Revisualizing Visual Culture

Edited by
CHRIS BAILEY
Leeds Metropolitan University, UK

HAZEL GARDINER
King's College London, UK

London 2010

ASHGATE

Chapter 4
Resource Discovery and Curation of Complex and Interactive Digital Datasets

Stuart Jeffrey

The realm of the visual record, like almost all forms of information, has long since divorced itself solely from the physical. Much material that began life in one format – oil on canvas, text on paper, silver nitrate film – has gone through the ubiquitous and, many would argue, liberating, transformational process of digitization into an intricately coded collection of binary data. In this form it can, in theory at least, be infinitely manipulated, infinitely copied and infinitely accessed. Any constraints on its transmission and accessibility are a function of its legal or social context, and not the technological limits of the digital form. Of course, a digital version of a physical object is, as it is correctly termed, a surrogate; it is not the duplicate of an original, and our modes of interacting with, analysing and enjoying the surrogate are very different from how we experience the original. At the same time as many physical records are making the transition to digital surrogate, even more material is being created initially in digital form.

These objects, both born-digital and digitized, are often gathered together for long-term storage and preservation in the context of archives, libraries and galleries. In practice the gathering process can be notional as the physical storage sites for digital media can be, some say should be,[1] geographically distributed. However, the user may perceive them as being a single location because their access point makes them appear accessible from a single web location, normally a search interface or catalogue. In fact for many users the phrase 'on the internet' almost implies the concept of a single place. This is testament to the power of the internet, particularly the World Wide Web, to present itself as a unified whole with the web browser apparently acting as a window on a single semi-coherent collection of information. Users show little tolerance for the difficulties that exist in making diverse datasets fully discoverable and cross-searchable. This is because conceptually the data appear to be already stored in the same place, i.e. on 'the internet' and in the same format, i.e. 'digital'. The user might then ask: surely it is an easy matter to search quickly and comprehensively for all things digital on the internet? In fact discovering what resources exist on the internet, either for the purposes of research

1 Distributed physical storage, at least of copies of material, is considered a prerequisite of safe archival practices on the understanding that catastrophic events, war, fire, earthquake, etc., are less likely to affect two remote geographical locations simultaneously.

or simply to consume and enjoy, is far from straightforward. The whole world of resource search and discovery is bound up with complex commercial, economic and social issues that for the moment allow the most simple search paradigms, such as the Google 'search box', to dominate. In fact, in terms of volume of searches the Google search engine itself dominates the internet, being by far the most used.[2] How this type of search works, essentially a brute-force free-text search, is somewhat naïve when compared to the sophisticated cataloguing and resource discovery mechanisms familiar from the world of libraries and archives. More importantly it does not (and does not claim to) search what is known as the 'deep web'. The deep web (sometimes called the invisible web) is not normally searchable by standard search engines such as Yahoo and Google and is generally thought to constitute the major part of the web, although putting an exact figure on how much is searchable and how much is 'deep' is virtually impossible as the border between the two is fluid and changes constantly with policy and technology changes. Some estimates assess the amount of information available on the web, but not accessible by search engines, and therefore 'deep', as many tens of times larger than the total amount of data on the surface web.[3] An example of deep web material might be data held behind a database interface or held on web servers that exclude access by Web Crawlers and Spiders[4] as a matter of policy. How much access a search engine has to this material can to a certain extent be dictated by the data provider. However, the question that library or archive-based data providers might find themselves asking is why they have built sophisticated catalogue systems, relying on finely tuned categorization schema and delivered by a bespoke database interface, when a standard generic search engine will simply discard all that richness in favour of a simple text search? For organizations with a commitment to providing access to their data to a broad range of audiences, and for organizations whose performance is judged directly on the level of web usage, it is entirely necessary to open up their resources to as much search engine cataloguing as possible. This, after all, is how the majority of users will find their way to the resources in practice. However, it is important that other forms of resource discovery are not discarded. Those forms of discovery that rely on rich categorization – rich metadata – to describe the resource, and sophisticated and reliable search mechanisms, have often arisen from, and remain part of, professional or academic practice, and should continue to be valued as long as they remain useful to those audiences.

---

2 Nielsen NetRatings Search Engine Ratings in July 2006 gave Google 49.2% of all search engine usage with the nearest rival on 23.8%: http://searchenginewatch.com/showPage.html?page=2156451, accessed July 2008.

3 For more detail on the deep web and useful references see Wikipedia: http://en.wikipedia.org/wiki/Deep_Web, accessed July 2008.

4 Web Crawlers and Web Spiders are the terms for the automated software agents employed by search engine companies, such as Google, Yahoo and Microsoft (sometimes known collectively as GYM), to 'crawl' the web and pre-index word occurrence on web pages.

In many arts and humanities disciplines the pros and cons of various forms of categorization and the value of categorization itself remain the subject of debate.[5] It is indeed a truism that categorization schemas are the manifestation of a particular view of the world and therefore necessarily privilege that originating world view over the myriad others that we know to exist. In critical thought, this is one of a number of fundamental notions that must be constantly held in mind: how a 'thing' is intellectually conceived, constructed, and referred to is negotiable and dynamic. The process of categorization necessarily stems this process, locking the 'thing' into a near-rigid intellectual framework. Understanding this means that while a categorization schema or ontology may indeed be a manifestation of a particular world view it need not necessarily be seen as a statement of the intrinsic 'rightness' of that particular way of seeing the world. It can in fact be viewed as a pragmatic solution to allowing the discovery of a specific item of interest from a vast, near-overwhelming volume of potentially relevant data.

The Archaeology Data Service (ADS)[6] holds a large collection of digital resources that are intended to act as source material for both the teaching of archaeology in Higher Education and further research in archaeology within the higher education sector. Many of the resources the ADS holds comprise multiple file forms such as text, still images, spreadsheets and audio-visual material. Increasingly frequently they comprise databases, 3D, VR and Geographic Information System (GIS) files. These file types are designed to have an element of interactivity and, in the case of databases and GIS files, are designed to allow user-defined queries in order to derive full benefit from them. Unlike many types of data they cannot simply be passively consumed.

Archaeology as a discipline has a long tradition of experimenting with novel methods of recording its primary data. Consequently, the curation and delivery methodologies of these data have grown up to reflect this tradition, allowing for the discovery and reuse of such resources as photographic images, text documents, maps, plans and geophysical survey plots. More recently, driven in no small part by the explosion of user-friendly application programmer's interfaces (APIs) to common programmes, newly adopted methodologies are producing outputs that incorporate a degree of interaction with the data that are actually integral to the understanding of the data themselves. Examples of these include complex database front-ends, virtual reality models and GIS data. Also included are more standard formats such as those produced in the audio and video recording of archaeological sites and archaeologists' practice. Many of these new outputs are both ground-breaking and aesthetically pleasing and sit at the cutting edge of what is possible and what can be imagined as a way of presenting data about our past and its inhabitants.

---

5 For an example of the nature of this debate in archaeology see A. Baines and K. Brophy, 'What's Another Word for Thesaurus? Data Standards and Classifying the Past', in P. Daly and T.L. Evans (eds), *Digital Archaeology: Bridging Method and Theory* (London, 2006), p. 237.

6 The Archaeology Data Service website: http://ads.ahds.ac.uk/, accessed July 2008.

Ever-increasing linkages between the data and the software applications that deliver it, and even the nature of the digital infrastructure by which it is served up, challenge the normal practice of digital curation where deconstruction of the resource to simple open-source formats for preservation has been the norm. The core of digital preservation approaches such as the Open Archival Information System (OAIS) reference model is that data held in preservation formats can be easily reconstructed into a delivery package which represents the best, most current way to deliver a particular type of data. ('best' here usually means the most accessible rather than the most elegant in pure informatics terms). How does such an approach function when the data cannot be logically broken down to preservation formats without rendering it meaningless, and the ways of describing degrees and forms of interactivity in this context have not been specified? This increasing integration of archaeological data with its mode of delivery throws up two major issues for those tasked with curating it for the long term. How do we allow resource discovery within (rather than simply of) time-based media and, perhaps more significantly, how do we describe the significant properties of interactive resources?

This richness in interactivity raises a number of challenges for organizations hosting and delivering such resources. Included amongst these are the following questions. Is additional software needed and available to access the files? Without downloading and experimenting with a large dataset how might the user be informed of the levels of interactivity available and/or required? Does the user require an understanding of the underlying processes and data structures when interacting with the data? If so, how are they to be informed? How can the user target the right data at the right level of detail for their purposes without downloading and examining it? All of the above issues are to some degree relevant for passively consumed data, but the impact and implications for resources that require some kind of interaction are far greater, especially when data volume and application specificity are considered. Potentially very useful resources will remain untapped because the likely user community is not aware of what they are and what they are capable of doing. In the following sections some of the above issues are explored and, although there is no obvious solution to many of them, future directions for facilitating the best possible access to this material are suggested.

The key tool in the resource discovery of a digital dataset, or indeed any dataset, is the concept of metadata, or data describing data. Metadata is used to describe a record or an archived resource in such a way, and using such descriptors, that a researcher can easily discover that it contains information relevant to their researches. Some metadata is extremely straightforward, to the point of being obvious. For example the 'title' of a resource, say the title of a book, is often a useful bit of information. If you are looking for a book on the theory and practice of archaeology, then the title *Archaeology, Theories, Methods and*

Practice[7] would look like an obvious choice and would almost certainly come up in a search for 'archaeological theory'. Similarly, if you know that Colin Renfrew writes on archaeological methods then searching on that author's name will also return the book *Archaeology, Theories, Methods and Practice*. Unfortunately not all titles are as helpful as this and a researcher does not always know the author or year or publication of all works that might be relevant. Here is where a more rigorous approach to resource discovery becomes useful. Dating back to 1873, one of the earliest and most widely adopted approaches that could be described as a metadata schema[8] is the Dewey Decimal Classification[9] system (DDC) used extensively in libraries. This system recognizes that it would be useful to be able to classify documents by their content as well as by their author, date, title and so on. DDC categorizes 'field of knowledge' broad headings and then develops them hierarchically so that at each level of the hierarchy the category is more focused. The following example[10] shows the classification for texts dealing with caring for cats and dogs:

600	Technology
630	Agriculture and related technologies
636	Animal husbandry
636.7	Dogs
636.8	Cats

There are many hundreds of categories and subcategories allowing books in a library to be discovered by topic. In the example above, the number 636.8, when assigned to a book on cat welfare, would allow for its discovery in searches for 'technology', 'agriculture' and 'animal husbandry'.

This form of metadata, which uses a controlled list of words to organize general knowledge can obviously be applied to digital objects also, although frequently a much more liberal approach to describing content is taken. The hierarchical approach of the DDC system is reflected in archaeological classification systems such as the Thesaurus of Monument Types (TMT),[11] which is used to assign a specific term to each of the huge and varied range of records about monuments

7 C. Renfrew and P. Bahn, *Archaeology, Theories, Methods and Practice* (London, 1991).

8 A metadata schema is a proscriptive, although possibly extensible, system for creating metadata.

9 A full list of the DDC categorization system is available on the OCLC website: http://www.oclc.org/dewey/resources/summaries/default.htm, accessed July 2008.

10 This example is drawn from the OCLC's own documentation available from the OCLC website referenced above.

11 The TMT was created by English Heritage National Monuments Record, it can be explored online at: http://thesaurus.english-heritage.org.uk/, accessed July 2008.

held in either national monuments records or historic environment records.[12] Because of this the TMT can be used to assign a specific metadata term relating to monument type in all forms of record that relate to that monument.[13] As well as performing a function describing the content of a digital resource, or indicating its associations in the case of the TMT, other metadata elements can deal with technical aspects of the resource and are intended to be of use to the archive rather than being a tool for resource discovery, although they are often used as such. An example of this type of metadata might be drawn from the Dublin Core metadata schema, which was specifically developed for archiving and discovery of a broad range of digital material. The 'core' of the Dublin Core[14] is fifteen properties of a resource that help to describe it. As well as the expected properties, such as 'title', 'creator' and 'publisher', there are also properties that allow for the description of a resource's file formats, data types and even aspects of copyright and intellectual property law relating to it. It should be obvious that the file format of digitally archived data is essential to its discovery as well as to its archiving. A researcher may be looking specifically for text, audio, video or a 3D model; the file format is a key indicator for this. The relationship between file type and resource discovery is an important one and the proliferation of sophisticated and highly specialized delivery mechanisms for media like digital video, but more particularly 3D models, digital panoramas, pseudo-immersive VR models and so on, means that a researcher may wish to narrow a search to the formats he or she can meaningfully access. When the choice is between image formats such as TIFF, JPEG and GIF then this is trivial. When it comes to differing (and proprietary) video 'Codecs' (coding and decoding algorithms)[15] the choice becomes harder. Ultimately when the distinction is between formats used to store and deliver highly interactive datasets, such as the pseudo-immersive VR mentioned above, the choice is often a very simple one, i.e. is there an application available to read this format and does the researcher have access to it?[16]

---

12 In the UK, National Monuments Records are curated by the appropriate national body for England, Scotland, Wales and Northern Ireland. Historic Environment Records (sometimes called Sites and Monuments Records) are held by local authorities and perform a role specific to local authority responsibility for the planning process. However, they also represent a significant HE research resource.

13 For example, a digitized journal article that refers to a particular site might have a term from the TMT used in its metadata.

14 The Dublin Core Metadata Element Set, Version 1.1 is available from the Dublin Core website: http://dublincore.org/documents/dces/, accessed July 2008.

15 The coding and decoding algorithms used in digital audio and video are often considered as key intellectual property of the companies that developed them and as a result are closely guarded secrets made available to other software developers only at significant cost, if at all.

16 It is major plank of good practice in digital archiving to recognize that software companies and the codecs they own the rights to should not be assumed to be permanently available.

There is another approach to metadata for digital resources, including time-based media such as audio and video. It is most apparent in so-called Web 2.0 contexts where the content of a website might be created by the people who are also using the website. This is commonly known as user-generated content or UGC. Far from their origins as a fringe activity, these types of site now hold very large volumes of data. For some, the explosion of user-generated content, whether it be still images, audio or video or more complex file types, has gone hand-in-hand with an entirely different attitude to resource discovery (and also technical metadata). Sites such as Napster,[17] Flickr[18] and YouTube[19] would seem at first blush to operate as a kind of archive for audio, images and videos. Leaving aside the numerous issues surrounding the legal, social, technical and ethical positions of these services as they are currently structured, what is of interest here is the notion of unstructured user-generated metadata. In Web 2.0 jargon this concept is most frequently referred to as tagging, social tagging or user-generated tags. How exactly tagging is implemented may vary from site to site, but the principle remains the same, the original depositor and then subsequent site visitors can choose and associate words and phrases to describe a resource. The terms that are most frequently given are assumed to be the most useful. Here is a partial list of tags drawn from an image of Stonehenge on Flickr:[20]

> Stonehenge, Wiltshire, UK, BRAVO, ABigFave, Superbmasterpiece, BlueRibbonWinner, ExcellentPhotographer, Awards, amazingamateur, ancient Mesolithic, bronze, stone, circles, sunset, landmark ...

As can be seen from this list, many of the tags act as relevant metadata for one purpose, say archaeological research, while some clearly refer to the subjective quality of the image ('Superbmasterpiece') and others refer to non-archaeological features of the image ('sunset') and finally, others appear to have no general relevance at all ('BRAVO, ABigFave'). Despite the shortcomings of this approach for resource discovery in a research context, it is undeniable that different conceptions of a monument might be elucidated by this type of tagging. While this is most likely a desirable outcome and would be generally welcomed in archaeology it is only ever likely to be complementary to more rigorous systems.

It is very important to remember that these sites, and even search engines such as those controlled by the market leaders, do not actually have as their highest

17 Napster is a peer-to-peer file sharing site that was subject to intense legal pressure over alleged copyright infringement relating to content when the company first started. It is owned by Roxio Inc.: http://trial.napster.co.uk/, accessed July 2008.

18 Flickr is a photo-sharing site owned by Yahoo!: http://www.flickr.com/, accessed July 2008.

19 YouTube is a video-clip-sharing site, owned by Google since 2006: http://www.YouTube.com/, accessed July 2008.

20 http://www.flickr.com/photos/nardip/1433903816/, accessed July 2008.

priority facilitating the easy and quick discovery of information or resources that might be relevant to a focused user query. All these services are commercial enterprises, irrespective of how they originally emerged. As such they are now driven primarily by the profit motive. The actual business models by which profit is delivered or is going to be delivered by these sites is still a somewhat unresolved question and new models emerge, are tested, and then are replaced by others on a frequent basis. Currently, the two dominant models are versions of the sale of advertising space on web pages. This is often augmented by tailoring advert presentation based on a user profile, and the sale of information on customer (user) behaviour on the internet. Clearly the first model echoes that of traditional media such as newspapers and commercial television and, in common with these media, the veracity, quality and discoverability of the content are only considered relevant if they have a direct impact on the levels of usage of the medium, i.e. reading a paper, watching television or using the internet. Therefore this type of site may hold vast amounts of material, it might even be tagged with significant amounts of useful pointers to its content, but this is ancillary to the function of the site, which is to sell advertising. The waters are muddied still further by a trend to disguise advertising as content, for example film trailers on YouTube, and the fact that actually making things harder to discover, say by presenting a search result in conjunction with other content that is associated in some way, or even only tangentially relevant, keeps a user on the site, and therefore exposed to advertising, for longer.

In addition to the direct use of audio and visual techniques for archaeological recording and owing to a broad general public interest in archaeology, the discipline has benefited over the years from a substantial volume of broadcast documentaries and reconstructions. Much of this material is of significant interest to the research community and is beginning to be offered to the relevant digital archives, such as the ADS. Metadata provision has been identified as one of the most significant stumbling blocks confronting archivists in the field of time-based media.[21] Historically, time-based media has not required widely usable metadata or metadata suitable for public searching or research purposes (i.e. resource discovery, rather than technical or management metadata) as they were until recently held only by the organizations that created and broadcast them. All general access to these media was dictated entirely by the broadcaster and there was little or no private consumption. Resource discovery was via a published schedule such as the *Radio Times*.[22] Recent trends in play-on-demand (over the internet), such as BBC

21 See Section 1.6 in A. Wilson et al., *AHDS Moving Images and Sound Archiving Study* (AHDS, London, 2006): http://ahds.ac.uk/about/projects/archiving-studies/index.htm, accessed July 2008.

22 *Radio Times*, BBC Publications, London: http://www.radiotimes.com/, accessed July 2008. The exception being video/DVDs. Here the description of the content, beyond technical information, BBFC Classification, etc. is intended primarily to sell the product.

iPlayer,[23] and the vast collections of user-generated content on YouTube and similar sites have served simply to highlight the deficiencies in resource description for users looking for specific elements of content. The problem associated with time-based media, the opaqueness of media content,[24] is well known. It is very hard to know what the content of a time-based media resource is without watching it all the way through. The approaches to resource discovery offered by these online services may be just about acceptable for content designed as entertainment, but where the content is relevant to research, more suitable ways of finding it, and what lies within it, are needed.

Organizations that archive and broadcast these media have developed techniques for describing such content. The most widely used is 'logging' whereby individual scenes inside a video (or audio clip) have their timestamp associated with their content. For audio and video, logged metadata is just as important as the classic forms of bibliographic and descriptive metadata detailed earlier. The simplest form of logging data is a transcription of the words heard in a soundtrack. Often this accompanied by the description of 'keyframes' or scenes within the video. A very simple example of this type of logging might look like this:

Table 4.1 An example extract from a frame logging document

In Point	Out Point	Name	Comment/Description
00:00:20	00:00:60	Introduction	Introduction to the film including an overview presented Jack Smith.
00:00:60	00:01:20	Scene One	Helicopter flypast of Stonehenge in winter with voiceover describing elements of the monument.

There are numerous schemas that perform this content description function, but, as already pointed out, there is no universally accepted way of doing this. Another problem with this approach is that, as in the example above, the actual description is not very helpful and in this case does not relate to any knowledge organization system that would facilitate meaningful discovery of, or searching within, the resource. The transcription of audio-only or video soundtracks provides a searchable text and is much more amenable to meaningful searching.[25] Some

23 The BBC iPlayer is a web based 'play on demand' service offering a selection of previously broadcast content: http://www.bbc.co.uk/iplayer/, accessed July 2008.

24 See K. Green, 'More-accurate Video Search', in *Technology Review*, 12 June 2007, http://www.technologyreview.com/read_article.aspx?ch=specialsectionsandsc=searchandid=18847anda=f, accessed July 2008.

25 A commercial example of this approach is the EveryZing system: http://www.everyzing.com, accessed 7 July 2008.

digital time-based formats, such as MPEG-7, are designed to encapsulate this type of logging. In addition, this type of content description can easily be expressed in an extensible mark-up language (XML)[26] and reused for both resource discovery and intra-resource searching. There is a strong possibility that speech recognition in tandem with natural language processing (NLP)[27] on the resulting text will actually deliver meaningful and highly structured metadata, however this will be contingent on appropriate ontological structures, such as the TMT in archaeology, being available to capture the semantics of the text. Clearly video that does not contain the spoken word or which does not have the appropriate quality of speech (speech that actually gives an indication of the visual content) is never going to succumb to an automated metadata extraction process relying on either speech recognition or NLP.

Earlier in this chapter a number of approaches to resource discovery and technical metadata that we see in digital archives and online libraries were outlined. Some systems, such as DDC, describe content in relation to a specific form of knowledge organization. Other systems, such as decimal classification (DC), describe content in a more open way, but usually still in relation to some form of knowledge organization. Further systems, such as tagging, exert no control over how a resource is described. In addition to the different approaches to formalizing the metadata into schema, the different approaches to creating the metadata – creator-generated, user-generated and automatically generated – have also been touched upon. However, there are other issues around the structure and creation of metadata, which concern the increasing complexity and interactivity of digital objects and their relationship to the techniques of data curation.

As discussed above, the ADS is a data archive, facilitating access to numerous data resources. This means that the mission of the ADS is not just about resource discovery but also about developing procedures and systems that keep these resources safe and accessible for the long term. This is not the place for a discussion on the fragility of digital data and the need for well thought through preservation strategies, as this is dealt with extensively in the literature,[28] but there are aspects of the preservation process that have a direct bearing on time-based media, complex digital objects and hardware-dependent objects. In order to draw

26 XML is a widely used general-purpose specification for creating custom mark-up languages; it allows the specification of custom tags, thus allowing an XML document to capture the quite specific features of such things as video logging schemas: http://www.w3.org/XML/, accessed July 2008.

27 An example of the use of NLP to extract keywords automatically from archaeological text to populate a pre-existing ontology and thereby automatically generate resource discovery metadata can be seen in the Archaeotools project: http://ads.ahds.ac.uk/archaeotools, accessed July 2008.

28 For example: F. Condron, J. Richards, D. Robinson and A.Wise, *Strategies for Digital Data* (Archaeology Data Service, York, 1999): http://ads.ahds.ac.uk/project/strategies/, accessed July 2008.

out the relationship between these types of archival object and the nature of the archival systems it is first necessary to understand some of the principles behind digital archiving in general.

The Open Archival Information System (OAIS)[29] reference model has, in the absence of any serious competition, established itself as the *de facto* standard for digital archiving in a number of sectors, such as UK higher education, including the ADS. Already widely used, OAIS, developed by the Consultative Committee for Space Data Systems, in fact became an ISO Standard in 2003. The standard itself is represented by a large and complex document, but in essence it defines a way of thinking about archiving digital material that ensures all the key issues in the process are addressed. OAIS identifies six major functions of the digital archive:

- Negotiate for appropriate deposits
- Obtain sufficient control of resources
- Determine scope of community
- Ensure independent utility of data
- Follow procedures for preservation
- Disseminate to the designated community

The point that is of most concern in the context of this discussion is the fourth one, 'ensure independent utility of the data'. Especially problematic is the word 'independent'. In implementing an OAIS-based system this core function is often tackled by migrating the submitted data to a number of different information 'packages'. These are termed:

- SIP – Submission Information Package
- AIP – Archival Information Package
- DIP – Dissemination Information Package

Each of these packages represents a different manifestation of the submitted data, each with a specific purpose. The SIP is the untransformed package initially submitted by a depositor to the archive. The AIP represents a form of that same data that is designed to assure independent utility of the data. The DIP is a form of the data that is most suitable for dissemination. An example of why these distinctions are necessary might be as follows:

29 OAIS became an ISO standard in 2003, ISO 14721:2003. The full OAIS specification is available as a PDF document from: http://public.ccsds.org/publications/archive/650x0b1.pdf, accessed July 2008.

SIP: A Microsoft Word document
Rationale: A very popular word processing format, this is how the document is submitted to the archive by the depositor. The SIP maintains this format unchanged and it is held in deep storage as an 'original'.

AIP: An ASCII[30] text file and associated TIF[31] images
Rationale: MS Word is a proprietary format and there is no guarantee that the software to access it will be freely available and supported in the future. Extracting the text into ASCII and the images into TIFs ensures their accessibility into the future in open formats. However, it renders them unsuitable for delivery in most contexts.

DIP: An Adobe Portable Document Format (PDF) document
Rationale: The archive can't rely on every potential user having access to Microsoft Office and would also like to distribute only an uneditable version of the document. The AIP version would require the user to piece the document together from its separately-archived elements. PDF uses a freely available reader and allows a document to be locked for editing.

It is clear from the above that the key to fulfilling the role of a digital archive is intimately bound up with the ability to move data freely from format to format, both to ensure its preservation in a neutral 'open' format, protecting it from the vagaries of a rapidly changing commercial software environment and to ensure its reusability as a disseminated item. Thus the traditional archival model clearly draws a distinction between the data and the mode of delivery. However, when we examine more complex digital objects this barrier begins to break down. Compromises have ultimately to be made where an unbreakable link exists between the data and either the software that created it or the software that delivers it or even between the data and the hardware by which it is delivered. An example of each of these three cases is given below.

1. The clearest and most common example of an unbreakable link between data and the software that created it is where that software is proprietary and there is no equivalent format, open or otherwise. If the data cannot be migrated to some accessible format it has no useful life when dissociated from its intended software. Unfortunately historical attempts to defend market share meant that commercial software developers on occasion deliberately made it difficult to convert their files to alternative formats.

30 ASCII stands for the American Standard Code for Information Interchange and dates back to the days when text was encoded for transmission along telegraph wires.

31 Tagged Interchange Format, or TIF, although not actually an open standard – it is still technically owned by Sun – is generally considered open as it is fully published and described in the literature.

For a number of technical and commercial reasons this is becoming less of a problem, but lack of access to original software remains a significant problem for digital archives dealing with legacy datasets.

2. An example of a situation where the relationship between data and the software that created it is crucial is Agent Based Modelling or ABM.[32] This approach involves the running and re-running of computer simulations containing models of both environments and agents within the environment. These simulations might be run for many cycles and then reset with new variables relating to environmental factors or agent behaviour and then rerun. It is the honing of these elements of the simulation that represent the research process just as much as the outcomes of the model. Indeed, the simulation is constructed from the interaction of the model elements, which might be expressed as 'data' and the algorithms embedded within the simulation software. In short, the raw data is only meaningful when used with one particular piece of software. The ADS has, as a matter of policy, not accepted software for archiving. This is because archiving software, with the associated issues of versioning, porting and commercial exploitation is felt to be outside the remit of the ADS. If ABM modelling material is to be archived then the ADS or a similar organization would have to accept that this is a more complex task than simply archiving a series of digital outputs.
3. Links have always existed between data, software and hardware. Even the simplest file type needs to be translated via an application into a readable form for display on a computer screen. In most situations for archiving purposes this does not present a significant issue. If the data can be made independent of the software it is normally automatically independent of the hardware. However where the means of display or interaction with the data go beyond simple pictures and sound via a screen and speakers this relationship can become problematic. In the world of head-mounted displays for fully immersive VR models or hemispherical displays or even haptic devices relaying tactile information about a model, the data may be so closely linked to the form of its ultimate dissemination that there is no other meaningful form in which it can be archived. The question then arises, is it worthwhile archiving it at all if its associated hardware is not likely to be accessible over the long lifetime of the archive?

The three scenarios above should be considered in combination with the resource discovery and intra-resource discovery problems discussed earlier in this chapter.

32 Agent Based Modelling is not new in archaeology, but is undergoing a renaissance owing to, amongst other things, the increased availability of high end computing. A current example of this type of project is the AHRC-EPSRC-JISC e-Science funded MWGrid: Medieval Warfare on the Grid project at the University of Birmingham. http://www.cs.bham.ac.uk/research/projects/mwgrid/, accessed July 2008.

There is one final, vital complication in creating meaningful metadata that should be noted. Novel and sophisticated modes of presentation such as Quick Time Virtual Reality (QTVR),[33] panoramas or object movies and Virtual Reality Modeling Language (VRML) models allow rich and varied modes of interaction with the data being presented (the panorama, the object or the VRML model). The levels of interactivity offered, and often how that interactivity is exploited, are not standard. For example, the frame-passing function of QTVR might be utilized to change lighting angles or lighting conditions on an object rather than the more usual approach of using each frame to change the angle of view[34] (see Plate 4.1). In these cases the functionality offered can be key to understanding the data, yet there is no standard way of describing this functionality. Metadata, as discussed above, might describe the object of the data, but it will not yet allow us to describe how we can interact with the data in a formal and universally understood fashion. A user might even understand the general levels of interactivity offered by QTVR, so the metadata element that indicates an object is of this file type should indicate to some extent how the data can be interacted with. However there is no metadata element or formal terminology that allows the archivist to say 'this particular file changes angle of lighting rather than angle of view', but this difference in functionality results in a significantly different resource. Although the above example might seem trivial, the problem it represents, the blurring of the division between data and delivery method and the inability to describe functionality for resource discovery purposes, raises some potentially serious problems for the digital archivist that will only increase with the complexity of the objects being deposited for archive.

Where then does this overview suggest we currently stand with regard to both finding and searching within complex digital media? Finding one's favourite web resource or digital object is not always as straightforward as we might like it to be. It would be good to think that this situation will resolve itself somewhat in the future by various means. For the majority of internet content right now, which is predominantly text and images (although often presented in sophisticated ways), this may well be the case.

The ADS is actively engaged in research and development activities designed to enhance researchers' abilities to discover relevant resources that we either hold ourselves or provide access to by aggregating resource discovery metadata. Where metadata schemas have been adhered to and where they are underpinned

33 QTVR is a proprietary from of pseudo-immersive virtual reality owned by Apple: http://www.apple.com/quicktime/technologies/qtvr/, accessed July 2008.

34 Examples of this type of unexpected use of interactivity can be seen in this night/day object movie of a standing stone from Machrie moor on Arran (see Plate 4.1) and the variable angle lighting of a medieval inscribed stone, both downloadable from the ADS, S. Jeffrey, *Three Dimensional Modelling of Scottish Early Medieval Sculpted Stones* (unpublished PhD Thesis, University of Glasgow, 2003): http://ads.ahds.ac.uk/catalogue/library/theses/jeffrey_2004/, accessed July 2008.

by rigorous (and rigorously managed) thesauri and word lists, this offers the possibility for faceted classification. Faceted classification is a very simple concept, but extremely powerful. Faceted classification browsing systems offer the most likely challenger to the broad-brush search approaches offered by 'type and hope' search boxes.[35] In essence a faceted classification browser allows a user to navigate a hierarchical knowledge organization structure, or tree, by clicking on the most relevant facets. For searching purposes in archaeology the facets that are most useful are: where, what and when. An archaeological example of a hierarchy of facets might look like this:

When – Medieval
Early Medieval

Where – United Kingdom
England
Yorkshire
York

What – Military and Defensive
Defended buildings
Castles

Each of the three facets can be selected by mouse clicks and, providing that the classification against the facts has been thorough, the user should have full confidence in the completeness and relevance of the returned results (in the case above, early medieval castles in York). This level of confidence is impossible if only using a text search box. The ADS created a proof of concept faceted classification browser interface with UK HE funding in 2004. Following workshops held by the AHRC ICT Methods Network in 2006, the ADS and the computer science department at the University of Sheffield gained funding from the e-Science Research Grants Scheme (funded by AHRC, EPSRC and JISC) for a project to bring to service a faceted classification browser based on archaeological monument inventory data. This system will be fully functional by early 2009.[36]

It is looking very likely that the internet of the future will be host to vast amounts of rich forms of data such as audio, video, 3D models, geographical data and databases, as well as forms and formats that we cannot yet imagine. As research material presented in these formats becomes standard in the arts and humanities, discovery mechanisms that have evolved to cope with text and images will struggle to maintain their usefulness for researchers' day-to-day use. The

35 A number of demonstrations of faceted classification approaches can be seen at Facetmap: http://facetmap.com/browse, accessed July 2008.

36 http://ads.ahds.ac.uk/project/archaeotools/, accessed July 2008.

current archival structures, and resource and object description schemas are simply not designed to cope with the explosion of time-based media, complex digital objects or software dependent digital objects. One thing is clear, however: just as the challenges of metadata creation for time-based media are beginning to be addressed,[37] new challenges requiring the ability to describe the levels and forms of interactivity offered by even more complex digital objects arise. The difficulties of archiving complex forms of data may require a more flexible approach to the notion of independent utility. Certainly, in the case of objects like the outputs from ABM and collections of data inextricably or very strongly linked with particular hardware suites, re-use cases for the data should be very clearly made before the data is considered for archiving in the usual way.

It may be argued that current resource discovery and archival approaches have plenty of time to adapt to changes in the nature of the resources they deal with. In fact, to date, all the 3D material archived with and disseminated by the ADS was created specifically as part of projects looking into the usage of 3D recording and modelling in archaeological practice or was created at such an early stage that no standard metadata schema for this form of material could be expected to have arisen. Despite this, the history of rapid media development, format development and even assorted broad paradigm shifts since the opening of the internet to the public and commerce suggests that now is the time to tackle the problems of dealing with resource discovery and curation of complex digital objects. Just as a core theme in digital archiving best practice is that archival strategies for data should be thought about at the very outset of a project, perhaps format and application developers in interactive media should be thinking more about discovery and description issues from the outset of the design process.

37 For example, recent work by the Technical Advisory Service for Images (TASI) includes reviews of moving image technologies and search engines for moving images and audio: http://www.tasi.ac.uk/, accessed July 2008.