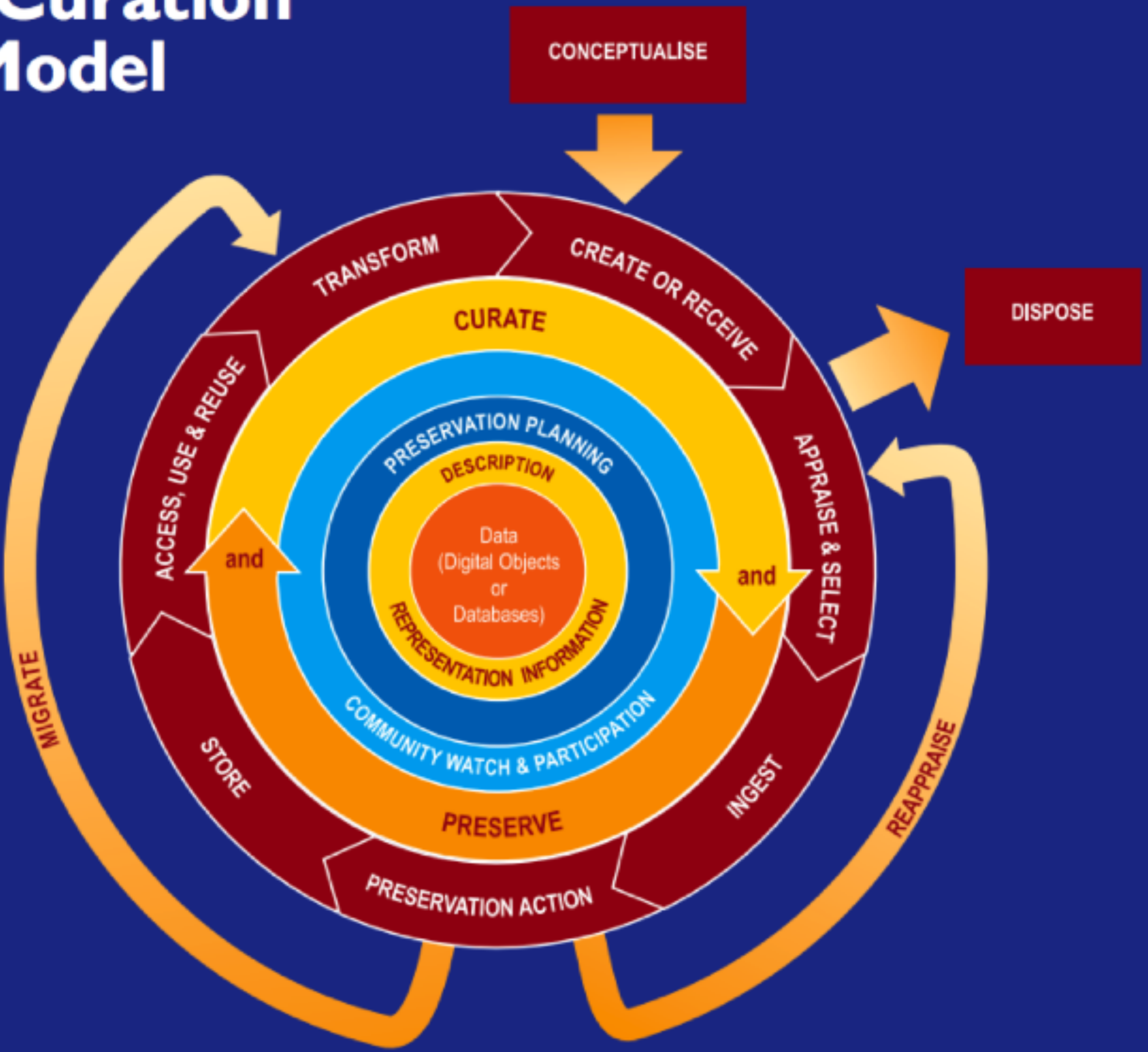
The Digital Curation Lifecycle

What is Digital Curation?

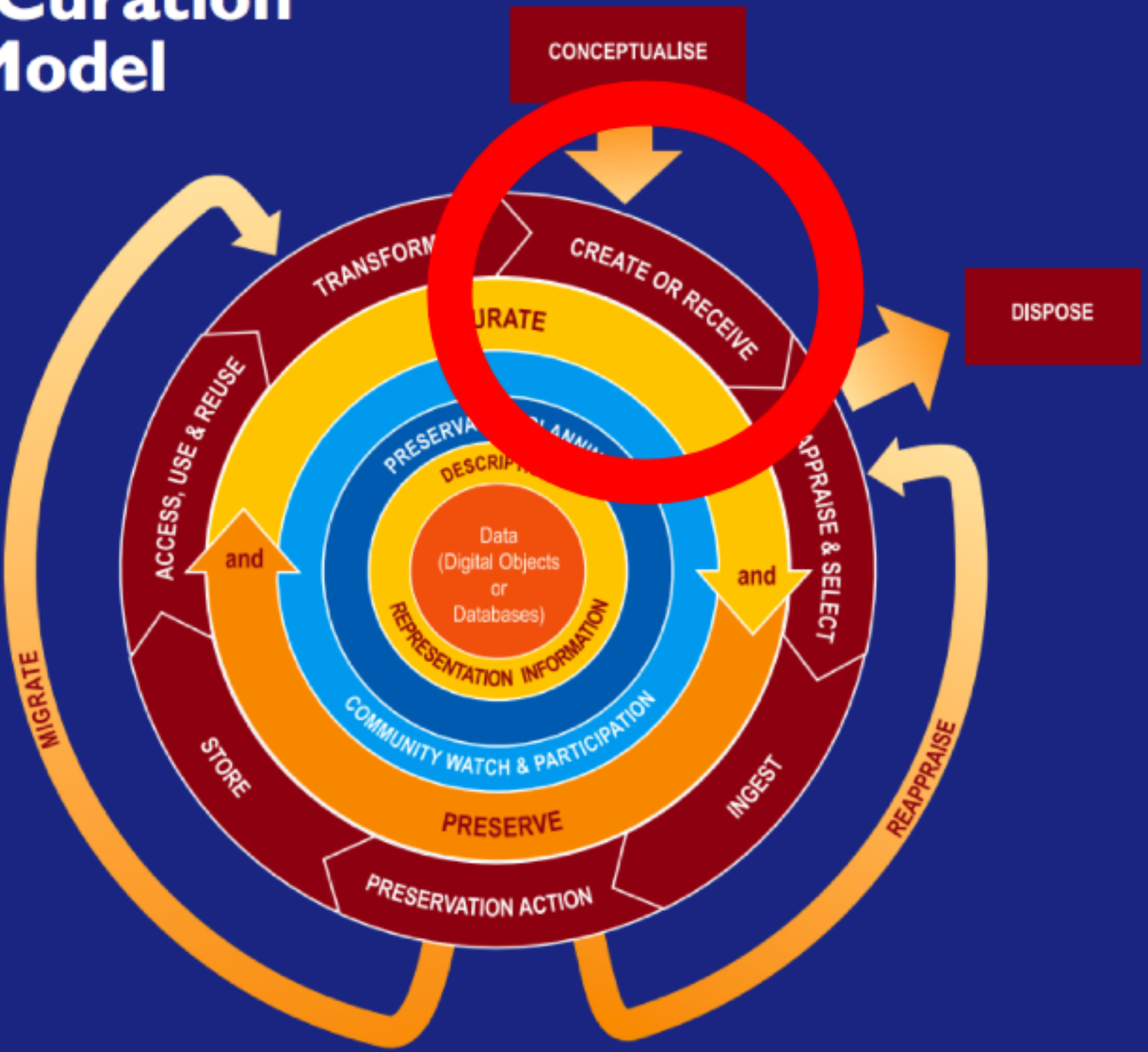
"Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle."

Digital Curation Centre. "What is digital curation?", http://www.dcc.ac.uk/digital-curation/what-digital-curation

The DCC Curation Lifecycle Model



The DCC Curation Lifecycle Model



Create or Receive

How do we get well formed data?

We create *curation* ready data!

Key Points

1. Develop, document, and apply policies about creating and receiving data

2. Influence data creators to create data that is curation friendly

3. Create data in standard data formats and file types that can be processed with open-source, well-documented programs

Key Points

4. Collect and keep documentation about the data, formats, software, agreements about its use, and provenance; and

5. Develop and implement procedures for receiving data

POLICY!

f4af8b5789576c000ce9105b25609bd6

Well formed data/objects?

What does that look like!?

0000000 643c 7669 6320 616c 7373 223d 6f62 6472

Openness

Portability

Quality

Sample file format policy

Audio

- flac
- wav

Image

- tif
- jp2

Video

- 8-10 bit uncompressed avi
- jp2

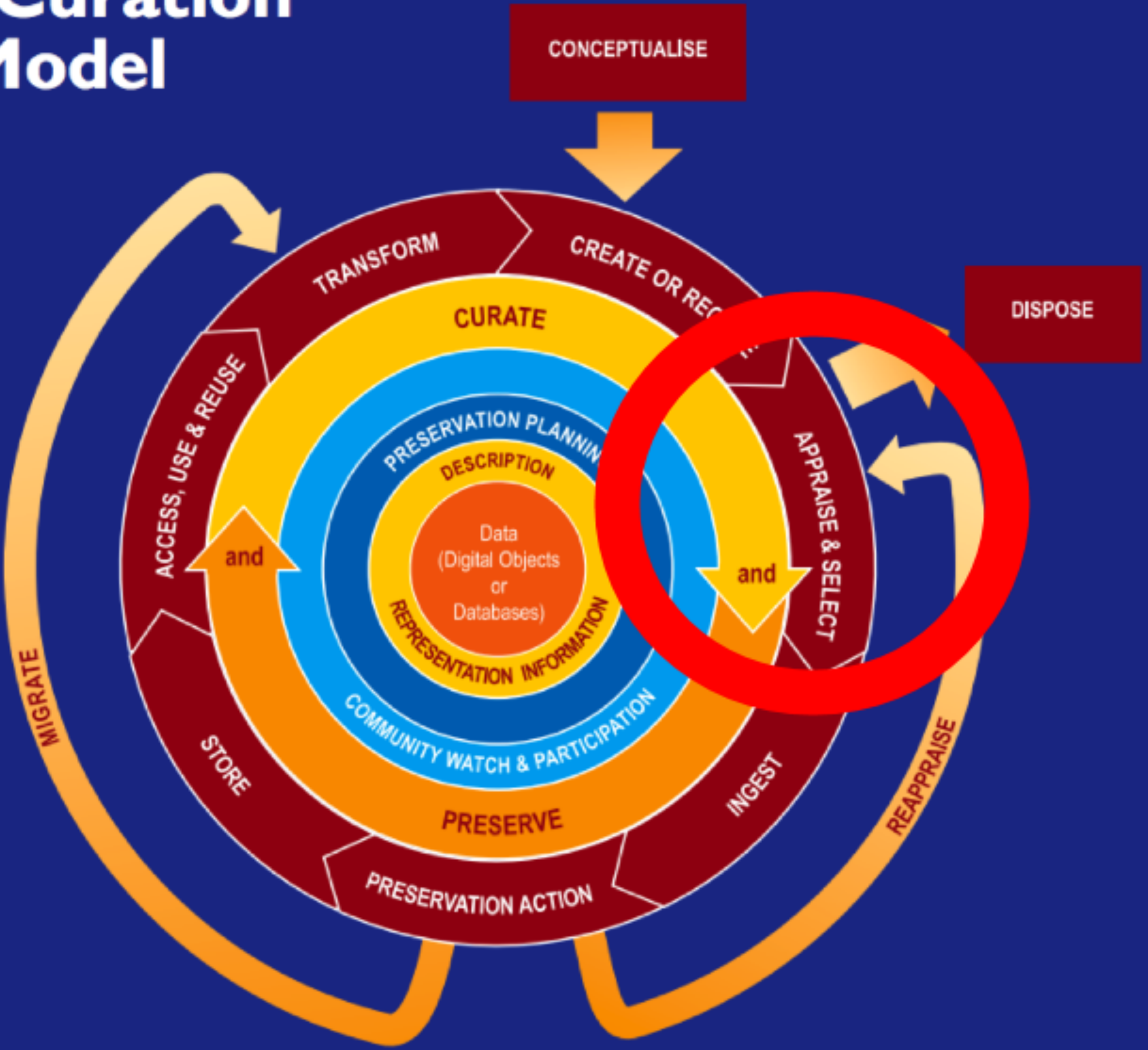
Text

- txt
- rtf
- pdf-a
- odf

Websites

- warc

The DCC Curation Lifecycle Model



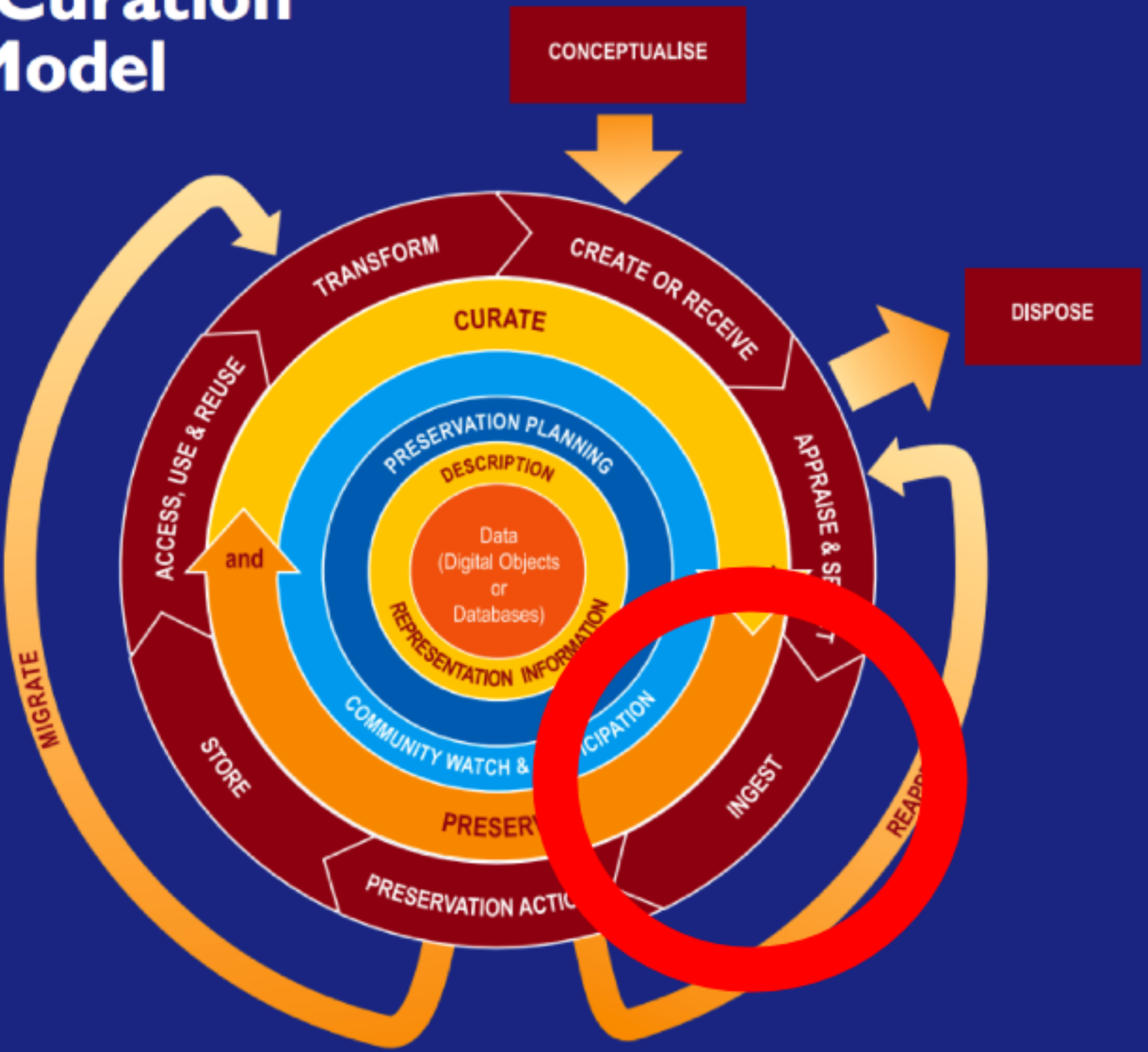
Everything that we choose to preserve means that something else won't be.

How do we decide?

- Needs of users (Designated Community)
- Feasibility of preservation
- Legal and IP rights
- Criticality of data
- Presence of associated data/metadata

POLICY!

The DCC Curation Lifecycle Model



SIPs to AIPs

Step 1: Submit

Step 2: Ingest

What's an AIP?

Grape AIP

1. Reference information

2. Provenance information

3. Context information

4. Fixity information

Process

1. Receive and accept SIP
2. Prepare SIP for storage and management
3. Perform quality assurance activities
4. Initiate format conversion
5. Generate AIP

POLICY!

Policy. Policy. Policy.

Sort & Identify

List data/objects

Uncompress

Virus & malware

Permanent identifiers

Acknowledge receipt

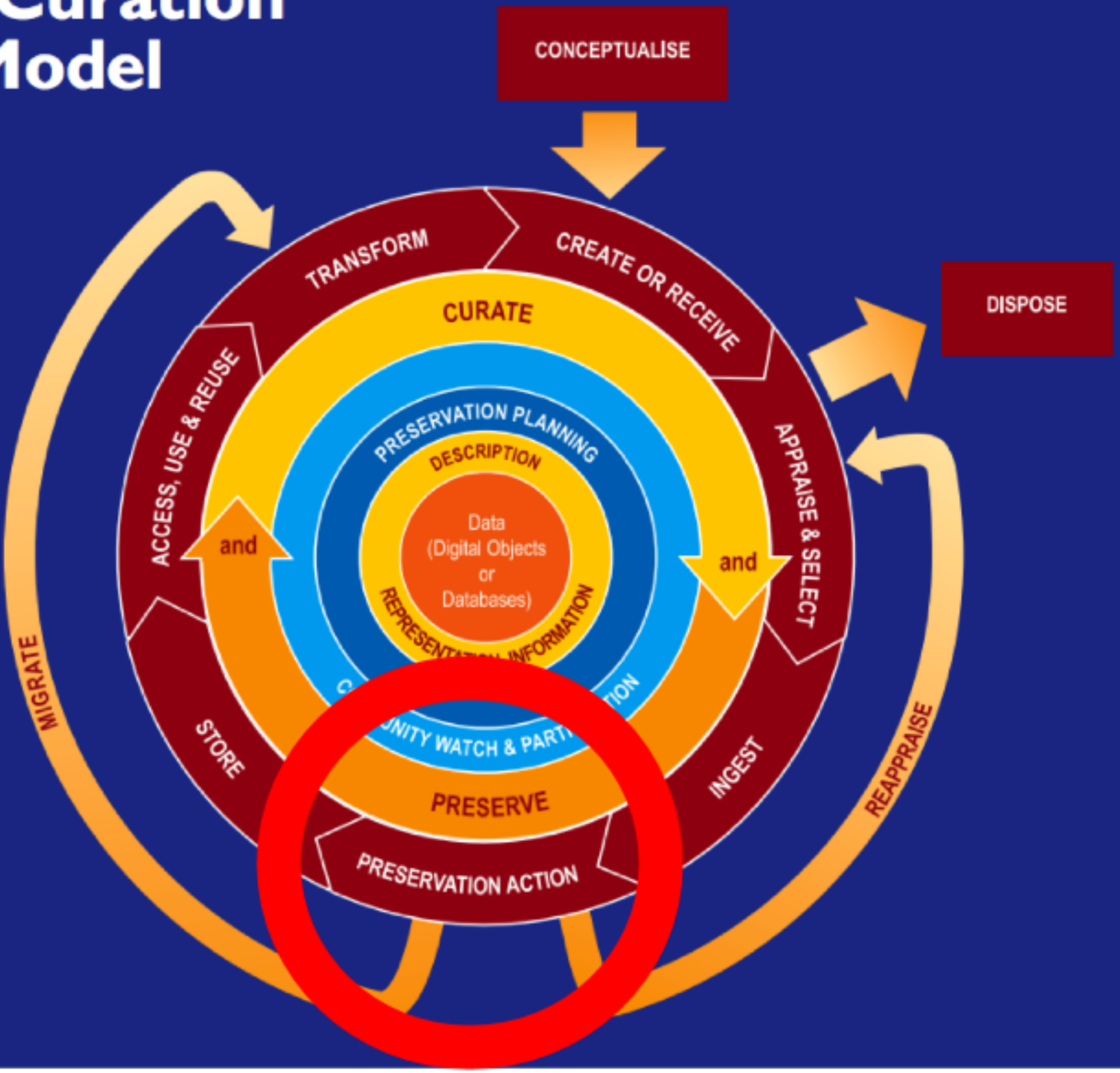
Depositor agreement

Unencrypt

Fixity

Transform/derivatives

The DCC Curation Lifecycle Model



Preservation Action

Anything that we do to maintain the

- Integrity
- Authenticity
- Usability

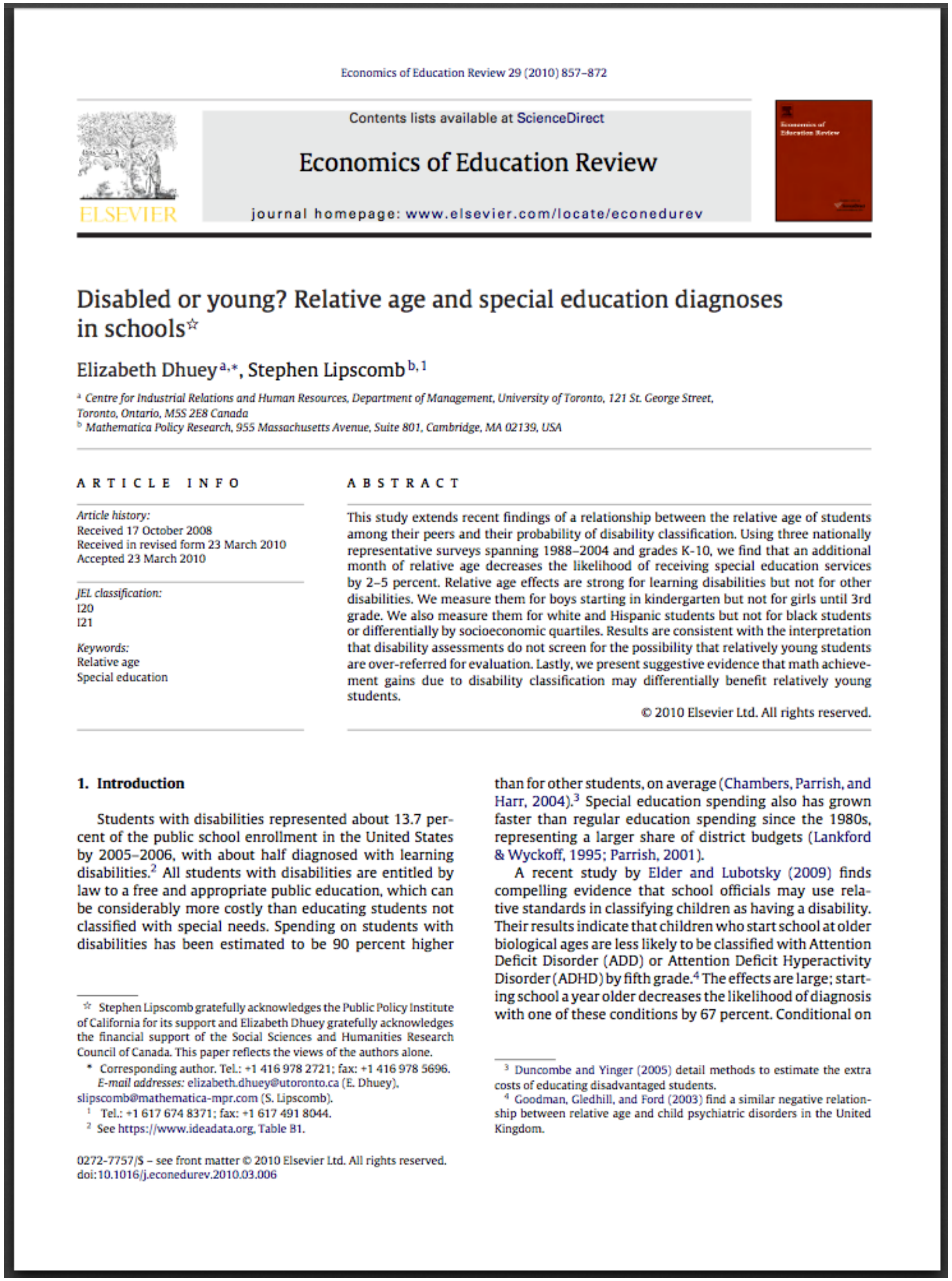
of our content.

Usability

3 main strategies:

- Technology preservation
- Technology emulation
- Information migration

Case 1:



Economics of Education Review 29 (2010) 857–872

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Disabled or young? Relative age and special education diagnoses in schools☆

Elizabeth Dhuey[a,*], Stephen Lipscomb[b,1]

[a] Centre for Industrial Relations and Human Resources, Department of Management, University of Toronto, 121 St. George Street, Toronto, Ontario, M5S 2E8 Canada

[b] Mathematica Policy Research, 955 Massachusetts Avenue, Suite 801, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 17 October 2008
Received in revised form 23 March 2010
Accepted 23 March 2010

JEL classification:
I20
I21

Keywords:
Relative age
Special education

ABSTRACT

This study extends recent findings of a relationship between the relative age of students among their peers and their probability of disability classification. Using three nationally representative surveys spanning 1988–2004 and grades K-10, we find that an additional month of relative age decreases the likelihood of receiving special education services by 2–5 percent. Relative age effects are strong for learning disabilities but not for other disabilities. We measure them for boys starting in kindergarten but not for girls until 3rd grade. We also measure them for white and Hispanic students but not for black students or differentially by socioeconomic quartiles. Results are consistent with the interpretation that disability assessments do not screen for the possibility that relatively young students are over-referred for evaluation. Lastly, we present suggestive evidence that math achievement gains due to disability classification may differentially benefit relatively young students.

1. Introduction

Students with disabilities represented about 13.7 percent of the public school enrollment in the United States by 2005–2006, with about half diagnosed with learning disabilities.[2] All students with disabilities are entitled by law to a free and appropriate public education, which can be considerably more costly than educating students not classified with special needs. Spending on students with disabilities has been estimated to be 90 percent higher than for other students, on average (Chambers, Parrish, and Harr, 2004).[3] Special education spending also has grown faster than regular education spending since the 1980s, representing a larger share of district budgets (Lankford & Wyckoff, 1995; Parrish, 2001).

A recent study by Elder and Lubotsky (2009) finds compelling evidence that school officials may use relative standards in classifying children as having a disability. Their results indicate that children who start school at older biological ages are less likely to be classified with Attention Deficit Disorder (ADD) or Attention Deficit Hyperactivity Disorder (ADHD) by fifth grade.[4] The effects are large; starting school a year older decreases the likelihood of diagnosis with one of these conditions by 67 percent. Conditional on

☆ Stephen Lipscomb gratefully acknowledges the Public Policy Institute of California for its support and Elizabeth Dhuey gratefully acknowledges the financial support of the Social Sciences and Humanities Research Council of Canada. This paper reflects the views of the authors alone.

* Corresponding author. Tel.: +1 416 978 2721; fax: +1 416 978 5696.
E-mail addresses: elizabeth.dhuey@utoronto.ca (E. Dhuey), slipscomb@mathematica-mpr.com (S. Lipscomb).

[1] Tel.: +1 617 674 8371; fax: +1 617 491 8044.

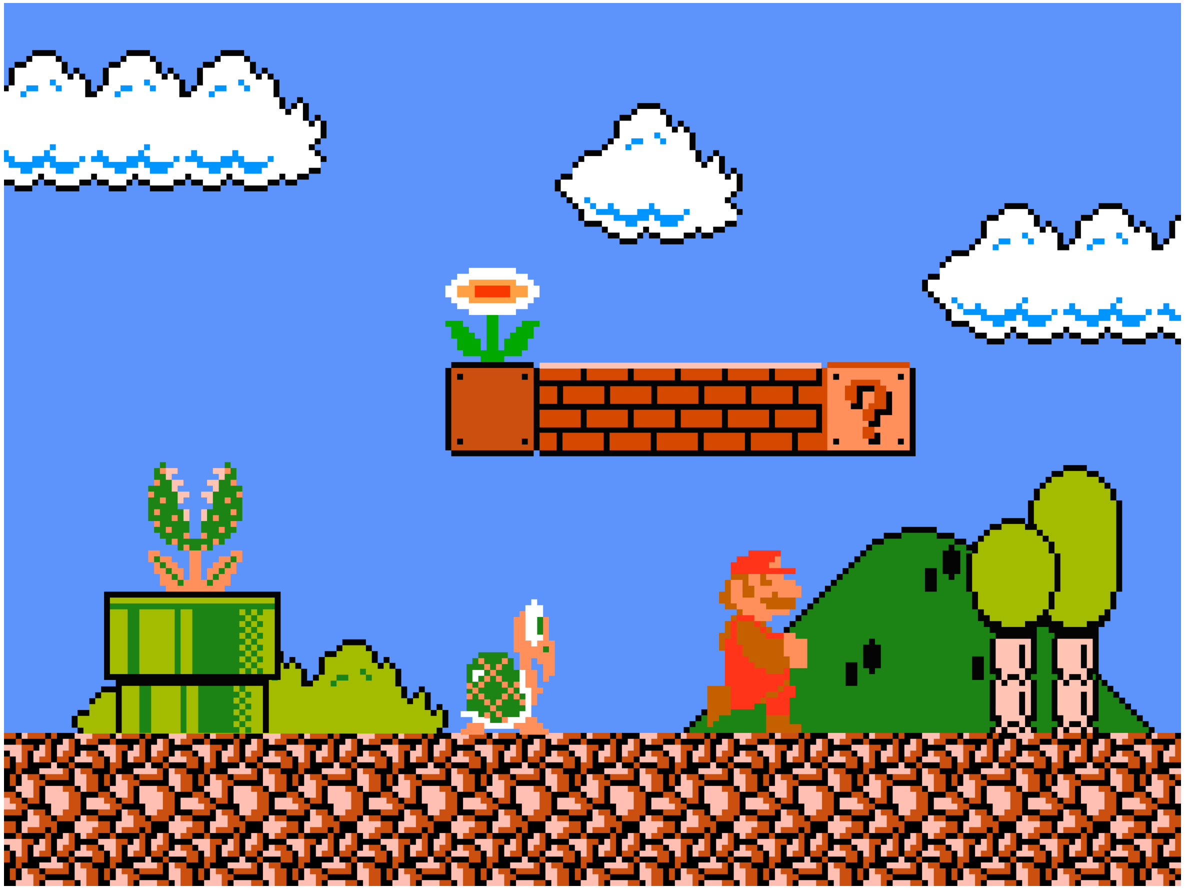
[2] See https://www.ideadata.org, Table B1.

[3] Duncombe and Yinger (2005) detail methods to estimate the extra costs of educating disadvantaged students.

[4] Goodman, Gledhill, and Ford (2003) find a similar negative relationship between relative age and child psychiatric disorders in the United Kingdom.

0272-7757/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.econedurev.2010.03.006

Case 2:



The 2 arrows

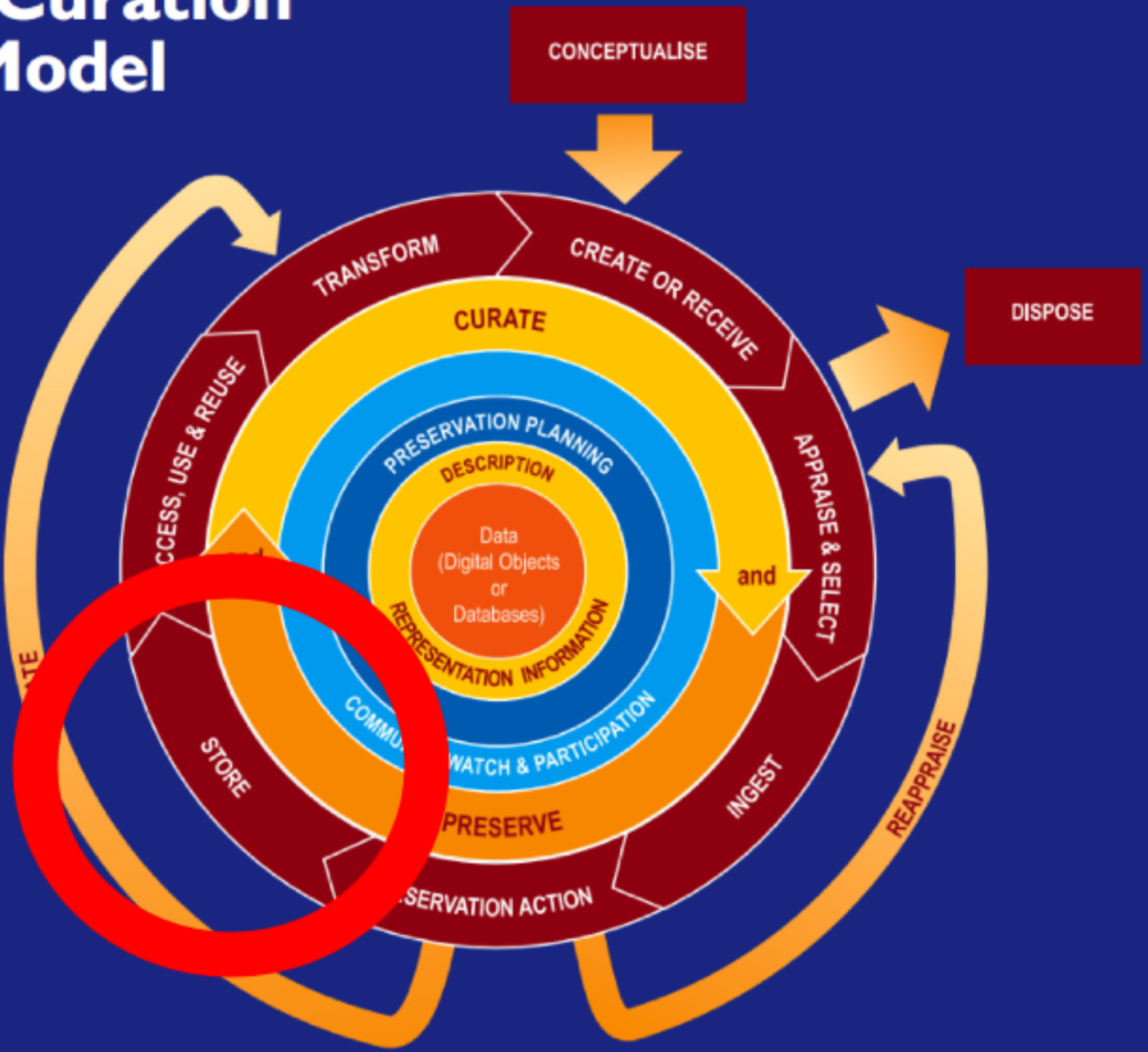
- Migration

- Reappraisal

POLICY!

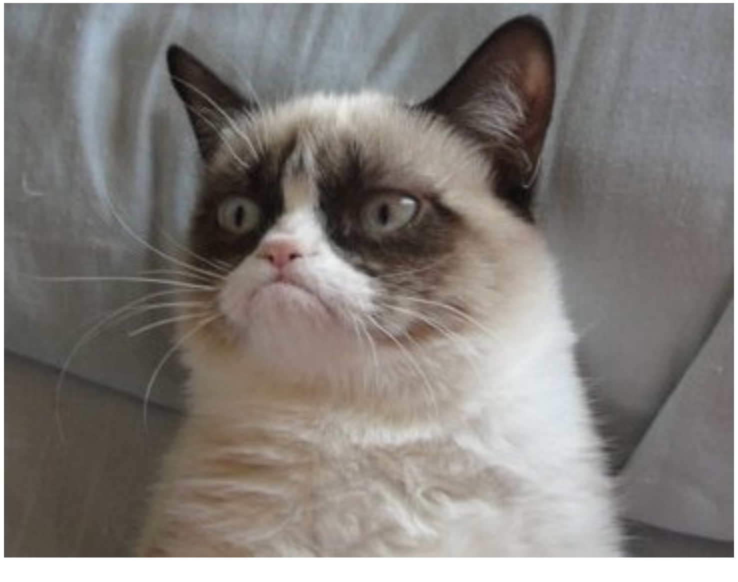
f4af8b5789576c000ce9105b25609bd6

The DCC Curation Lifecycle Model



Store! Store! Store!

Backups
are *NOT*
Digital Preservation



Policy!

f4af8b5789576c000ce9105b25609bd6

ensure that sufficient description and representation information is stored with data

**use a reliable storage medium
and
geographically distributed
backups systems**

monitor events that might trigger other preservation actions

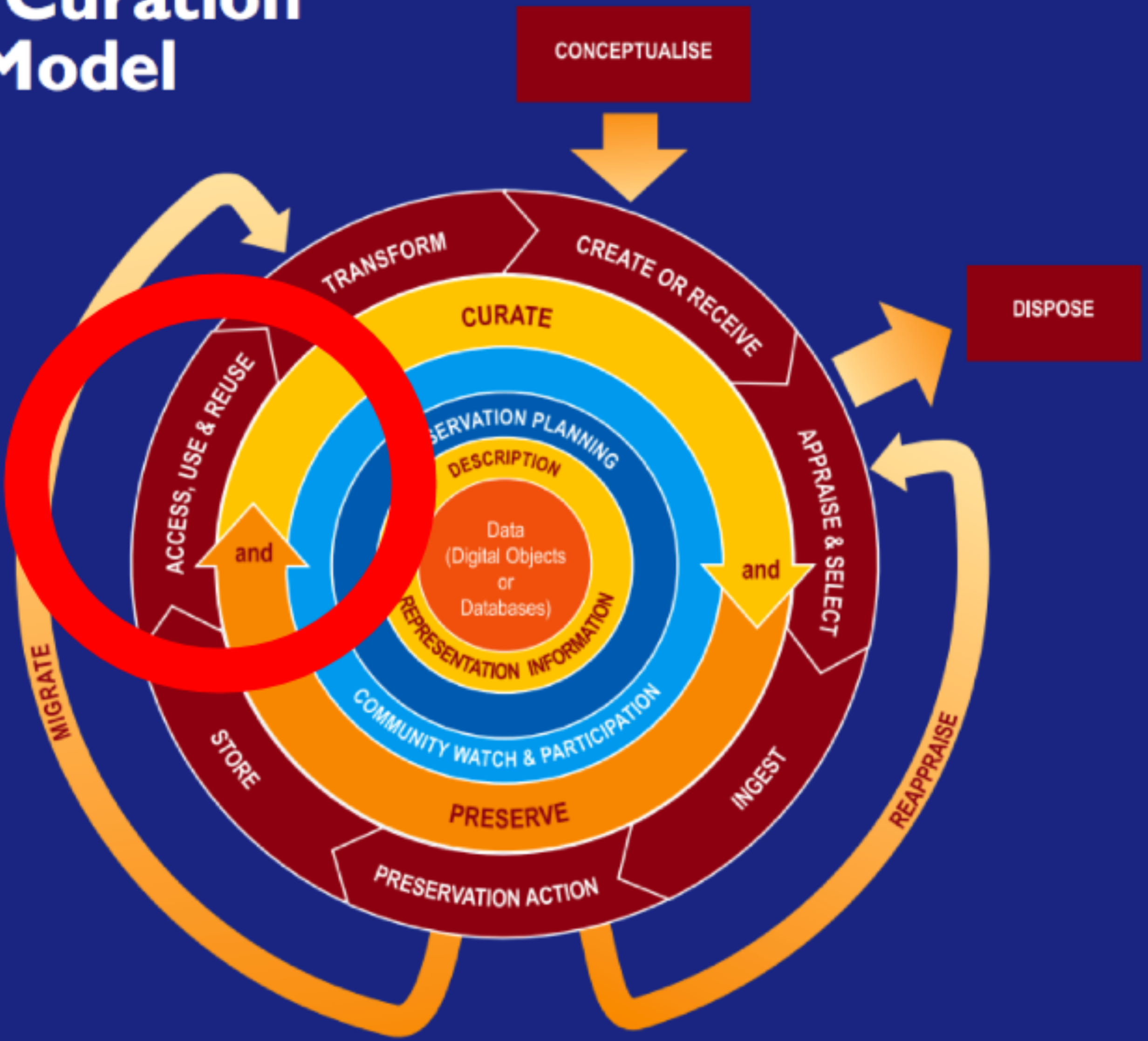
regularly check to ensure the integrity of the stored data and their description and representation information

ensure system and physical security

maintain and replace the technical infrastructure as necessary

develop, and administer as necessary, data recovery procedures

The DCC Curation Lifecycle Model

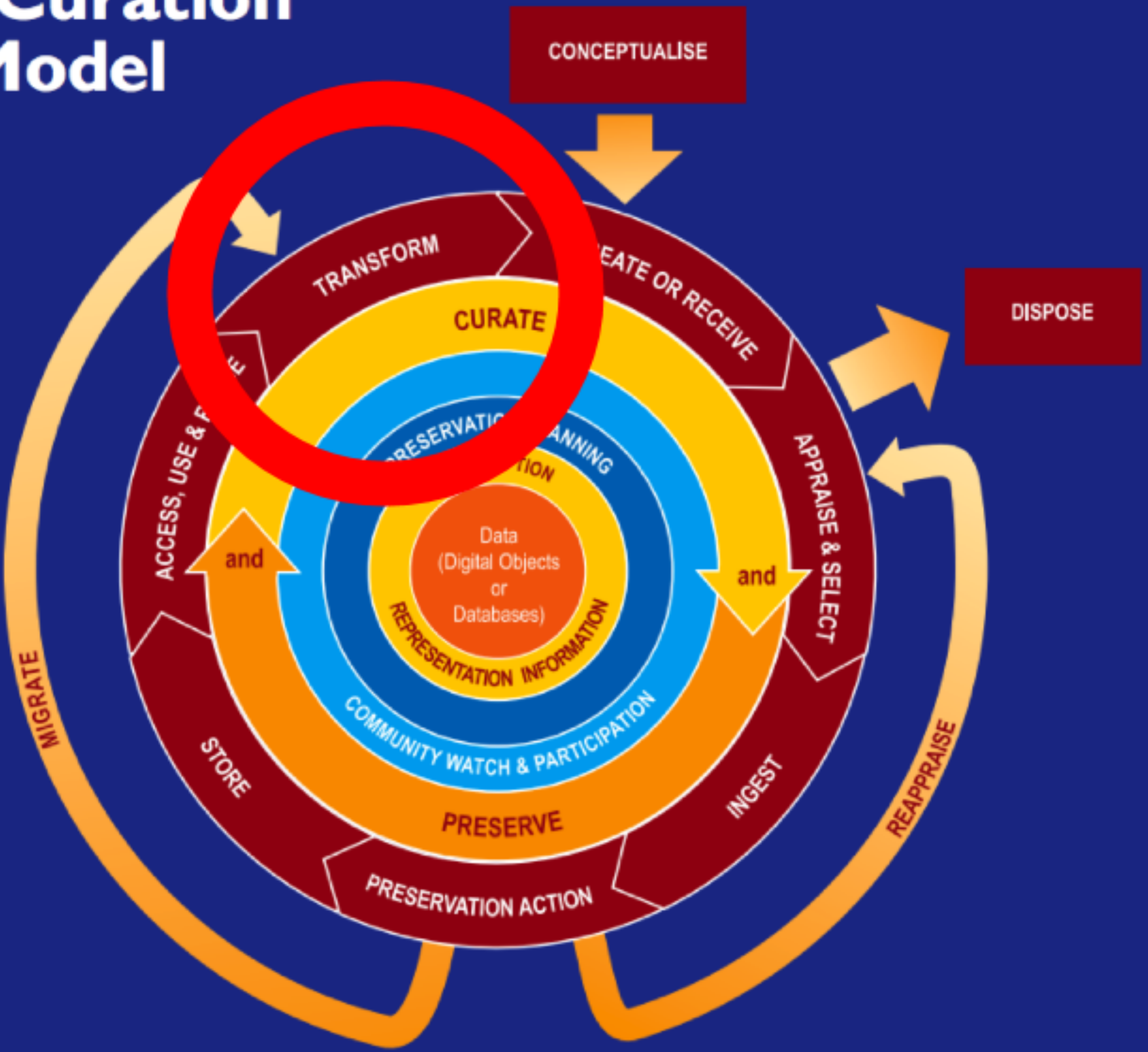


There is NO preservation without access.

What's required?

- Appropriate metadata to ensure data can be located.
- Appropriate legal permissions to ensure data can be (re-)used.
- Tools to allow the use of data.
- Access controls.

The DCC Curation Lifecycle Model



Transform

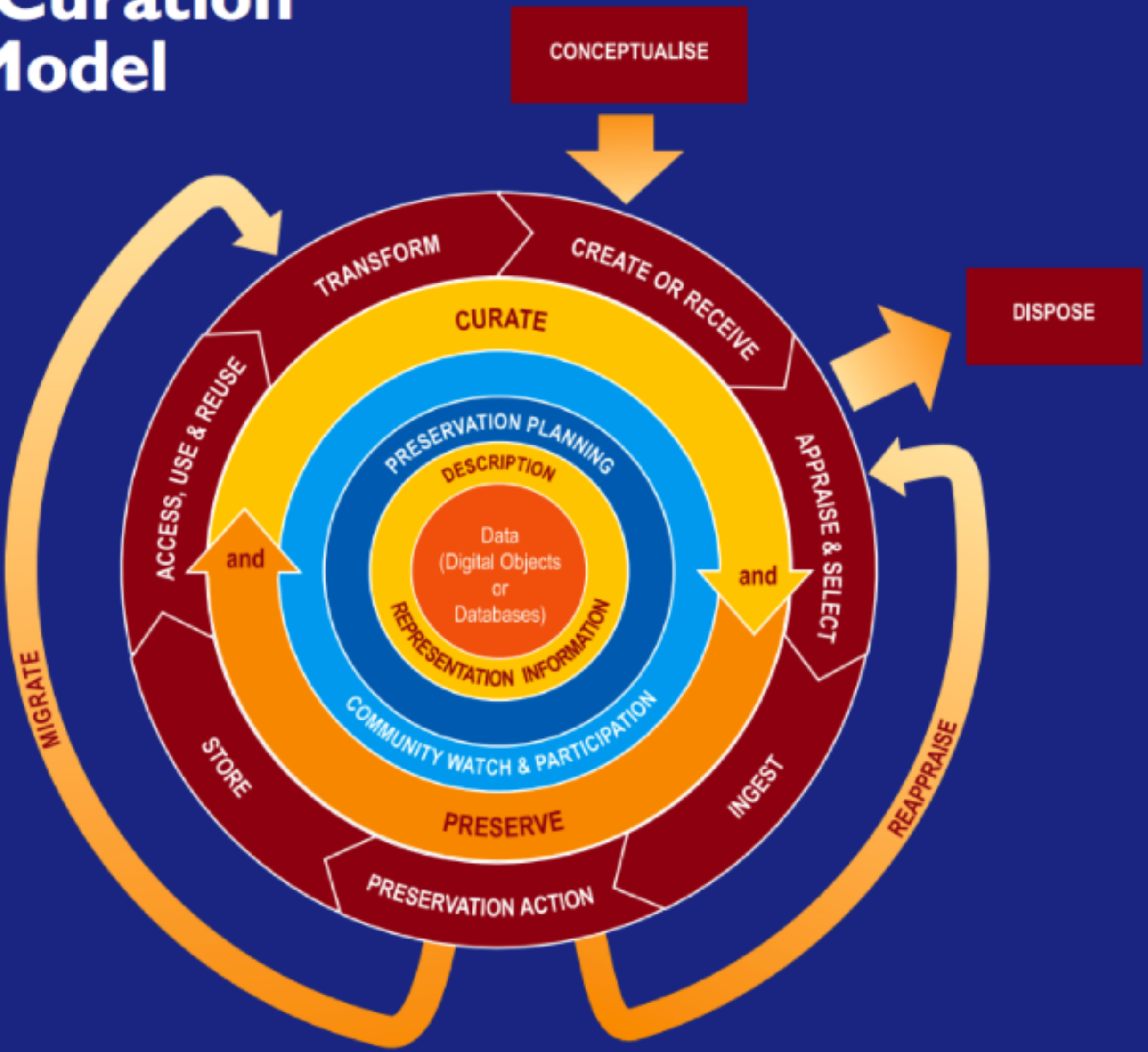
Transformation

- Can be invoked:
  - At time of access
  - As a preservation action

- At access: preservation formats are not always suitable to user needs.

- As preservation: mainly associated with 'Information migration' preservation strategy.

The DCC Curation Lifecycle Model



CONCEPTUALISE

DISPOSE

CREATE OR RECEIVE

APPRAISE & SELECT

INGEST

PRESERVATION ACTION

STORE

ACCESS, USE & REUSE

TRANSFORM

CURATE

and

PRESERVE

and

PRESERVATION PLANNING

COMMUNITY WATCH & PARTICIPATION

DESCRIPTION

REPRESENTATION INFORMATION

Data (Digital Objects or Databases)

MIGRATE

REAPPRAISE

Six very important takeaways from:

Sustainable Economics for a Digital Planet

Sustainability

Recognition of the benefits of preservation by decision makers

Process for ***Selecting*** digital materials with long-term value

Incentive for decision makers to preserve in the public interest appropriate ***Organization*** and ***Governance*** of digital preservation activities mechanisms to secure an ongoing, efficient ***Allocation*** of ***Resources*** for digital preservation activities

timely actions to ensure access

¡Yo quiero sostenibilidad!

1. Recognition of the benefits of preservation by decision makers
2. Process for Selecting digital materials with long-term value
3. Incentive for decision makers to preserve in the public interest
4. Appropriate organization and governance of digital preservation activities
5. Mechanisms to secure an ongoing, efficient allocation of resources for digital preservation activities
6. Timely actions to ensure access

1. ***Recognition*** of the benefits of preservation by decision makers

Do you have a statement emphasizing the importance of preservation as a part of your mission?

Are the benefits of preservation recognized at your location institution?

How many of you have preservation explicitly written into your strategic plan?

2. ***Process*** for **Selecting** digital materials with long-term value

How many of you have a committee, or group that evaluates digitization projects?

Do you have written policies in place that explicitly define your collection policy? If so, what kind of criteria does they use?

.

3. ***Incentive*** for decision makers to preserve in the public interest

Does your administration (at all levels) truly understand the responsibility in stewarding cultural heritage?

Is there funding available?

How might we structurally (hint: policy!) create incentives?

How do you, or how can we make preservation of public materials more feasible for us institutionally, or consortially?

If there is (there probably is!) a “shortfall” in digital curation, how do we address it?

4. Appropriate ***Organization*** and ***Governance*** of digital preservation activities

Does your organization have engagement mechanisms around digital preservation (e.g., committees, working groups)? Do these groups have a mandate to influence digital preservation activities?

Are these groups representative of all your institution’s stakeholders?

Does your organization have concrete policy around digital curation activities?

5. Mechanisms to secure an ongoing, efficient ***Allocation*** of ***Resources*** for digital preservation activities

Do you have an ongoing budget line for digital preservation?

Is the budget line from just the library budget, or do you have institutional support?

Do you used something along the lines of DuraCloud to preserve your items, a homegrown solution, a campus partnership (with central IT), or what if a consortial solution existed?

6. Timely actions to ensure access

Do you have appropriate platforms to provide access to your curated digital content? If there are gaps, what are they?

Is there a defined process by which content you are collecting/creating is made available? Is this process automated?

Does your content have access restrictions? Are you capable of enforcing them?

Curation Scenarios Exercise

In this exercise, you will be evaluating one of the following scenarios in terms of its place in the DCC Lifecycle and in terms of its sustainability, per the recommendations of the Blue Ribbon Task Force report. Your group should choose one institution present at your table to serve as the subject of this scenario.

On the provided worksheets, please answer the following questions for **each** step that you identify as being important to the scenario.

For each stage of the DCC Lifecycle model[1] that you identify, please answer the following questions:

DCC QUESTION:

- Broadly speaking (please don't get too bogged down in specific/technical detail if you can help it), how would you do this task? Outline the process in as many steps as you need.

DCC FOLLOW-ON:

- Do you have the expertise to do this locally? Are you aware of other OCUL schools/organizations that have this expertise? Think about potential partnerships, or ways you could leverage these relationships to build your own capacity.

BLUE RIBBON QUESTION:

- How sustainable are these activities? If you're doing them internally, is the support stable? If in partnership, are the partnerships stable?

BLUE RIBBON FOLLOW-ON:

- How can you move these activities toward sustainability? Think about activities you might undertake under the BRTF's recommended activities, and whether you could do them independently, or whether you would need support (personnel, infrastructure, knowledge/expertise) from OCUL or other member institutions.

[1] Create or Receive, Appraise & Select, Ingest, Preservation Action, Store, Access/Use/Reuse, Transform

Scenarios

- You work at a Map & Geospatial Information Center on campus. You have been made aware of an opportunity to digitize and preserve a collection of government topographic maps which have recently moved into the public domain. The copies of the maps you have available for digitization are oversized and fragile. You believe that these maps would be of interest to most if not all schools in Ontario.

  Some issues to consider: digitization of large, fragile materials; access issues; specialized metadata; possible large files, access issues

- A faculty member on your campus researches Internet culture around elections. Much of the original material she uses is in the form of websites that are created and maintained around the time of elections, but which tend to rapidly disappear after Election Day. For an upcoming election cycle, this researcher wants to save and archive as many websites as possible relating to a specific candidate. In addition to a more traditional qualitative method of study, this researcher is also interested in the potential for text-mining the corpus of documents collected to study how word usage changes with proximity to the election. Any archived pages collected as part of this initiative need to be not only human-readable, but also accessible to the researcher's specialized tools, whether this is on her computing environment or a shared one.

  Some issues to consider: web harvesting tools; content analysis; metadata

- Your special collections department has come into the possession of a large (500+ items) collection of wax cylinder records. A condition of the donation is that the library would digitize and make available the recordings. You currently have no in-house expertise in audio digitization. You have a manifest that lists the items donated (performer/recording) and their condition. There is no information about the copyright status of this material.

  Some issues to consider: Specialized digitization processes; access issues; copyright status; possible large files; use of digitization vendors

- A large international research program has selected your university as a custodian for the research data outputs of its funding program. The program in question ran for a single year, and involved 50 different research groups working across the country. At the time of application for funding, applicants were required to commit to archiving the data at an archive of the program's choosing, as well as making it publicly available where possible. As the funding program was multi-disciplinary, the data is quite heterogeneous and includes numeric data (some of it quite large), images, interviews, and some video. Early estimates of the aggregate size of the data are in the neighborhood of 50TB. Your university is asking the library what role they would like to play in managing this research collection.

Some issues to consider: large size of data; access vs. non-access issues; privacy issues around interviews; specialized metadata; liaising with researchers; data management planning; collections development priorities

- A large international research program has selected your university as a custodian for the research data outputs of its funding program. The program in question is ongoing, and involves as many as 50 grantees per year in different research groups working across the country. At the time of application for funding, applicants are required to commit to archiving the data at an archive of the program's choosing, as well as making it publicly available where possible. As the funding program is multidisciplinary, the data can be quite heterogeneous and may include numeric data, images, interviews, video, or anything else. Given that this is the first year of this mandate, the volume of data may be difficult to predict. Your university is asking the library what role they would like to play in managing this research collection.

  Some issues to consider: access vs. non-access issues; potential large size of files, coupled with unpredictable growth; privacy, ethics concerns; collection development priorities; ingest tools

- Your university archives department has received a gift of the original digital footage for an Oscar-winning movie by a famous director who is also an alumni. The total size of the gift is around 50 terabytes, and the content currently resides in a variety of places; some content is the commercial cloud storage platforms, and other material is on a number of hard drives that were provided by the donor.

  Some issues to consider: huge file sizes; major concerns around copyright; file formats; provenance & transfer

- The provincial association of government document librarians is recommending the digitization of the province's historical collection of agricultural pamphlets. These agricultural pamphlets were distributed for ~100 years to every GovInfo library in the province that wanted them. As a result, many different libraries have a collection of these pamphlets, although it is not always clear who has what, and nobody seems to have the whole thing. These pamphlets are oddly shaped and some of the older ones are quite fragile. In some cases, libraries have bound the pamphlets into volumes, usually by year.

  Some issues to consider: collection spread across multiple institutions; diverse digitization needs; crown copyright; access issues

- Your university archives needs to start processing electronic records from retiring faculty. These include: emails, websites, research data, drafts of published works, etc. Currently, the archives has some legacy policy but little in the way of infrastructure or procedures in place to handle such materials. In the interest of cost savings, it has been suggested by your campus administration that the archives share infrastructure costs with the library in processing and preserving this material.

  Some issues to consider: web harvesting tools; content analysis; disk imaging; access issues; differences between library and archives policy

Curation Scenarios Exercise

In this exercise, you will be evaluating one of the following scenarios in terms of its place in the DCC Lifecycle and in terms of its sustainability, per the recommendations of the Blue Ribbon Task Force report. Your group should choose one institution present at your table to serve as the subject of this scenario.

On the provided worksheets, please answer the following questions for **each** step that you identify as being important to the scenario.

For each stage of the DCC Lifecycle model[1] that you identify, please answer the following questions:

DCC QUESTION:

- Broadly speaking (please don't get too bogged down in specific/technical detail if you can help it), how would you do this task? Outline the process in as many steps as you need.

DCC FOLLOW-ON:

- Do you have the expertise to do this locally? Are you aware of other OCUL schools/organizations that have this expertise? Think about potential partnerships, or ways you could leverage these relationships to build your own capacity.

BLUE RIBBON QUESTION:

- How sustainable are these activities? If you're doing them internally, is the support stable? If in partnership, are the partnerships stable?

BLUE RIBBON FOLLOW-ON:

- How can you move these activities toward sustainability? Think about activities you might undertake under the BRTF's recommended activities, and whether you could do them independently, or whether you would need support (personnel, infrastructure, knowledge/expertise) from OCUL or other member institutions.

[1] Create or Receive, Appraise & Select, Ingest, Preservation Action, Store, Access/Use/Reuse, Transform

Scenarios

- You work at a Map & Geospatial Information Center on campus. You have been made aware of an opportunity to digitize and preserve a collection of government topographic maps which have recently moved into the public domain. The copies of the maps you have available for digitization are oversized and fragile. You believe that these maps would be of interest to most if not all schools in Ontario.

  Some issues to consider: digitization of large, fragile materials; access issues; specialized metadata; possible large files, access issues

- A faculty member on your campus researches Internet culture around elections. Much of the original material she uses is in the form of websites that are created and maintained around the time of elections, but which tend to rapidly disappear after Election Day. For an upcoming election cycle, this researcher wants to save and archive as many websites as possible relating to a specific candidate. In addition to a more traditional qualitative method of study, this researcher is also interested in the potential for text-mining the corpus of documents collected to study how word usage changes with proximity to the election. Any archived pages collected as part of this initiative need to be not only human-readable, but also accessible to the researcher's specialized tools, whether this is on her computing environment or a shared one.

  Some issues to consider: web harvesting tools; content analysis; metadata

- Your special collections department has come into the possession of a large (500+ items) collection of wax cylinder records. A condition of the donation is that the library would digitize and make available the recordings. You currently have no in-house expertise in audio digitization. You have a manifest that lists the items donated (performer/recording) and their condition. There is no information about the copyright status of this material.

  Some issues to consider: Specialized digitization processes; access issues; copyright status; possible large files; use of digitization vendors

- A large international research program has selected your university as a custodian for the research data outputs of its funding program. The program in question ran for a single year, and involved 50 different research groups working across the country. At the time of application for funding, applicants were required to commit to archiving the data at an archive of the program's choosing, as well as making it publicly available where possible. As the funding program was multi-disciplinary, the data is quite heterogeneous and includes numeric data (some of it quite large), images, interviews, and some video. Early estimates of the aggregate size of the data are in the neighborhood of 50TB. Your university is asking the library what role they would like to play in managing this research collection.

Some issues to consider: large size of data; access vs. non-access issues; privacy issues around interviews; specialized metadata; liaising with researchers; data management planning; collections development priorities

- A large international research program has selected your university as a custodian for the research data outputs of its funding program. The program in question is ongoing, and involves as many as 50 grantees per year in different research groups working across the country. At the time of application for funding, applicants are required to commit to archiving the data at an archive of the program's choosing, as well as making it publicly available where possible. As the funding program is multidisciplinary, the data can be quite heterogeneous and may include numeric data, images, interviews, video, or anything else. Given that this is the first year of this mandate, the volume of data may be difficult to predict. Your university is asking the library what role they would like to play in managing this research collection.

  Some issues to consider: access vs. non-access issues; potential large size of files, coupled with unpredictable growth; privacy, ethics concerns; collection development priorities; ingest tools

- Your university archives department has received a gift of the original digital footage for an Oscar-winning movie by a famous director who is also an alumni. The total size of the gift is around 50 terabytes, and the content currently resides in a variety of places; some content is the commercial cloud storage platforms, and other material is on a number of hard drives that were provided by the donor.

  Some issues to consider: huge file sizes; major concerns around copyright; file formats; provenance & transfer

- The provincial association of government document librarians is recommending the digitization of the province's historical collection of agricultural pamphlets. These agricultural pamphlets were distributed for ~100 years to every GovInfo library in the province that wanted them. As a result, many different libraries have a collection of these pamphlets, although it is not always clear who has what, and nobody seems to have the whole thing. These pamphlets are oddly shaped and some of the older ones are quite fragile. In some cases, libraries have bound the pamphlets into volumes, usually by year.

  Some issues to consider: collection spread across multiple institutions; diverse digitization needs; crown copyright; access issues

- Your university archives needs to start processing electronic records from retiring faculty. These include: emails, websites, research data, drafts of published works, etc. Currently, the archives has some legacy policy but little in the way of infrastructure or procedures in place to handle such materials. In the interest of cost savings, it has been suggested by your campus administration that the archives share infrastructure costs with the library in processing and preserving this material.

  Some issues to consider: web harvesting tools; content analysis; disk imaging; access issues; differences between library and archives policy