

Clipper Project Phase 3 ‘Snapshot’ Case Studies: Time-Based Media as Research Data

Author: John Casey

Date: 13 September 2016

Status: Working Draft

Disclaimer: This represents the views and impressions of the author only and not any of the participating organisations – and should not be construed as such.

A Preamble here about the project from existing resources (PowerPoint, brochure etc.)

Contents

The Roslin Institute (Bio Sciences) V2	3
Key Data Workflow Stages.....	3
Roslin: Sharing and content management requirements.....	3
Sharing.....	3
Admin and Editorial Roles.....	4
Clipper Export Options.....	4
<i>User generated Clipper data</i>	<i>4</i>
<i>Compiling user generated Clipper data together with the source media files.....</i>	<i>4</i>
Issues to consider	5
Integration with The Current Research Information System (CRIS) ‘PURE’	5
Integration with existing UoE Data storage solutions and possible hybrid cloud solutions	5
Internal / External users of Clipper at Roslin	6
Use of DOI.....	6
Graphical Representation of RDM uses of Clipper at Roslin.....	6
Understanding typical research work flows – for video.....	6
Existing UoE / Roslin Research Data Storage Solutions	6
<i>Roslin local computers and network drives.....</i>	<i>6</i>
<i>DataStore.....</i>	<i>6</i>
<i>UoE Data vault.....</i>	<i>6</i>
<i>Pure (CRIS).....</i>	<i>7</i>
<i>DataShare.....</i>	<i>7</i>
<i>Media Hopper</i>	<i>7</i>
Background Context(s) and Drivers.....	8
<i>Institutional Context (Roslin).....</i>	<i>8</i>
<i>Wider Context and Drivers.....</i>	<i>8</i>
Royal Scottish Conservatoire	8
EUSCREEN	8
BUFVC.....	8

Draft V1

The Roslin Institute (Bio Sciences) V2

Based on a series of discussions about research workflows and research data management needs

Key Data Workflow Stages

Creating / Using / Collaborating / Managing / Sharing / Publishing / Curation / Preservation

Think about mapping this to the Jisc RDM lifecycle graphic...? **Yes to do**

Roslin: Sharing and content management requirements

The points described below are in addition to the basic feature set of the Phase 2 Clipper prototype - which includes the ability to create projects to contain 'virtual' clips together with their annotations and collate them into cliplists and to be able to share them individually via URIs.

The following scenarios / requirements / use cases have been identified and are in the process of being clarified:

Sharing

- Flexibility and control over the management and sharing of the data are paramount all the way through the research process this parallels the needs in learning and teaching that have been emerging through discussion in the community workshops
- It is important to understand that researchers from different institutions are increasingly collaborating together on projects (partly through funder requirements). At the same time these researchers are often in competition with each other to be first to publish etc. Therefore, there is a requirement to have quite granular control over **what** is shared, **who** it is shared with and **when** it is released (embargo), the **duration** of sharing and the **conditions** (licence) governing what can be done with the shared item. For instance this means that some Clips / Cliplists may be shared but their related annotations may not be.
- Here are the draft requirements for sharing:
 - Projects - who, when, duration, licence
 - Cliplist – who, when, duration, licence
 - Clips – who, when, duration, licence
 - Annotation – who, when, duration, licence
- In practice this means a project (with multiple cliplists) might be shared with others but that some of the content of that project (cliplists, clips, annotations might be held back)
- Sharing of the different items needs to be able to be controlled with variable degrees of openness and the system will need to accommodate this. We will use the 'onion skin' metaphor here, to describe the levels of sharing that are needed - ranging from very specific to totally open. The system needs to be able to accommodate the definition of additional levels / groups. Users will also need to be able to change the shared status of their own items at will. Initial suggested access levels are listed below:
 - Just me – private

- Named individuals
- Group(s) – could be a research group composed of members drawn from different institutions or from within one institution
- Department / Faculty of an institution
- An institution
- Open to the web
- Sharing controls will need to include the ability to control export of the project contents in different editable formats (e.g. Json, CSV, Zipped web site etc.) to:
 - Named individuals
 - Groups
 - Department / Faculty of an institution
 - An institution
 - Open to the Web

Admin and Editorial Roles

- The sharing requirements require some kind of administration / editorial roles and functions in the system, to be able to control what users can do with the items they have access to via sharing – when installed in an institutional setting:
 - Super User / Admin role(s) who can:
 - Create / import user accounts and close / disable accounts
 - Assign privileges to users
 - Create user groups
 - Lock / Unlock projects
 - Group Admin who can
 - Manage the members of a group (in / out / disable edit functions i.e. view only)
 - Control what items a user can edit (useful for workflow management)
 - Can ‘roll back’ selected group projects to an earlier state (assumes a version control system is in place – planned for later development)
 - Can control the ability to control export out of the group data in different formats

Clipper Export Options

User generated Clipper data

Export selected projects, cliplists, clips, and annotations in a range of different formats (without the source media files):

- Self contained websites (Json, html, css) as zip files
- CSV
- Json
- XML (converted from Json)
- Others

Compiling user generated Clipper data together with the source media files

Be able to export the user-generated data together with the source media files in order to make the Clipper documents ‘unbreakable’ by ensuring the links to the

parent resources cannot break. This has been a common call at the workshops. We see two main compile options:

1. Export with the original source files in their entirety – with a choice of media transcoding options for the source files to cater for different platforms and choice of resolutions
2. Export with ‘physical’ excerpts from the original source files that match the clip specifications for start and end points – with a choice of media transcoding options for the source files to cater for different platforms and choice of resolutions

Issues to consider

Integration with The Current Research Information System (CRIS) ‘PURE’

There will be a need to integrate the Clipper toolkit into the Edinburgh CRIS (Current Research Information System) and more investigation needs to be done about this. However initial discussions have identified some of the admin and sharing features that Clipper will need to support that may be useful for enabling this. Including the ability to ‘compile’ a Clipper project with its source video files and export to data repositories. This has come up at several workshops.

At the moment PURE is the location where research papers and metadata about the research is kept and it is not intended to contain research data. From an institutional perspective, PURE is intended to be both the central location for research information management and the public facing site to promote and enable open access to the research outputs of the University. Bearing this in mind it might be an option to use PURE and the DOI system to help record and manage the location of the public and private data sets that Clipper users are referencing.

A simple solution to this might be to use a simple web page with a DOI, the web page could list the publicly available locations of the research data together with any Clipper document URIs that reference the original research. In this scenario the CRIS and its research data / Clipper page and DOI would act as a data ‘clearing house’ that would link out to the data repository(s) to where the larger audio-visual files and their related Clipper documents are stored. The use of a simple web page and a DOI would also greatly help in discoverability and hence aid in improving the research impact for the REF. Use of the DOI reference in the Media Hopper metadata / item description would be advisable and could link viewers through to the research (and vice versa).

Integration with existing UoE Data storage solutions and possible hybrid cloud solutions

See below for UoE solutions. Due to the size of some of the video data it may be worth looking at partial or total cloud solutions especially for video data that has a requirement to be made publicly available. This may include the use of Content Delivery Networks (such as Azure or Amazon) to serve the video data to the public (or even ‘free’ social media services like YouTube / Vimeo for lower costs) – these could be used in conjunction with DOIs and UoE of storage locations .

Internal / External users of Clipper at Roslin

Discussions so far indicate a clear distinction between using Clipper as a research data analysis tool on 'live' research projects (mostly closed) and as a dissemination tool to publicise the results of research. For dissemination in REF terms this would include in assisting in the 'impact' factor and to provide public research data to satisfy funders and publishers open research data mandates. Clipper might be first used at UoE for the outside function.

Use of DOI

There are a number of possible places to store and manipulate and possibly collaborate with video at UoE as can be seen in the section below on existing research data storage locations. For both public and private video data we discussed the possibility of Roslin 'minting' its own DOIs to record definitive information about the videos – with the DOIs resolving to a simple web page that listed the information and locations of the videos (open and closed), this could be very useful in helping to comply with funding and publisher mandates.

Graphical Representation of RDM uses of Clipper at Roslin

A mind map / venn diagram to show:

- Uses / users
- Open / closed services

Understanding typical research work flows – for video

We need to get an idea of the steps involved and at what point the video could be shared publicly and what needs to be done with it to make it shareable – jc note really me getting a couple of stories down on paper so I can understand it – the size issue, transcoding and steps to share etc.

Existing UoE / Roslin Research Data Storage Solutions

Notes: the size of some of the video files presents problems to the existing provision described below

On-site in the institution

Roslin local computers and network drives

- – not public

DataStore

- intended for managing live research data across the University- not public
– first 0.5Tb is free - <http://www.ed.ac.uk/information-services/research-support/data-management/data-storage>

UoE Data vault

- – for the preservation of research data from previous research - not public <http://www.ed.ac.uk/information-services/research-support/data-management/data-vault> - not sure about costs still a service in development

Pure (CRIS)

- - the UoE research management information system eprints repository for research information management and reporting needs research papers etc. see the quotes below – public but not really intended for research data deposit <http://www.ed.ac.uk/governance-strategic-planning/research/pure> - more about holding research papers and published results etc. However links from here to video storage locations and the use of DOIs would help
 - “Information held in PURE relates to research-active staff and their projects and collates research activity information (mainly publications but also other types of output such as professional activities, impacts and press clippings).”
 - “The default position in PURE is that all content is publicly visible via the Research Explorer (except for In Press or In Preparation research outputs).”

DataShare

- - is the UoE IS solution for researchers who want to deposit / share their data (in the long term) – public - <http://datashare.is.ed.ac.uk/> - current limit is 10Gb per file (<http://www.ed.ac.uk/information-services/research-support/data-library/data-repository/checklist>) - it does support MP4 and it is possible to embargo a deposit. Note the advice to researchers is to try and ‘future proof’ their content by choosing open / common file formats such as MP4 (a sensible choice in our opinion for both video and audio).
 - Initial tests suggests this could be a possible source for Clipper documents – however the deposit would need to be in appropriate formats - needs further investigation and testing

Media Hopper

- - <http://www.ed.ac.uk/information-services/learning-technology/media-hopper/introduction> - This is a new service is based on the open source Kaltura platform. It is intended to support video (and audio) across the university:

“Media Hopper is a new service that puts video at the heart of teaching, learning, research and public engagement. The goal is to use video to enrich learning, communicate more effectively to external audiences and further enhance the institution’s research activities.

This has the potential to play a part in Roslin’s use of video both for live internal ‘closed’ research and for public dissemination. From initial discussions, (for live research) the possibility of using high definition video research data in conjunction with Clipper for international collaboration is an attractive one. Especially with the parallel option to share ultra-high resolution still frames via the internet to access all the source data when needed.

Background Context(s) and Drivers

Institutional Context (Roslin)

Roslin is a large internationally recognised Research institute, part of The Royal (Dick) School of Veterinary Studies that resides within the College of Medicine and Veterinary Medicine at the University of Edinburgh.

Much of the research at Roslin generates very large data sets, such as:

- Gene sequencing 'runs' that can generate many terabytes of data
- Very high resolution time lapse video taken by specialist microscopes with data at several gigabytes per-frame

The very high size of the data sets involved presents problems for storage and management of the data inside the UoE systems

- The university wants to be able to monitor and record what digital research data is being generated, where, by whom and what it is being used for

Wider Context and Drivers

- Collaboration between institutions and researchers is a growing requirement from funders yet competition between researchers and institutions can be intense – the system needs to be able to accommodate controls over what is shared with whom etc. as described above
- Publishers Open data mandates – i.e. to provide research data under an open licence to meet the academic criteria of proving the reproducibility of research findings asserted in journal articles)
- Research funding now often includes open data mandates - i.e. to make research data openly available for others to use)

Royal Scottish Conservatoire

EUSCREEN

BUFVC