CURATING COMPLEXITY:
A sociomaterial framework for software preservation in research libraries

Dr. Alexandra Chassanoff
CREOS Lunchtime Lecture,
MIT Libraries
June 25, 2019
The scholarly record is increasingly computational.
Preserving our scientific and cultural heritage
What’s different about born-digital?
Why is preserving software difficult?

Conceptual challenges
- What constitutes software?
- Artifact vs process

Technical challenges
- Dynamic and complex
- Obsolete/fragile media
- Imperceivable(!)
The research library as scholarly steward

- Collecting, preserving, and providing access to resources for the pursuit of knowledge has been a core mission since 1850s
- Computational resources require active caretaking
- Understanding significant characteristics for access and use, over time
Digital curation models for archiving
Modeling the ecosystem(s) of research software

What significant ecosystem characteristics support long-term access to and use of research software for different communities of practice?
Research methodology

- Survey and literature review
  - Identified dimensions for software curation
  - Identified prototypical scenarios
- Five case studies of institutionally significant software
- Refinement of model / triangulation
A sociomaterial framework for software curation

Activities

Documentation

Boundary conditions

Purpose

Carriers

Scenario
# Modeling software curation dimensions

## TABLE 1. Modeling software curation.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Driving question</th>
<th>Example values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activities</td>
<td>What potential activities might designated communities want to do with software?</td>
<td>Aggregate, analyze, cite, create, deposit, migrate, transform, publish, version, reuse</td>
</tr>
<tr>
<td>Boundary conditions</td>
<td>What characteristics of the software experience emphasize the transmission of information or assist with its comprehension?</td>
<td>Renderability; file tree navigation; contextual metadata; installability; rights management</td>
</tr>
<tr>
<td>Carriers</td>
<td>In what file and media formats are the resources of interest instantiated? What risks to future understanding are posed by these formats?</td>
<td>(Media) Removable magnetic media; LTO-9 tape; Paper printout: dot-matrix/ASCII; cloud-storage: S3-bucket (File) Java bytecode v1.2; Fortran 90 source; PDF/A-3</td>
</tr>
<tr>
<td>Documentation</td>
<td>What existing information documents design choices, intended uses, and methods of operations—and how can these be used to support choices made by curators or end-users?</td>
<td>Readme files; metadata; codebooks; methodology; scripts; correspondence</td>
</tr>
<tr>
<td>Purpose(s)</td>
<td>Was there a specific intended task gap that needed to be filled?</td>
<td>to validate or test existing claims; to generate a research outcome; to document research process</td>
</tr>
<tr>
<td>Scenario(s)</td>
<td>What potential future scenarios could support each desired activity? Who are the stakeholders in each scenario, and how/why do they interact?</td>
<td>Cross-cultural heritage artifact; amateur hobbyist project; legal evidence; part of a virtual experience; as a research tool;</td>
</tr>
</tbody>
</table>
Case: GRAPPLE-ing with legacy software artifacts


### Implementation pathways for research libraries

<table>
<thead>
<tr>
<th>Research software stewardship phase</th>
<th>Dimensions</th>
<th>Decision points for stewardship organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software creation</strong></td>
<td>Boundary conditions; Carriers; Documentation; Purposes</td>
<td>Who are the key stakeholders who create, use, provide access to and preserve research software at your institution? What best practice training and workshops can libraries offer to aid in creating digital resources that are preservation-ready?</td>
</tr>
<tr>
<td><strong>Software selection and appraisal</strong></td>
<td>Activities; Boundary conditions; Purposes; Scenarios</td>
<td>What kinds of software does your institution want to collect, preserve, and provide access to? What copyright/intellectual property issues are associated with collecting this software? Is there specific language that should be included in deposit agreements?</td>
</tr>
<tr>
<td><strong>Software acquisition and ingest</strong></td>
<td>Activities; Boundary Conditions; Carriers; Documentation</td>
<td>What changes need to be made to existing workflow to acquire software? Where will acquired materials be stored pre-ingest? What kinds of quality assurance should be done after acquisition? What precautions need to be made for handling obsolete/at-risk media?</td>
</tr>
<tr>
<td><strong>Software description and access</strong></td>
<td>Activities; Documentation; Purposes; Scenarios</td>
<td>How should existing software in hybrid collections be described? What metadata schemas and standards are best suited for describing software? How will sensitive/copyright materials be flagged?</td>
</tr>
<tr>
<td><strong>Software preservation and storage</strong></td>
<td>Boundary conditions; Carriers; Purposes; Scenarios</td>
<td>What preservation strategies best fit institutional commitment to software (e.g., migration, emulation, normalization, archival storage). What components of software do we want to preserve? Should original media carriers be preserved? Are there different priority levels that can be assigned based on media risk?</td>
</tr>
</tbody>
</table>
Decision points for caretakers

- **Software selection and appraisal**
  What kinds of software does MIT want to collect, preserve, and provide access?

- **Acquisition**
  What precautions (if any) need to be made for handling obsolete/at-risk media?

- **Access and Description**
  How will MIT Libraries provide access to collected software?

- **Preservation & Storage**
  What preservation strategies best fit institutional commitment to software in its different forms? (e.g., migration, emulation, documentation?)
Future activities:
“The hard work of software history”

- Extracting and analyzing born-digital content
- Capturing and preserving dynamic and complex information
- Characterizing meaningful access for different communities
THANK YOU!

Email: amchassanoff@gmail.com
Project Blog: http://softarchitectures.wordpress.com
Twitter: @amchass