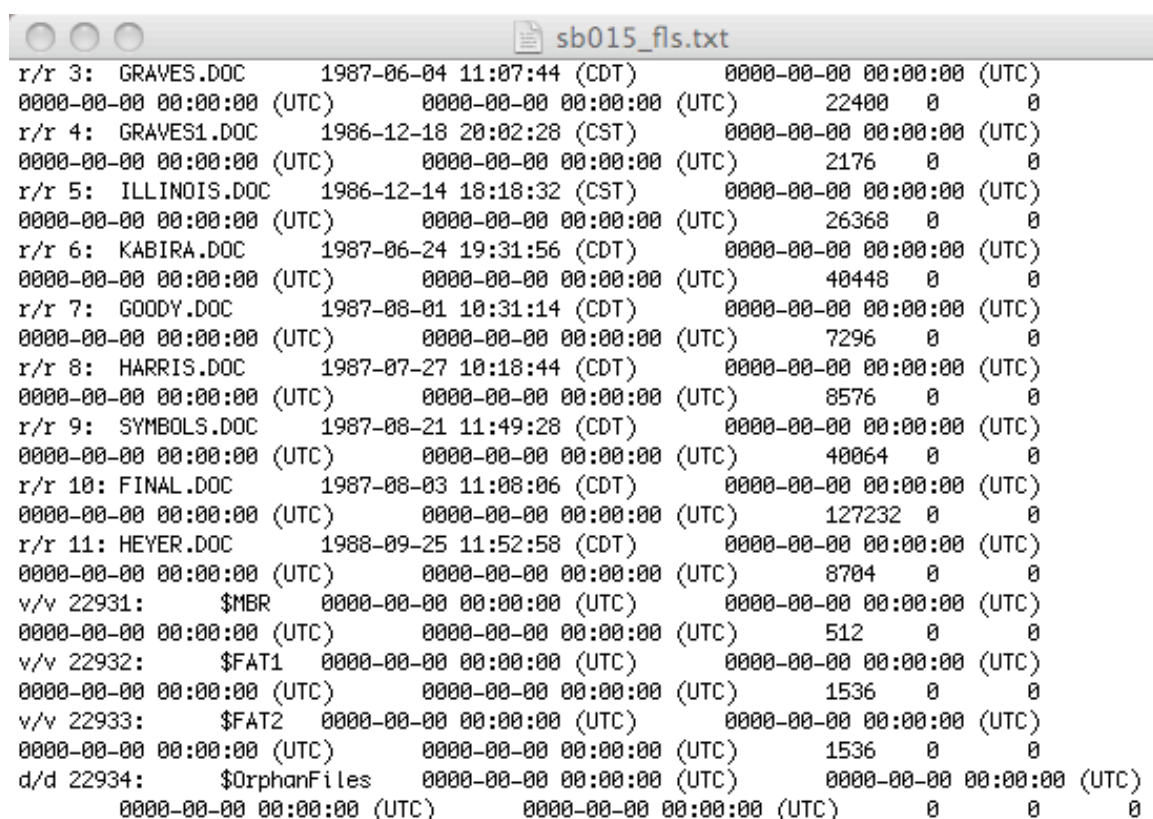
Above: output file from combined **disktype** and **hash** commands. The file system is analyzed at the top of the log file. At the bottom of the log file is the checksum of the physical disk followed by the checksum of the disk image. They are identical. Checksums calculated for working images were also identical to the disk images, and therefore to the original disks.

Below: results of a ClamAV virus scan showing the Laroux virus found in an executable program file.

```
/media/floppy0/Humans.EXE: XM.Laroux.A-gen FOUND

----- SCAN SUMMARY -----
Known viruses: 943305
Engine version: 0.95.3
Scanned directories: 1
Scanned files: 1
Infected files: 1
Data scanned: 4.88 MB
Data read: 1.37 MB (ratio 3.57:1)
Time: 3.609 sec (0 m 3 s)
```

Figure 7



```

r/r 3: GRAVES.DOC      1987-06-04 11:07:44 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      22400  0  0
r/r 4: GRAVES1.DOC    1986-12-18 20:02:28 (CST)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      2176  0  0
r/r 5: ILLINOIS.DOC   1986-12-14 18:18:32 (CST)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      26368  0  0
r/r 6: KABIRA.DOC     1987-06-24 19:31:56 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      40448  0  0
r/r 7: GOODY.DOC      1987-08-01 10:31:14 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      7296  0  0
r/r 8: HARRIS.DOC     1987-07-27 10:18:44 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      8576  0  0
r/r 9: SYMBOLS.DOC    1987-08-21 11:49:28 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      40064  0  0
r/r 10: FINAL.DOC     1987-08-03 11:08:06 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      127232  0  0
r/r 11: HEYER.DOC     1988-09-25 11:52:58 (CDT)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      8704  0  0
v/v 22931: $MBR        0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      512  0  0
v/v 22932: $FAT1       0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      1536  0  0
v/v 22933: $FAT2       0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      1536  0  0
d/d 22934: $OrphanFiles 0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)
0000-00-00 00:00:00 (UTC)      0000-00-00 00:00:00 (UTC)      0  0  0

```

Results of an **fls** command carried out on a disk image, showing file names and dates of last modification. Dates of creation were rarely found in this set of disks.

Figure 8

Directory structure for batch ingest, showing collection-level directories (sb001, sb002) containing item-level directories, as well as the contents of an item representing a disk image and an item representing an extracted files.

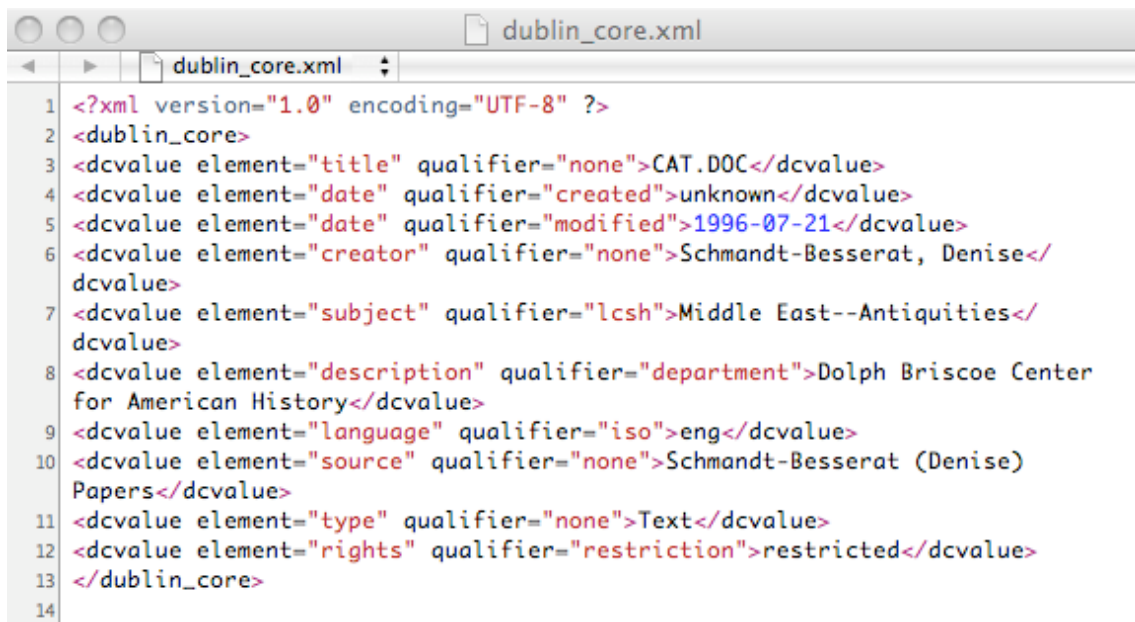
Filename	Size	Modified
▼ batch_ingest	4.0 KB	5/1/11 11:04 PM
▶ sb001	4.0 KB	4/27/11 12:25 PM
▶ sb002	4.0 KB	5/1/11 4:54 PM
▶ sb003	4.0 KB	4/20/11 11:04 AM
▶ sb004	4.0 KB	4/27/11 3:17 PM
▶ sb005	4.0 KB	4/17/11 5:39 PM
▶ sb006	4.0 KB	4/20/11 11:09 AM
▶ sb007	4.0 KB	5/1/11 9:09 PM
▼ sb008	4.0 KB	4/29/11 7:55 PM
▼ item_001	4.0 KB	4/29/11 7:55 PM
contents	116 B	4/27/11 3:55 PM
dublin_core.xml	1.1 KB	4/29/11 7:55 PM
imaging_session030211.log	269 B	4/27/11 3:19 PM
sb008.img	1.4 MB	4/20/11 11:10 AM
sb008.img_fls.txt	1.2 KB	4/27/11 3:19 PM
sb008.img.md5	44 B	4/27/11 3:19 PM
sb008.md5	46 B	4/27/11 3:19 PM
scansession_041711.log	223 B	4/27/11 3:19 PM
▶ item_002	4.0 KB	4/27/11 1:15 PM
▼ item_003	4.0 KB	4/27/11 1:16 PM
contents	35 B	4/27/11 1:16 PM
dublin_core.xml	810 B	5/1/11 5:27 PM
human figurines.xls	31.5 KB	4/20/11 11:10 AM
▶ sb009	4.0 KB	4/20/11 11:19 AM
▶ sb010	4.0 KB	4/17/11 5:44 PM
▶ sb011	4.0 KB	4/27/11 1:50 PM
▶ sb012	4.0 KB	4/17/11 5:45 PM
▶ sb013	4.0 KB	4/20/11 11:31 AM
▼ sb014	4.0 KB	4/29/11 8:05 PM
▶ item_001	4.0 KB	4/29/11 8:05 PM
▶ item_002	4.0 KB	4/27/11 2:27 PM
▶ item_003	4.0 KB	4/27/11 2:28 PM
▶ item_004	4.0 KB	4/27/11 2:28 PM
▶ item_005	4.0 KB	4/27/11 2:29 PM
▶ item_006	4.0 KB	4/27/11 2:29 PM
▶ item_007	4.0 KB	4/27/11 2:30 PM
▶ item_008	4.0 KB	4/27/11 2:30 PM
▶ item_009	4.0 KB	4/27/11 2:31 PM
▶ item_010	4.0 KB	4/27/11 2:31 PM
▶ sb015	4.0 KB	4/17/11 5:49 PM

Figure 9

Partial FITS output for an extracted file, showing tentative file format identifications.

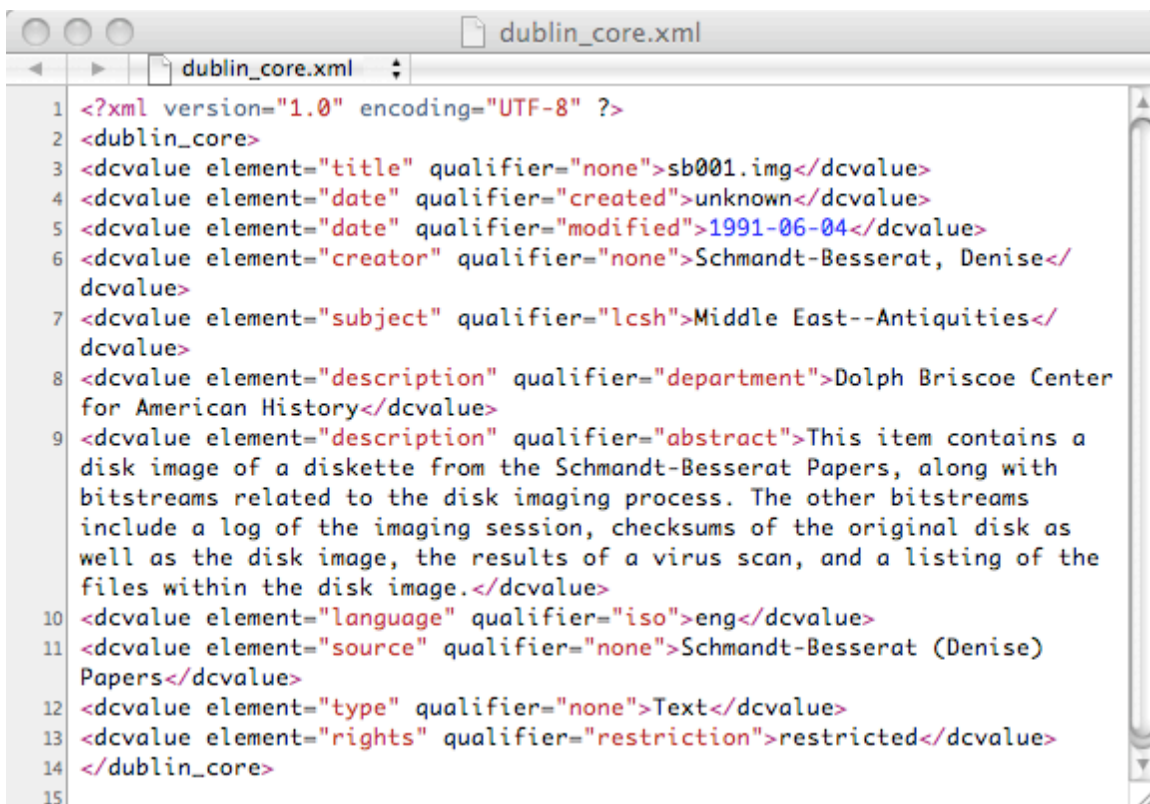
```
<tool name="Droid" version="3.0">
  <FileCollection xmlns="http://www.nationalarchives.gov.uk/p
    <DROIDVersion>3.0</DROIDVersion>
    <SignatureFileVersion>13</SignatureFileVersion>
    <DateCreated>2011-04-20T11:44:04</DateCreated>
    <IdentificationFile IdentQuality="Tentative">
      <FilePath>/media/floppy0/cat.doc</FilePath>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>Microsoft Word for Macintosh Document</Name>
        <Version>6.0</Version>
        <PUID>x-fmt/2</PUID>
        <MimeType />
      </FileFormatHit>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>Wordperfect Secondary File</Name>
        <Version>5.0</Version>
        <PUID>x-fmt/42</PUID>
        <MimeType />
      </FileFormatHit>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>Wordperfect Secondary File</Name>
        <Version>5.1/5.2</Version>
        <PUID>x-fmt/43</PUID>
        <MimeType />
      </FileFormatHit>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>WordPerfect for MS-DOS/Windows Document</Name>
        <Version>6.0</Version>
        <PUID>x-fmt/44</PUID>
        <MimeType />
      </FileFormatHit>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>Microsoft Word for Macintosh Document</Name>
        <Version>X</Version>
        <PUID>x-fmt/129</PUID>
        <MimeType />
      </FileFormatHit>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>Stationary for Mac OS X</Name>
        <PUID>x-fmt/131</PUID>
        <MimeType />
      </FileFormatHit>
      <FileFormatHit>
        <Status>Tentative</Status>
        <Name>Microsoft Word for MS-DOS Document</Name>
        <Version>3.0</Version>
        <PUID>x-fmt/273</PUID>
        <MimeType>application/msword</MimeType>
      </FileFormatHit>
    </IdentificationFile>
  </FileCollection>
</tool>
```


Figure 10



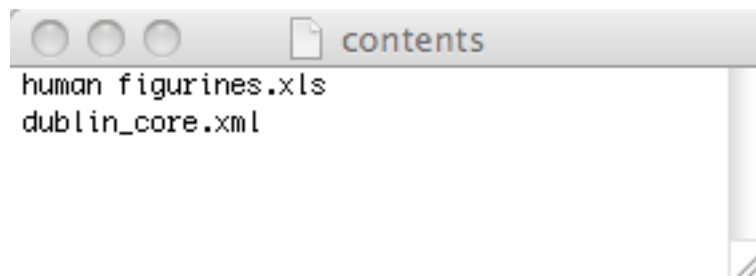
```
1 <?xml version="1.0" encoding="UTF-8" ?>
2 <dublin_core>
3 <dcvalue element="title" qualifier="none">CAT.DOC</dcvalue>
4 <dcvalue element="date" qualifier="created">unknown</dcvalue>
5 <dcvalue element="date" qualifier="modified">1996-07-21</dcvalue>
6 <dcvalue element="creator" qualifier="none">Schmandt-Besserat, Denise</
dcvalue>
7 <dcvalue element="subject" qualifier="lcsch">Middle East--Antiquities</
dcvalue>
8 <dcvalue element="description" qualifier="department">Dolph Briscoe Center
for American History</dcvalue>
9 <dcvalue element="language" qualifier="iso">eng</dcvalue>
10 <dcvalue element="source" qualifier="none">Schmandt-Besserat (Denise)
Papers</dcvalue>
11 <dcvalue element="type" qualifier="none">Text</dcvalue>
12 <dcvalue element="rights" qualifier="restriction">restricted</dcvalue>
13 </dublin_core>
14
```

Above: xml file for an extracted file. Below: xml file for a disk image, describing the entire item, including ancillary files produced during the imaging and ingest process.



```
1 <?xml version="1.0" encoding="UTF-8" ?>
2 <dublin_core>
3 <dcvalue element="title" qualifier="none">sb001.img</dcvalue>
4 <dcvalue element="date" qualifier="created">unknown</dcvalue>
5 <dcvalue element="date" qualifier="modified">1991-06-04</dcvalue>
6 <dcvalue element="creator" qualifier="none">Schmandt-Besserat, Denise</
dcvalue>
7 <dcvalue element="subject" qualifier="lcsch">Middle East--Antiquities</
dcvalue>
8 <dcvalue element="description" qualifier="department">Dolph Briscoe Center
for American History</dcvalue>
9 <dcvalue element="description" qualifier="abstract">This item contains a
disk image of a diskette from the Schmandt-Besserat Papers, along with
bitstreams related to the disk imaging process. The other bitstreams
include a log of the imaging session, checksums of the original disk as
well as the disk image, the results of a virus scan, and a listing of the
files within the disk image.</dcvalue>
10 <dcvalue element="language" qualifier="iso">eng</dcvalue>
11 <dcvalue element="source" qualifier="none">Schmandt-Besserat (Denise)
Papers</dcvalue>
12 <dcvalue element="type" qualifier="none">Text</dcvalue>
13 <dcvalue element="rights" qualifier="restriction">restricted</dcvalue>
14 </dublin_core>
15
```


Figure 11



Above: contents file for an extracted file item. Below: contents file for a disk image item.

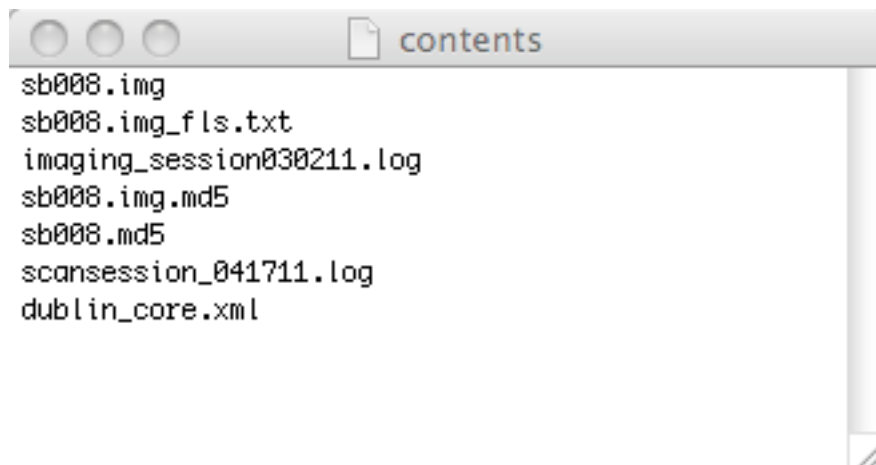
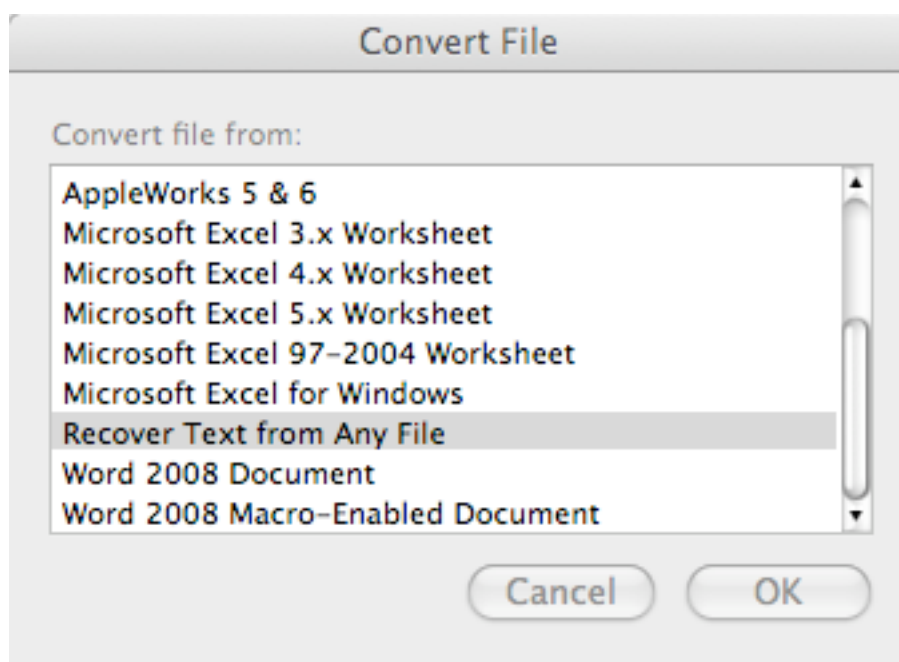


Figure 12



Top: Convert file dialog box in Word for Mac 2004.

.<??C
C:\WORD31\NORMAL.STY
IBMPRO

Above: Recovered header information for a .doc file. Below: Recovered footer information for a .doc file.

e1`-\EWèSYN[ÈJ†
`ô□W%□S'□N]□J†
>□wY□rb□nx□i-□e=
'F□\f□W...
nân'ncèn2Ènj
n©Èn...Èn
nâ□nY□n"□n\$□nZ
n
08/29/9006/18/90k

Analysis and Recommendations (Meredith Bush, Mark Firmin, & Javier Ruedas)

Future student teams working towards batch ingest would benefit from the establishment of procedures for automating metadata and contents file production. In previous semesters, students have written PERL scripts to automate the production of metadata files and contents files, but these were not easily available to spring 2011 students. Our team obtained a PERL script for automating contents files, but were unable to run it on our Mac and Windows systems. We tried several automation procedures, but always found that debugging the scripts and making them work on our particular computers was very time-consuming. In each case, we calculated that we would finish the project more quickly if we manually constructed xml and contents files. The manual creation of xml and contents files was time-consuming and left room for human error. An automated process for creating these files would improve the quality of metadata and facilitate any corrections needed for batch ingest. In spring 2011, the Digital Archaeology Lab team thoroughly clarified the use of UNIX commands for all phases of disk imaging and disk image analysis and processing, greatly facilitating the batch ingest process for our team and other teams. We recommend that in a future iteration of the course, students with scripting skills be assigned to develop xml and contents file automation scripts to be run on the DAL computers, clarifying and facilitating the batch ingest process in a similar way to the spring 2011 DAL team's efforts to clarify the imaging procedures.

Our team encountered a difficulty in constructing the batch ingest structure when we learned that two stakeholders in our project—the iSchool institutional repository and the UTDR—used different structures for batch ingest. UTDR employs a simpler structure, in which extracted files are placed as bitstreams together with the disk images from which they were extracted. However, UTDR employs a more complex structure for contents files. Once we learned about this difference, we had to devote time and effort to clarifying the structure to be used prior to proceeding. We recommend that the existence of different structures for each repository be made explicit earlier in the project, in cases where UTDR and iSchool are both stakeholders. It is possible that scripting procedures, as recommended above, could facilitate the simultaneous production of both structures, but it would be preferable if there were only one accepted structure.

The batch ingest process is complex; students facing the probability of batch ingest would benefit from analyzing the steps to batch ingest and developing a proposed workflow early in the semester. An assignment similar to the management plan assignment, in which students are asked to examine the steps to be taken and visualize the workflow from beginning to end, could be helpful in helping students manage their efforts throughout the semester.

Future projects to recover data from the computer tapes in the Schmandt-Besserat collection should keep in mind that it is the donor's explicit wish that the data on the tapes be made publicly accessible to a wide audience of scholars. These data are different from the

textual material, including teaching materials and recommendations, found on the disks. The tapes are likely to contain large volumes of scientific data that should be made available to the designated user community rather than being restricted to the CAH reading room.

Groups that are likely to conduct oral history interviews should be encouraged to familiarize themselves with the iSchool audio recording equipment early in the semester. Waveform Audio File Format, or WAV, is the current standard for uncompressed archival-quality spoken voice recordings. It would be desirable for students who are going to conduct oral history interviews to practice with the recording equipment in order to produce high-quality, uncompressed archival sound files for long-term preservation, even if compressed files are created for easier access.