THE BARTLETT
INSTITUTE FOR SUSTAINABLE HERITAGE

Best practice and tools for use and re-use
[of scientific data in archaeology and heritage]

Dr Scott Allan Orr
Lecturer in Heritage Data Science

SSHOC/SEADDA/E-RIHS Workshop on
Use and Re-Use of Scientific Data in Archaeology and Heritage

2 April 2020

Outline

- Data in heritage and archaeological contexts
- Re-purposing data
- Best practice
- 5-star Open [Linked] Data
- Analysis

Data stewardship versus re-use

- Inherently related
- Data management for **re-use** oriented towards comprehension
  - Accurate
  - Efficient
  - Transparent
- ‘Think about your future self…’

What is specific to scientific data?

- Produced by computer, but interpreted by humans and computers
- Wide range of formats, usually structured
- Can be of staggering scale
- Power is in interconnectivity
- *Analysis*

Data in heritage and archaeology

- Scientific imaging
- Analytical techniques
- Environmental monitoring
- Collection records
- Born-digital assets
- n-D representations

Data in heritage and archaeology

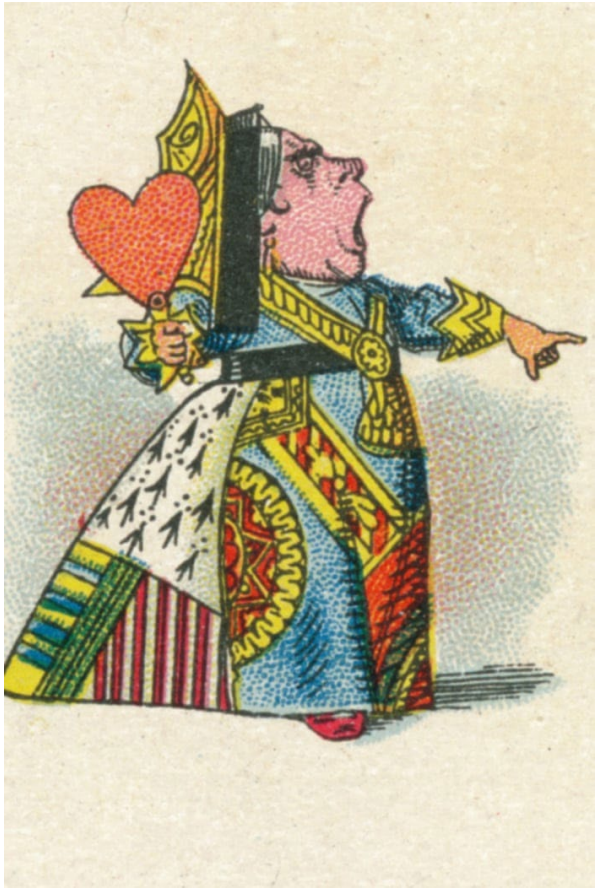
- Metadata
  - Structural, administrative, rights, management
- Paradata
  - Process by which the data was collected
  - Can be integral to informing re-use by understanding intentions, bias, and gaps

Potential re-uses of heritage and archaeological scientific data

- Conservation and management
- Interpretation
- Engagement

Examples of best practice

- File formats (proprietary versus plain)
- Headers
- Eternal columns
  - Is there a reason to keep them? Software, human assumption, etc.

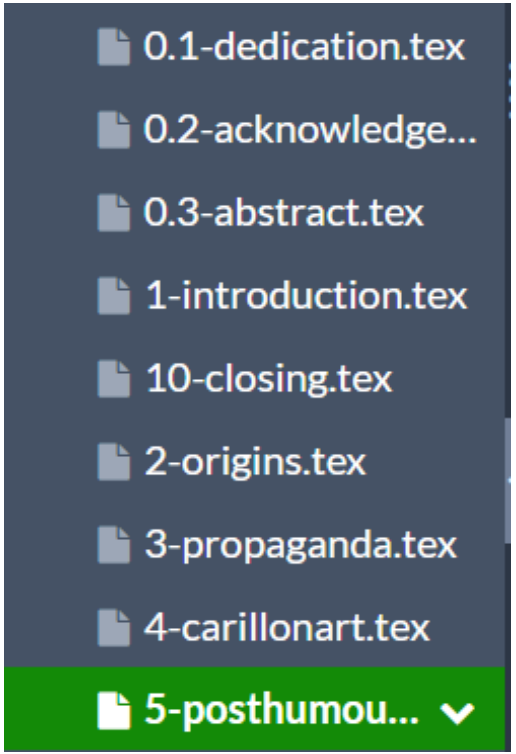


Examples of best practice

- Archive unprocessed data
  - Enables the widest variety of types of re-use

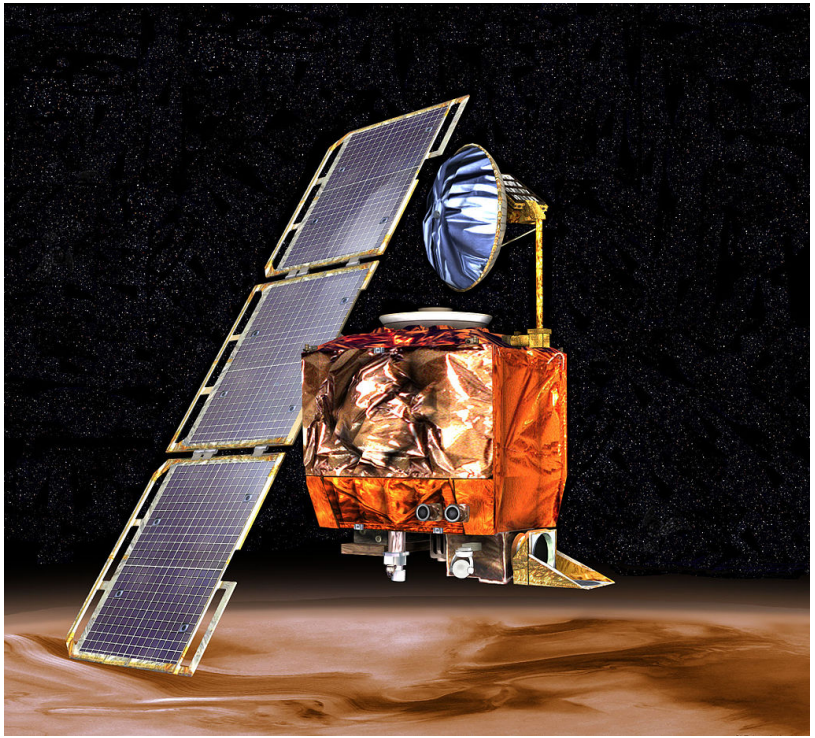
Naming conventions

- Dates: YYYYMMDD
  - Chronological ordering
- Leading zeroes
- Alternatives to spaces
  - file_underscore
  - CamelCase or camelCase
  - file-dashed



Examples of best practice

- Floating points and units
  - Cultural differences
  - Disciplinary variation



Examples of best practice: file formats

- *Lossless* file formats

Data type	Lossless	Lossy
Image	RAW, TIFF (usually), PNG (reversible), BMP (proprietary)	JPG, JPEG, GIF (for more than 256 colours)
Audio	WAV	MP3, AAC, OGG
Video	Few, due to size	H.264, H.265, MPEG4

5-star Open Data

★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

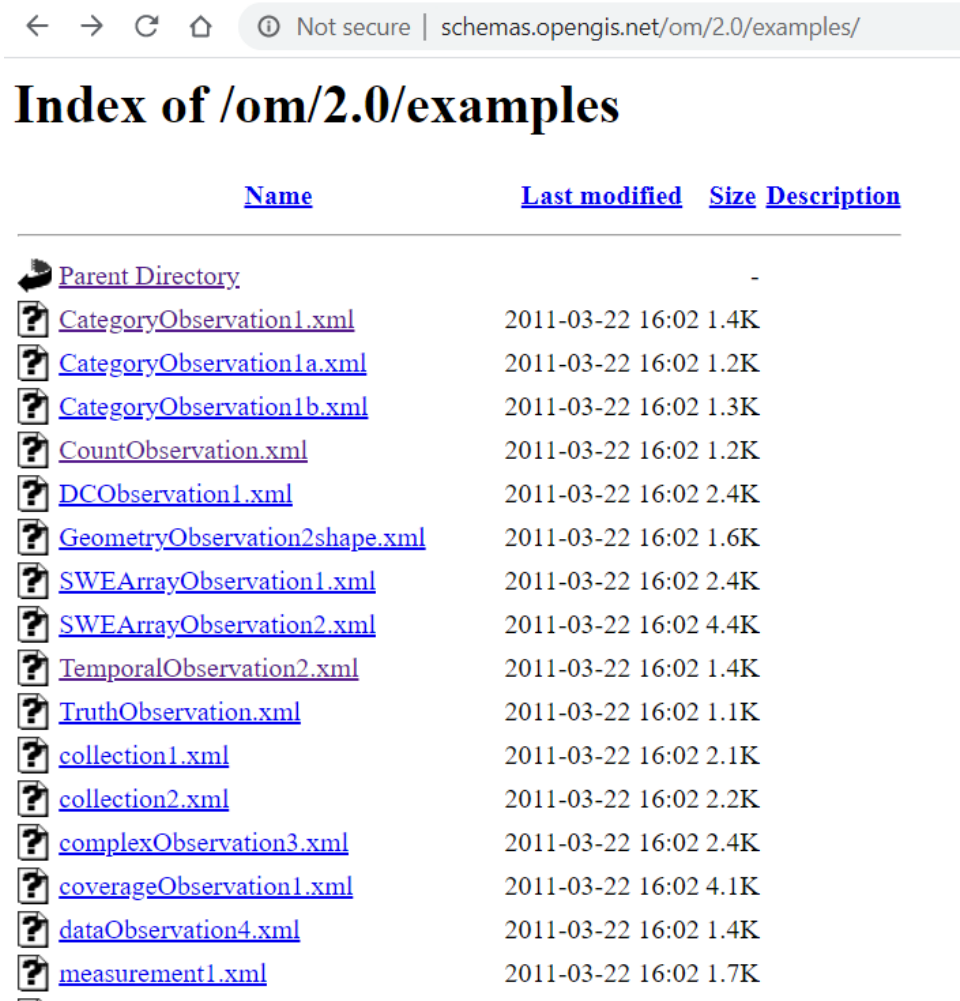
★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]

https://5stardata.info/en/

XML Data Formats on the Web

- XML implementation for the OGC and ISO Observations and Measurements (O&M) conceptual model (OGC Observations and Measurements v2.0 also published as ISO/DIS 19156)



Example: True or false?

```
<om:OM_Observation xmlns:om="http://www.opengis.net/om/2.0"
                   xmlns:xs="http://www.w3.org/2001/XMLSchema"
                   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                   xmlns:xlink="http://www.w3.org/1999/xlink"
                   xmlns:gml="http://www.opengis.net/gml/3.2" gml:id="obsTest2t"
                   xsi:schemaLocation="http://www.opengis.net/om/2.0
                                       http://schemas.opengis.net/om/ 2.0/observation.xsd">
   <gml:description>Observation test instance: truth test</gml:description>
   <gml:name>Observation test 2t</gml:name>
   <om:type xlink:href="http://www.opengis.net/def/observationType/OGC-OM/2.0/OM_TruthObservation"/>
   <om:phenomenonTime>
      <gml:TimeInstant gml:id="ot2t">
         <gml:timePosition>2005-01-11T17:22:25.00</gml:timePosition>
      </gml:TimeInstant>
   </om:phenomenonTime>

   <om:resultTime xlink:href="#ot2t"/>
   <om:procedure xlink:href="http://www.example.org/register/party/abc99"/>
   <om:observedProperty xlink:href="urn:example:Truth"/>
   <om:featureOfInterest xlink:href="http://wfs.example.org?request=getFeature featureid=Statement37f "/>
   <om:result xsi:type="xs:boolean">false</om:result>
</om:OM_Observation>
```

http://schemas.opengis.net/om/2.0/examples/TruthObservation.xml

Example: True or false?

Product that has a price of $200” The following would be valid XML:

```
<item>
   <title>AWorkOfArt</title>
   <creator>NameOfArtist</creator>
</item>
```

Another valid XML could be:

```
<item title=”AWorkOfArt”>
   <creator>NameOfArtist</creator>
</item>
```

Modeling this same data in RDF would only have one way of representing it:

```
ex:item1 rdf:type ex:artwork .
ex:item1 ex:title “AWorkOfArt” .
ex:item1 ex:creator “NameOfArtist” .
```

URI

Adapted from: https://www.dataversity.net/introduction-to-rdf-vs-xml/

Analysis: code, algorithms

- Code and algorithms
- Algorithms are not just code – also procedures
  - Analog and digital
  - Manual and automatic

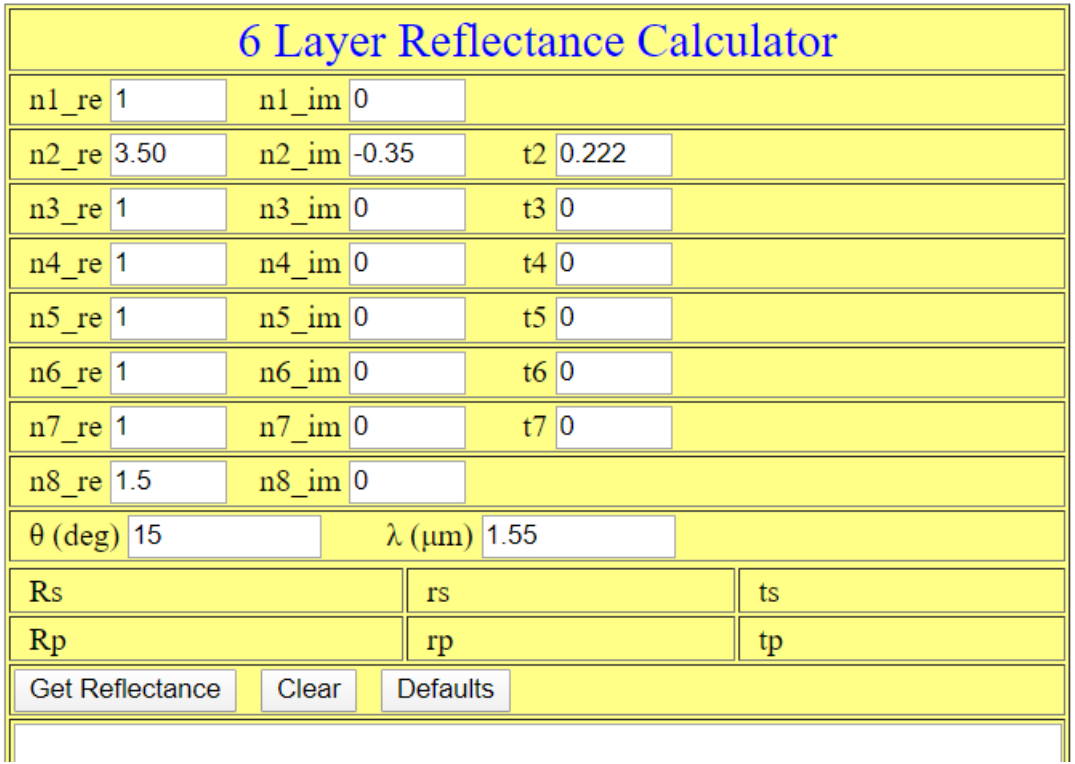
Code documentation

- Sample use

```
1  # READ.MOIST350B
2  # SCOTT A ORR
3  # 19 MARCH 2017
4
5  # dir <- "/Users/orrscott/Desktop"
6  # filename <- "GraniteWalls_Set5b2.hfs"
7  #
8  # setwd(dir)
9  # filepath <- paste0(dir, "/", filename)
10 # d <- readLines(filepath)
11 #
12 # arrs.t <- hfRead(d)
13
14 ######
15
16 # FUNCTIONS ########
17
18 hfRead <- function(filepath) {...}
40
41 makeArray <- function(line.ref, layer.starts) {...}
51
52 makeLayer <- function(line.ref, dims, y.vals, x.vals) {...}
63
64 norm <- function(matrix) {...}
69
```

Code documentation

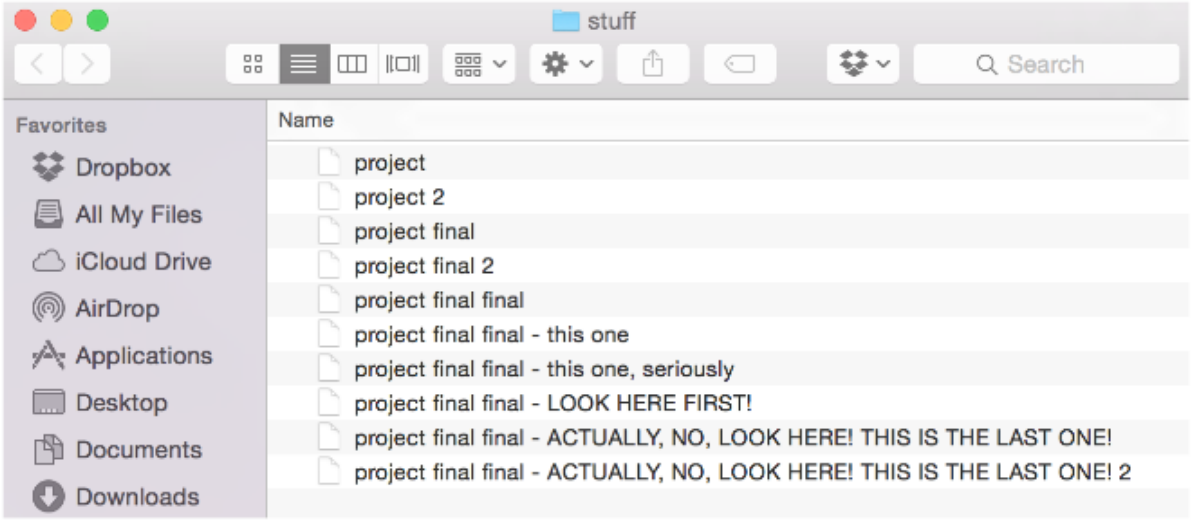
- Markup



```
var Nsqr = [0, 0, 0, 0, 0, 0, 0, 0] ;
var Nsqi = [0, 0, 0, 0, 0, 0, 0, 0] ;
var ksqr = [0, 0, 0, 0, 0, 0, 0, 0] ;   // initialize various arrays for 8 regions
var ksqi = [0, 0, 0, 0, 0, 0, 0, 0] ;
var kxr = [0, 0, 0, 0, 0, 0, 0, 0] ;
var kxi = [0, 0, 0, 0, 0, 0, 0, 0] ;


k0 = 2*Math.PI/lam;     // in um-1
k0sq = k0*k0;
//kz = k0*N1r*Math.sin(theta);      // propogation constant along layers (same in each layer .... basically Snell's law of refraction) .. real for real angle of incidence
//kzsq = kz*kz;
kz_r = k0*N1r*Math.sin(theta);      // kz may be complex if incident medium is complex
kz_i = k0*N1i*Math.sin(theta);
kzsq_r = Csqr(kz_r, kz_i );
kzsq_i = Csqi(kz_r, kz_i );



for(i=0; i<8; i++) {
 Nsqr[i] = Csqr(Nr[i], Ni[i]) ;
 Nsqi[i] = Csqi(Nr[i], Ni[i]) ;
 ksqr[i] = k0sq*Nsqr[i] ;
 ksqi[i] = k0sq*Nsqi[i] ;
 kxr[i] = Crootr(ksqr[i] - kzsq_r,  ksqi[i] - kzsq_i);
 kxi[i] = Crooti(ksqr[i] - kzsq_r,  ksqi[i] - kzsq_i);
 if (Ni[0] == 0 &&  kxi[i] > 0) {       // if incident medium is lossless, normal components of propog. const. must all be negative; otherwise select other root
    kxr[i] = -kxr[i] ;
    kxi[i] = -kxi[i];
  }
}
```

Version control

- Version control, Git (e.g. Github, etc.)



Structure

- Each script is one step, operating on a single subject
- One master file to run them analysis
- Keep multi-use components in a central place
  - All config set-ups go in here

```
1  # READ.MOIST350B
2  # SCOTT A ORR
3  # 19 MARCH 2017
4
5  # dir <- "/Users/orrscott/Desktop"
6  # filename <- "GraniteWalls_Set5b2.hfs"
7  #
8  # setwd(dir)
9  # filepath <- paste0(dir, "/", filename)
10 # d <- readLines(filepath)
11 #
12 # arrs.t <- hfRead(d)
13
14 ######
15
16 # FUNCTIONS ########
17
18 hfRead <- function(filepath) {⟺}
40
41 makeArray <- function(line.ref, layer.starts) {⟺}
51
52 makeLayer <- function(line.ref, dims, y.vals, x.vals) {⟺}
63
64 norm <- function(matrix) {=⟺}
69
```

“Put your faith in what you most believe in…”

- Don’t trust someone else’s code
- Never trust your own code !

Additional tips for code

- Save the intermediate result of each script**
  - Useful if the processing takes a long time
  - Not necessarily that important if analysis is quick
  - Storage and memory limitations apply as well
- Plot everything

Delete superfluous files

.DS_Store	17/08/2016 11:38	DS_STORE File	7 KB
.Rhistory	01/08/2019 12:54	RHISTORY File	1 KB
FLIR1760.csv	16/08/2016 16:31	Microsoft Excel Co...	515 KB
FLIR1760.jpg	16/08/2016 16:20	JPG File	722 KB
FLIR1760.txt	16/08/2016 16:36	Text Document	516 KB
FLIR1814- photo.jpg	16/08/2016 16:21	JPG File	74 KB
FLIR1814.csv	16/08/2016 17:09	Microsoft Excel Co...	515 KB
FLIR1814.jpg	16/08/2016 16:21	JPG File	673 KB
IRFit_fromT.R	16/08/2016 18:53	R File	6 KB
Untitled.R	16/08/2016 17:42	R File	6 KB
Untitled2.R	17/08/2016 12:11	R File	7 KB

Types of analytical tools

- Consider a wide range of users (or your future self)
- Open source (Python vs. MATLAB)
  - Need to balance with existing packages and tools

© Patrick JEAN / muséum d'histoire naturelle de Nantes

Final points

- Disclaimer: do as I say, not as I do
- Best practice and tools for re-use
    - Meet funder requirements
    - Foster future use (you would hope!)
    - Improve the robustness of current work
    - Support collaboration
    - Save time (now, and in future; and for all involved)
    - Avoid awkward (and sometimes serious) mistakes

THE BARTLETT
INSTITUTE FOR SUSTAINABLE HERITAGE

Best practice and tools for use and re-use
[of scientific data in archaeology and heritage]

Dr Scott Allan Orr
Lecturer in Heritage Data Science

SSHOC/SEADDA/E-RIHS Workshop on
Use and Re-Use of Scientific Data in Archaeology and Heritage

2 April 2020