Supporting Data Management for 3D and Raster Data: Lessons learned from the DataPool Project.

Gareth Beale, Steve Hitchcock, Hembo Pagi, Richard Boardman, Graeme Earl; University of Southampton; Southampton, UK

Abstract

Research institutions, funding bodies and researchers themselves are becoming increasingly aware of the need to manage imaging and 3D data. At an institutional level data management policies are playing an increasingly significant role laying down plans for the provision of infrastructure, policy and guidance. Drawing upon the preliminary results of the University of Southampton's JISC funded DataPool project, this paper will gauge the extent to which institutional policy development might be supplemented or even enhanced by an increased awareness of localised responses to the challenges of imaging and 3D data management. The paper will review approaches to data management that have been adopted by individuals and research groups and will propose that in many cases these developments might be pivotal in defining the form of institutional data management policy should take.

Introduction

Raster Image data and 3D data occupy an increasingly central position in research of all types. New applications of 2D and 3D imaging data are constantly being developed by or made available to research communities. Media which were recently highly specialized are now being produced and used in a wide range of research contexts from physical and applied sciences all the way through to fine art and the humanities. As the volume of data produced expands and the communities involved in the production of these data diversify, the requirement for varied and responsive approaches to data management has become more pressing than ever.

These changes are reflected in the increased emphasis placed by universities and research institutions on the development of coherent and co-ordinated approaches to research data management. The number of UK universities producing institutional data management guidelines for researchers has increased constantly during recent years [1]. At the same time UK funding bodies are increasingly requiring researchers to consider the management of their research at each stage of the data management lifecycle and to build these considerations into their workflow [2]

The development of institutional data management policy is frequently characterised as representing a challenge to the research sector [3, 4]. It can however represent an opportunity. The most obvious benefit to the research institutions lays in the prospect of data which is both better managed and more accessible, a particular attraction as research councils increasingly demand the open publication of research data [5]. But as well as producing valuable outputs, the design of an institutional data management policy can also be an enlightening process in itself. The research and design which goes into the development of an institutional data management policy represents a unique opportunity to investigate and ultimately to better understand the various ways in which people have responded to the challenges of data management.

The drive towards institutional data management at the University of Southampton has been led by The Institutional Data Management Blueprint project (IDMB) which took place between 2010 and 2012 [6]. The importance of imaging data of all types was recognised from the o2utset as the project mapped the future of institutional data management at Southampton and developed a plan for its implementation. Following on from this the JISC funded DataPool project has begun the process of implementation. IDMB stated from its inception that any meaningful data management policy should be coherent at an institutional level but that it should also reflect the needs of specific researchers and research groups these goals are reflected in the work of the DataPool project.

The research described here represents one strand of the DataPool Project. It has identified groups at the University who use 3D and Raster Image data and it seeks to better understand how they have responded to the challenges of data management a local level. The project aims to reveal these local examples of highly developed and innovative solutions to data management and to ensure that these examples help to drive the development of institutional data management at the University.

Methodology

The project was divided into stages, the first stage sought to locate external sources of guidance relating to data management best practice for users of 3D and raster image data. Directories of resources are to be compiled and made available to research staff within the University. These resources will help researchers to better understand the issues of data management and to design data management strategies appropriate to their research area.

The second phase involved locating facilities and equipment which were either exclusively or partly used for the creation or processing of 3D or imaging data. Data gathered during this phase of the project is to be made available through the University of Southampton's Open Data Service. This phase of the research was carried out in order to add to a pre-existing source of open data which will inform the planning of on-going research and spending within the University.

The final and most substantial phase of the project involved making contact with researchers who create or work with 3D data or raster image data and investigating the ways in which they currently manage data. A community of 30 researchers was assembled. Half of these were regular creators or users of 3D data and half were regular creators or users of raster image data. However as the research progressed it became apparent that there were frequent overlaps between these groups. Participants were distributed across the University of Southampton in disciplines including Fine Art, Physics, Archaeology, Geography, the Library Service, Engineering and Electronics and Computer Science.

Participants were asked to fill in an online questionnaire which invited them to describe key aspects of their standard data management processes. The results of this questionnaire were then used to inform a semi structured interview with the participant. Occasionally two or more participants would be interviewed simultaneously where they worked closely together.

Interviews were designed in order to identify key aspects of researcher's working practice and sought to better understand how researchers approached the challenges of data management. Resulting data allowed us to identify examples of innovation, best practice and creative data management as well as identifying several areas which were problematic across a range of disciplines. Some of the key findings will be summarised below.

Results

One of the trends with most quickly began to emerge from questionnaire results and interviews with researchers was the growing awareness which researchers had of the need to manage their data more effectively. Almost all of the researchers contacted were aware of the potential hazards which might arise from the inappropriate management of research data and had made changes to their data management practice in recent years to mitigate against these risks.

The risks most commonly cited by researchers were; the risk of lost data due to reliance upon inappropriate storage media, inaccessibility of data following staffing changes and associated loss of expertise or knowledge and the dangers associated with reliance on outdated machinery or technology for which limited support was available. This result was not surprising given that participants had been selected based upon the centrality of data to their work. These points all arise from the threat of disruption to core working practise and so are what might be termed *fundamental risks*. These risks are applicable to research data of all types. However, as shall become clear below, responses were specific both to the media being used and to the working practice of the researcher.

Strategies adopted by researchers to cope with these fundamental risks were diverse but fell into three broad categories. Some researchers had failed to address these issues and were reliant upon hard disk storage in a single desktop computer or other inappropriate media. Most however had adopted informal data management solutions which made frequent use of cloud based media sharing platforms in addition to University storage. Three research groups out of the 20 groups with which we came into contact had implemented locally developed data management systems which incorporated infrastructure and formalised policies for data management.

Informal systems

Many researchers drew upon a range of commercially produced technologies in order to back their data up and to perform data curation and publication tasks. Cloud based media sharing platforms such as Google Drive, Drop Box and Flickr were often cited as playing a central role in the preservation of data, particularly during the process of creation analysis and use. This was often though not always) in addition to the use of more conventional University supplied storage [7]. Participants described a range of secondary benefits derived from the use of these platforms in addition to their use as additional storage, these related especially to the curation and dissemination of data.

Researchers from one fieldwork based research project frequently used a dedicated project Flickr profile to upload and store images. At the most fundamental level the service offered a means of guarding against the loss of images while cameras and other imaging devices were in the field. This backup mitigated against the risk of data being lost if equipment was lost or damaged. The use of Flickr was also seen to have the secondary benefit of allowing the immediate sharing and dissemination of content amongst project members based in a number of countries. It allowed content to be highly searchable and structured with the addition of a few key words. These secondary benefits were used differently by different project members. They allowed research data management practice to evolve in response to specific circumstances and challenges as they emerged. Of particular interest to project members had been the use of variable access permissions and the use of Creative Commons licensing when images were made public. This allowed control to be exercised over who could see content and what the legal status of the images was. In this way a media sharing platform played a dual role as a means of storing, disseminating and effectively publishing content. These settings were used to great effect when the project in question became the subject of media interest, allowing images to be shared and made available to media outlets in real time.

Another example of the creative application of media sharing platforms lay in the use of a researcher in the arts and humanities. Part of the professional practice of the researcher revolved around photography and the production of visual media. The production and manipulation of images was not conducted exclusively during working hours. The need to differentiate between public and private content and the desire to be able to access this archive easily from a number of locations was important. A flexible system of access management was required in order to differentiate between material which was unequivocally personal and that which was integral to research, it was also important that this differentiation be flexible. The use of permissions settings built into media sharing platforms (in this case Drop Box) overcame the need for an arbitrary separation between personal and professional media storage and allowed the creative and dynamic use of this content.

The use of these proprietary systems offered functionality which researchers found to be invaluable to their research practice. The flexibility which these toolsets leant to the research process was believed by the participants to have improved efficiency, time management and data security. Crucially, the ability of researchers to adopt or to reject specific aspects of functionality meant that data management practice evolved in response to the requirements of the situation.

It is highly significant that among users of 3D data this informal adaptation of existing media sharing technologies was not nearly so widespread. Where it was used it tended to be used as a form of temporary storage. None of the users of 3D data used a system which allowed the online viewing and manipulation of 3D data. A range of responses are offered by researchers who use 3D data when asked why they did not make use of these additional functions. The most common response was that the sharing of 3D data using media sharing platforms was not nearly so highly developed as the sharing of image data. Consequently platforms did not offer a quick and easy method of uploading and viewing content. Furthermore, the complexity of 3D assets, which frequently consist of more than one inter-dependent file, meant that the correct configuration of files so that they could be properly viewed on the web was deemed too problematic to be worthwhile.

Collaboration

It was notable that several of the data management strategies developed were based upon collaborations with other disciplines, other research institutions or organisations from a different sector.

One example of this practice was the development of an integrated project data management system which was designed to manage research data from the point of capture to the point of publication. The system allowed the quantitative and qualitative analysis of large data sets which included large numbers of images, large 3D data sets and 3D models created for animation and publication. It also contained facilities for producing diagrammatic representations based upon quantitative data. This system was developed in collaboration with a commercial organisation working in the same sector. The development work was funded by a grant from a Research Councils UK funding body and the resultant system has been released as an open source project. Collaboration between a research and a private organisation was highly significant in helping to ensure that the tools developed would be of use across the field and not merely of benefit to those working in an identical area to the researchers. The system has since been promoted by the private organisation and has been taken up by several academic research projects, as a result the investment made in the system can be seen to have had significant impact on the field beyond the limits of the funded project.

Similar collaborations have led to systems which are widely used in the fields for which they were developed. Another research group at the University of Southampton have invested in the development of a system specifically for the curation and long term management of image data. Development in collaboration with an EU based partner university has ensured that the result could be funded to a higher level than would have been possible for either institution if working alone. In this case international development of this system is highly significant. The resulting system is multi-lingual and has a far wider audience than might otherwise have been possible.

Collaborations between organisations have numerous benefits. Greater efficiency through collaboration is a significant factor but the added value which comes from co-development can ensure that outputs are more versatile, more accessible and have a much longer life. Collaborations of this type perhaps offer a blueprint for a model of development which might be effectively employed at an institutional level. Collaborations occurred in these cases because the interests of the partners were co-incident on many levels. It is likely that these shared objectives and aspirations may also be found between faculties, research groups or even individuals. Furthermore these examples, particularly the first, offer an insight into the long term impacts which well-designed data management products with appropriately organised support structures can have.

Infrastructure

As well as developing informal approaches to data management as outlined above, the research identified examples of formalised data management strategies implemented at a local level. These strategies are notable for the fact that they have in many cases been comprehensive implementations involving localised policy, support and infrastructure.

There are two instances within the research of groups who have implemented significant local infrastructure in order to cope with the challenges of Data Management. In both cases these systems have been developed in order to cope with the storage of large data sets. One of the groups in question (Group B) was producing more than two terabytes of 3D and 2D data per day. The group produces data for external clients as well as for internal research purposes and consequently transparency and simplicity were integral features of their data management plan.

The group initially looked into using centralised University storage but found that the storage of big data in this way did not suit their requirements. The database system which they have developed is constantly adapted according to the needs of researchers and clients. Local hosting means that development work can take place at any time and is not restricted by the need for liaison or scheduling with those administering centralised storage. The costs associated with high security data storage are also a factor in deciding upon appropriate media and storage location. Centralised institutional repositories which are of sufficient scale to host archives of several hundred terabytes are generally designed with preservation in mind and consequently have additional costs associated with security. Storage of an archive of this size at the University of Southampton would simply not have been financially viable with annual costs of 1 terabyte of storage standing at £1000 in 2011/2012 [7]. The group estimate that storage in this way would amount to 50% of the cost of carrying out the recording work. These costs were believed by researchers to be at a similar level to many other institutions offering similar services.

Significant economic benefits were to be found in the extent to which storage of data allowed researchers to plan investment and efficient use of resources. Group B statistically analyse metadata relating to stored data sets and metadata produced by capture devices. These analyses allow the group to more efficiently timetable the use of capture devices and also to plan investment in storage media and replacement parts for capture devices.

Communication

The kind of creative approaches to data management which are described above all contain useful insights into how working practice can be improved. However, excellence at a localised level can only ever represent part of the picture. In order for localised best practice and innovation to have an impact at an institutional level it is important that channels of communication are in place. These must allow good ideas to spread and to be adopted elsewhere.

The research we conducted found that there were few channels available through which researchers could share their experiences of data management or disseminate their data management strategies to a wider audience. Furthermore, researchers tended to have a very limited awareness of the data management support that was available to them from services such as the University of Southampton Library's Data Management Guidance Service, the Software Sustainability Institute or the University of Southampton Library's Digitisation Unit.

Conclusions

The results gathered as part of this research project depict a research community in at a point of pivotal change. The researchers questioned affirmed that as research data continued to grow it would be necessary to consider how these data were curated and preserved.

Numerous examples of good practice were highlighted through the questionnaires and interviews. Researchers had frequently responded creatively to the challenges and opportunities which emerge from working with increasingly large collections of imaging data. Commercially available cloud based media sharing platforms were seen as being particularly useful for the curation of collections of images although uptake was far less prevalent where 3D data were concerned, perhaps due to the diversity of formats and data structures which were in use in this area.

There can be little doubt that local innovation was far more effective at managing the curation and publication than the preservation of data. This is perhaps not surprising when we consider the costs and the technical challenges of implicit in secure long term preservation of digital assets. Long term preservation requires investment in facilities and benefits from the provision of expert advice. Researchers frequently felt unable to deal with these challenges at a local level.

In addition, it was also noted that while localised innovation was frequently highly sophisticated, there were very few instances of research groups sharing expertise or consulting available expert guidance in order to refine their approaches. This did not appear to be due to a lack of willingness but was related to uncertainty as to where they might look to receive or to offer help. Researchers often suggested that they would be willing to share their ideas, methodologies and workflows but that they were unsure how to make contact with those to whom this advice might be applicable.

Exceptions to this tended to be collaborations which had been sought with external partners. Researchers who were involved in these collaborative approaches to data management saw several benefits to working in this way. The most immediate benefit was efficiency. Financial investment by two partners led to projects which were more fully realised than would have been possible if a single partner had undertaken the work alone. In addition to this pragmatic benefit researchers also highlighted the fact that collaboratively developed tools tended to incorporate a wider range of requirements and consequently be more versatile in their potential use. As mentioned above, the research group which collaborated with a commercial organization in the production of an open source content management system found that their co-investment had produced a resource with applications far beyond the scope of the project. These observations perhaps offer a model for the development of data management practice at all levels. The propagation of best practice is entirely reliant upon the efficient use of available resources and the effective communication and sharing of ideas.

If the development of institutional data management policy is to have a meaningful impact upon workflows and research practice it is imperative that existing innovations are used as an engine with which to drive this process. The work we carried out revealed a research community who were aware of the value of the content which they were producing and who were keen to increase the impact of the work they were carrying out. In order though for localized excellence to lead to institutional excellence we must build mechanisms through which best practice can is encouraged and where it occurs is shared.

References

[1] Data Curation Centre, UK Institutional Data Policies, Retrieved February 3, 2013, from Data Curation Centre: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies/uk-institutional-data-policies

[3] Ronald Yanosky, Institutional Data Management in Higher Education (ECAR Key Findings), Retrieved February 3, 2013, from Educause: http://net.educause.edu/ir/library/pdf/EKF/EKF0908.pdf

[2] Data Curation Centre, Overview of Funder's Data Policies, Retrieved February 3, 2013, from Data Curation Centre: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

[4] A. Hey and E Trefethen, The data deluge: An e-science perspective, In F. Berman, G. C. Fox, & A. J. G. Hey (Eds.), Grid computing: Making the global infrastructure a reality, page 809–824, (Wiley and Sons, 2003)

[5] Research Councils UK, RCUK Policy on Open Access, Retrieved February 3, 2013, from RCUK: http://www.rcuk.ac.uk/research/Pages/outputs.aspx

[6] K. Takeda, M. Brown, S. Coles, L. Carr, G. Earl, J. Frey, P. Hancock, W. White, F. Nichols, M. Whitton, H. Gibbs, C. Fowler, P. Wake, S. Patterson, Data Management for All - The Institutional Data Management Blueprint project, Proc. International Digital Curation Conference, pg. 12. (2010).

[7] University of Southampton, Research Data Storage Options, Retrieved February 3, 2013, from University of Southampton Libraries: https://www.southampton.ac.uk/library/research/researchdata/storage_options.htm l

Author Biography

Gareth Beale is a PhD student in Archaeology based at the Archaeological Computing Research Group at the University of Southampton in the UK. His research focusses on the archaeological applications and management of 3D data capture and physically accurate light simulation.