Kapturing the Right Software for the Arts

Carlos Silva
KAPTUR Technical Manager
Visual Arts Data Service
University for the Creative Arts

Screenshots from the key software assessed through the KAPTUR project, for managing research data in the visual arts.

Background

KAPTUR is funded by JISC through the Managing Research Data programme (2011-13), and aims to discover, create and pilot a sectoral model of best practice in the management of research data in the visual arts. The project is led by the Visual Arts Data Service (VADS) a Research Centre of the University for the Creative Arts, in collaboration with four institutional partners: Goldsmiths, University of London; The Glasgow School of Art; University of the Arts London; and the University for the Creative Arts itself.

Whilst it is accepted that each institution is unique, the aim of KAPTUR was to work collaboratively to create and refine a model that will be effective within the partner institutions, as well as for other specialist arts institutions or multidisciplinary institutions with art departments.

The Technical Infrastructure work package led by the Technical Manager built upon the previous Environmental Assessment work package during which the partner institutions had begun to investigate the current state of visual arts research data. The Technical Manager continued this work with a series of interviews with the four KAPTUR Project Officers and with IT staff at each partner institution, with the purpose of creating a user requirement to inform the selection of research data repository.

The technical analysis report (Garrett, L. et al., 2012) encompassed both the user requirement as well as a study of seventeen potential systems. Using a scoring mechanism, based on one point per requirement, three of these systems were identified as potential solutions for the KAPTUR project. Following consultation with the project partners it was decided to further investigate figshare, DataStage and EPrints. In October 2012, following the project partners' agreement, the piloting of figshare and DataStage came to an end and a new potential solution was introduced - CKAN. CKAN was scored using the same matrix from the technical analysis report to ensure consistency.

This case study was requested by the KAPTUR partner institutions as an additional output, because they expressed that the technical analysis work had been a valuable outcome of the project and that it was hoped that by disseminating this more widely it would save other institutions time and resources.

Expectations

From the beginning of the technical work package, it was clear that there was no single product that could completely fulfill all the requirements of the KAPTUR project partners. This was later confirmed by the technical analysis where three contenders were identified in the first round and a fourth contender at a later stage during the project.

EPrints was already in use at the partner institutions, and was both graded and ratified by the Project Officers in each partner institution as the most viable option which fulfilled most of the requirements of the project. However EPrints was not a clear-cut winner as the grading by the partner institutions was very close, and there were elements of figshare, DataStage and CKAN, which fulfilled some of the requirements that the EPrints software is not able to perform, or which would require additional development work. Examples of these requirements are: a 'local' file management environment; improved visualization of documents and multimedia, and a user-friendly upload feature. Specifically the project team began to view the need for a two stage set-up, with one installation for managing active research data and a second repository to deposit the completed research data. Due to the partners' familiarity with EPrints for research outputs it was decided this could fufill the role of the second repository; therefore this has been a constant throughout the KAPTUR project. In order to manage active research data, three systems have been considered during the project: figshare; DataStage; and since October 2012 - CKAN:

- an integration of EPrints with figshare;
- a separate piece of work linking DataFlow's DataStage with EPrints;

- and finally an integration of CKAN with EPrints.

Approach

Technical infrastructure analysis

The technical analysis was framed around the research question: which technical system is most suitable for managing visual arts research data?

The first stage involved a literature review including information gathered through attendance at meetings and events, and Internet research, as well as information on projects from the previous round of JISCMRD funding (JISC, 2009-11).

During the second and third month of the work package, the Technical Manager carried out interviews with the four KAPTUR Project Officers and also met with IT staff at each institution. This led to the creation of a user requirement document, which was then circulated to the project team for additional comments and feedback.

The selection criteria were agreed with appropriate representatives from the four institutional partner institutions and used to evaluate potential software solutions, bearing in mind the scope and resources of the project. Throughout this stage of the project the team identified five key requirements:

Solution Type
Research data management software costs vary widely but generally fall into one of two main types: open source or commercial software. The partners expressed a strong preference for open source software, which was fortunate as project resources were limited.

Storage
The project team identified a requirement for the research data management solution to be able to handle a variety of different types of data, from simple and small text items to large complex multimedia items with the flexibility or potential to include unusual file formats.

Interface
The solution must also comply with W3C standards; provide quality assurance features; and provide a user-friendly and engaging upload service.

System
This selection criterion related to system requirements such as operating systems, virtual servers and cloud storage environments, which any potential solution might need to address. Consideration was also given to defined limits for data upload and the ability to integrate the software with tools and other software currently in use by the partner institutions.

Institutional
Institutional requirements included specific requirements from each partner institution in terms of workflow, statistical reporting, legal compliance, preservation and disposal of data.

Methodology

Seventeen systems were identified as having potential to meet the requirements of the project. From the total, six were selected as the most suitable for use with visual arts research data. Each of these was then considered by the team during the selection process with reference made to issues facing the visual arts (Garrett, L. et al, 2012).

Following the initial selection of five potential solutions, a further review, using a matrix of priorities defined by the Project Officers, was undertaken; this returned the following scores, in order of

usefulness to the visual arts: EPrints (184.00); DSpace (180.00); figshare (171.75); DataFlow (171.00); and Fedora (159.00). When CKAN was later scored using the same system its score was 176.00. EPrints was graded and verified by the Project Officers as the most viable option because it fulfilled most of the requirements of visual researchers and their host institutions. However it was also acknowledged that the scoring of all the solutions was extremely close and there were elements that the EPrints software was not able to perform without further development work, in particular a local file management environment for managing active research data.

Therefore to fully appreciate and understand how best to meet the research data management requirements of researchers and their institutions it was agreed that two software pilots would be considered: an integration of DataFlow's DataStage with EPrints and figshare with EPrints. Following the investigation of the first two pilots which were discounted by the project partners, a third pilot was investigated: CKAN with EPrints. Significantly all three set-ups have been tested by the Project Officers and in some cases by other staff at their institution.

Three pilot research data management systems

1. DataStage to EPrints

DataStage, part of the DataFlow project, is an open source software which is currently developing and promoting a free-to-use cloud-hosted system for management, preservation and publication of research datasets (University of Oxford, 2012).

The project is based on the prototype developed by the JISC-funded ADMIRAL project (University of Oxford, 2009-11) which developed a two-tier federated data management infrastructure for use by life science researchers. This provides services to meet researchers' local data management needs for the collection; digital organisation, metadata annotation and controlled sharing of research datasets; and provides an easy and secure route for archiving annotated datasets to an institutional repository, The Oxford University Data Store. The Data Store assigns Digital Object Identifiers (DOIs) and uses Creative Commons licensing, and it also enables long-term preservation and access to research data.

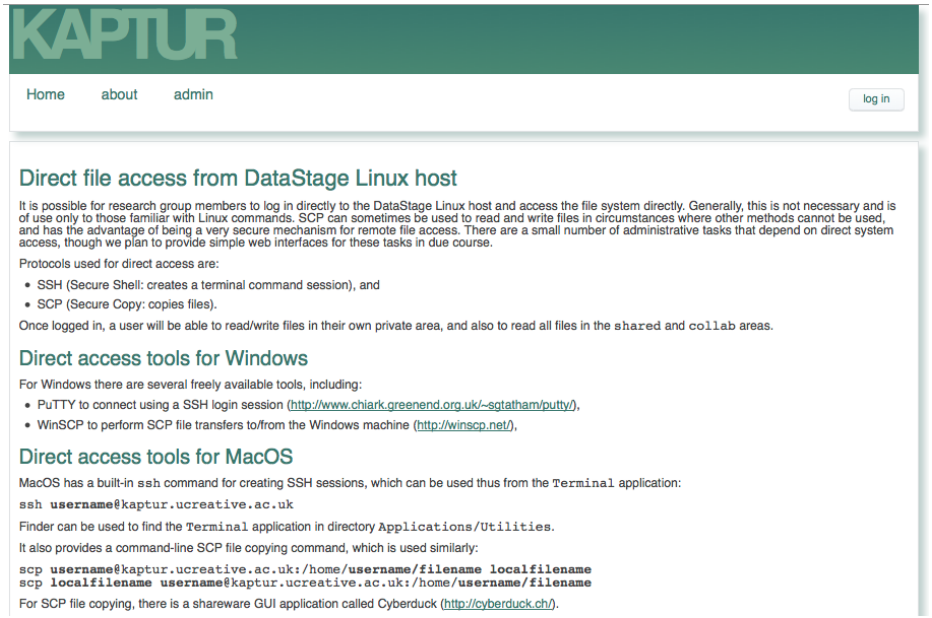

Fig.1 – DataStage setup for KAPTUR

During the KAPTUR project, DataStage was used and tested as originally proposed on the technical analysis review. The software offered a simple deposit interface managed by an administrator or by the researchers themselves, a structured metadata collection interface and a

popular storage approach similar to that of Dropbox. The software was tested and analysed by the Project Officers together with the Project Manager and feedback was received. However, DataStage is currently under development and although it has been releasing development versions of the software for both its DataBank and DataStage solutions, its current version is not yet ready for public release and the production environment. A problem with the protocol used to transfer content from DataStage into EPrints prevented the pilot from carrying forward and the development work stopped until further notice.

2. EPrints

EPrints was developed at the University of Southampton and is freely available as open source software (University of Southampton). Originally designed for creating and managing open access institutional repositories of digital research papers and publications, EPrints is now used to store and manage a much broader range of content types and data.

Led by the University of Southampton, the JISC funded Kultur project (2007-09) piloted a model for repositories suitable for the specialist needs of arts researchers, and founded start-up repositories for research outputs at University of the Arts London and University for the Creative Arts (University of Southampton, 2007-9).

EPrints can accommodate different types of workflows. These can be edited to provide different options such as sending email notifications to administrators and editors. Its content can be stored in any file format as designated by the repository manager during configuration and multiple representations of the same content are permitted.

With the release of EPrints version 3.3 in September 2011, repository managers can install applications, plugins and updates with the EPrints Bazaar, which can be downloaded and installed without affecting the core configuration and settings of the repository, and applications can also be easily disabled or deleted.

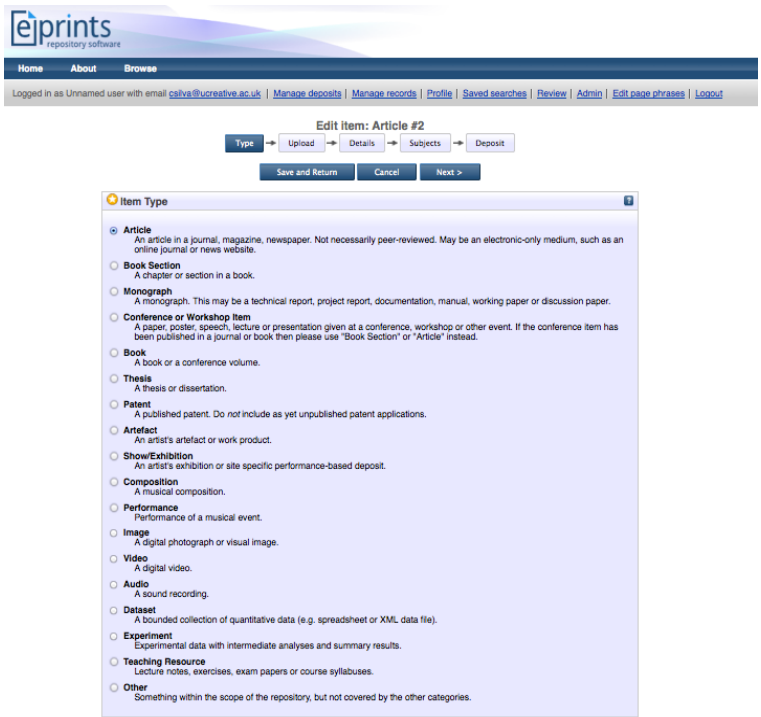

Fig.2 – EPrints 3.3 with 'Kultur' tools embedded

The KAPTUR project has been testing different types of research data in a pilot EPrints 3.3 repository whilst decisions were made about the system for managing active research data.

3. figshare to EPrints

figshare is a web-based platform aimed at addressing the needs of individual researchers. It was originally developed as an 'open science project' by Mark Hahnel whilst he was completing his PhD at Imperial College, University of London; it is now supported by Digital Science (from September 2011) and was re-launched with improved functionality in January 2012.

Researchers are encouraged to publish all their research data online, including negative data and unpublished data. Persistent identifiers are provided by the Handle System; Creative Commons licences are used; and there are tools to enable searching and sharing of data.

The idea behind using figshare as part of the KAPTUR pilot was that it offered a simple deposit interface managed directly by the researchers themselves. In addition figshare offers an interactive and easy to use interface where any published data is presented according to its file type.

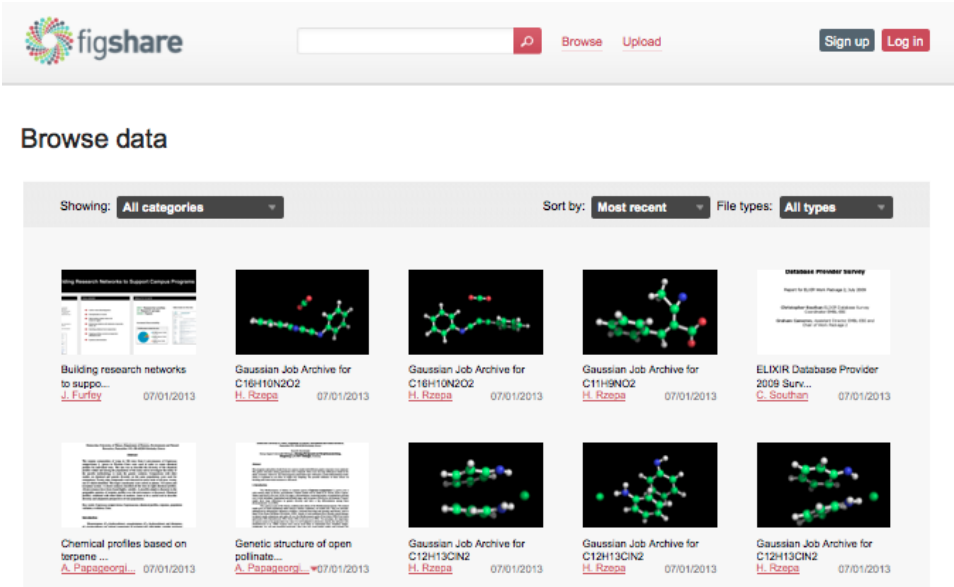

Fig.3 – figshare's search engine and current output display

Additionally the Technical Manager worked closely with the project team to propose development work to enable figshare to be integrated with EPrints. There were reservations about the cost of the development work and this was compounded by the discovery that figshare expected all datasets to be made available under a Creative Commons CC0 licence. Unfortunately the licence set-up was non-negotiable and restricted the partner institutions from continuing the pilot with figshare. This was not only due to the different issues with visual arts research data which can be commercially viable or ethically sensitive, but also due to the importance of encouraging the proper citation of visual materials (CC0 requires no attribution). Finally the other issue identified with CC0 licensing is that once a CC0 licence has been granted it cannot be revoked.

3. CKAN to EPrints

CKAN is a web-based system for the storage and distribution of data, and the contents of databases supported by the Open Knowledge Foundation. It is inspired by the package management capabilities common to open source operating systems like Linux.

The system is used both as a public platform on thedatahub.org and in various government data catalogues, such as the UK's data.gov.uk, the Dutch National Data Register, and The United States government's data.gov 2.0.

CKAN was introduced to the KAPTUR project following a demonstration from the JISC funded Orbital project at the JISCMRD programme workshop at Nottingham in October 2012. CKAN is being used by Orbital to manage and store research data, with the project developing a 'bridge' between CKAN and their EPrints repository.

Following a review, a complete assessment against the user requirements, tests and feedback from the project partners and group, the Technical Manager together with the Project Manager and Project Director/Principal Investigator made the recommendation to pilot CKAN as part of the KAPTUR project.

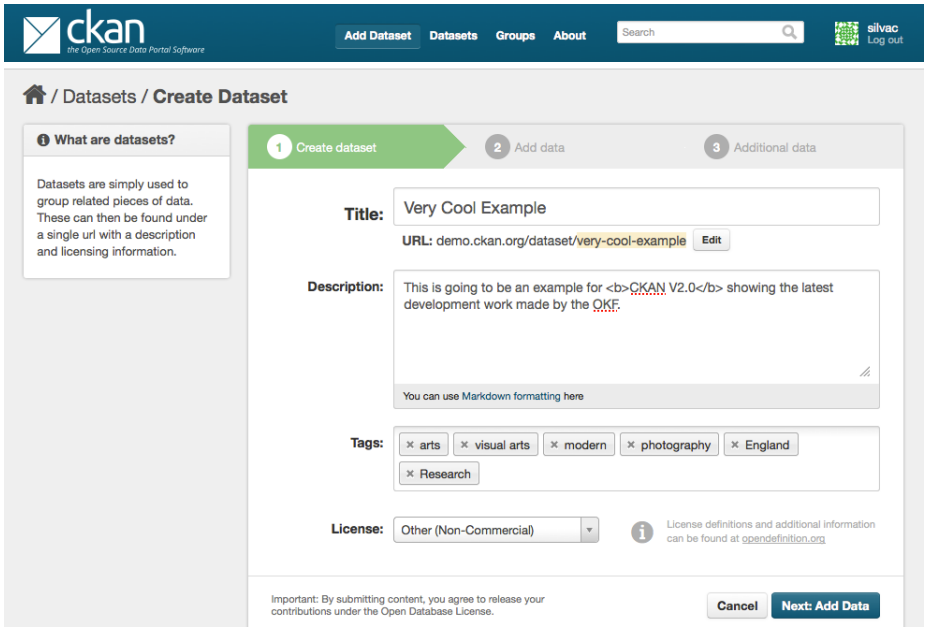

Fig.4 – Simple and user-friendly upload interface on CKAN 2.0

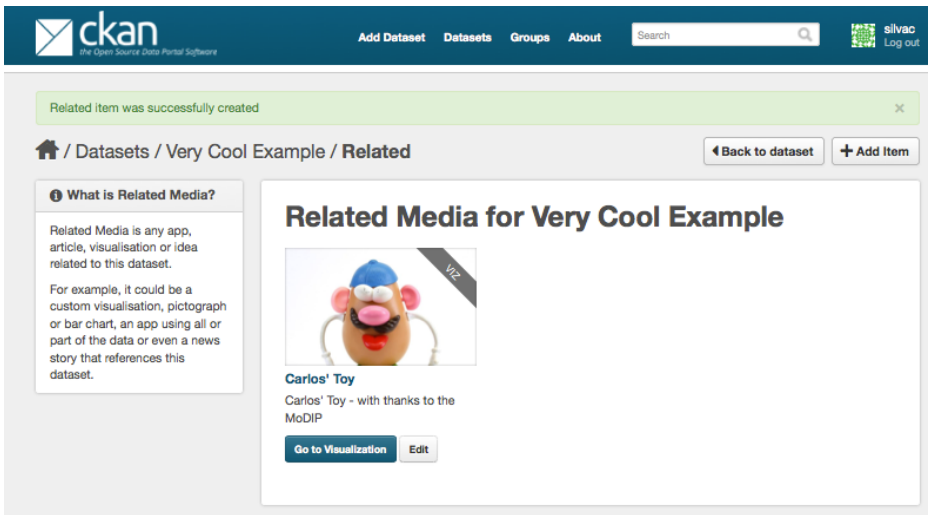

Fig.5 – Adding related media files or additional resources to existing data on CKAN 2.0

Overall, CKAN offers a data management platform capable of handling versioning, role management, data previews, multiple metadata formats, dataset history, comprehensive search, plugins and application programming interfaces (APIs) amongst other functionality. These, together with the support given by the Open Knowledge Foundation and government bodies across the world make CKAN a sustainable product capable of addressing some of the outstanding needs of the KAPTUR project. Project Officer feedback on CKAN has been positive based upon ease of use, interface design, and the available features. The Technical Manager has

met with representatives from the Orbital project as well as the Open Knowledge Foundation. The only drawback discovered so far has been the issue with versioning: CKAN is being developed by the Orbital project as version 1.7 and so in order to make use of their 'bridge' software this is also the version we need to work with; however the latest version of CKAN contains several significant improvements which are recommended by the Open Knowledge Foundation.

Next Steps

Given the nature of the requirements involved to completely fulfill the management of research data, it has not been possible to identify a single system capable of being installed across the four partner institutions; however the partner institutions have been willing to explore alternatives and have consequently provided valuable input into this work package. Additionally, Goldsmiths, University of London have already installed an EPrints 3.3 repository to store their research data; this will be investigated by the institution over the next year as mentioned in their case study..

Work in the area of research data management has been pioneered by other institutions and recently the common interest in using CKAN as part of a long term solution to address research data management means that not only the visual arts research audience can benefit from it, but also a much broader audience. The CKAN4RDM workshop, led by the Orbital project and including various JISCMRD programme funded projects such as KAPTUR, was organised with the aim of detailing community requirements for managing research data using CKAN (Silva, C., 2013).

The KAPTUR project has provided tools for the partner institutions, and available to the wider sector; the project has already addressed not only the technical issues and possible solutions for visual arts research data management, but also created momentum at a senior management level to continue with further development and an adoption of a tool to successfully manage their research data.

Conclusions and Recommendations

At this point in time there is no single solution, which can completely fulfill all the requirements of researchers and research teams, and their host institutions in the visual arts. The piloting of EPrints, as the preferred choice, with the addition of features from three of the other systems has allowed the project team to investigate, test, document and identify a more comprehensive and viable research data management system for the visual arts.

The latest adoption of CKAN and its continuous support from the Open Knowledge Foundation means that CKAN is placed as a much stronger contender compared with other software solutions, particularly in terms of sustainability, which is significant when looking in the medium to long term scenario. Collaboration has played a key role during this project and has driven the project forward.

A lesson learnt during this project was that there are various considerations when working with beta versions of software due to the time needed to create, test and fix any issues with the software directly, which means that delays in third party software could impact directly on the outcome of the project.

Finally, from the visual arts community perspective, although CKAN can’t currently address all the requirements from our user requirements list, there is scope for further development as outlined in the recent CKAN4RDM workshop (Silva, C., 2013).

Key Points

- Collaboration has been central to the investigation of the appropriate technical solution, whether at the level of the partner institutions or within each institution across IT, Library and Research Office departments.

- Effective re-use of content including software and infrastructure to support research data can save time and resources.

- Software costs and sustainability are important considerations and particularly when dealing at senior management level within institutions.

- Community-based software and services have advantages and disadvantages; in theory the software should be sustained in the longer term, however there can also be significant delays with development of working versions of the software.

- It is important to consider new software in terms of its integration with existing systems and services both within and outside the institution; there are significant benefits to saving users time by linking with appropriate systems.

References

figshare, 2013. figshare. Available from: http://www.figshare.com [Accessed 22 February 2013].

Garrett, L., Silva, S., and Gramstadt, M., 2012. Kaptur Technical Analysis Report. Available from: http://vads.ac.uk/kaptur/outputs/Kaptur_technical_analysis.pdf [Accessed 22 February 2013].

JISC, 2009-11. JISC Managing Research Data (JISCMRD). Available from: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx [Accessed 22 February 2013].

Silva, C., 2013. CKAN4RDM workshop. Available from: http://kaptur.wordpress.com/2013/03/02/ckan4rdm-workshop [Accessed 05 March 2013].

University of Oxford, 2012. The ADMIRAL Project. Available from: http://imageweb.zoo.ox.ac.uk/wiki/index.php/ADMIRAL [Accessed 22 February 2013].

University of Oxford, 2012. DataFlow. Available from: http://www.dataflow.ox.ac.uk [Accessed 22 February 2013].

University of Southampton. EPrints. Available from: http://www.eprints.org [Accessed 05 March 2013].

University of Southampton (2007-9). Kultur Project. Available from: http://kultur.eprints.org [Accessed 05 March 2013].

Contact

Carlos Silva
Planning and Development Manager
Visual Arts Data Service (VADS) Research Centre
University for the Creative Arts
csilva@ucreative.ac.uk