Appendix E: Sample Processing Plans

1. University of Hull: Stephen Gallagher Processing Plan

UNIVERSITY OF Hull

Processing Plan

Acc No: 2010/15 Ref: U DGA

born-digital archives

OVERVIEW			
Collection Title:	Stephen Gallagher		
Creator / Depositor:	Stephen Gallagher		
Related Material at HUA: Paper archives already deposited - 2008/10 (42 boxes) – mainly paper with a few boxes of publications, copies of DVDs etc - 2010/14 (12 boxes) – further publications (foreign editions etc) and production material Not tackled – blog / website (possibly recommend the British Library Web Archive) and email			
Brief Description of the material: Material relates to his writing, (short-stories, novels, radio and screen) including research process, drafts etc. Also material relating to his blog / website with some publicity/promotional material. There are only isolated email messages (no mailboxes).			
Extent:	**13.6 GB**	**No of files:**	**14,320 ***
Comments re extent: There are also 39 3" Amstrad discs			
ARCHIVAL DESCRIPTION			

Proposed level of archival description to be applied:

- Primarily at series level

Justification:

Stephen Gallagher considers each piece of work as a discrete project. Interest in the material is likely to be on two accounts:

- writing process following a particular story from idea through research, drafts, pitching and completion (whether publication of novel or filming of screenplay etc)
- a particular piece of work

This means that if describe the project we do not necessarily need to describe particular content

Cataloguing Priority for this accession:

1.	Research potential	**3**	5. Education potential	**2**	**Priority Score:**
2.	HHC specialist area	**3**	6. Community/outreach potential	**1**	**18** / 24
3.	Topicality / time crucial	**1**	7. Summary list is sufficient	**3**	
4.	UoH teaching potential	**2**	8. Complexity of cataloguing	**3**	

Scoring: 3 = high, 2 = medium, 1 = low 0 = no potential

APPRAISAL

Is appraisal necessary? ~~**Yes**~~ **| No |** ~~**N/A**~~

Potential for appraisal?

Initial investigations identified very little material that could or should be appraised

ARRANGEMENT

Integrate with existing arrangement? **Yes |** ~~**No**~~ **|** ~~**N/A**~~

Does the current arrangement include b-d material? ~~**Yes**~~ **| No |** ~~**N/A**~~

Justification:

There is considerable overlap between paper and born-digital material

Potential arrangement issues?

- Paper files being catalogued at file level – need to consider implications for discovery & access
- To not try to describe each born-digital item but include an overview of born-digital material within the series description

Any restricted / sensitive content?

- Some personal material (e.g., references for 3rd parties) that should be closed
- Suggest that most recent work (i.e., last x years) should be closed [discuss this with SG]
- *ResearchDocs* folder (1226 files in 87 folders, 14.5MB) material is mostly saved web-pages – need to consider arrangement /access issues
- *MyRadio* folder (44 files, 1.85GB) recorded broadcasts can be included in the archives but are subject to copyright so should not be made available online via repository

PRESERVATION

Media issues:

- Main body of material was selected by SG and transferred via external hard drive
- There are 39 3" Amstrad discs that cannot be read with current hardware

Content issues:

- 291 files in *FinalDraft* format (*.fdr) contact Mary-Jane Dickenson (Drama) to use their copy of *FinalDraft* – looked at files (June/July 2011) and created PDF copies for access
- How to present the old website content to users as web pages (via a web browser etc) rather as individual unlinked pages

Proposed preservation actions:

Import the FinalDraft PDFs and attach to the original *.fdr file

Plan produced by: Simon Wilson **Date: 13th Sept 2011**

Suggested Review Date:

2. Stanford University: Gould Processing Plan

Stephen Jay Gould papers.

Bio/Scope & Content:

Influential American paleontologist, evolutionary biologist and historian of science, Gould began his career at Harvard University in 1967 and worked until his death in 2002. One of the most popular science writers of our time, he is the author of 22 books, 479 peer-reviewed scholarly papers, 300 essays, and 101 reviews.

Scenario in 2009:

At the time of the AIMS grant, the Gould collection consisted of 8 accessions acquired between 2004 and 2010. Totaling over 500 linear feet of material, the collection contains specimens and legacy computer media. Items (159) of computer media were "recorded" during the accessioning process. Since then, we have uncovered more computer media (21 more sets of computer punch cards) in the 2008 accession and odds and ends scattered within folders throughout the accessions.

Media enumerated initially consisted of: 60 5.25-inch floppy diskettes, 81 3.5-inch floppy diskettes, two cartons of computer punch cards and 3 computer tapes from 1987, 1988, and 1994. The diskettes contain bibliographic databases and working drafts of many of Gould's publications. The punch cards and the data tapes appear to contain datasets used in his evolutionary biology research.

There are no online guides to any of the collection although rough container lists were created when the collection was packed up initially. The papers, audio & video are being processed concurrently.

Catalog record states: "Collection in process but open for research. Some materials may not be available. Preliminary container list available."

Trials/Actions taken:

Capture:

8 sets of punch cards (from one carton) were migrated by Computer History Museum, Mountain View, California and stored on DVD. This DVD was labeled Computer Media #144. One small set of punch cards was unreadable because there was no sorting key. Three computer tapes and 6 cartons of punch cards have not been migrated at this time (approx. 24 sets). Diskettes were labeled and numbered beginning with "Computer Media 001" or cm01. Photographic images of the diskettes and existing labels were taken for subsequent access by users.

First trial:

Disk images of floppy diskettes were created using ImageTool and a Catweasel in FRED. [A Catweasel is just an interface card for computer which don't have a floppy interface in the motherboard. Write-blocking is enabled by putting a tape at the "write-protect" area in a 5.25 inch floppy disk.] However, ImageTool did not generate an audit log file to confirm successful imaging nor a file listing of the disk contents. Our second attempt utilized an old personal computer with on-board floppy disk controller was used to image the diskettes using free software called FTK Imager. Outputs from FTK Imager include: disk images, audit log files to confirm successful imaging and file listings of the diskette contents. [Peter could not find a motherboard with floppy disk controller and interface on sale in May 2010 when I tried to do the imaging. So he brought his old computer in to do the imaging. He discovered the Gigabyte motherboard which had a floppy disk interface in Feb 2011 and built the capture station that winter.]

These were stored in a stand-alone personal computer. After detecting and cleaning computer viruses using Sophos Anti-Virus, the files were transferred to Stanford Powervault (a secured server with regular backup schedule). Only "cm94" (a high-density, 3.5-inch diskette) contained a virus which was removed.

Unreadable media (loss was 6%): CM001-CM003 (single-sided single-density 5.25-inch diskette) unreadable with existing equipment; no files copied. CM035 (double-sided high-density 5.25-inch diskette) sustained physical damage before transfer to Stanford and no files copied.

Processing

First Trial - Processing using Windows Explorer:

Quickview Plus was used to view the content of the files. Folders were created that mirrored "series" and "subseries" in EAD and files were moved from original media folder into appropriate place using Windows Explorer. This however changed metadata associated with the files – such as original file path, etc. Adobe Acrobat Professional was used to convert files in obsolete file formats such as WordPerfect, MS DOS Word, etc. to PDF/A for access. The PDF/A version of the original files provide files with current format which can be accessed with current software. This version of the original files do not contain the original file creation dates. The file creation dates of the PDF/A files are the dates when the files were converted. The conversion also alter the last accessed dates of the original files.

Second Trial - Processing using AccessData FTK:

Logical images were created the second time around. After hearing from the curator that Creeley had deleted files on purpose that he did not want kept, Peter created logical images of the files on the floppy diskettes.

FTK extracted technical metadata (file size, creation, last modification and last accessed dates, file format, checksum, etc.) of the files in the disk images loaded. "File Category" provided a summary of how many files are in different file formats. The interface to hide the duplicate files was activated so that users are working on unique files (FTK uses the checksums of the files to identify duplicate files.). Restricted content such as credit cards, social security number, student grades, etc. were identified using the pattern & full-text searches functions. The files identified were flagged as "Privileged". Although the search may not find ALL "Restricted" contents, it is a much better alternative to read all files. Bookmarks were created with names mirrored "series" and "subseries" in EAD. The embedded viewer (reads over 200 file formats) was used to view files with obsolete file formats. Files are then assigned to bookmarks according to intellectual contents individually or in batch. Although FTK did not forbidden the assignment of one file to more than one bookmarks, the system would change the color of the file name and its associated metadata from black to purple after the file was assigned to one bookmark. This could act as a reminder that which files had been assigned to bookmarks. "Labels" were used to represent access restrictions, document types, computer media type, and subject headings. Reports in XML/HTML format are generated to export files to access repository (Hypatia). The files carried the bookmarks, labels, privileged flag, and technical metadata with them.

EAD draft excerpts (see below)

Outstanding AIMS Project work:

- Data modeling for Gould data and metadata including EAD
- Complete EAD description for b-d materials (currently listed as Series VI – Scope & Content, Arrangement and Physical Description notes only)

- Determine delivery of b-d material, possibly by file format? – files and vehicle (Hypatia)
    - Text: manuscript writings, correspondence
    - Data sets: will be described as part of the Gould finding aid in a separate series [?] and include a live link to their digital surrogates, which will be deliverable as individual file downloads.
- Determine use/delivery of photographic images of original media labels if any
- Publish online guide in September along with paper components
- Awaiting capture of last batch of punch cards from CHM
- FOUND: 5 more cartons of punch cards as of June 2011 in 2008 accession – need to codify methodology for reading punch cards – either 1) work out exchange with CHM & quicker turn around, 2) use CHM equipment to read ourselves, or 3) costs for outsourcing
- A selection of born-digital materials will be delivered via Hypatia (demo instance)
- The catalog record will be updated with links to online guides and born-digital instance

Non-AIMS Updates

- Gould's papers will be fully processed by 8/31/11 – including all artifacts and specimens.
- The online guide to the papers will be posted online at Stanford and the Online Archive of California with series level description re born-digital materials and link to:

Series VI: Stephen Jay Gould Born-Digital Material

PHYSICAL DESCRIPTION: 52 megabytes (1,180 files)

FILE TYPES AND FORMATS

File Types: Computer Program; Data set; Document; Speadsheet. File Formats: ASCII Text; WordPerfect 4.2, 5.0, 5.1, 6.0, 6.1; Microsoft Word 2.0, 6.0, 97, 2000; Microsoft RTF; Microsoft Excel 4.0; Lotus 1-2-3 2.0

FINDING AID LINK: To cite or bookmark this finding aid, use the following address:
http://hdl.handle.net/10079/fa/

Access

Collection is open for research; digital material is available online; other materials must be requested at least 48 hours in advance of intended use.

File types and formats

File Types: Computer Program; Data set; Document; Speadsheet. File Formats: ASCII Text; WordPerfect 4.2, 5.0, 5.1, 6.0, 6.1; Microsoft Word 2.0, 6.0, 97, 2000; Microsoft RTF; Microsoft Excel 4.0; Lotus 1-2-3 2.0

Scope and Contents

This series consists primarily of the born digital material from the Stephen Jay Gould (SJG) papers. The born digital material was stored in floppy diskettes, tapes and punch cards. The original labels, if any, on the computer media are

in many cases too brief to identify the contents of the diskettes. The processor viewed the contents of each file to determine to what category the file belonged. Since SJG divided his works into "Articles", "Abstracts, Reviews, Letters, etc.", "Natural History Column", and "Books" in his bibliography, the processor followed this arrangement and added "Bibliography & Curriculum Vitae", "Teaching", "Rare Books", "Punch Cards", "Misc.", and "Computer Media Photos" as other subseries.

Details of the ten categories of files are as follows (these are added as LABELS in FTK and will display as FACETS in Hypatia):

- Articles (99 files)
- Abstracts, Reviews, Letters, etc. (107 files)
- Natural History Columns (171 files)
- Books (drafts of 12 books written by SJG in 404 files):
    - The Structure of Evolutionary Theory,
    - Full House,
    - The Book of Life,
    - Triumph and Tragedy in Mudville,
    - Dinosaur in a Haystack,
    - The Burgess Shale and the Nature of History,
    - Time's Arrow, time's Cycle,
    - The Lying Stones of Marrakech,
    - Eight Little Piggies,
    - Hidden Histories of Science,
    - The Hedgehog, the Fox, and the Magister's Pox
    - The Mismeasure of Man
- Bibliography & Curriculum Vitae (44 files)
- Teaching (12 files)
- Rare Books (28 files)

- Data Sets (11 files)
  Re: computer programs and data migrated from one box of punch cards. Data in another box of punch cards is not migrated. [21 more sets discovered in 2008 addenda; unread]
- Miscellaneous (18 files) - divided into 3 sub-groups:
  - National Science Foundation (NSF)
  - Paleontological Society
  - Miscellaneous
- Computer Media Photos (165 files)

Processing Information:
Logical images of the files in floppy diskettes were created using FTK Imager and stored in a standalone personal computer. After detecting and cleaning computer virus using Sophos Anti-Virus, the cleaned files were transferred to Stanford Powervault (a secured server with regular backup schedule).

FTK Toolkit was used to assign access rights, identify restricted materials, assign series subseries information and other descriptive metadata, and generate technical metadata (MD5 checksum, file format, etc.) The files with all the metadata have been transferred to Hypatia (Hydra Platform for Access To Information in Archives).

All files will be ingested into the Stanford Digital Repository (SDR; a dark digital archive) for long term preservation. One box of punch cards was migrated by Computer History Museum, Mountain View, California, USA and stored in DVD. The DVD is assigned as Computer Media #144. One small set of punch cards was unreadable because the sorting order of the cards were mixed up. Three computer tapes and one other box of punch cards have not been migrated at this time.

Unreadable media: Computer Media #1-3 (Single sided single density 5.25 inch. floppy) unreadable with existing equipment; no files copied. Computer Media #35 (Double sided high density 5.25 inch. floppy) physical damage; no files copied. Computer Media #39 (Double sided double density 5.25 inch. floppy) blank diskettes. Computer Media #60, 134, 135 (High density, 3.5 inch. floppy) blank diskettes. Computer Media #94 (High density, 3.5 inch. floppy) contained virus and was cleaned using Sophos Anti-Virus.

3. University of Virginia: Cheuse Papers Processing Plan

University of Virginia
Processing Plan
Collection 10726, The Papers of Alan Cheuse

Collection Name:	The Papers of Alan Cheuse
Collection Date:	Ca. 1950 – 2009
Collection Number:	10726; accessions _ through al
Extent (pre-processing):	83 disks (3.5" and CD) approx. 5.31 MB; ca. 80 linear feet
Types of materials:	3.5" disks and CDs, video cassettes and DVDs, paper manuscripts
Custodial History:	Alan Cheuse placed the papers on loan to the Library beginning in 1987. Earlier accessions were then purchased in 2003 with a commitment to purchase further groups.
Restrictions from Donors:	Explicit digital rights have yet been discussed. Four series (Accessions 17, 18, 20, and 21) are restricted from access until 2012.
Separated Materials:	Disks have been separated from the manuscript drafts and are stored with the other media and a/v.
Related Materials:	None
Preservation Concerns:	None
Languages other than English:	None
Overview of Contents:	This collection consists of the papers of the American author, book reviewer, and George Mason University professor, Alan Cheuse. These papers include manuscripts for articles, speeches, interviews, and short stories; book reviews; screen plays; cassette tape recordings; computer disks; video cassette & DVD; printed material; contracts and royalties; passports; photographs and drawings; correspondence; research material; short stories by other authors; appointment calendars; short stories and book manuscripts.
Existing Order and description:	Sixteen of the thirty-two accessions have been processed separately, as per institutional practices. They are described in both EAD finding aids and MARC records. They are each organized by type of writing (correspondence, topical files, novel manuscripts, review manuscripts, etc.) to the folder level. The other 16 accessions are recorded in MARC records at varying degrees of detail, some with no more than a title, date, and generic note. All computer media has been separated, numbered, and is referenced in finding aids and records, but has mostly not been processed. The contents of some disks were printed and filed with paper manuscripts. Seven of the accessions contain computer disk materials. Only one of these accessions has been described in an EAD finding aid.

Desired Processing:	All computer media should be processed. Additionally, all accessions should be combined into a single finding aid. Where EAD exists, these records will be combined into a single <archdesc> and <dsc> with each accession being represented as a series. The accessions represented by MARC records will be converted to series components. In addition, subject headings, which were not included in the original EAD, should be added from all MARC records. No further work will be done with paper materials at this time. The processor will create disk images of the disks and then process using FTK. Disks containing commercial works that were used for research purposes should not be imaged or stored at this time. Individual files will be labeled with the disk number so that they may later be associated with the correct container element in the EAD. Titles of individual works will be added to the finding aid so that some reference to the works available on the disks is present. This is to match the level of processing of the paper manuscripts, which are indicated by name within the collection descriptions. Files containing confidential information will be completely restricted at this time. Obsolete file formats will not be migrated at this time, but this work should be considered in the future. Access to materials on the disk will be at the individual file level. After imaging the disk a copy of the image will be transferred to the StoreNext preservation store. Copies of the unrestricted files will be added to the Hypatia repository for public access. The disk images will be referenced by identifier number within the ead. They will exist as individual subcomponents of the accession or sub-series (if it exists) and the disk number will be referenced in a "unitid" attribute. The finalized finding aid will also be uploaded into the Hypatia repository and the individual files will be linked to the accession or container they belong to.
Next steps	Reprocessing all accessions into one collection arranged intellectually, rather than intellectually within individual accessions, is recommended for the future when the collection is deemed "complete." As technology and infrastructure develop, migration of obsolete formats and redaction within restricted files in order to make them available should also be undertaken.
Notes to Processors:	Examine the contents of the CDs later in the series to determine which are simply copies of commercially produced works and do not need to be imaged.
Anticipated Time for Processing:	5 days

4. Yale University: Tobin Collection Processing Plan

Processing Work Plan

Institution: MSSA
Archivist: Mark A. Matienzo
Date: June 7, 2011
Collection title: James Tobin papers
Creator: Tobin, James
Current call number(s): MS 1746, **Accession 2004-M-088**
Provenance: Gift of Elizabeth Tobin, 2004.
Extent: 8.75 linear feet; 27 3.5" inch diskettes (35.7 MB)

Overview:
Research strengths: correspondence regarding professional activities; working and final drafts of conference papers, periodical columns, and other publications.

Types of electronic records present: Correspondence (e-mail and computer-written letters); writings; spreadsheets and graphs; office files (biographical statements, calendars, publication lists, etc.), course materials. Files are primarily WordPerfect and Lotus 1-2-3; some Quicken files exist; e-mail is in text form, either in Eudora mailboxes individually saved text files.

Significant preservation concerns: See file formats above. Most significant concern is Lotus 1-2-3 files; several should be considered compound objects with graphs and formatting information.

Description:

Current: Minimal. Labels from individual diskettes have been transcribed as component titles within finding aid.

Proposed enhancement: Description should follow executed organization as specified below.

Recommended description work for later: see under organization.

Organization:

Current: Hard to determine. Paper records do not seem to have a coherent overall organization, with the exception of the correspondence; however, correspondence is still scattered between "Letters to Jim," "Professional Correspondence", "Nobel Prize Correspondence," and "Personal Correspondence." Writings are very disorganized;

Diskettes appear to be used as transfer media for files between his office, his home, and his cottage in Wisconsin. A few disks, or sets thereof, show some grouping based on type of records, such as "office files" (publication lists, telephone lists/address books) and letters that Tobin wrote in WordPerfect. Writings are not grouped together thematically.

Proposed arrangement: Arrangement should be based on record types. Within the electronic records for this accession, logical groupings and subgroupings are as follows:

- Correspondence, 1992-2001 and undated
    - Correspondence written using WordPerfect, 1992-2000
    - E-mail, 1996-2001 and undated
- Course materials for Economics 480B, 1998
    - Lotus 1-2-3 spreadsheets, 1992-1997

    - "Primer" spreadsheets and graphs, 1996-1997
- Office files, 1995-2001
    - Biographical statements
    - Calendars
    - Lists of Tobin's publications
    - Quicken files
    - Recommendation letters and lists of recommendations
    - Telephone lists
- Writings, 1992-2001

Of all groupings, the Writings grouping would need the most considerable organization and description. In the short term I recommend either not listing individual files, or listing individual files with filename and date only.

Recommended arrangement work for later: Combine paper records and electronic records into a common arrangement. Considerable attention to Tobin's personal papers is needed, especially those related to his military service. Arrange writings alphabetically by title, identify explicit drafts, and reconcile against publication lists included in this accession as available from the Cowles Foundation. In the long term, we should plan to process the collection as a whole and integrate all the accessions into a common arrangement.

Appraisal:

Diskettes 1-3, 11, and 17 should be discarded; #1-3 contain printer drivers; #11 contains modem software; and #17 contains many deleted files and is mostly blank.

Some of Tobin's "office files" are of uncertain or low research value, such as the Quicken files, biographical statements and telephone lists. The publication lists are of questionable value as the Cowles Foundation has a detailed publication list in PDF form; however, Tobin has some topic-specific publication lists that may be helpful. Some of the office files also appear to be inventories of paper files, which may or may not be reflected in the paper records previously acquired.

Restrictions:

Other (paper) correspondence within this accession is restricted. E-mail contains both personal and professional correspondence; personal/family correspondence includes reference to health issues. Consider restricting e-mail under similar conditions. Most letters written using WordPerfect are professional in nature. Recommendation letters and Quicken files (which deal with Tobin's personal finances) should be restricted.

Preservation:

Proposed action now: Investigate migration options for Lotus 1-2-3 files, particularly those that reference graphs.

Recommended for later: Migrate WordPerfect files to PDF/A; migrate e-mail to a different format.

Access:

See Preservation. Files should be extracted into a storage option such as the YUL Rescue Repository so they can be paged on request. This collections does not have a high level use, so there is probably not an immediate need to create use copies.