Fedora and the Preservation of University Records Project

2.2 Ingest Projects

Version
1.0

Date
September 2006

Digital Collections and Archives, Tufts University
Manuscripts & Archives, Yale University

An Electronic Record Research Grant funded by the
National Historical Publications and Records Commission

Digital Collections and Archives
Tisch Library Building
Tufts University
Medford, Massachusetts 02155
http://dca.tufts.edu

Manuscripts and Archives
Yale University Library
Yale University
P.O. Box 208240
New Haven, Connecticut 06520-8240
http://www.library.yale.edu/mssa/

Co-Principle Investigators
Kevin Glick, Yale University
Eliot Wilczek, Tufts University

Project Analyst
Robert Dockins, Tufts University

This document is available online at
http://dl.tufts.edu/view_pdf.jsp?urn=tufts:central:dca:UA069:UA069.004.001.00007
(September 2006)

Fedora and the Preservation of University Records Project Website at
http://dca.tufts.edu/features/nhprc/index.html

Funded by the
National Historical Publications and Records Commission
Grant Number 2004-083

Fedora and the Preservation of University Records Project

PART ONE: INTRODUCTION

PART TWO: INGEST

PART THREE: MAINTAIN

PART FOUR: FINDINGS

TABLE OF CONTENTS

OVERVIEW

Although the Ingest Guide is a prescriptive guide for managing an ingest process, archives can implement the Guide in a variety of ways, from an entirely manual to an extensively automated process. Archives can use the Guide to manage a wide range of accessions: small or large acquisitions, complex or simple collections, single or recurring accessions. Both large and small archives in a variety of industries can use the Guide. In order to explore and illustrate how archives could implement the Ingest Guide, the project team undertook three Ingest projects.

The Ingest projects were

1. Website of an ad-hoc committee on undergraduate life
2. Working papers of the Board of Trustees saved as desktop applications on CDs
3. Library administration records saved as desktop applications stored in an instance of SharePoint Team Services, a web-based share-space environment.

These three projects are only examples of how archives *could* use the Ingest Guide; they are *not* instructions for how archives must use the Guide.

METHODOLOGY

Each Ingest project's description consists of an:

- Ingest Project Narrative
- Survey Report
- Ingest Guide Steps: Archive and Producer Actions

Ingest Project Narrative

The Ingest Project Narrative provides a descriptive summary of the actions the project team actually undertook during each Ingest project. Because the project team undertook the projects before the Guide was finalized, the actions they undertook do not precisely follow all the steps of the Guide. In addition, the Ingest Guide calls for archives to support their Ingest processes with a variety of resources that the project team did not have and their development was beyond the scope of this research project. Therefore the project team had to undertake these projects without the benefit of these resources.

Survey Report

While the survey reports in this document are based on the actual actions undertaken by Tufts and Yale during each of their six Ingest projects, project staff did not create these reports while undertaking their projects. The survey reports found below are theoretical explorations of how Tufts or Yale might construct their reports for a fully operational Ingest process described by the Ingest Guide.

The Ingest Guide calls for the Archive to start a survey report early in Section A of the Guide, Negotiate Submission Agreement, and to continue to add information to the survey throughout many of the Steps in Section A. The Ingest Guide defines a survey report as:

> A Report that identifies the records an Archive should accession during the Ingest Project. Survey Reports can vary greatly in detail, from a general description of the records the Archive should accession to an item-level inventory of those records. An Archive may create an early working draft of the Report in Step A2.1, after it and the Producer agree on the scope of the records that will be surveyed. In Steps A3.1 and A3.2, the Archive describes in the Survey Report the records it surveyed to the level of detail it requires. In Steps A3.3 and A3.4, the Archive documents in the Report its decisions on which, if any, records in the survey it should accession and what essential elements of these records it needs to preserve. To guide the Archive's appraisal decisions in Steps A3.3 and A3.4 and to be useful in Parts A3 through A10, the Survey Report needs to identify the records' Producer, Record Types, format type, file size, confidentiality requirements, copyright status, and any Producer-created identifiers.

The survey reports in these Ingest Project descriptions consist of five parts:

1 *Background*
This gives a brief overview of the Ingest project.

2 *Description of Records Surveyed*
This contains a general description the records and identifies the Producer, the functions of the records, the record types of the records, their format types, file size, any Producer identifiers, any confidentially and access restriction requirements, the copyright status, and listing (to varying degrees of detail) of the records within the purview of the Ingest project.

3 *Evaluation of Recordkeeping System*
This contains a description of how well the recordkeeping system storing and managing the records enables the feasible and scaleable transfer of those records to the Archive. This portion of the survey report also indicates if the records are managed according to the rules of the recordkeeping system.

4 *Evaluation of Authenticity*
This presents the Archive's judgment of the authenticity of the records involved in the Ingest Project. This section describes the reasoning behind the Archive's judgment.

5 *Appraisal Decision*
This documents the Archive's appraisal decision on the records in the Ingest Project. Each type of record listed the "Description of Records Surveyed" section has a disposition decision; a reason for disposition that usually references a retention schedule, collection policy, or other warrant for the decision; and a description of the essential elements of the records.[1]

Ingest Guide Steps: Archive and Producer Actions
These three-column spreadsheets describe the actions Tufts and Yale might take for each step of the Ingest Guide with a fully operational Ingest process described by the Ingest Guide. Like the survey reports, these spreadsheets are theoretical explorations of what Tufts and Yale might do, not what they actually did in the actual Ingest Projects.

The first column lists the steps of the Ingest Guide, the second column describes the actions Tufts or Yale would undertake for their Ingest projects, and the third column describes the actions of the Producer. A step from the Guide that requires no actions in a particular Ingest Project is not listed in that Project's spreadsheet. Terms in bold in the Archive Action or Producer Action columns refer to Resources, Products, or Documentation in the Ingest Guide.[2]

Many of the Archive's actions refer to an "ingest application." This application refers generically to an application that an archive would implement to handle many of the tasks described in Part

[1] The essential elements analysis, which considers documentary form, annotations, context, and medium, is largely based on the InterPARES, "Authenticity Task Force Report," *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, <http://www.interpares.org/book/interpares_book_d_part1.pdf>
[2] See the "Components, Resources, Products, and Documentation" section of the Ingest Guide.

B, Transfer and Validate, of the Ingest Guide. These tasks include transfer, validation, format transformation, and turning SIPs into AIPs.[3]

[3] Tufts University developed the Tufts Ingest Prototype System (TIPS) as an initial attempt to develop such an application that would help an Archive semi-automate the steps in Part B of the Ingest Guide. See 2.3 Ingest Tools for a description of the application.

INGEST PROJECT ONE
Task Force on the Undergraduate Experience

Ingest Project Narrative

In 2001 Tufts University created the Task Force on the Undergraduate Experience to study undergraduate life at Tufts University. The Taskforce completed its mission in June 2003. In April 2003 the Task Force Project Coordinator contacted the Digital Collections and Archives (DCA) concerning the records of the Task Force. The DCA accessed paper records and several CDs from the Task Force in May 2003. However, the Task Force also created a website as part of its business. The DCA did not accession the website and only gave it a cursory appraisal at the time. The Task Force officially disbanded at the end of June 2003.

In October 2005 the project team decided to use the website of the Task Force as one of our Ingest Projects because it raises a number of interesting appraisal issues and represents a situation that most archives at colleges and universities are likely to face. There are important issues regarding how to view the website as a record, the essential elements of the website, and the way to run an ingest project in the absence of the original producer. Because the website is still live on a public webserver, the project team was able to access the website at any time.

To create the SIP, the project team used the wget utility[4] to fetch a copy of the website to a local filesystem. The team then manually extracted records from the website that the team judged to have enduring value. These records composed the website which included PDFs of reports and some of the text of the HTML files that made up the site. The project team then ran the HTML of the website through the W3C HTML tidy utility[5] to produce nearly-valid XHTML. Then the team packaged both the original and tidied versions of the entire site into separate ZIP files. The team packaged together all of the records into a single SIP.

The project team submitted this SIP to the Tufts Ingest Prototype System (TIPS).[6] This application performed all the checks of Part B.2 from the Ingest Guide (albeit with minimal validation) and then created one AIP for each record, with the tidied and untided versions of the whole website as one AIP. The ingest application was able to submit many of the AIPs to Tufts' Fedora 1.2.1 repository. However, the largest record (containing the contents of the website), caused the ingest system to crash. The problem is related to the way SOAP bindings to Fedora 1.2.1 are handled. Fixing this problem will require upgrading the Fedora SOAP interface to use SAAJ or MOTM.

The project team skipped the difficult task of building resources. We did not create record type records, format type information, or producer records and none already existed to utilize. How to represent, manage, and acquire this information remains an open question. This resulted in us

[4] Wget is a command-line tool for UNIX-like systems which retrieves files using the HTTP, HTTPS, and FTP protocols. See http://www.gnu.org/software/wget/wget.html for additional information on GNU wget.

[5] Tidy is a command-line tool which parses HTML and produces normalized HTML or XHTML. It is carefully written to maintain the visual appearance of the HTML as much as possible. See http://tidy.sourceforge.net/.

[6] See 2.3 Ingest Tools for a description of the application.

creating a number of records in the Fedora repository with dummy links to stub resources—essentially empty Fedora objects—that stood in for various resources.

Survey Report

Background

Eliot Wilczek, University Records Manager, Digital Collection and Archives (DCA) met with Armand Greene, Project Coordinator of the Task Force on the Undergraduate Experience on 04/10/03 to conduct a records survey. He sent a Records Survey Report to Armand Greene on 04/17/03. The report essentially deferred on the appraisal of the Task Force's website. The report said in part:

> "The website essentially provides access to a variety of documents. The DCA would focus on preserving these documents rather than saving the whole website as a website. Documents and records on the website include:
>
> President Bacow's charge
> Progress information which includes a meetings list
> Questions of the week
> Listing of people on the Task Force
> Reports and proposals
> Bibliography of the Task Force in the news with PDF news clippings
>
> "Transfer all to the Digital Collections and Archives except the PDF news clippings. The DCA has these publications so destroy the clippings when they are no longer needed. Transfer to the DCA the bibliography listing the articles concerning the Task Force.
>
> "Many of the documents on the website exist in electronic and/or paper format elsewhere and therefore the copies on the website may not need to be transferred to the DCA. This should be discussed in further detail."

The Task Force transferred records to the DCA on 05/22/03 in accordance with the records survey. These were paper records and electronic records in desktop application formats on CDs. The website or components of the website were not transferred to the DCA, although the reports on the website were transferred to the DCA in paper and desktop application format.

Armand Greene verbally informed Eliot Wilczek on 05/22/03 the Task Force website would remain at http://ugtaskforce.tufts.edu for sometime into the future, although Mr. Greene did not specify a time period. Eliot Wilczek verbally indicated to Armand Green that the DCA will look at accessioning the website or components of the website sometime in the future. Armand Green accepted this proposed effort.

On 06/30/03 the Task Force on the Undergraduate Experience concluded its activities and ceased as a unit of Tufts University in accordance with its mandate from the President of Tufts University.

On 10/12/05 Eliot Wilczek conducted a records survey of the Task Force website located at http://ugtaskforce.tufts.edu as part of a separate Ingest project.

Description of Records Surveyed

General Description

This is the website of the Task Force on the Undergraduate Experience. The website is located at http://ugtaskforce.tufts.edu. It is a relatively small and simple website. It has the following sections:

Home page
President's Charge
Progress
In the News
People
Contact Us

Producer

Task Force on the Undergraduate Experience

Producer Role(s) Creator

Function(s)

The website serves as the main vehicle for the Task Force to disseminate its findings and information about its activities to the Tufts community.

Record Type(s)

Charges
Reports
Event Records
News Clippings
Subject Files
Publications

Format Type(s)

Web pages composed in html.

For most pages in `<head></head>`

```
<meta name="GENERATOR" content="Microsoft FrontPage 5.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
```

Not valid HTML, no `DOCTYPE` found, attempted validation HTML 4.01 Transitional.
Validation used: W3C Markup Validation Service v0.7.0
http://validator.w3.org accessed 10/13/05.

http://ugtaskforce.tufts.edu/contactus.html notes

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
```

Not valid HTML 4.0 Transitional.
Validation used: W3C Markup Validation Service v0.7.0
http://validator.w3.org accessed 10/13/05.

PDF files.

No validation performed.

JPG image files.

No validation performed.

File Size

Approximately 5 to 10 MB. A substantial portion of this is the pdfs of reports and news clippings.

Producer Identifier

All of the website and all of its components do not have producer identifiers that need preservation.

Confidentiality Requirements/Access Restrictions

All of the website and all of its components have no requirements for access restrictions. They should all have Category 2: Universal Distribution.

Copyright Status

Copyright of website as a whole and individual reports held by Tufts University.

Copyright of *Tufts Daily* articles in the news clippings held by the *Tufts Daily*.

Copyright of other student publications in the news clippings held by the author.

Copyright of Tufts University publications in the news clippings held by Tufts University.

List of Records Surveyed

President's Charge

Description	University President's charge to the Task Force.
Record Type	Charges
Dates	2001
Format	HTML and PDF

Various Reports

Description	Various interim, status, and final reports created by the Task Force.
Record Type	Reports
Dates	2001 through 2003
Format	HTML and PDF

Outreach Activities List

Description	List of outreach activities undertaken by the Task Force.
Record Type	Event Records
Dates	2003
Format	HTML

Links to News Stories

Description	List of links to online news stories concerning the Task Force.

Record Type	News Clippings
Dates	2003
Format	HTML

News Stories

Description	Digitized print news stories concerning the Task Force; provides bibliographic citations of the news stories.
Record Type	News Clippings
Date	2003
Format	PDF

Membership List

Description	List of Task Force members.
Record Type	Subject Files
Dates	2003
Format	HTML

Benchmarking Studies List

Description	List of links to studies concerning undergraduates at other institutions the Task Force used as benchmarks.
Record Type	Subject Files
Dates	2003
Format	HTML

Website

Description	The Task Force website as a whole.
Record Type	Publications
Dates	ca. 2001 through 2003
Format	HTML, PDF, JPG

Evaluation of Recordkeeping System

The PDF, image, and HTML files are stored on a public webserver accessible at http://ugtaskforce.tufts.edu. These files appear to be stored on the webserver in normal manner although the Digital Collections and Archives did not find any procedures or rules on the management of these files. Acting as both the Archive and the Producer, the DCA does not have direct access to the server and must access the files through a web browser. This is a trustworthy but not scaleable transfer process. However, because the website is small and the DCA will only have to capture files from the site once because the Task Force no longer exists and the website will not change, the DCA will make the effort to manually capture the files.

Evaluation of Authenticity

The Digital Collections and Archives judges the Task Force on the Undergraduate Experience website at http://ugtaskforce.tufts.edu and all of its component parts that it has evaluated in this survey to be authentic for the following reasons:

- At the 05/22/03 survey interview, Armand Green declared that the Task Force created the website in the normal course of its business.
- At the 05/22/03 survey interview, Armand Green identified the website at http://ugtaskforce.tufts.edu as the Task Force website.
- During the 10/12/05 survey, Eliot Wilczek determined the website at http://ugtaskforce.tufts.edu was the same website that Armand Green identified as the website on 05/22/03 based on a brief visual review of the website's appearance and content.
- The probability someone would maliciously alter the website—particularly its content—and try to hide that content alteration is extremely low because no one has a reasonable motivation to undertake such an action.

Appraisal Decision

President's Charge				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen071
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *Verbatim* Webpage Structure: *General Appearance* PDF Structure: *General Appearance*
			Signs	None
		Annotations		None
		Context		*Information* that delivered as webpage and PDF document via HTTP protocol with no restrictions
		Medium		*Information* that HTML and PDF
Various Reports				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen065
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *Verbatim* Webpage Structure: *General Appearance* PDF Structure: *General Appearance*
			Signs	None
		Annotations		None
		Context		*Information* that delivered as webpage and PDF document via HTTP protocol with no restrictions
		Medium		*Information* that HTML and PDF
Outreach Activities List				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen034

	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *General Appearance* Webpage Structure: *General Appearance*
			Signs	None
		Annotations		None
		Context		*Information* that delivered as webpage document via HTTP protocol with no restrictions
		Medium		*Information* that HTML
Links to News Stories				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen024
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *General Appearance* Webpage Structure: *General Appearance*; *Information* of hyperlink URL, hyperlink functionality not needed
			Signs	None
		Annotations		None
		Context		*Information* that delivered as webpage document via HTTP protocol with no restrictions
		Medium		*Information* that HTML
News Stories				
	Disposition			Destroy when no longer needed. [Because it is difficult to remove these records from the website when the DCA captures the Website as a whole, the DCA will probably transfer the News Stories to the DCA. However, the DCA would make no effort to preserve the News Stories.]
	Reason for Disposition			Although Retention Schedule gen024 applies, the DCA already holds the publications that contain these News Stories.
Membership List				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen010
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *General Appearance* Webpage Structure: *General Appearance*
			Signs	None

Annotations	None
Context	*Information* that delivered as webpage document via HTTP protocol with no restrictions
Medium	*Information* that originally in HTML
Benchmarking Studies List	
Disposition	Transfer to DCA
Reason for Disposition	Retention Schedule gen010
Essential Elements	
Documentary Form	
Content	*Verbatim*
Presentation	Textual Structure: *General Appearance* Webpage Structure: *General Appearance*; Information of hyperlink URL, hyperlink functionality not needed
Signs	None
Annotations	None
Context	*Information* that delivered as webpage document via HTTP protocol with no restrictions
Medium	*Information* that originally in HTML
Website (as a whole)	
Disposition	Transfer to DCA
Reason for Disposition	Retention Schedule gen015
Essential Elements	
Documentary Form	
Content	*Verbatim*
Presentation	Textual Structure: *General Appearance* Webpage Structure: *General Appearance*; *Information* of internal and external hyperlink URL, hyperlink *functionality* not needed. [Preserving *functionality* of internal hyperlink because it is an easier preservation strategy than preserving *information* of internal hyperlink is ok]
Signs	None
Annotations	None
Context	*Information* that delivered as website via HTTP protocol with no restrictions
Medium	*Information* that originally in HTML

Ingest Guide Steps: Archive and Producer Actions		
Steps	**Archive Actions**	**Producer Actions**
A Negotiate Submission Agreement		
A1 Establish Relationship		
A1.1 Initiate Contact/Ingest Project	The Archive had previously worked with the Producer on another Ingest Project. At that time, the Archive and Producer informally agreed that after the Producer ceases operations, the Archive would act as both the Archive and the Producer in this Ingest Project.	The Producer had previously worked with the Archive on another Ingest Project. At that time, the Archive and Producer informally agreed that after the Producer ceases operations, the Archive would act as both the Archive and the Producer in this Ingest Project.
A1.2 Identify Producer	The Archive identifies the Task Force on the Undergraduate Experience as the Producer.	
A1.3 Has the Archive already defined its relationship with the Producer?	Yes, the Archive already defined its relationship with the Producer when it ingested the paper records of this Task Force. *The Archive skips Steps A1.4 through A1.6.*	
A2 Define Project		
A2.1 Identify records at issue, agreeing upon scope of survey	The Archive and Producer (during previous Ingest Project) identify the contents of the Task Force website as the records of this Ingest Project.	The Producer and Archive (during previous Ingest Project) identify the contents of the Task Force website as the records of this Ingest Project.
A2.2 Does the Producer have custody/authority over the records identified in A2.1?	Yes, the Archive determines that the Task Force has the appropriate authority over the records identified in Step A2.1. *The Archive skips Steps A2.3 through A2.6.*	
A3 Collect Information and Assess Value of Records		
A3.1 Conduct Records Survey, note attributes of records	The Archive conducts a survey and produces a **Survey Report**.	The Archive as the Producer allows access to the records for the survey.
A3.2 Judge authenticity of records	The Archive examines content of records and speaks to the original Task Force Project Coordinator (during previous Ingest Project) and determines that they are very probably authentic, and updates the **Survey Report**.	
A3.3 Should the Archive accession at least some of the records?	Yes, the Archive determines that the content of the website as a whole and the reports contained therein have archival value, and judges them to be authentic. The Archive updates the **Survey Report** to reflect this decision.	

A3.4 Determine the essential elements of the records that should be accessioned	The Archive examines records and determines that the essential elements of the reports include the content of the text and the formatting. It determines the essential elements of the website include the content of the pages and evidence of the way in which the Task Force presented itself via the site. The Archive adds its determinations to the **Survey Report**.	
A4 Assess Record Type		
A4.1 Are all records identified as a Record Type?	Yes, the Archive determines that the reports are of record type *Report*, and the website is of record type *Publication*. These record types are already known. The Archive creates a **Record Type List** which references **Record Type Records.** *The Archive skips Step A4.2*.	
A5 Assess Formats		
A5.1 Are any records in file formats that are not a preservation format?	Yes, the Archive determines the following: The website is non-valid HTML, which is not a preservation format; the reports are in PDF format, which is a preservation format; the images are in GIF and JPEG formats, which are preservation formats. For the non-valid HTML files the Archive produces a **Transformation Plan**.	
A5.2 Should Archive transform or natively handle these formats?	The Archive will transform the non-valid HTML. *Archive skips Steps A5.3 through A5.4.*	
A5.5 Choose appropriate format	The Archive will transform the non-valid HTML into valid XHTML. The Archive updates its **Format Transformation Plan** to reflect this decision.	
A5.6 Is format chosen in A5.5 a Preservation Format the Archive already uses?	Yes, the Archive determines that valid XHTML is a preservation format based on its **Formats Standards Policy**.	
A6 Assess Identifier Rules		
A6.1 Is there a Producer naming/ identification scheme that needs accommodation?	No, the Archive determines there is no **Producer Naming/ Identification Scheme** that it needs to preserve or accommodate. The *Archive skips Steps A6.2 through A6.3.*	
A6.4 Determine appropriate naming/ identification scheme(s).	The Archive decides to use the standard **Archive Naming/ Identification Scheme**. The Archive records its **Naming/Identification Scheme Decision**.	

A7 Assess Copyright		
A7.1 Determine copyright status of records in Ingest Project	The Archive determines that the records are under copyright of Tufts University. The Archive records the **Copyright status**.	
A7.2 Does the Archive need to acquire copyright or license for records?	No, the Archive determines that the University holds copyright. *Archive skips Steps A7.3 through A7.8.*	
A8 Assess Access Rights		
A8.1 Determine records' Records Security Profile	The Archive determines that Task Force previously made the records available to the general public and has should have a **Records Security Profile** of open access. The Archive documents this decision as a **Records Security Profile Decision** which references the **Records Security Profile**.	
A8.2 Does current security component meet the access control needs of the records?	Yes, the Archive determines the security component of the Preservation System meets the needs of the **Records Security Profile**. *Archive skips Steps A8.3 through A8.7.*	
A9 Assess Recordkeeping System		
A9.1 Has the Archive documented recordkeeping system as supporting feasible and trustworthy transfer?	No, the Archive has not previously examined the webserver that stores and serves the website.	
A9.2 Can recordkeeping system support feasible and trustworthy transfer?	The Archive determines that the recordkeeping system can support the trustworthy transfer of records to the Archive, but not in a scalable manner. It can support a feasible transfer if the volume of records is low. The Archive produces a **Recordkeeping System Report.** *The Archive skips Step A9.3.*	
A9.4 Is Archive or Producer willing to take extraordinary measures to transfer records?	The Archive determines that it is willing to manually transfer the records from the recordkeeping system to the Archive because of the small volume of records involved in the Ingest Project. The Archive produces **SIP Creation Procedures** for moving the records from the recordkeeping system to the Archive. *The Archive Steps A9.5 through A9.7.*	
A10 Assess Feasibility		

A10.1 Can the Archive feasibly accession the records?	Yes. The Archive determines that it is feasible to manage and preserve the records and its requirements without extraordinary effort or special accommodation. The Archive produces a **Preservation System Availability Statement**. *The Archive skips Steps A10.2 through A10.5.*	
A11 Finalize Submission Agreement		
A11.1 Add description of Metadata Encoding Rules to Submission Agreement.	The Archive chooses standard Dublin Core metadata encoding and documents its **Metadata Encoding Rules Decision**.	
A11.2 Add description of Transfer Procedures to Submission Agreement	The Archive determines that it will retrieve the contents of the site via the HTTP protocol from the public webserver that currently hosts the website. The Archive documents its **Transfer Procedures Decision** in the Submission Agreement.	
A11.3 Add description of Validation Procedures to Submission Agreement	The Archive documents the validation procedures for HTML, JPEG, GIF and PDF files in the **Validation Procedures Decision**.	
A11.4 Add Transfer Schedule to Submission Agreement	The Archive selects an indefinite schedule, and documents this decision in the **Transfer Schedule Decision**.	
A11.5 Add SIP Creation Procedures to Submission Agreement	The Archive chooses its standard SIP format and documents its **SIP Creation Procedures Decision**.	
A11.6 Finalize Submission Agreement	The Archive draws up the **Draft Submission Agreement** based on its previous decisions.	
A11.7 Does Archive and Producer agree to and approve the Submission Agreement?	Yes, the Archive and Archive-as-Producer agree and produce the **Finalized Submission Agreement**. The Archive submits finalized submission agreement to the ingest system. The Archive's ingest application accepts and validates a machine-readable version of the **Finalized Submission Agreement.** *Archive skips Steps A11.8 through A11.10.*	Yes, Archive-as-Producer and Archive agree and produce the **Finalized Submission Agreement**.
B **Transfer and Validation**		
B1 **Create and Transfer SIPs**		
B1.1 Producer prepares SIP according to Submission Agreement		The Archive-as-Producer retrieves the content of the website to a workspace under Archive control. There the Archive manually extracts the subparts of the website and constructs the **SIP**. The Archive-as-Producer signs the SIP with its own digital signature.

B1.2 Producer transfers the SIP to the Archive		The Archive-as-Producer places the SIP in the ingest application drop-box.
B2 **Validate**		
B2.1 Archive receives SIP from Producer	The Ingest application accepts the SIP from the drop-box and produces **Documentation of Receipt** and delivers it to the Archive-as-Producer.	The Archive-as-Producer receives **Documentation of Receipt**.
B2.2 Is SIP well formed?	The ingest application checks the SIP format. The SIP is well formed. The Application updates the **SIP validity statement.**	
B2.3 Does SIP contain malicious code?	The ingest application scans the SIP components for viruses and other malicious code. All SIP components are clean. The application updates the **SIP validity statement**.	
B2.4 Is the submitter authorized to submit SIP to the Archive?	The ingest application validates the SIP signatures and validates identities against its database of certificates. The SIP is signed by an authorized person. The application updates the **SIP validity statement**.	
B2.5 Does SIP contain all necessary records components?	The ingest application checks all included records for completeness. All records in the SIP are complete. The application updates the **SIP validity statement**.	
B2.6 Do the record components in SIP validate?	The ingest application tests the record components for validity, where necessary. All record components validate. Application updates the **SIP validity statement**.	
B3 **Transform and Attach Metadata**		
B3.1 Do any of the records in SIP require transformation?	Yes, the Archive determines that the HTML files require tidying and transformation to valid XHTML according to the **Format Transformation Plan** in the Submission Agreement.	
B3.2 Perform transformation on records that require transformation	The Archive tidies and transforms the HTML files into valid XHTML, producing **Transformed Records**.	
B3.3 Attach to records metadata inferred from Submission Agreement	The ingest application attaches stock metadata from Submission Agreement, **Producer Record**, and **Record Type Records**, creating **Records with Attached Metadata**.	
B3.4 Attach to records the Records Security Profile defined by Submission Agreement	The ingest application attaches the **Record Security Profile** identified in the Submission Agreement, creating a **Records with Security Profile**.	
B4 **Formulate AIPs**		

B4.1 Formulate AIPs	The Ingest application creates an **AIP** for each record.	
B5 **Assess AIPs**		
B5.1 Are all of the records in the AIP part of accession described by Submission Agreement?	The Archive verifies that all of the records to be accessioned come from the website and produces an **AIP Validity Statement**. The ingest application accepts the **AIP Validity Statement**. *The Archive skips Steps B5.2 through B5.3.*	
B5.4 Is proper metadata attached to records in the AIP?	The Archive verifies that all of the records have sufficient and correct metadata and updates the **AIP Validity Statement**. The ingest application accepts the **AIP Validity Statement**. *The Archive skips Step B5.5.*	
B6 Formally Accession		
B6.1 Submit AIPs into Preservation Repository.	The ingest application submits the AIPs to the Preservation Repository.	
B6.2 Formally notify Producer that Archive has accepted and accessioned records described by Ingest Project.	The ingest application generates a **Transfer Notice** for the Producer and an entry in the **Accession Log**.	The Producer receives the **Transfer Notice**.

INGEST PROJECT TWO
Board of Trustees

Ingest Project Narrative

The Board of Trustees of Tufts University meets three times a year. Before each of these meetings, the Office of the Board of Trustees compiles a set of materials relevant to the upcoming meeting, including agendas, reports, minutes of past meetings, and other meeting records. In the past, staff from the Office of the Trustees assembled a ring binder containing these materials for each board member, mailing it to him or her several weeks before the meeting. However, in 2003, the Board began distributing these materials electronically, by burning CDs containing electronic versions of the working papers rather than filling binders.

The project team decided to make these meeting packet CDs one of ingest projects because it represents an interesting case of dealing with physical media, because the meeting records of the Board of Trustees clearly have enduring value, and because the project team had easy access to copies of the media. The organization of files on the CD represents an interesting appraisal situation by itself. Records are organized into directories by committee, and they sometimes have additional subdirectories. Most text documents have the original MS Word file along with a PDF version of the same file. Finally, there is an additional video on each CD, in which the University President describes events occurring around the Board meeting and points out items of particular note on the meeting agenda.

We decided to preserve each packet of documents that compose a record of a Board meeting—really a series of meetings of the full Board and its various committees held over the course of two days—as a single complex object containing all the data on the CD. The project team made this decision because:

1. The filesystem on the CD represented a significant element of the original order of the materials
2. The filesystem standard itself (ISO9660) is widely used, standardized, and quite preservable
3. There is little need to format-shift the bulk of the materials on the CD, because they are represented in both their original formats (MS Word) and in the more preservable PDF format.

As the project team had little experience with video formats, it felt it could not take any meaningful preservation action for the video at the time of the ingest project.

The project team took bit-for-bit copies of the CDs, using standard facilities found in Mac OS X. The team loaded these images from disk to ensure that they were not corrupted during the copy. The team packaged each individual filesystem image into a SIP and submitted to the Tufts Ingest Prototype System (TIPS),[7] which them performed the checks from Part B.2 of the Ingest Guide (however, with very minimal verification) and then created AIPs. However, the project team encountered the problem that Fedora would not accept records components over a certain size.

[7] See 2.3 Ingest Tools for a description of the application.

The meeting packets are all about 300 MB in size, and all caused ingest problems. Thus, we were unable to complete the ingest of any of the packets.

Survey Report

Background

Eliot Wilczek, University Records Manager, Digital Collection and Archives (DCA) met with Lydia Evans, Secretary of the Faculty, on 04/22/05 to discuss a variety of records management issues concerning Board of Trustees records. Lydia Evans indicated that the Office of the Board of Trustees has stored meeting records from 2000 through 2004 on CDs.

Lydia Evans agreed to give Eliot Wilczek two CDs with two sets of meeting records covering Trustees meetings from May 2003 and February 2004 so the Digital Collections and Archives (DCA) could produce this survey report on the Trustees' meeting records stored on CDs. She gave Eliot Wilczek the CDs on the day of their meeting.

Description of Records Surveyed

General Description
These are the meeting records of Board of Trustees of Tufts University. Each CD contains a website that helps a Trustee navigate through the meeting records, a video of the University President welcoming the Board members and giving an overview of the upcoming meetings, an agenda and schedule of the Board meetings and events (a Board "meeting" is really composed of several meetings and events that usually occur over two days), contact information for the Trustees and background information about Tufts.

Producer
Office of the Board of Trustees
Producer Role(s) Creator

Function(s)
The Office of the Board of Trustees sends Board members a set of meeting records. From 2000 to 2004 the Office mailed each Board member a CD of the meeting records for an upcoming meeting. The meeting records give a Board member the information he or she needs to make votes at Board meeting in an informed manner.

Record Type(s)
Meeting Records

Format Type(s)
Web pages composed in html.
For most pages in `<head></head>`

```
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-
1">
<meta name="GENERATOR" content="Mozilla/4.79 [en]C-CCK-MCD {C-UDP;
Tufts University granite 2/11/2002}  (Windows NT 5.0; U) [Netscape]">
<title>index</title>
```

Valid HTML 4.0 Transitional.
Validation used: W3C Markup Validation Service v0.7.0
http://validator.w3.org accessed 04/29/05.

PDF files.
No validation performed.

JPG image files.
No validation performed.

Microsoft Word Documents.
No validation performed.

MOV video file.
No validation performed.

Microsoft Excel Worksheet file.
No validation performed.

Executable programs—installers.
No validation performed.

File Size
February 2004 meeting: 294 MB.
May 2003 meeting: 257 MB.

Producer Identifier
All of meeting records and all of their components do not have producer identifiers that need preservation.

Confidentiality Requirements/Access Restrictions
All of the records have requirements for access restrictions. They should all have Category 1: Confidential University Records.

Copyright Status
Copyright of all records held by Tufts University.

List of Records Surveyed
Board of Trustees Meeting Records

Description Documents Trustees receive before Board meetings to help they make informed decisions. Meeting records are composed of a variety of other record types, usually reports, meeting agendas, and meeting minutes. Records of a Board meeting—really a series of meetings of the full Board and its various committees held over the span of two days—are bundled together on a CD. In this context, these records compose meeting records. The meeting records have two copies of all text documents, one copy in Word format and one copy in PDF.

Record Type Meeting Records

Dates 2003-2004

Format	HTML, PDF, WORD, JPEG, XSL, MOV

Evaluation of Recordkeeping System

All files are stored on CDs. The Office of the Board of Trustees distributes use copies to members of the Board and senior administrators. The Office maintains record copies. The Office carefully creates and manages these records but does not have written procedures for its recordkeeping of these records.

Evaluation of Authenticity

The Digital Collections and Archives judges the meeting records and all of the components of the May 2003 and February 2004 Board meetings on the CDs the Office of the Board of Trustees gave to the DCA to be authentic for the following reasons:

- At the 04/22/05 meeting, Lydia Evans declared that the Office of the Board of Trustees created the May 2003 and February 2004 meeting records in the normal course of its business.
- At the 04/22/05 meeting, Lydia Evans declared that from 2000 to 2004 the Office of the Board of Trustees stored meeting records on CDs in the normal course of its recordkeeping activities.
- Shortly after the 04/22/05 meeting, Eliot Wilczek briefly reviewed the files on the CDs that Lydia Evans gave to him and determined that they contained the May 2003 and February 2004 meeting records as she claimed.
- The probability someone would maliciously alter the records—particularly its content—and try hide that content alteration is extremely low because the opportunity and motivation to undertake such an action are both low.
- The DCA should be able presume the authenticity of meeting records the Office of the Board of Trustees directly gives to the DCA on CDs.

Appraisal Decision

Meeting Records				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen091
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *Verbatim*
				Video Structure: *General Appearance*
				PDF Structure: *General Appearance*
				Word Structure: *None*
				Excel Structure: *General Appearance—no calculations*
				HTML Structure: *General Appearance*
			Signs	None
		Annotations		None
		Context		*Information* that delivered on CDs to Board members before board meetings

Medium	*Information* that HTML, PDF, MOV, Word, and Excel stored on CDs

Ingest Guide Steps: Archive and Producer Actions		
Steps	**Archive Actions**	**Producer Actions**
A Negotiate Submission Agreement		
A1 Establish Relationship		
A1.1 Initiate Contact/Ingest Project	The Archive had previously worked with the Producer on a variety of Ingest Projects and other work. The Archive contacts the Producer about Trustee meeting records stored on CDs.	The Producer had previously worked the Archive on a variety of Ingest Projects and other work. The Archive contacts the Producer about Trustee meeting records stored on CDs.
A1.2 Identify Producer	The Archive identifies the Office of the Board of Trustees as the Producer.	
A1.3 Has the Archive already defined its relationship with the Producer?	Yes, the Archive already defined its relationship with the Producer during previous Ingest Projects. *The Archive skips Steps A1.4 through A1.6.*	
A2 Define Project		
A2.1 Identify records at issue, agreeing upon scope of survey	The Archive and Producer identify the meeting records of the Full Board of Trustees and its various committees for its May 2003 and February 2004 meetings as the records of this Ingest Project.	The Archive and Producer identify the meeting records of the Full Board of Trustees and its various committees for its May 2003 and February 2004 meetings as the records of this Ingest Project.
A2.2 Does the Producer have custody/authority over the records identified in A2.1?	Yes, the Archive determines that the Office of the Board of Trustees has the appropriate authority over the records identified in Step A2.1. *The Archive skips Steps A2.3 through A2.6.*	
A3 Collect Information and Assess Value of Records		
A3.1 Conduct Records Survey, note attributes of records	The Archive conducts a survey and produces a **Survey Report**.	The Producer provides the Archive with copies of the CDs.
A3.2 Judge authenticity of records	The Archive examines content of records and received the CDs directly from the Producer and determines that they are very probably authentic, and updates the **Survey Report**.	
A3.3 Should the Archive accession at least some of the records?	Yes, the Archive determines that the meeting records have archival value, and judges them to be authentic. The Archive updates the **Survey Report** to reflect this decision.	

A3.4 Determine the essential elements of the records that should be accessioned	The Archive examines records and determines that the essential elements of the meeting records are the content and formatting of the PDF, Word, and Excel files; the content, formatting, and functionality of the HTML wrapper; and the content of the video. The Archive adds its determinations to the **Survey Report**.	
A4 Assess Record Type		
A4.1 Are all records identified as a Record Type?	Yes, the Archive determines that records surveyed include meeting records. This record type is already known to the Archive. The Archive creates a **Record Type List** in the Submission Agreement which references **Record Type Records.** *The Archive skips Step A4.2*.	
A5 Assess Formats		
A5.1 Are any records in file formats that are not a preservation format?	Yes, the Archive determines the following: The meeting record consists of PDF, Word, Excel, MOV formats and the HTML wrapper consists of valid HTML and JPEG images. According to the Formats Standards Policy, PDF, JPEG, and HTML are preservation formats, while Word, Excel, and MOV are not. For the Word, Excel, and MOV files the Archive produces a **Format Transformation Plan**.	
A5.2 Should Archive transform or natively handle these formats?	The Archive will transform the Excel files, and handle the Word, MOV, PDF, JPEG, and HTML files natively. *Archive skips Steps A5.3 through A5.4.*	
A5.3 Identify needed Representation Information for new preservation format	The Archive will add the MOV image format to its **Formats Standards Policy** and create **Representation Information** for the format, adding it to its **Format Representation Information System**. Although it will handle the Word files natively, it will no make no effort to preserve these files because the records that exist in Word format also exist in PDF format. Therefore, the Archive will not add Word to its **Formats Standards Policy** or create **Representation Information** for the format.	
A5.5 Choose appropriate format	The Archive will transform the Excel files into comma-separated values format (CSV) files. The Archive updates its **Format Transformation Plan** to reflect this decision.	
A5.6 Is format chosen in A5.5 a Preservation Format the Archive already uses?	Yes, the Archive determines that the CSV format is a preservation formats based on its **Formats Standards Policy**.	
A6 Assess Identifier Rules		

A6.1 Is there a Producer naming/ identification scheme that needs accommodation?	No, the Archive determines there is no **Producer Naming/ Identification Scheme** that it needs to preserve or accommodate. The *Archive skips Steps A6.2 through A6.3.*	
A6.4 Determine appropriate naming/ identification scheme(s).	The Archive decides to use the standard **Archive Naming/ Identification Scheme**. The Archive records its **Naming/Identification Scheme Decision**.	
A7 Assess Copyright		
A7.1 Determine copyright status of records in Ingest Project	The Archive determines that the records are under copyright of Tufts University. The Archive documents the **Copyright status**.	
A7.2 Does the Archive need to acquire copyright or license for records?	No, the Archive determines that the University holds copyright. *Archive skips Steps A7.3 through A7.8.*	
A8 Assess Access Rights		
A8.1 Determine records' Records Security Profile	The Archive determines that the records are administrative records and are closed for 75 years from the date of creation and should have a **Records Security Profile** of closed access. The Archive documents this decision as a **Records Security Profile Decision** which references the **Records Security Profile**.	
A8.2 Does current security component meet the access control needs of the records?	Yes, the Archive determines the security component of the Preservation System meets the needs of the **Records Security Profile**. *Archive skips Steps A8.3 through A8.7.*	
A9 Assess Recordkeeping System		
A9.1 Has the Archive documented recordkeeping system as supporting feasible and trustworthy transfer?	No, the Archive has not previously examined the CDs and the Producer's method of managing the records on the CDs.	
A9.2 Can recordkeeping system support feasible and trustworthy transfer?	The Archive determines that the recordkeeping system can support the trustworthy transfer of records to the Archive, but not in a scalable manner. It can support a feasible transfer if the volume of records is low. The Archive produces a **Recordkeeping System Report.** *The Archive skips Step A9.3.*	
A9.4 Is Archive or Producer willing to take extraordinary measures to transfer records?	The Archive determines that it is willing to manually transfer the records from the recordkeeping system to the Archive because of the small volume of records involved in the Ingest Project. The Archive produces **SIP Creation**	

	Procedures for moving the records from the recordkeeping system to the Archive. *The Archive Steps A9.5 through A9.7.*	
A10 Assess Feasibility		
A10.1 Can the Archive feasibly accession the records?	Yes. The Archive determines that it is feasible to manage and preserve the records and its requirements without extraordinary effort or special accommodation. The Archive produces a **Preservation System Availability Statement**. *The Archive skips Steps A10.2 through A10.5.*	
A11 Finalize Submission Agreement		
A11.1 Add description of Metadata Encoding Rules to Submission Agreement.	The Archive chooses standard Dublin Core metadata encoding and documents its **Metadata Encoding Rules Decision**.	
A11.2 Add description of Transfer Procedures to Submission Agreement	The Producer has provided copies of the CDs to the Archive. The Archive will create the SIPs from the CDs at a convenient time. The Archive documents its **Transfer Procedures Decision** in the Submission Agreement.	The Office of the Board of Trustees has provided CDs to the Archive.
A11.3 Add description of Validation Procedures to Submission Agreement	The Archive documents the validation procedures for HTML, JPEG, PDF, Word, Excel, and MOV files in the **Validation Procedures Decision**.	
A11.4 Add Transfer Schedule to Submission Agreement	The Archive selects an indefinite schedule, and documents this decision in the **Transfer Schedule Decision**.	
A11.5 Add SIP Creation Procedures to Submission Agreement	The Archive chooses its standard SIP format and documents its **SIP Creation Procedures Decision**.	
A11.6 Finalize Submission Agreement	The Archive draws up the **Draft Submission Agreement** based on its previous decisions.	
A11.7 Does Archive and Producer agree to and approve the Submission Agreement?	Yes, the Archive and Producer agree and produce the **Finalized Submission Agreement**. The Archive submits finalized submission agreement to the ingest system. The Archive's ingest application accepts and validates a machine-readable version of the **Finalized Submission Agreement.** *Archive skips Steps A11.8 through A11.10.*	Yes, the Producer and Archive agree and produce the **Finalized Submission Agreement**.
B **Transfer and Validation**		
B1 **Create and Transfer SIPs**		

B1.1 Producer prepares SIP according to Submission Agreement		The Producer gathers the CDs containing meeting records and ensures that the CDs contain the appropriate records according to the **SIP Creation Procedures Decision.**
B1.2 Producer transfers the SIP to the Archive		The Producer manually delivers CDs containing meeting records to the Archive according to the **Transfer Procedures Decision**.
B2 **Validate**		
B2.1 Archive receives SIP from Producer	The Archive accepts the SIP from the Producer and extracts the data from the CDs as TAR files and places the files into the ingest application. The Archive then produces a **Documentation of Receipt** and delivers that to the Producer.	The Producer receives the **Documentation of Receipt**.
B2.2 Is SIP well formed?	The ingest application checks the SIP format. The SIP is well-formed. The Application updates the **SIP validity statement.**	
B2.3 Does SIP contain malicious code?	The ingest application scans the SIP components for viruses and other malicious code. All SIP components are clean. The application updates the **SIP validity statement**.	
B2.4 Is the submitter authorized to submit SIP to the Archive?	The ingest application validates the SIP signatures and validates identities against its database of certificates. The SIP is signed by an authorized person. The application updates the **SIP validity statement**.	
B2.5 Does SIP contain all necessary records components?	The ingest application checks all included records for completeness. All records in the SIP are complete. The application updates the **SIP validity statement**.	
B2.6 Do the record components in SIP validate?	The ingest application tests the record components for validity, where necessary. All record components validate. Application updates the **SIP validity statement**.	
B3 **Transform and Attach Metadata**		
B3.1 Do any of the records in SIP require transformation?	Yes, the Archive determines that the Excel files require transformation to CSV files according to the **Format Transformation Plan** in the Submission Agreement.	
B3.2 Perform transformation on records that require transformation	The Archive transforms the Excel files to CSV producing **Transformed Records**.	
B3.3 Attach to records metadata inferred from Submission Agreement	The ingest application attaches stock metadata from Submission Agreement, **Producer Record**, and **Record Type Records**, creating **Records with Attached Metadata**.	

B3.4 Attach to records the Records Security Profile defined by Submission Agreement	The ingest application attaches the **Record Security Profile** identified in the Submission Agreement, creating a **Records with Security Profile**.	
B4 **Formulate AIPs**		
B4.1 Formulate AIPs	The Ingest application creates an **AIP** for each record.	
B5 **Assess AIPs**		
B5.1 Are all of the records in the AIP part of accession described by Submission Agreement?	The Archive verifies that all of the records to be accessioned come from the CDs and produces an **AIP Validity Statement**. The ingest application accepts the **AIP Validity Statement**. *The Archive skips Steps B5.2 through B5.3.*	
B5.4 Is proper metadata attached to records in the AIP?	The Archive verifies that all of the records have sufficient and correct metadata and updates the **AIP Validity Statement**. The ingest application accepts the **AIP Validity Statement**. *The Archive skips Step B5.5.*	
B6 Formally Accession		
B6.1 Submit AIPs into Preservation Repository.	The ingest application submits the AIPs to the Preservation Repository.	
B6.2 Formally notify Producer that Archive has accepted and accessioned records described by Ingest Project.	The ingest application generates a **Transfer Notice** for the Producer and an entry in the **Accession Log**.	The Producer receives the **Transfer Notice**.

INGEST PROJECT THREE
University Library Council and University Library Council Teams

Ingest Project Narrative

The University Libraries Council (ULC) at Tufts University is composed of the directors of the six libraries at the Tufts. The ULC establishes and manages University-wide library policies. The ULC also oversees a number of inter-library teams that attend to a variety of library functions and issues. These teams include Staff Development, Fair Use, and Copier Contract Compliance, among others. The teams store their working documents, meeting minutes, bylaws, policies, and other documents in a web collaboration environment called SharePoint® Team Services (http://lib.tufts.edu), which acts as a recordkeeping system. The documents stored in lib are mostly MS Word documents with occasional Excel spreadsheets, HTML files, and plain text files. ULC and team members have read/write access to the records in SharePoint.

The project team chose this accession because it presents issues with troublesome format types and with records stored in a troublesome recordkeeping system. This is a situation common to recordkeeping systems at many universities.

The project team contacted Carol Johnson—director of University Library Technology Services, which manages the SharePoint instance for the ULC—about participating in the grant project. She was amenable to the pilot project. In order to accession the records of the lib system, the project team set up a comparable hardware and OS environment to the system which runs lib. Then the team developed tools which would allow us to extract records from the lib system. The team aimed to replicate lib onto its own hardware before running the extraction utilities to avoid running relatively untested code on their production server. It turned out that replicating the SharePoint instance was not easy. Furthermore, SharePoint stores almost no metadata, so preparing records for ingest required significant metadata reconstruction.

The file formats in this accession present a difficult but all-too-common problem; the data is largely stored in a proprietary format with no easy transformation route. Because the project team wished to model scalable processes, manual transformation (via loading into MS Office and saving in a different format) was not an option. Furthermore, such transformations usually lose information. The OpenOffice.org (OOo) project is able to correctly open and render a large class of MS Office documents and save them in a well-documented format, but again the project team wanted to avoid manual transformations. However, OpenOffice.org has a programming interface that allows interaction-free access to loading and saving capabilities. Unfortunately, this interface is clumsy and OOo was never designed to be used in an interaction-free way. All this make the process of using OOo difficult, but it appears to be the best option.

The project team acquired the data from the SharePoint instance and loaded it onto a server under our control. From there the team extracted the records of interest using custom-built tools to create a number of SIPs. Because the SharePoint system does not maintain metadata that is necessary to handle the records in an archival setting, the SIP extraction process involved applying rules to create metadata.

The project team fed these SIPs into the Tufts Ingest Prototype System (TIPS)[8] which is able to accession the records into a Fedora system in a limited way. The team used the OpenOffice suite to perform office document transformation. However, the version OpenOffice which the team configured to use with TIPS is an older version which does not support the target format (OpenDocument). Instead the team transformed the Word and Excel documents into PDF documents. Furthermore, because of limitations in Fedora 2.0, the project team had to encode all complex objects in Base64 and wrap them in an xml tag. This makes retrieving the files from Fedora cumbersome.

[8] See 2.3 Ingest Tools for a description of the application.

Survey Report

Background

Eliot Wilczek, University Records Manager, Digital Collection and Archives (DCA) called Carol Johnson, Director, University Library Technology Services (ULTS) on 06/08/05 to discuss the records on the SharePoint Team Services site at http://lib.tufts.edu. Lib serves as the recordkeeping system for the records of the University Library Council (ULC) and its teams. Lib also serves as the recordkeeping system for some ULTS records.

During this conversation Carol Johnson verbally agreed to allow the DCA to survey the records on Lib and transfer any records that it appraises to have archival value to the DCA.

On 10/28/05 Eliot Wilczek conducted a survey of the records on Lib. Mr. Wilczek excluded the ULTS records from the survey because they appeared to be a small and incomplete set of ULTS' records.

Description of Records Surveyed

General Description

SharePoint Team Services is the recordkeeping system for the records of the ULC and its teams, which include:

Collections and Licensing
Copier Contract and Compliance
Electronic Dissertations & Theses
Fair Use
Integrated Library System Implementation
Publicity and Marketing
Staff Development
Services Steering

The ULC and the teams go through a self-selection process when they post records to Lib. A high percentage of the records in Lib have archival value, but the records in Lib do not represent all of the records of the ULC and its teams.

Producer

University Library Technology Services

Producer Role(s) Custodian

Notes The *Creator* of these records is the University Library Council and its teams listed above. The ULC create and direct the teams—the teams are not independent entities. ULTS manages the records on behalf of the ULC and its teams and has the authority to execute their disposition. *Also note* as Director of the ULTS, Carol Johnson is a member of the ULC.

Function(s)

The records on Lib document the important decisions and actions of the ULC and its teams. Some records also serve as working files for the teams and ULC. Posting the records onto Lib allows all members of the teams and the ULC to share these records with each other.

Record Type(s)
Charges
Reports
Meeting Minutes
Subject Files

Format Type(s)
SharePoint Team Services presented as web pages via http protocol.

For all pages—based on small sample of pages

```
<html dir="ltr" xmlns:o="urn:schemas-microsoft-com:office:office">
```

and

```
<HEAD>
<META Name="GENERATOR" Content="Microsoft FrontPage 5.0">
```

No validation performed.

MS Word files.
No validation performed.

MS Excel files.
No validation performed.

Web pages composed in html. This does not include external web pages not on Lib.
No validation performed.

Executable programs—installers.
No validation performed.

File Size
Most of the records are smaller than 1mb. Many of them are significantly smaller than 1 MB. However, Lib has a few executable files (installers) that are 20 to 40 MB in size.

Producer Identifier
The records do not have producer identifiers that need preservation.

Confidentiality Requirements/Access Restrictions
All of the records have requirements for access restrictions. They should all have Category 3: General University Records. Creators and the Producer can have immediate access.

Copyright Status
Copyright of all the records are held by Tufts University.

List of Records Surveyed
Charges

Description	The ULC’s charge to the various teams. All charges should be in “Charges” directories.

Record Type	Charges
Dates	1999 through 2005
Format	MS Word

Reports

Description	Mostly periodic but some subject-oriented reports created by the ULC and its teams. Most reports should be in “Annual Reports” directories, some are in “Shared Documents” or other directories.
Record Type	Reports
Dates	1999 through 2005
Format	MS Word

Meeting Minutes

Description	Written records of decisions made and actions taken by the ULC and its teams at their respective meetings. All meeting minutes should be in “Minutes” directories.
Record Type	Meeting Minutes
Dates	1999 through 2005
Format	MS Word

Subject Files

Description	A wide variety of records that often serve as the working files of the ULC and its teams. Most subject files should be in “Shared Documents” directories.
Record Type	Subject Files
Dates	1999 through 2005
Format	MS Word, MS Excel, HTML, Executable programs—installers

SharePoint Team Services Instance

Description	A web-based file-sharing system used as a recordkeeping system for records created by the ULC and its teams.
Record Type	Recordkeeping System
Dates	2002 through 2005
Format	SharePoint Team Services

Evaluation of Authenticity

The Digital Collections and Archives judges the records on the SharePoint Team Services instance at http://lib.tufts.edu, which it has evaluated in this survey, to be authentic for the following reasons:

- At the 06/08/05 survey interview, Carol Johnson indicated that ULTS installed and maintains the SharePoint instance on behalf of the ULC and its teams in the normal course of their business.
- At the 06/08/05 survey interview, Carol Johnson identified the SharePoint instance at http://lib.tufts.edu as the recordkeeping system for the ULC and its teams.

- During the 10/28/05 survey, Eliot Wilczek determined the SharePoint instance at http://lib.tufts.edu was the same that Carol Johnson identified on 06/08/05 based on a brief visual review of the SharePoint instance, its directory of files and the content of a sample of the records it contained.
- The probability someone would maliciously alter the SharePoint instance or any of the individual records managed by that instance—particularly their content—and try to hide that content alteration is extremely low because no one has a reasonable motivation to undertake such an action.

Appraisal Decision

For all records that have a disposition of "Transfer to DCA," it is critical to preserve the information of the file path. The file structure—which serves as a de facto taxonomy—will provide essential information for post-processing descriptive metadata work, particularly assigning records to the correct collections and series.

This instance of SharePoint is an active system; appropriate staff members add records to Lib continuously and can update records already in Lib at anytime. The records are arranged with no chronological cut-offs. Appraisal of these records will have to occur at systematic time intervals so the DCA can track which records it has previously appraised and accessioned.

Charges				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen042
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *Verbatim* MS Word Structure: *General Appearance*
			Signs	None
		Annotations		None
		Context		*Information* that managed and shared in SharePoint Team Services instance via http delivered as MS Word documents with restrictions limiting access to ULC and Team members.
		Medium		*Information* that MS Word
Reports				
	Disposition			Transfer to DCA
	Reason for Disposition			Retention Schedule gen065
	Essential Elements			
		Documentary Form		
			Content	*Verbatim*
			Presentation	Textual Structure: *Verbatim* MS Word Structure: *General Appearance*
			Signs	None

Annotations	None
Context	*Information* that managed and shared in SharePoint Team Services instance via http delivered as MS Word documents with restrictions limiting access to ULC and Team members.
Medium	*Information* that MS Word
Meeting Minutes	
Disposition	Transfer to DCA
Reason for Disposition	Retention Schedule gen025
Essential Elements	
Documentary Form	
Content	*Verbatim*
Presentation	Textual Structure: *Verbatim* MS Word Structure: *General Appearance*
Signs	None
Annotations	None
Context	*Information* that managed and shared in SharePoint Team Services instance via http delivered as MS Word documents with restrictions limiting access to ULC and Team members.
Medium	*Information* that MS Word
Subject Files	
Disposition	Transfer to DCA. Many Subject Files do not have permanent value but are so thoroughly mixed in with subject files of archival value that all subject files should be transferred to the DCA. The DCA will make no effort to preserve the Installers because none of them have archival value.
Reason for Disposition	Retention Schedule gen024
Essential Elements	
Documentary Form	
Content	*Verbatim,* includes *Data Integrity* for records in MS Excel
Presentation	Textual Structure: *General Appearance* MS Word Structure: *General Appearance* MS Excel Structure: *General Appearance* HTML Structure: *General Appearance* *Information* of hyperlink URL, hyperlink functionality not needed Installers: *None*
Signs	None
Annotations	None
Context	*Information* that managed and shared in SharePoint

	Team Services instance via http delivered as MS Word documents with restrictions limiting access to ULC and Team members.
Medium	*Information* that either MS Word, MS Excel, HTML, or executable applications—installers.
Post-Processing Notes	Determine if DCA needs to remove and destroy non-archival records from subject files.
SharePoint Team Services instance	
Disposition	Retire when no longer needed and all records it contains are transferred to the DCA or other disposition has been executed. Destroy all data when instance retired. However, retain *information* that the Producer stored and managed the records described above in a SharePoint Team Services instance at http://lib.tufts.edu with read/write access for members of the ULC and its teams.
Reason for Disposition	The SharePoint instance in and of itself is not a record and its functionality is not an essential element of any of the record it contains.
Essential Elements	
Documentary Form	
Content	None
Presentation	None
Signs	None
Annotations	None
Context	I*nformation* that Producer stored and managed surveyed records in a SharePoint Team Services instance at http://lib.tufts.edu with read/write access for members of the ULC and its teams.
Medium	None

Ingest Guide Steps: Archive and Producer Actions

Steps	Archive Actions	Producer Actions
A Negotiate Submission Agreement		
A1 Establish Relationship		
A1.1 Initiate Contact/Ingest Project	The Archive contacts the Director of the University Library Technology Services (ULTS) to initiate an Ingest Project concerning the records of University Library Council and its teams stored and managed on an instance of SharePoint Team Services (lib.tufts.edu).	The Director of the University Library Technology Services (ULTS) responds to the Archive's contact and agrees to initiate an Ingest Project.
A1.2 Identify Producer	The Archive identifies the ULTS as the Producer.	
A1.3 Has the Archive already defined its relationship with the Producer?	No, the Archive has not defined its relationship with the Producer.	
A1.4 Is this the Appropriate Archive?	Yes, the Archive determines it is the appropriate Archive for the ULC, its teams, and the ULTS. *The Archive skips Step A1.5.*	
A1.6 Collect and document information about Producer	The Archive collects information about ULC, its teams, and the ULTS and creates a **Producer Record** for each entity, and produces a **Producer Entry** in the Submission Agreement which references the appropriate **Producer Record**.	
A2 Define Project		
A2.1 Identify records at issue, agreeing upon scope of survey	The Archive and Producer identify the records of the ULC and its teams in the SharePoint instance at http://lib.tufts.edu managed by the ULTS as the records of this Ingest Project.	The Archive and Producer identify the records of the ULC and its teams in the SharePoint instance at http://lib.tufts.edu managed by the ULTS as the records of this Ingest Project.
A2.2 Does the Producer have custody/authority over the records identified in A2.1?	Yes, the Archive determines that the ULTS has authority over the records identified in Step A2.1 because it manages the records on behalf of the ULC and its teams and has the authority to execute their disposition. *The Archive skips Steps A2.3 through A2.6.*	
A3 Collect Information and Assess Value of Records		
A3.1 Conduct Records Survey, note attributes of records	The Archive conducts a survey and produces a **Survey Report**.	The Producer gives the Archive access to the SharePoint instance so it can conduct its survey.

A3.2 Judge authenticity of records	The Archive examines content of records in the SharePoint instance and determines that they are very probably authentic, and updates the **Survey Report**.	
A3.3 Should the Archive accession at least some of the records?	Yes, the Archive determines that the ULC and team records have archival value, and judges them to be authentic. The Archive updates the **Survey Report** to reflect this decision.	
A3.4 Determine the essential elements of the records that should be accessioned	The Archive examines records and determines that the essential elements of the minutes, reports, subject files, and charges are their content, textual formatting, and the information that they were managed in a SharePoint Team Services recordkeeping system with read and write access to members of the ULC and its teams. The Archive adds its determinations to the **Survey Report**.	
A4 Assess Record Type		
A4.1 Are all records identified as a Record Type?	Yes, the Archive determines that records surveyed include Meeting Minutes, Charges, Reports, and Subject Files. All are record types already known to the Archive. The Archive creates a **Record Type List** in the Submission Agreement which references **Record Type Records.** *The Archive skips Step A4.2*.	
A5 Assess Formats		
A5.1 Are any records in file formats that are not a preservation format?	Yes, the Archive determines the following: MS Word, MS Excel, and non-valid HTML are not preservation formats according to the **Formats Standards Policy**. For the Word, Excel, and non-valid HTML files the Archive produces a **Format Transformation Plan**.	
A5.2 Should Archive transform or natively handle these formats?	The Archive will transform the non-valid HTML, Word, and Excel files. *Archive skips Steps A5.3 through A5.4.*	
A5.5 Choose appropriate format	The Archive will transform the non-valid HTML into valid XHTML; it will transform the Word into PDF; it will transform the Excel into PDF. The Archive updates its **Format Transformation Plan** to reflect this decision.	
A5.6 Is format chosen in A5.5 a Preservation Format the Archive already uses?	Yes, the Archive determines that valid XHTML and PDF is a preservation format based on its **Formats Standards Policy**.	
A6 Assess Identifier Rules		

A6.1 Is there a Producer naming/ identification scheme that needs accommodation?	No, the Archive determines there is no **Producer Naming/ Identification Scheme** that it needs to preserve or accommodate. The *Archive skips Steps A6.2 through A6.3.*	
A6.4 Determine appropriate naming/ identification scheme(s).	The Archive decides to use the standard **Archive Naming/ Identification Scheme**. The Archive records its **Naming/Identification Scheme Decision**.	
A7 Assess Copyright		
A7.1 Determine copyright status of records in Ingest Project	The Archive determines that the records are under copyright of Tufts University. The Archive records the **Copyright status**.	
A7.2 Does the Archive need to acquire copyright or license for records?	No, the Archive determines that the University holds copyright. *Archive skips Steps A7.3 through A7.8.*	
A8 Assess Access Rights		
A8.1 Determine records' Records Security Profile	The Archive determines that the records are administrative records and are closed for 20 years from the date of creation and should have a **Records Security Profile** of closed access. The Archive documents this decision as a **Records Security Profile Decision** which references the **Records Security Profile**.	
A8.2 Does current security component meet the access control needs of the records?	Yes, the Archive determines the security component of the Preservation System meets the needs of the **Records Security Profile**. *Archive skips Steps A8.3 through A8.7.*	
A9 Assess Recordkeeping System		
A9.1 Has the Archive documented recordkeeping system as supporting feasible and trustworthy transfer?	No, the Archive has not previously examined the http://lib.tufts.edu instance of SharePoint and the Producer's method of managing the records on that SharePoint instance.	
A9.2 Can recordkeeping system support feasible and trustworthy transfer?	The Archive determines that the recordkeeping system can support the trustworthy transfer of records to the Archive, but not in a scalable manner. It can support a feasible transfer if the volume of records is low. The Archive produces a **Recordkeeping System Report.** *The Archive skips Step A9.3.*	
A9.4 Is Archive or Producer willing to take extraordinary measures to transfer records?	The Archive determines that it is willing to manually transfer the records from the recordkeeping system to the Archive because of the small volume of records involved in the Ingest Project. The Archive produces **SIP Creation**	

	Procedures for moving the records from the recordkeeping system to the Archive. *The Archive Steps A9.5 through A9.7.*	
A10 Assess Feasibility		
A10.1 Can the Archive feasibly accession the records?	Yes. The Archive determines that it is feasible to manage and preserve the records and its requirements without extraordinary effort or special accommodation. The Archive produces a **Preservation System Availability Statement**. *The Archive skips Steps A10.2 through A10.5.*	
A11 Finalize Submission Agreement		
A11.1 Add description of Metadata Encoding Rules to Submission Agreement.	The Archive chooses standard Dublin Core metadata encoding and documents its **Metadata Encoding Rules Decision**.	
A11.2 Add description of Transfer Procedures to Submission Agreement	The Producer has allowed the Archive to create an inactive copy of the http://lib.tufts.edu SharePoint instance on a server controlled by the Archive. The Archive will create the SIPs from the SharePoint instance it manages at a convenient time. The Archive documents its **Transfer Procedures Decision** in the Submission Agreement.	The Producer allows the Archive to create an inactive copy of the http://lib.tufts.edu SharePoint instance on a server controlled by the Archive.
A11.3 Add description of Validation Procedures to Submission Agreement	The Archive documents the validation procedures for HTML, Word, and Excel files in the **Validation Procedures Decision**, which becomes part of the **Submission Agreement**.	
A11.4 Add Transfer Schedule to Submission Agreement	The Archive selects an indefinite schedule, and documents this decision in the **Transfer Schedule Decision**.	
A11.5 Add SIP Creation Procedures to Submission Agreement	The Archive chooses its standard SIP format and documents its **SIP Creation Procedures Decision**.	
A11.6 Finalize Submission Agreement	The Archive draws up the **Draft Submission Agreement** based on its previous decisions.	
A11.7 Does Archive and Producer agree to and approve the Submission Agreement?	Yes, the Archive and Producer agree and produce the **Finalized Submission Agreement**. The Archive submits finalized submission agreement to the ingest system. The Archive's ingest application accepts and validates a machine-readable version of the **Finalized Submission Agreement.** *Archive skips Steps A11.8 through A11.10.*	Yes, the Producer and Archive agree and produce the **Finalized Submission Agreement**.
B **Transfer and Validation**		

B1 **Create and Transfer SIPs**		
B1.1 Producer prepares SIP according to Submission Agreement		The Producer allows the Archive to make an inactive copy the Producer's http://lib.tufts.edu instance of SharePoint on a server managed by the Archive. The Archive then extracts records from the SharePoint instance it controls and constructs the SIPs. The Archive signs the SIP with its own digital signature.
B1.2 Producer transfers the SIP to the Archive		The Archive places the SIP in the Ingest application drop-box.
B2 **Validate**		
B2.1 Archive receives SIP from Producer	The Ingest application accepts the SIP from the drop-box and produces **Documentation of Receipt** and delivers that to the Producer.	The Producer receives the **Documentation of Receipt**.
B2.2 Is SIP well formed?	The ingest application checks the SIP format. The SIP is well-formed. The Application updates the **SIP validity statement.**	
B2.3 Does SIP contain malicious code?	The ingest application scans the SIP components for viruses and other malicious code. All SIP components are clean. The application updates the **SIP validity statement**.	
B2.4 Is the submitter authorized to submit SIP to the Archive?	The ingest application validates the SIP signatures and validates identities against its database of certificates. The SIP is signed by an authorized person. The application updates the **SIP validity statement**.	
B2.5 Does SIP contain all necessary records components?	The ingest application checks all included records for completeness. All records in the SIP are complete. The application updates the **SIP validity statement**.	
B2.6 Do the record components in SIP validate?	The ingest application tests the record components for validity, where necessary. All record components validate. Application updates the **SIP validity statement**.	
B3 **Transform and Attach Metadata**		
B3.1 Do any of the records in SIP require transformation?	Yes, the Archive determines that the HTML files require tidying and transformation to valid XHTML and the Word and Excel files require transformation to PDF according to the **Format Transformation Plan** in the Submission Agreement.	
B3.2 Perform transformation on records that require transformation	The Archive tidies and transforms the HTML files into valid XHTML and transforms the Word and Excel files into PDF files, producing **Transformed Records**.	

B3.3 Attach to records metadata inferred from Submission Agreement	The ingest application attaches stock metadata from Submission Agreement, **Producer Record**, and **Record Type Records**, creating **Records with Attached Metadata**.	
B3.4 Attach to records the Records Security Profile defined by Submission Agreement	The ingest application attaches the **Record Security Profile** identified in the Submission Agreement, creating a **Records with Security Profile**.	
B4 **Formulate AIPs**		
B4.1 Formulate AIPs	The Ingest application creates an **AIP** for each record.	
B5 **Assess AIPs**		
B5.1 Are all of the records in the AIP part of accession described by Submission Agreement?	The Archive verifies that all of the records to be accessioned come from the SharePoint instance at http://lib.tufts.edu and produces an **AIP Validity Statement**. The ingest application accepts the **AIP Validity Statement**. *The Archive skips Steps B5.2 through B5.3.*	
B5.4 Is proper metadata attached to records in the AIP?	The Archive verifies that all of the records have sufficient and correct metadata and updates the **AIP Validity Statement**. The ingest application accepts the **AIP Validity Statement**. *The Archive skips Step B5.5.*	
B6 Formally Accession		
B6.1 Submit AIPs into Preservation Repository.	The ingest application submits the AIPs to the Preservation Repository.	
B6.2 Formally notify Producer that Archive has accepted and accessioned records described by Ingest Project.	The ingest application generates a **Transfer Notice** for the Producer and an entry in the **Accession Log**.	The Producer receives the **Transfer Notice**.