Preservation

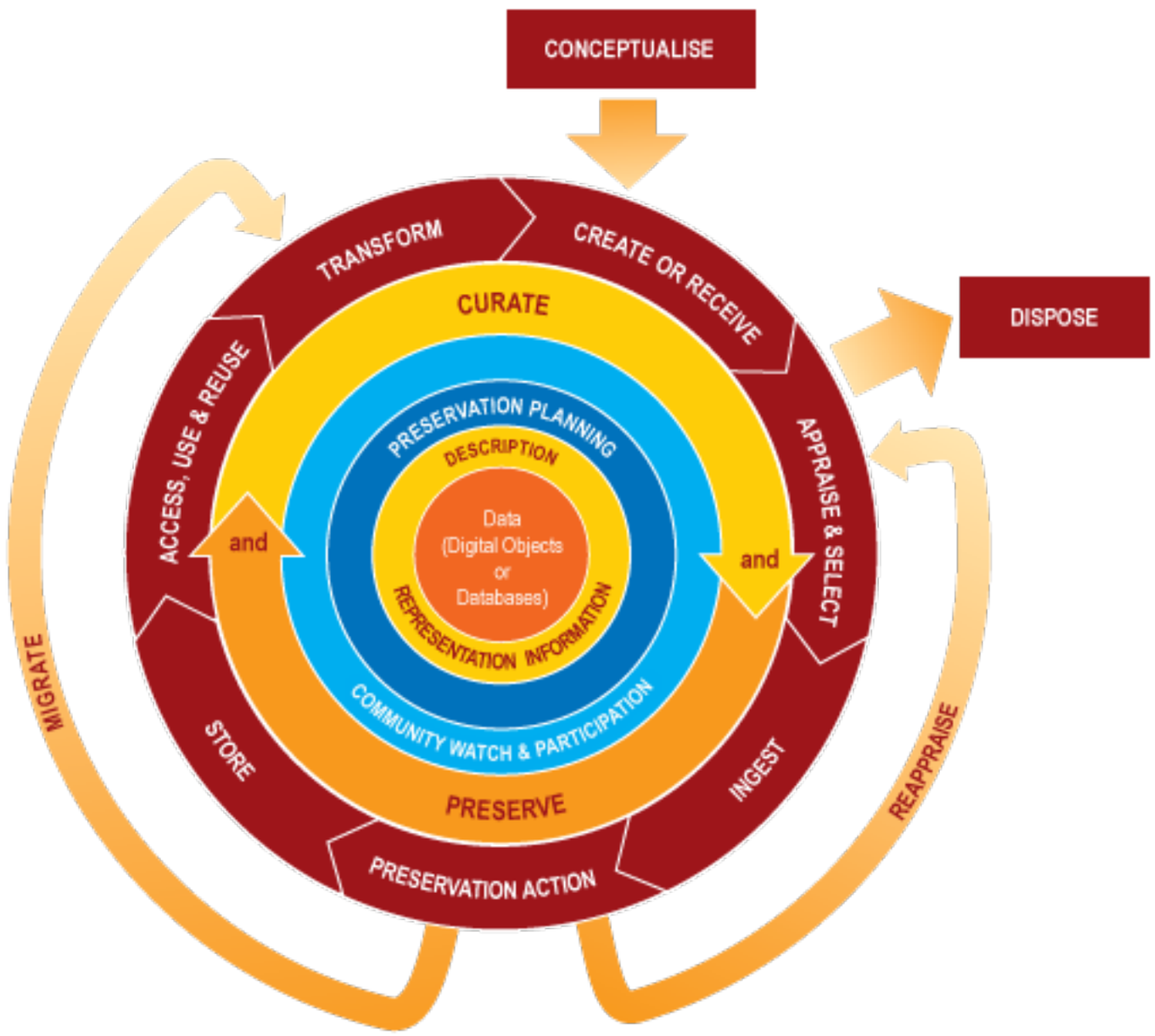
INFX 551
Winter 2018

Course Outline

Week 3
(final week in this module)

Data	Data Systems	Policy, Privacy, & Ethics
Types & Roles	Repositories	Policy
Documentation	Preservation	Privacy
Standardization	Cost Models	Ethics

…active and ongoing management throughout a lifecycle of use, including reuse in unanticipated contexts.



1. Digital Objects - formats and encodings

2. Managing Digital Objects - storage and integrity

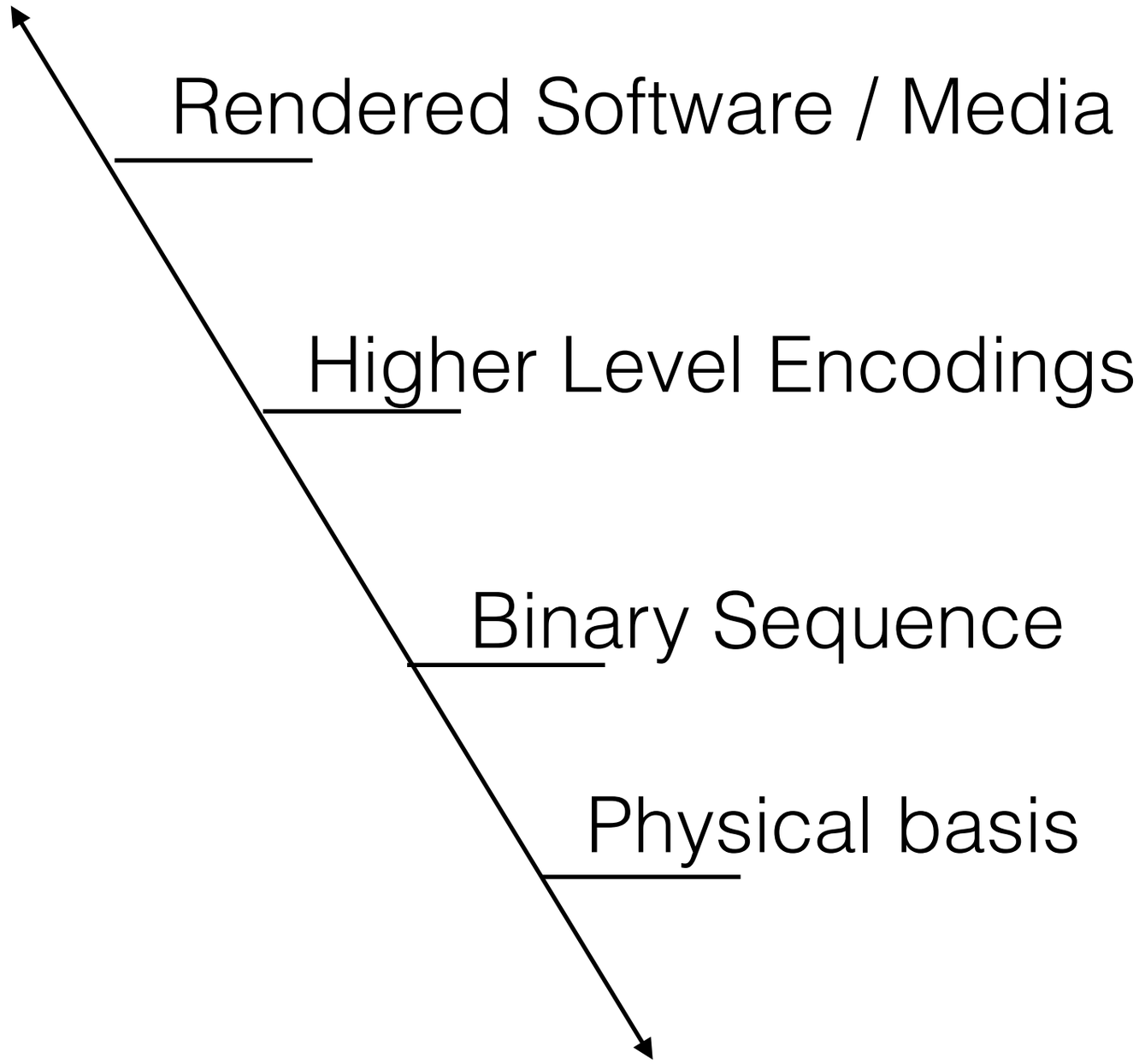
3. Preservation Actions - Migration and Emulation

4. Preservation Environments - the OAIS model of a trusted digital repository

5. Levels of Service - how to differentiate storage, from reliable preservation

1. Digital Objects

- Born digital or digitized content.
- All digital objects are a bit stream made of up of 1's and 0's
- An encoding of a digital object translates those 1's and 0's into a higher level sequence that is able to be rendered in our computer
- All file formats have certain encoding information - most simply a header and a footer sequence that tells us - this is an MP3, or this is a PDF, or this is a JPEG



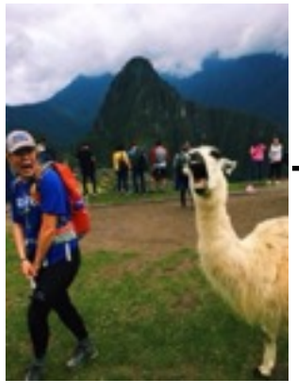
Ladder of Abstraction

Digital image

Encoding (hexadecimal)

```
0008810: 2828 2828 2828 2828 2828 2828 2828 28ff  ((((((((((((((.
0008820: c000 1108 0379 024f 0301 1100 0211 0103  .....y.O........
0008830: 1101 ffc4 001c 0000 0203 0101 0101 0000  ................
0008840: 0000 0000 0000 0003 0401 0205 0600 0708  ................
0008850: ffc4 004a 1000 0202 0103 0303 0203 0604  ...J............
0008860: 0306 0402 0b01 0203 1100 0412 2105 3141  ............!.1A
0008870: 1322 5106 6132 7181 0714 2342 91a1 52b1  ."Q.a2q...#B..R.
0008880: c1d1 1533 f024 3462 72e1 f108 1643 8235  ...3.$4br....C.5
0008890: 5392 b217 2573 a2c2 d236 4463 ffc4 001a  S...%s...6Dc....
00088a0: 0100 0301 0101 0100 0000 0000 0000 0000  ................
00088b0: 0000 0102 0304 0506 ffc4 0033 1100 0202  ...........3....
00088c0: 0202 0104 0201 0304 0301 0002 0300 0102  ................
00088d0: 1103 2112 3141 0413 2251 3261 7105 1442  ..!.1A.."Q2aq..B
00088e0: 2333 8191 15a1 b152 62f0 c1d1 e1ff da00  #3.....Rb.......
00088f0: 0c03 0100 0211 0311 003f 00dd 9b4e dae8  .........?...N..
0008900: 77f4 b8cb c2b2 9500 c772 283c 8b3f ea33  w........r(<.?.3
```

AlpacaScream.jpg



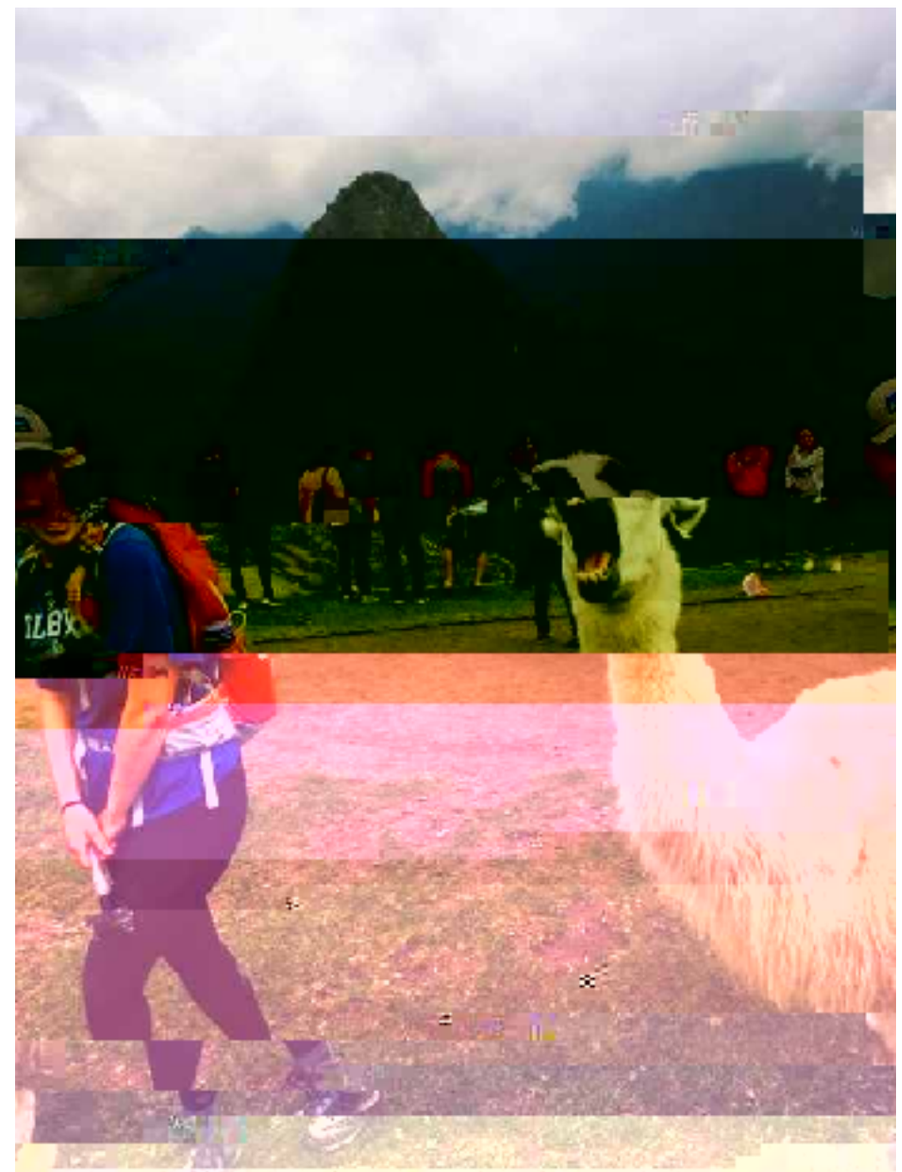
Is this a Jpeg?

Yes

2. Managing Digital Objects

- Digital objects are stored on media
    - Hard-drive, tape reels, spinning disk, etc.
- Media break down - but they also become unstable (or lose integrity)
- Check-sum - a courtesy call to our digital objects - where we ask - are you still the same digital object that I thought you were? This is also referred to as a **Hash Value.**
- A check-sum is calculates a unique string of digits that are logged within our computers - and we can repeatedly “check” this “sum” to make sure it remains the same over time.
- We refer to digital objects as having **integrity** if the check-sum remains the same over time. Sometimes this is referred to as **fixity.** Meaning the digital object is “fixed” on our media and is not drifting or becoming unstable. (I’ll show an example at the end)

Data Integrity: the quality of a bitstream being unaltered or changed through loss, tampering, or corruption



3. Preservation Actions

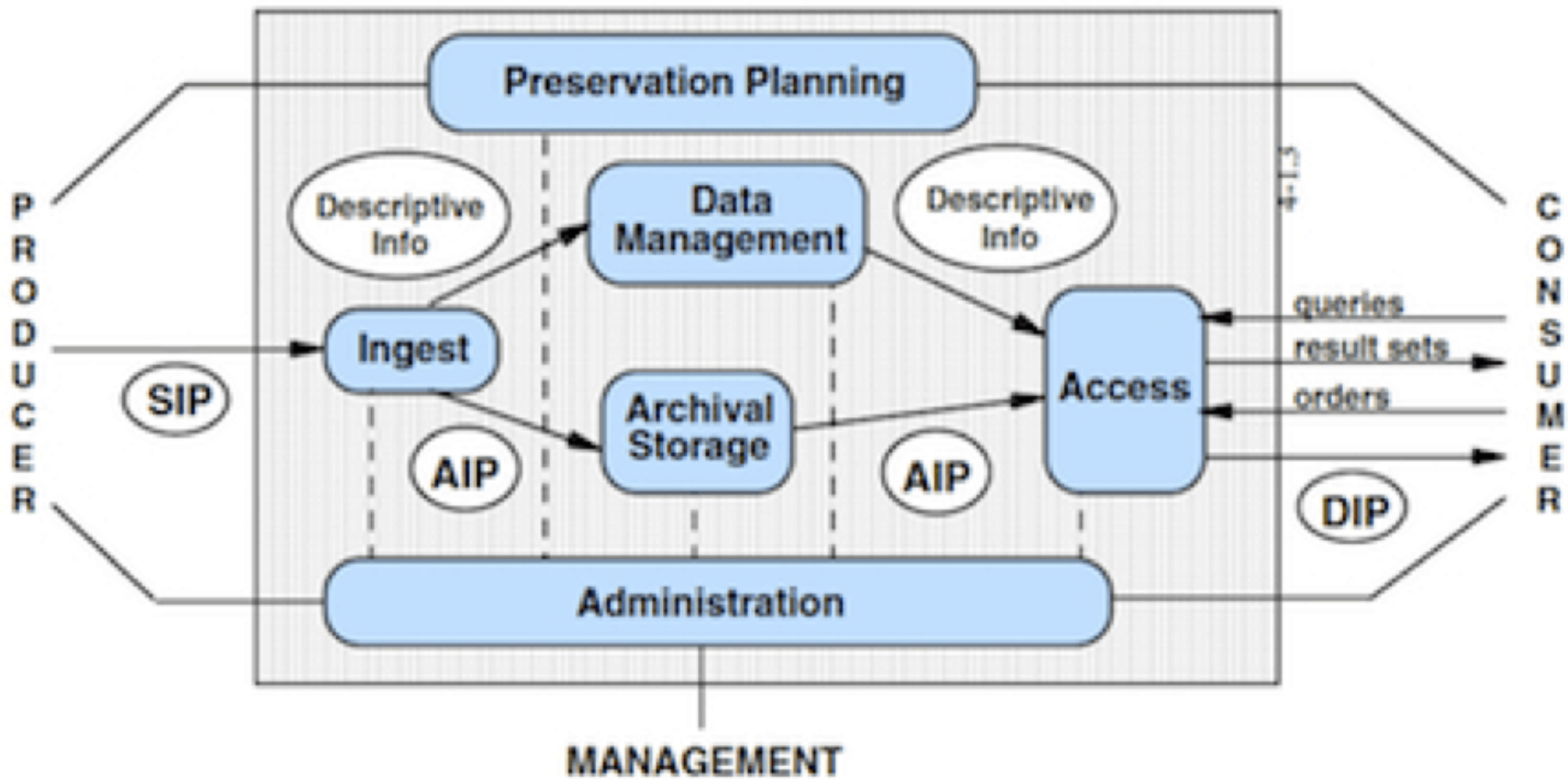
- What happens if we determine a digital object we are managing becomes unstable (loses fixity)?
- We need to take a ‘preservation action’
- 2 most common forms:
    - Migration - moving content from one media, file, or file format to another.
    - Emulation - recreating the environment in which a stable item is rendered.

Data Storage

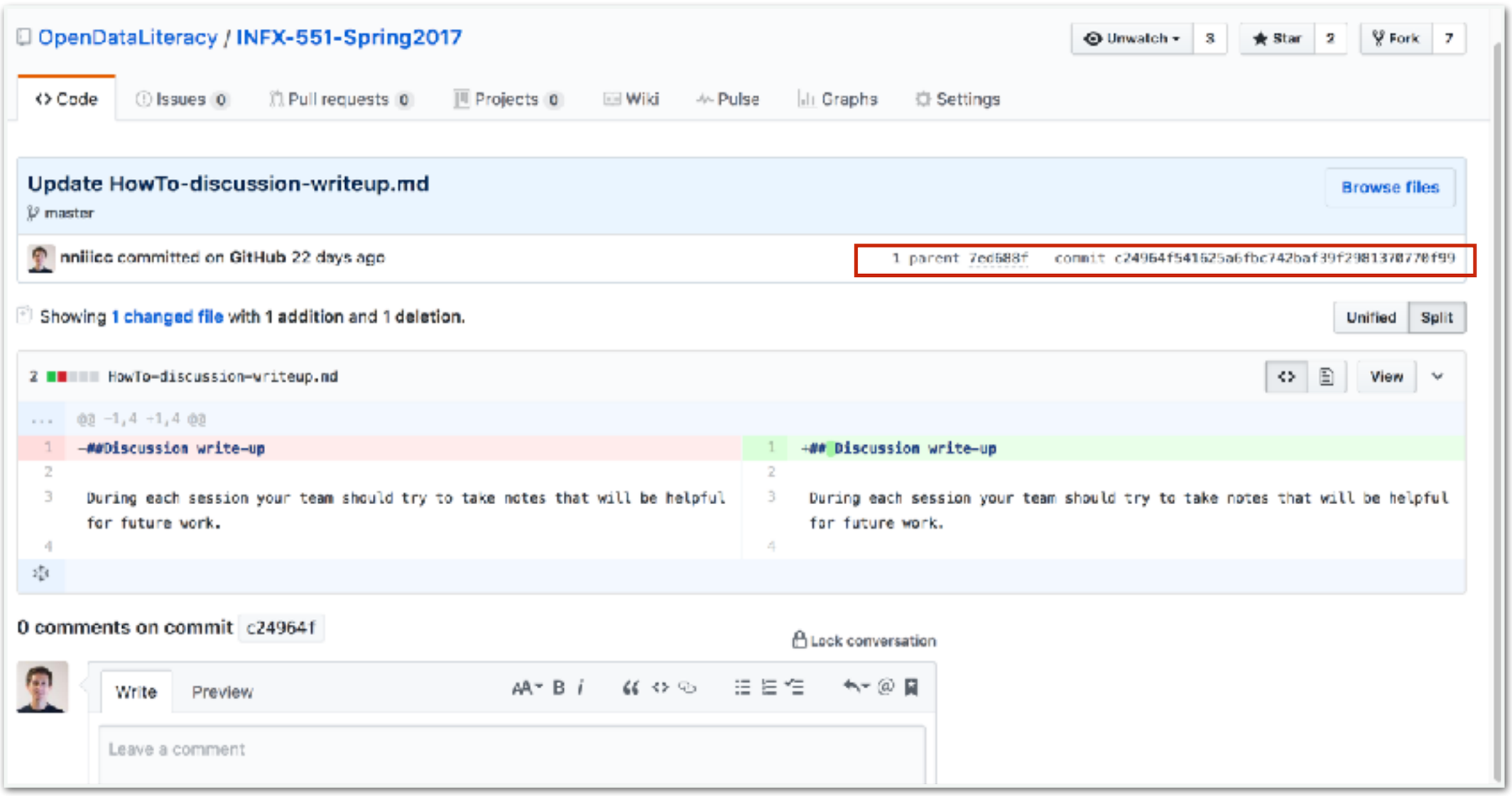
- The whole point of using dedicated software frameworks - like a data repository - is that these are environments that allow us to automate, and reliably perform data management actives like…
    - Ingest
    - Preserve
    - Publish

4. Digital Preservation Environments

- Storage of digital objects in systematic way
- Creates a way to **Ingest**, **Archive**, and **Disseminate** digital objects
    - Ingest -> Submission Information Package
    - Archive -> Archival Information Package
    - Publish -> Dissemination Information Package


Preservation Planning
Descriptive Info
Data Management
Descriptive Info
PRODUCER
Ingest
SIP
Archival Storage
AIP
AIP
Access
queries
result sets
orders
DIP
CONSUMER
Administration
MANAGEMENT

“Infrastructure is often invisible until breakdown” (or it gets pointed out to us.



Dataverse

About Guides Support Sign Up Log In

Supplementary Information to 'GWISFI: a Universal GPU Interface for Exhaustive Search of Pairwise Interactions in Case-Control GWAS in Minutes' (Revised Version)

Wang, Qiao; Shi, Fan; Kowalczyk, Andrew; Campbell, Richard M; Goudey, Benjamin; Rawlinson, David; Harwood, Aaron; Ferra, Herman; Kowalczyk, Adam, 2017, "Supplementary Information to 'GWISFI: a Universal GPU Interface for Exhaustive Search of Pairwise Interactions in Case-Control GWAS in Minutes' (Revised Version)", doi:10.7910/DVN/OPLUKK, Harvard Dataverse, V1

Cite Dataset

Learn about Data Citation Standards.

Description

This is the supplementary material (revised version) for GWISFI paper. The original source is available at https://sourceforge.net/projects/gwisfi/

NB: NICTA has transformed to CSIRO Data61, and consequently the official download link bioinformatics.research.nicta.com.au/gwisfi is no longer accessible. A new development of this work is in progress.

Subject

Medicine, Health and Life Sciences; Computer and Information Science

Related Publication

Qiao Wang, Fan Shi, Andrew Kowalczyk, Richard M Campbell, Benjamin Goudey, David Rawlinson, Aaron Harwood, Herman Ferra and Adam Kowalczyk (2014) "GWISFI: a Universal GPU Interface for Exhaustive Search of Pairwise Interactions in Case-Control GWAS in Minutes" doi: 10.1109/BIBM.2014.6999192

Dataset Version: 1.0

Files Metadata Terms Versions

Search this dataset... Find

1 File

gwisfi_supplement.pdf

Adobe PDF - 152.6 KB - Apr 27, 2017 - 0 Downloads

MD5: 76ead1ec75ec2ba8d2277f0311b80926

Download

Digital Preservation in Context

	Level One (Protect your data)	Level Two (Know your data)	Level Three (Monitor your data)	Level Four (Repair your data)
Storage and geographic location	▪Two complete copies that are not collocated ▪For data coming in on heterogeneous media (optical disks, hard drives, floppies) get the digital content off the medium and into your storage system	▪Three complete copies ▪At least one copy in a different geographic location ▪Document your storage system(s) and storage media and what you need to use them	▪At least one copy in a geographic location with a different disaster threat ▪Start an obsolescence monitoring process for your storage system(s) and media	▪All copies in geographic locations with different disaster threats ▪Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems.
File Fixity and Data Integrity	▪Check fixity on ingest if it has been provided with the content ▪Create fixity info if it wasn't provided with the content	▪Check fixity on all ingests ▪Use write-blockers when working with original media ▪Virus-check high risk content	▪Check fixity on all transformative acts ▪Check fixity of sample files/media at fixed intervals ▪Maintain logs of fixity info; supply audit on demand ▪Ability to detect corrupt data ▪Virus-check all content	▪Check fixity of all content in response to specific events or activities ▪Ability to replace corrupted data
Information Security	▪Identify who has read, write, move, and delete authorization to individual files ▪Restrict who has those authorizations to individual files		▪Maintain logs of who has accessed individual files	▪Maintain logs of who performed what actions on files, including deletions and preservation actions ▪Perform audit of logs
Metadata	▪Inventory of content and its storage location ▪Ensure backup and non-collocation of inventory	▪Store administrative metadata ▪Store transformative metadata and log events	▪Store standard technical and descriptive metadata	▪Store standard preservation metadata
File Formats	▪Encourage use of limited set of known and open file formats and codecs	▪Inventory of file formats in use	▪Validate files against their file formats ▪Monitor file format obsolescence threats	▪Perform format migrations, emulation and similar activities

- Review
    - Digital objects have different levels of abstraction - we need to go down a level for preservations sake
    - Fixity and Integrity are two critical ideas that we focus on in ensuring that our data remain useful and usable over the long term