Hands-on Exercises on Data Transfer, Calculating Checksum, and Metadata Extraction

Ritu Arora

Email: rauta@tacc.utexas.edu

October 27, 2014

Objectives

- You will learn about
  - Different data transfer protocols
  - Utility of file metadata including checksums

- What will you do
  - Connect to Stampede
  - Copy data from a source to destination using some of the protocols discussed today
  - Install a software named DROID on Stampede
  - Run DROID for metadata extraction
  - Run scripts for checksum calculation

Exercise 0 : Accessing Stampede & the Sample Dataset

- Log on to Stampede using **your_login_name**

This is the username for Stampede that was provided to you on the sign-up sheet

- Uncompress the tar file, **workshop_data.tgz** , that is located in the **~train00** directory into your $SCRATCH directory

```
ssh <your_login_name>@stampede.tacc.utexas.edu

cds

tar -xvzf ~train00/workshop_data.tgz
```

- This dataset is around 7.7 GB in size, it might take 20 – 30 minutes to copy and extract and hence let us move to the next part while this is going on
  - Will come back to check on the transfer and use the transferred data

Data Transfer

- What did we just do?
  - Copy the data from one account on Stampede to another
  - This is also a form of data transfer but from one file-system to another on Stampede
  - The dataset had the appropriate permissions set so that you could copy it form another account on Stampede (`~train00` to yours)
- If you are getting started on Stampede for conducting your data management functions, as a **first step, you would need to get your data from some other resource to Stampede for any processing or analyses**
- Different protocols exist for data transfer between remote sites, *e.g.*,
  1. Linux command-line utilities **scp** & **rsync** (will be practicing these)
  2. Globus Connect (will be covered in the demo by Vas Vasiliadis)
  3. Globus' globus-url-copy command-line utility (will not be covered today)

Data Transfer Using **scp**

- If your local computer is a Mac or a Linux laptop, you can use the `scp` commands to transfer data to and from a remote resource like Stampede

```
localhost% scp filename
username@stampede.tacc.utexas.edu:/path/to/project/di
rectory
```

- If you are using a Windows computer, you can download and use the WinSCP application (GUI-based) or download and use Cygwin (command-line based, can run the aforementioned commands)

Data Transfer Using `rsync` (1)

- The `rsync` command is another way to transfer data and to keep the data at the source and destination in sync
- If transferring the data for the first time to a remote resource, `rsync` and `scp` might show similar performance except when the connection drops
  - If a connection drops, upon restart of the data transfer, `rsync` will automatically transfer only the remaining files to the destination, it will skip the already transferred files
- `rsync` transfers only the actual changed parts of a file (instead of transferring an entire file)
  - this selective method of data transfer can be much more efficient than `scp` because it reduces the amount of data sent over the network

Data Transfer Using **`rsync`** (2)

- The following example demonstrates the usage of the `rsync` command for transferring a file named `myfile.c` from the current location on Stampede to Lonestar's - another supercomputer a TACC - `$WORK` directory

```
login1$ rsync myfile.c
username@lonestar.tacc.utexas.edu:/work/01698/username/
data
```

Data Transfer Using **rsync** (3)

- Transferring an entire directory from Stampede to Lonestar
  - To preserve the modification times use the `-t` option
  - To preserve symbolic links, devices, attributes, permissions, ownerships, etc. transfer in the archive mode using the `-a` option
  - To increase the amount of information displayed during transfer use the `-v` option (verbose mode)
  - To compress the data for transfer, use the `-z` option
  - The following example demonstrates the usage of the `-avtz` options for transferring a directory named `gauss` from the present working directory on Stampede to a directory named `data` in the `$WORK` file system on Lonestar

```
login1$ rsync -avtz ./gauss
username@lonestar.tacc.utexas.edu:/work/01698/username/data
```

Exercise 1: Using `rsync` for data transfer

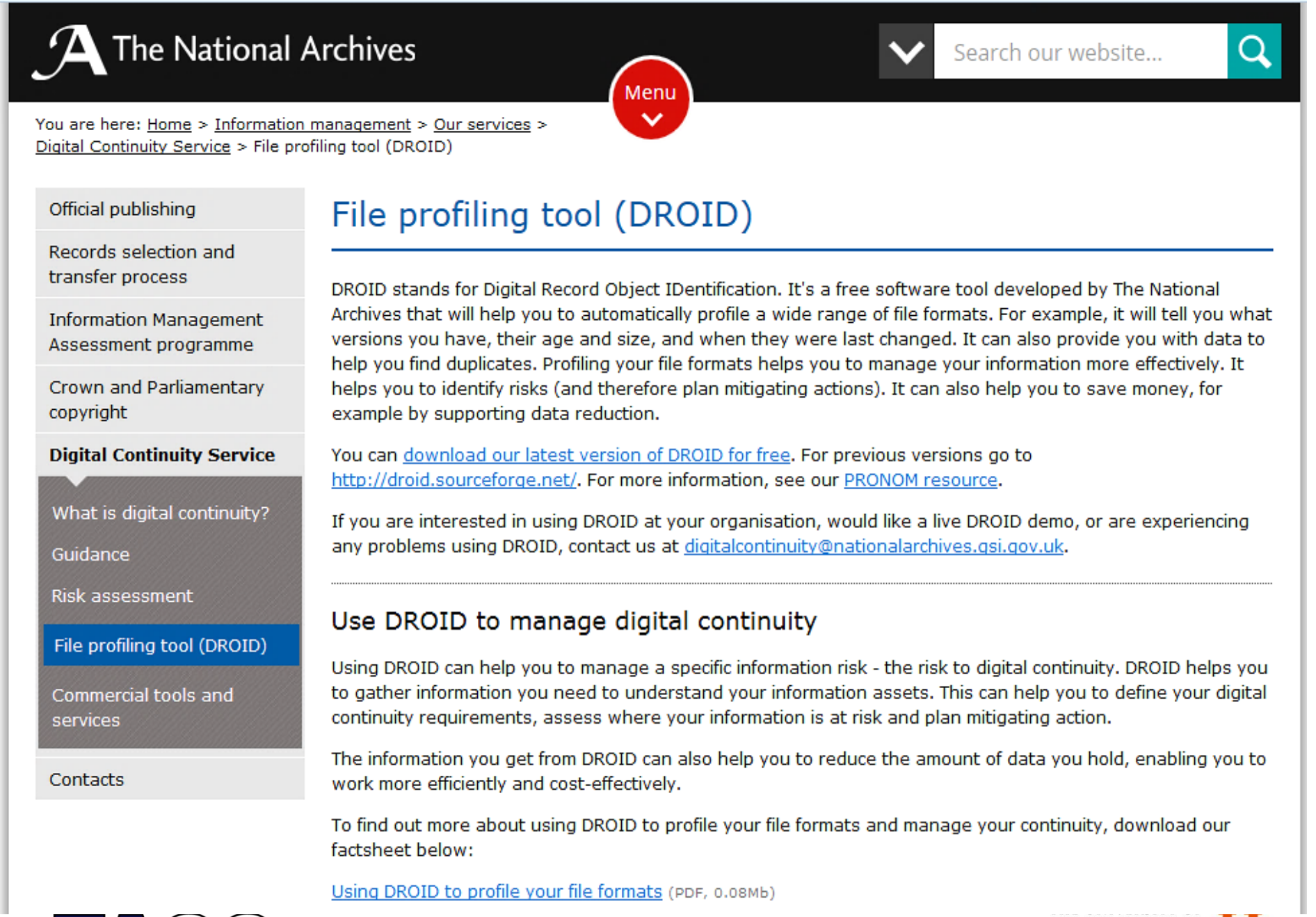
- Transfer a file from Stampede to Hopper
  - Connect to Stampede using SSH
    - Hostname: **stampede.tacc.utexas.edu**
    - Username and password: provided to you on the sign-up sheet
  - Connect to Hopper using SSH
    - Hostname: **hopper.nersc.gov**
    - Username and password: provided to you on the sign-up sheet
  - Create a file on Stampede – name it as "testTransfer.txt"

```
staff$ cat > testTransfer.txt
Hi I am testing data transfer from Stampede to
Hopper
<press ctrl +D>
```

  - Transfer the file to Hopper using `scp` and `rsync`

File Metadata

- Metadata is a well-defined or formal data that can be used for managing an information source
  - Useful in the search and retrieval of relevant information
  - Useful in clustering of related data
  - Useful for archiving and preservation purposes
- Some of the metadata elements that are used for file identification are file-name, file-type, file-formats, date of modification, checksum *etc.*
- One of the goals of digital preservation is to ensure that the digital files are accessible even in future – the metadata needed for this purpose is known as preservation metadata
  - *viz.* file-format and file-version
  - the DROID tool can be used to obtain this information


The National Archives
Search our website...
Menu
You are here: Home > Information management > Our services > Digital Continuity Service > File profiling tool (DROID)
Official publishing
Records selection and transfer process
Information Management Assessment programme
Crown and Parliamentary copyright
Digital Continuity Service
What is digital continuity?
Guidance
Risk assessment
File profiling tool (DROID)
Commercial tools and services
Contacts
File profiling tool (DROID)
DROID stands for Digital Record Object IDentification. It's a free software tool developed by The National Archives that will help you to automatically profile a wide range of file formats. For example, it will tell you what versions you have, their age and size, and when they were last changed. It can also provide you with data to help you find duplicates. Profiling your file formats helps you to manage your information more effectively. It helps you to identify risks (and therefore plan mitigating actions). It can also help you to save money, for example by supporting data reduction.
You can download our latest version of DROID for free. For previous versions go to http://droid.sourceforge.net/. For more information, see our PRONOM resource.
If you are interested in using DROID at your organisation, would like a live DROID demo, or are experiencing any problems using DROID, contact us at digitalcontinuity@nationalarchives.gsi.gov.uk.
Use DROID to manage digital continuity
Using DROID can help you to manage a specific information risk - the risk to digital continuity. DROID helps you to gather information you need to understand your information assets. This can help you to define your digital continuity requirements, assess where your information is at risk and plan mitigating action.
The information you get from DROID can also help you to reduce the amount of data you hold, enabling you to work more efficiently and cost-effectively.
To find out more about using DROID to profile your file formats and manage your continuity, download our factsheet below:
Using DROID to profile your file formats (PDF, 0.08Mb)

Please Note

- The command-line version of DROID does not provide checksums
- You would need a different script for checksum
- If DROID does not work once, try again!

Checksum Calculation

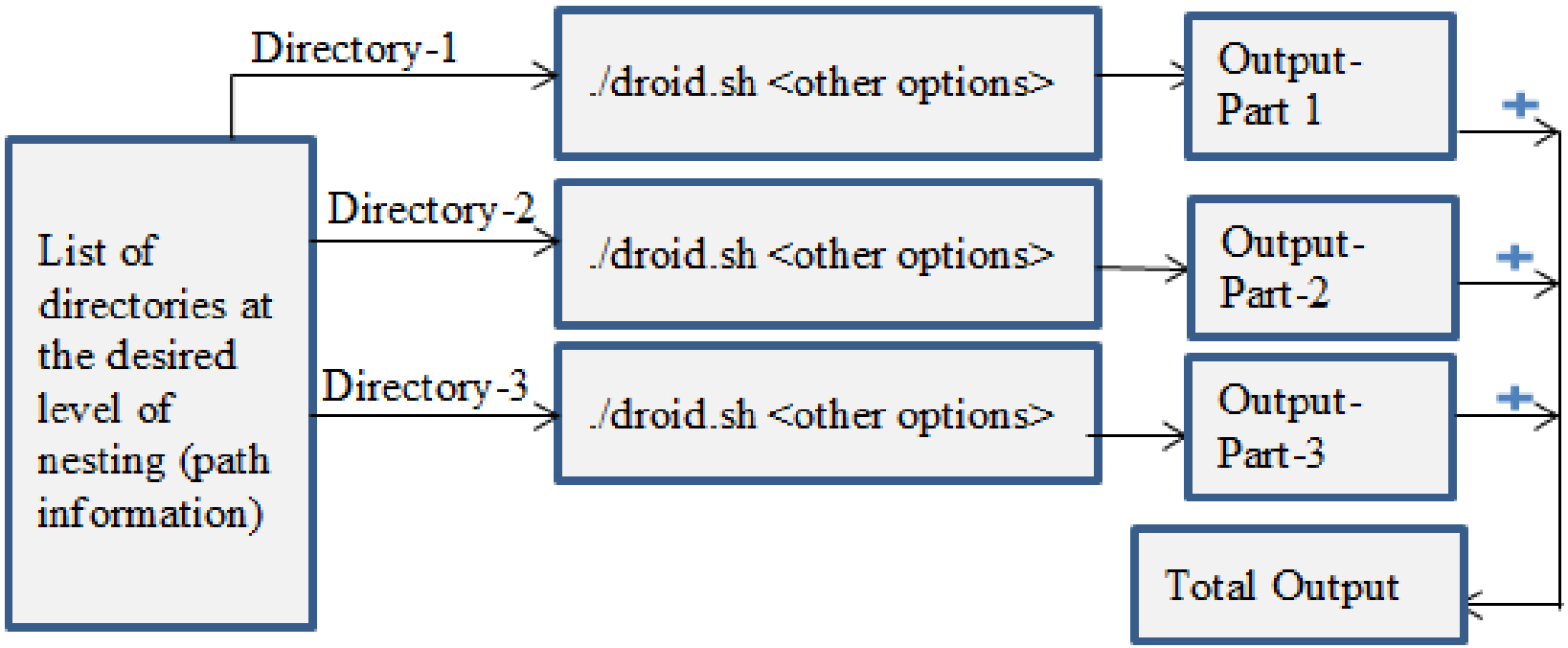
- A checksum or hash sum is a small-size datum from an arbitrary block of digital data for the purpose of detecting errors during its transmission or storage
- By themselves checksums are often used to verify data integrity but can be used for detecting duplicate files and removing or reorganizing them in a data collection
- Common algorithms for calculating checksums: md5sum, sha1sum

Get the Link from the Linkedin.com group

Switch to Google Doc:

https://docs.google.com/document/d/10ojgr-jw5v0EG6iJsV1_m7t_jNbxCpk8fDSafVIgHtU/edit?usp=sharing

Running DROID in Parallel



1. **We ran 31 parallel instances of DROID after coarsely dividing the workload (directories) for file-profiling into 31 directories**
2. **The entire work-flow (work-load distribution, submitting multiple DROID jobs on the supercomputer, combining the results of multiple DROID runs, etc.) is mostly automated, few manual steps though**

Time-taken when Different Number of DROID Instances are Used Simultaneously