Preservation

INFX 551
Spring 2017

- Metadata wrap-up
- Will Saunders, OCIO Open Date Lead at the State of Washington
- Lecture & Discussion
- Activity: Core and Extended Application Profiles

Some leftover metadata concepts

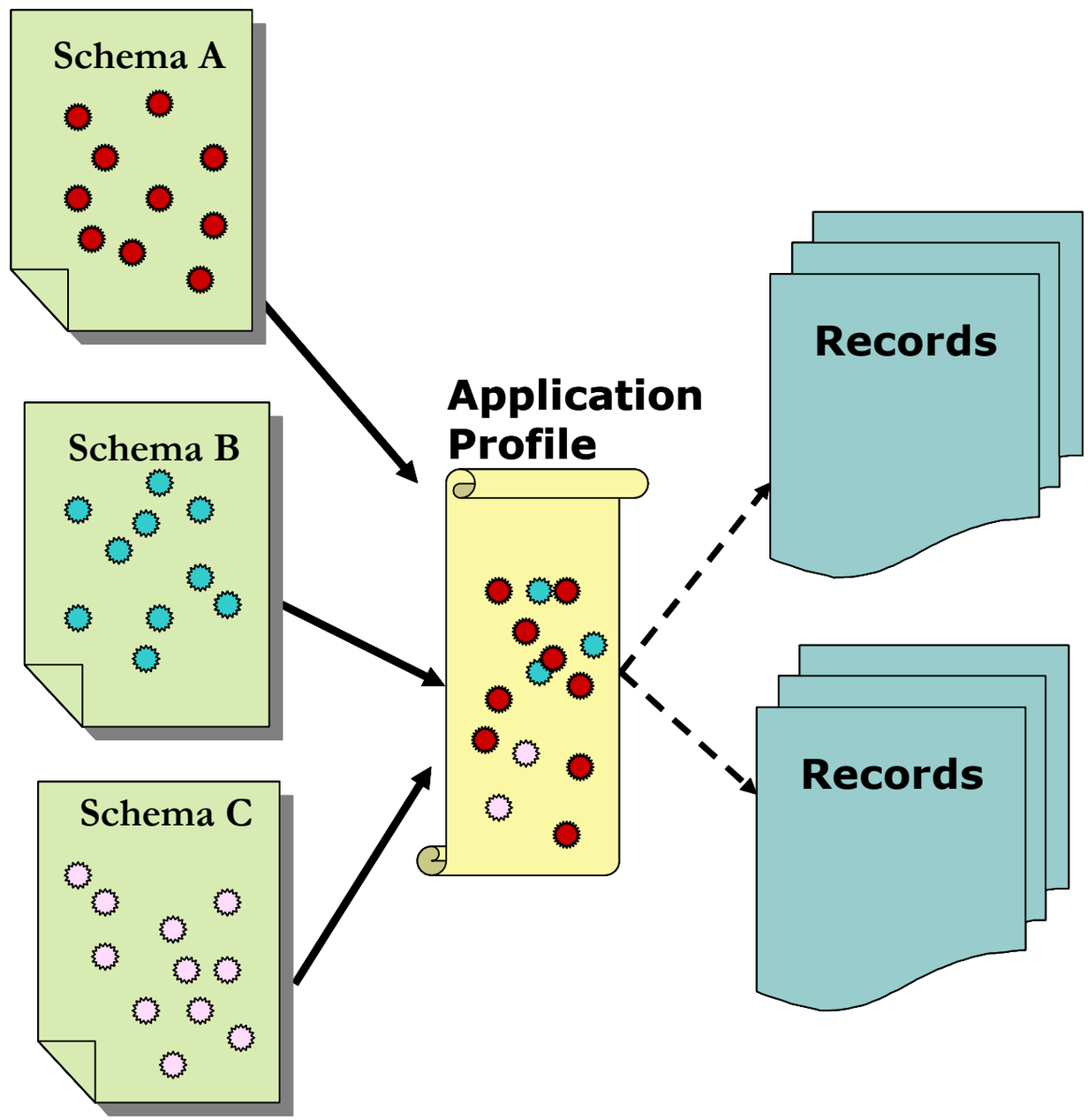
- Attribute (element) and Value Pairs
    - Title: Moby Dick
- Machine Readable Encoding
    - <dc:title>Moby Dick</dc:title>
- Vocabulary Encoding Scheme
    - indicates that the value of an element is taken from a controlled vocabulary (or concept-space), such as the Library of Congress Subject Headings.
    - "Resource X has the topic: 'Textile design--China--History'. This topic is a member of the Vocabulary Encoding Scheme 'Library of Congress Subject Headings'."

```xml
<?xml version="1.0" encoding="UTF-8"?>

<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/">
<dc:title>Alpaca Scream</dc:title>
<dc:creator>The Internet</dc:creator>
<dc:subject xsi:type="dcterms:AAT">Memes</dc:subject>
<dc:subject xsi:type="dcterms:LCSH">Internet Memes</dc:subject>
<dc:description>A picture of an Alpaca screaming at a tourist. </dc:description>
<dc:publisher>imu.gr</dc:publisher>
<dc:contributor>u/duckwtihdacks</dc:contributor>
<dc:date xsi:type="dcterms:W3CDTF">04/28/2017</dc:date>
<dc:type xsi:type="dcterms:DCMIType">image</dc:type>
<dc:format xsi:type="dcterms:IMT">jpeg</dc:format>
<dc:identifier>ecb6fbe3458y243958797345fcd821ae3</dc:identifier>
<dc:identifier>i.reddituploads.com/aaf03e3bwoer910239u</dc:identifier>
<dc:source>reddit.com</dc:source>
<dc:language xsi:type="dcterms:ISO639-2">en</dc:language>

</metadata>
```

- Namespace
  - Declarations in machine readable form.
  - Sets out rules for validity
- Application profile
  - “are schemas which consist of data elements drawn from one or more namespaces, combined together by implementers, and optimized for a particular local application.”(Heerv & Patel, 2000)
  - In other words - a combination of different elements from different namespaces that is unique to your org.context.


Schema A
Schema B
Schema C
Application Profile
Records
Records

What if there are no standards for *my* metadata needs?

"If an implementor wishes to create 'new' elements that do not exist elsewhere then (under this model) they must create their own namespace schema, and take responsibility for 'declaring' and maintaining that schema."

Heery and Patel (2000)

Digital Preservation

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

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

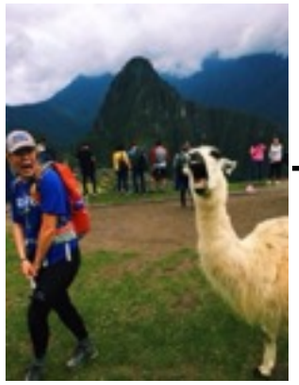
1. Digital Objects

- But how do we know encodings, and file formats?
- Lots of ways:
    - File extensions: MyAwesomeThesis.PDF
    - Tools: extract an encoded sequence of bits and tells us “This thing is probably a PDF”
- But, how do tools know?
    - Extract: Sequences, and Technical Metadata (more on that in a minute)
    - Compares this to format registries (think of this as a reference library for file formats)

Preservation

```
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)

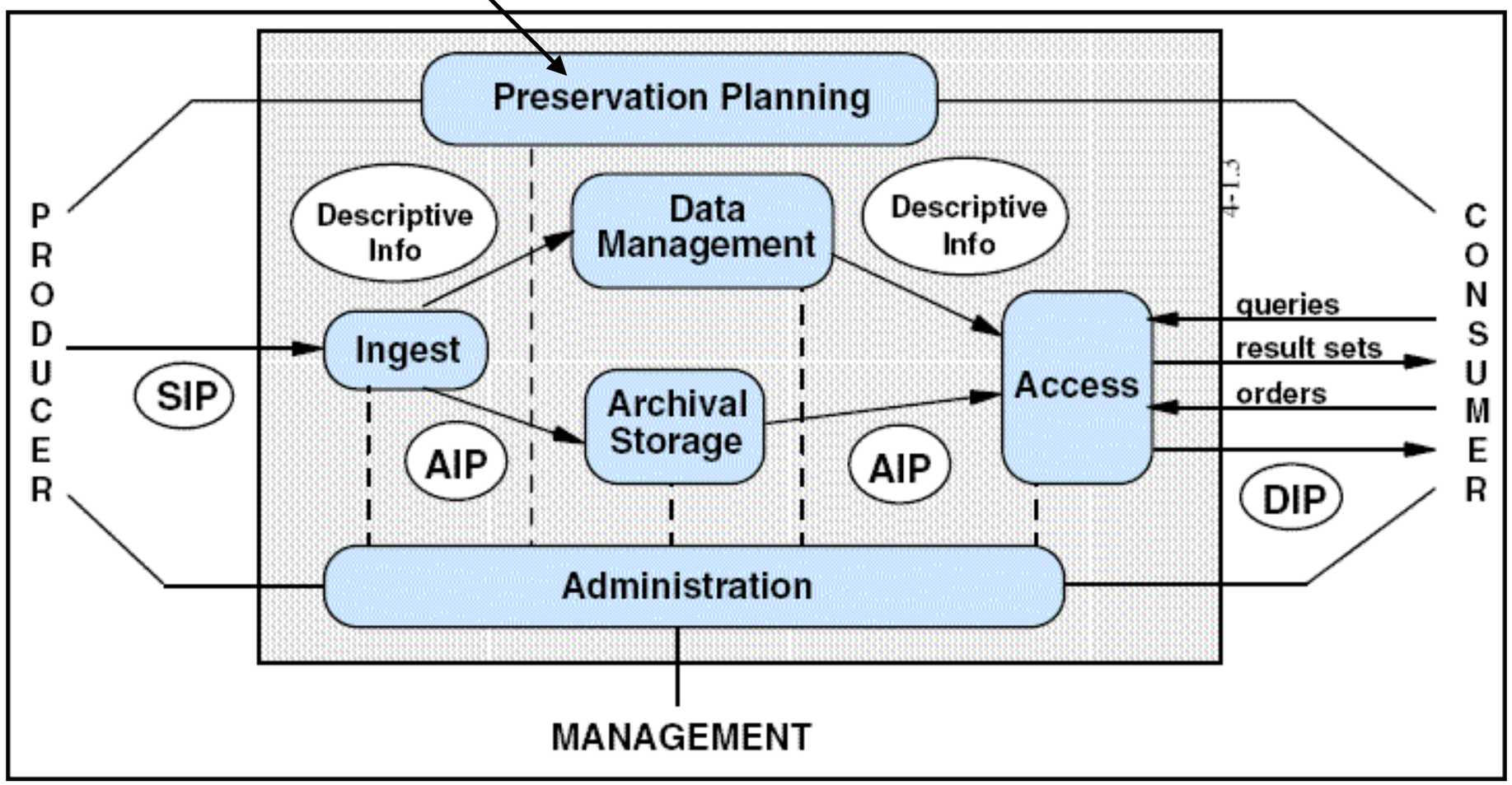
3. Preservation Actions

- 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.

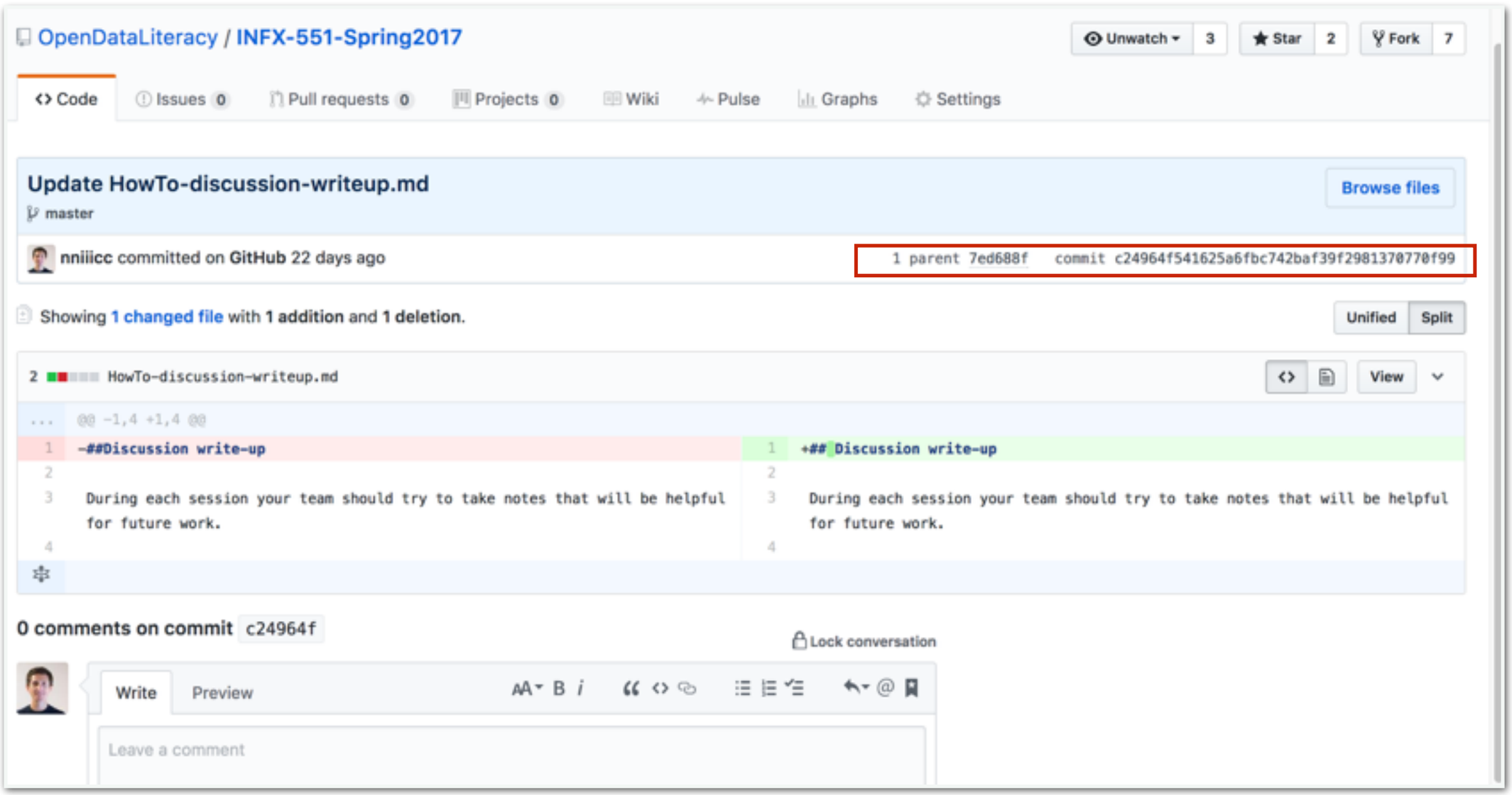
4. Digital Preservation Environments

- 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
    - Disseminate -> Dissemination Information Package

Hey, look


Preservation Planning
Descriptive Info
Data Management
Descriptive Info
PRODUCER
Ingest
SIP
AIP
Archival Storage
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 - 152.8 KB - Apr 27, 2017 - 0 Downloads
MD5: 76eadfec75ae2ba8d2277f9311b80926

Download