

# Digital Repositories and Archives Inventory: Conclusions and Recommendations

Daisy Abbott, HATII, University of Glasgow  
Sheila Anderson, King's College London

<b>1</b>	<b>OVERVIEW</b> .....	<b>1</b>
1.1	Summary.....	1
1.2	DRAI Project: Overview .....	2
<b>2</b>	<b>ANALYSIS AND RESULTS</b> .....	<b>3</b>
2.1	Data Limitations.....	4
2.2	Collections Data .....	5
2.2.1	Subject Coverage .....	7
2.2.2	Access: methods and controls.....	10
2.2.3	Use Rights.....	13
2.2.4	Ingest Policies .....	14
2.2.5	Preservation.....	15
2.2.6	Relationships between collections .....	15
2.3	Collection Owners.....	17
<b>3</b>	<b>CONCLUSIONS AND RECOMMENDATIONS</b> .....	<b>18</b>
3.1.1	Number of digital collections within the scope of DRAI .....	18
3.1.2	Availability of information about collections .....	19
3.1.3	Acknowledging the complexity of the field.....	19
3.1.4	Administrative and management relationships.....	20
3.1.5	Digital resource provision for HE.....	20
3.1.6	Usage rights .....	20
3.1.7	Preservation environment .....	21

## 1 Overview

### 1.1 Summary

The Digital Repositories and Archives Inventory (DRAI) has produced data which offers a view of the current landscape of digital resource provision and preservation in the UK. Overall, the picture is of a huge and tremendously complex and varied field. The scope of DRAI was to aggregate content from various existing sources into one inventory, however there are many digital collections which are not included in major information sources and therefore may have been omitted from the DRAI Project. Collections tend to be produced individually on an ad hoc or 'one-off funding' basis which leads to a great deal of fragmentation within the field as a whole. Different collections can have extremely complex relationships with each other and with managing institutions or 'parent' repositories. Collection owners' relationships are similarly complex.

It proved difficult and time-consuming to discover detailed technical information about collections as this information tends to either be unavailable on the Web, or is

hidden in a mass of other information. This holds true for the very existence of digital content itself – although over 120,000 sites are referenced by the Intute portal, even this is not a comprehensive record of every collection within the scope of the DRAI Project, and the great majority of these links are not for providers of digital content but for organisations, and simple Web pages outside the scope of the project. DRAI identified a difficulty in retrieving information for the purposes of this specific project, however the confusion in searching for and within digital resources can also be applied to their general use for content discovery. The LAIRAH Report<sup>1</sup> suggests that 30% - 35% of digital resource in the arts and humanities are not used indicating that searching and browsing for content needs to be more sophisticated and made easier for users.

Analysis of the preservation information gathered for the DRAI project also identifies a significant concern in terms of the current UK preservation environment. There are few examples of best practice for preservation and rights management policies – these tend to only be available for national repositories with clearly defined missions, for example the AHDS, ESDS, and TNA.<sup>2</sup> Additionally, the vast majority of resources appear to have no plans for sustainability and long-term access for users which poses a significant risk to the immense intellectual and financial value of current digital assets; a problem which has recently been exacerbated by the withdrawal of funding to the AHDS.

This report presents an overview of the DRAI project, some initial analysis of the data gathered, and some conclusions and recommendations based on this analysis.

## **1.2 DRAI Project: Overview**

HE institutions, education and research organisations, and research groups in the UK have created and provide access to a wide range of electronic content for use in learning, teaching and research. Access to this content is provided through a number of different routes including web accessible digital archives, open access repositories, and web-based collections. These repositories and archives are hosted by departments, institutions, consortia, and national bodies such as research councils, learned societies and other publicly funded organisations. Some registries and services aggregate these sources of content, but there is currently no single place where these sources are listed.

There has been a clearly articulated need for a “one stop shop” for information discovery across different digital collections.<sup>3</sup> The catalogue of resources created during the Digital Repositories and Archives Inventory (DRAI) project updates and complements previous aggregation efforts and provides more specific information about the preservation of each collection (which has not been part of the scope of previous portals). This information is crucial to understand the current preservation environment in the UK and will build on previous work by the AHDS and JISC (amongst others) in building strategies for digital preservation.

---

<sup>1</sup> <http://www.ucl.ac.uk/slais/claire-warwick/publications/LAIRAHreport.pdf>

<sup>2</sup> <http://ahds.ac.uk/>, <http://www.esds.ac.uk/>, <http://www.nationalarchives.gov.uk/>

<sup>3</sup> <http://www.ahds.ac.uk/performingarts/pubs/scoping-study-2006.pdf>

The overall approach was focussed on interoperability with the JISC Information Environment Service Registry (IESR)<sup>4</sup> and in future the DRAI data can be incorporated with the minimum of effort. Phase 1 of the DRAI Project aggregated and classified nearly 2,000 records from a variety of important existing sources and reports, including checking over 60,000 of Intute's links to digital resources.<sup>5</sup> A wide variety of information on the access to and preservation of these digital collections was recorded into a database and XML document, analysis of which forms the basis of this report. The project also delivered the MySQL database data, the data in XML format, the conversion script, a document showing mapping to the JISC IESR, and full documentation for the project.

The aims of the project as outlined in the project plan were as follows:

1. To gain an understanding of digital resources available to UK Higher Education.
2. To produce a comprehensive XML catalogue of digital resources freely available for educational use in the UK.
3. To examine the overall provision of digital resources across subject areas and formats.
4. To discover and analyse the preservation environment for these digital resources.
5. To report on the catalogue, providing conclusions and recommendations.
6. To produce, either as part of the catalogue or documentation, secondary outputs such as information on digital collections which do not form part of the main catalogue, and the relationships between different collections.
7. To fully document the project in order that it can be easily repeated in the future.

Full details of the project methodology and its implementation can be found in the JISC Final Report, and it is not intended to repeat that information here. Readers are strongly recommended to read this report in conjunction with the *JISC Final Report: DRAI Project* in order to gain a full overview of the DRAI Project and its outcomes. Some sections of that report have been repeated here where relevant to understanding of the results.

## 2 Analysis and Results

What became immediately clear during the initial stages of the DRAI Project was the complexity and diversity of 'repositories and archives' as defined by the JISC in the Invitation to Tender and the associated difficulty of collecting comprehensive information within the timescale of the project. Heery and Anderson in *Digital Repositories Review* suggest that:

Increasingly widespread use of a term goes hand in hand with increasing diversity of meanings. Repositories are 'collections of digital objects' but what makes repositories different from other collections of digital objects such as directories, catalogues and databases?<sup>6</sup>

---

<sup>4</sup> <http://iesr.ac.uk/>

<sup>5</sup> <http://www.intute.ac.uk/>

<sup>6</sup> [http://www.jisc.ac.uk/uploaded\\_documents/digital-repositories-review-2005.pdf](http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf)

This question runs throughout this analysis. The scope of the DRAI Project included digital archives, open access repositories, content services requiring registration, and web-based collections of images, audio and other types of content. This document uses the generic term ‘collection’ or ‘resource’ to refer to digital content for which records were produced in order to avoid imposing the potentially more narrow definition of ‘repository’ or ‘archive’ onto content which it does not necessarily describe.

It is clear from the evidence gathered by the project that a significant amount of digital content is delivered via a Web interface as a ‘standalone’ collection. Most are collections of digital objects that make sense to be grouped together and delivered as a set of related content, but equally they are also embedded within databases, and made available under a single title. Other collections are included within a managed repository framework, many with associated policies to guide their curation, preservation and future accessibility. Even limited to resources which are freely available for educational use and hosted in the UK (the scope defined by JISC), the number and diversity of collections (and their creators and owners) should not be underestimated. See **Section 2.2 Collections Data** and **Section 2.3 Collection Owners** below for more detail.

## **2.1 Data Limitations**

One of the discoveries of the DRAI Project is that it is extremely time-consuming to attempt to discover information such as back-end access controls and methods and preservation policies through resources’ public Websites. In general, only organisations with significant and specific expertise in digital resource provision (or those specifically based on the details of digital access or preservation, such as AHDS or Opendoar)<sup>7</sup> make this kind of information readily available. For most collections this information is either deeply buried or not publicly available. Therefore it is extremely difficult to discover some technical details about collections from the information available on their public Websites. Properties which proved very time-consuming and difficult to discover included:

- Delivery Software
- Access Control
- Access Method
- Metadata Used
- and perhaps surprisingly, Temporal (i.e. what time period the content of the resource refers to).<sup>8</sup>

The results and conclusions of the DRAI Project are based on the data which it was possible to discover. However, the very fact that most digital collections do not make this sort of information readily available must be acknowledged. It is a reasonable assumption to state that those sources which make their technical infrastructure openly available are likely to be more engaged with the challenges of digital resource provision and digital preservation. Therefore, the technical data collected has value despite the large number of null fields in some of the properties as, for example,

---

<sup>7</sup> <http://www.opendoar.org/>

<sup>8</sup> For exact definitions of these properties, please refer to the schema provided in *the JISC Final Report: DRAI Project*.

whilst a preservation policy not being easily identified from the Website does not prove that a collection does *not* have one, it does provide an indication of the level of importance attached to digital preservation by that collection or owner.

Due to the wide scope of the project, and the very time-consuming process of classification for inclusion and information discovery, it was not possible to catalogue every digital resource that may fulfil the criteria. All sources mentioned in the project plan were fully aggregated, except for the very large links portal Intute, which holds around 120,000 links, about 60% of which were checked and aggregated into the database where they satisfied the inclusion criteria. However, we can nevertheless glean useful information has been collected and collated, and an initial analysis is presented here from the information gathered so far, acknowledging limitations in the coverage of the data.

JISC has already provided funds for Phase 2 of the DRAI Project to continue the data collection exercise in order to gain a more comprehensive picture of the scale of digital content available for use. Phase 2 Outputs will be available around July 2008.

It should be noted that the scope of the DRAI Project is to aggregate collection data from existing sources. Whilst providing a valuable snapshot, this will not present a completely comprehensive picture of digital resource provision in the UK as there are hundreds of small projects which have produced valuable digital collections but are not referenced by any central source. University libraries, research consortia, and individual departments can have small collections buried deep within their Web pages, and those outside the academic sector will be even less likely to have submitted their collections to hubs or sources of the type used for the basis of aggregation in this project. Therefore, whilst the number of collections aggregated provides statistical validity to the conclusions drawn here, there may be unacknowledged biases in the data based on the sources from which it was aggregated.

## **2.2 Collections Data**

The DRAI Project captured technical information alongside content information for 1,924 UK-based digital resources of use to the HE sector and as such provides the largest detailed inventory of collections yet produced. This figure is much higher than anticipated at the start of the project, in part due to the wide definition of 'digital resource'.

The graph below shows the types of content held by each collection, using categories defined by JISC's IESR schema.<sup>9</sup> Many digital resources hold more than one type of content with text, predictably, being the most pervasive format with 87.4% of resources containing text content in some form.<sup>10</sup> Nearly half of all collections aggregated hold images (not necessarily in a structured format,

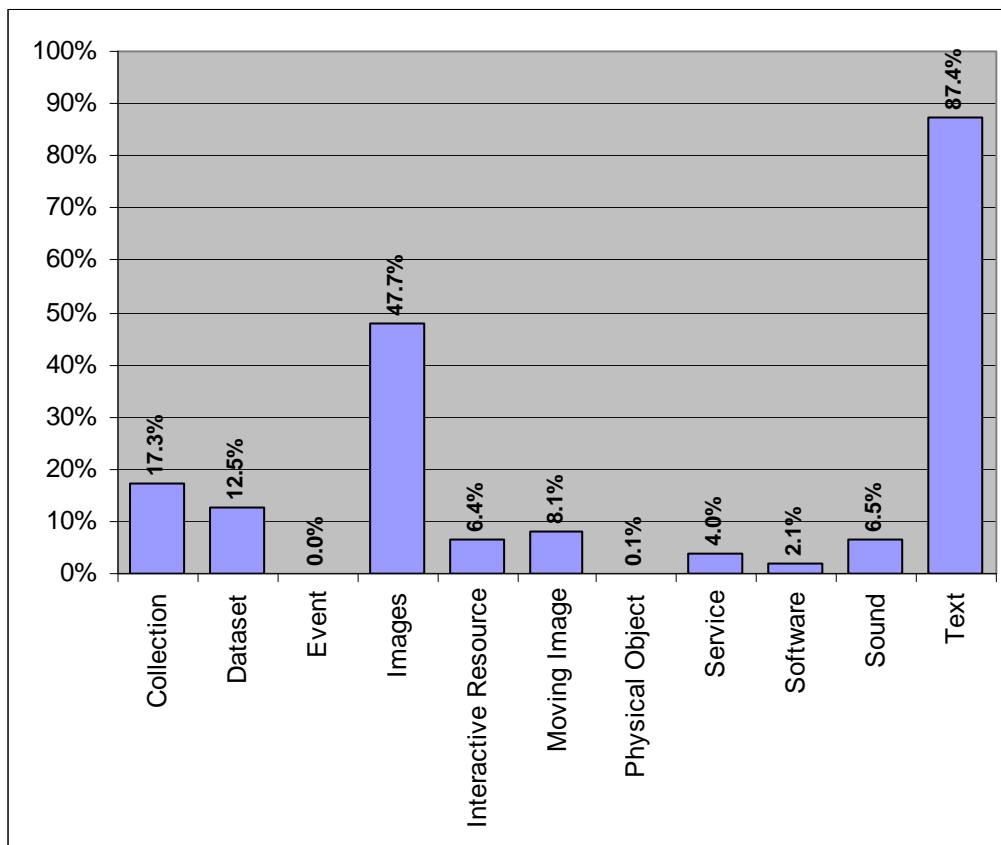
---

<sup>9</sup> <http://iesr.ac.uk/guidelines/content/collection/item-type/>. The DRAI Project also collected information about the specific filetypes of each type of content held. More detailed information on specific file formats can be gained by further analysis of the XML data provided.

<sup>10</sup> Note that 'text' here refers to actual text content such as articles or Webpage content, not text used merely as Webpage introductions, metadata or as short fields within datasets.

therefore images presented using html are included alongside image archives if they can be considered to constitute a 'collection' rather than simply a Web page). Datasets are the third most popular content type with 12.5% of collections containing a dataset. Dataset here excludes metadata catalogues and finding aids for the collection itself instead referring to discrete, self-contained datasets held within (and often forming the only item in) a collection.

Figure 1 - Item types



Moving images are held in 8.1% of collections proving more popular than sound without images (6.5%). 'Interactive resource' refers to meaningful interaction with digital objects (excluding simple hypertext navigation, using search forms and the like). The vast majority of collections were delivered through Web pages so it is perhaps surprising that so few include interactive resources such as search tools. This could be because of the expense and level of expertise necessary to produce this type of resource. Nevertheless, in terms of long-term access to resources, it is actually quite promising that so few resources are interactive as these interfaces (whilst often very engaging and attractive) tend to be at odds with the recommendations of best practice for storing and delivering digital resources, due to their reliance on proprietary and often quickly changing software.

It can be seen that, roughly in line with expectation, content types are dominated by text and images, which are usually relatively uncomplicated in terms of file format (although a moderate number of collections did hold text in complex formats such as Microsoft Word, plain text, marked-up text, or pdfs were more common).

The collection content type indicates that a resource contains other discrete collections within it (for example, an overall resource which acts as the parent for sub-collections). With over a sixth of resources referencing sub-collections, this demonstrates an unexpectedly high level of interconnections between different individual collections. It was expected that the majority of these ‘parent’ collections are high level, managed repositories such as the AHDS, or large institutions with a well-established track record in the creation and preservation of digital content, such as the British Library, BBC, and English Heritage. However, a surprising number of smaller resources also hold sub-collections. It is surmised that this is as a result of small-scale digitisation activity funded in sporadic bursts (therefore producing multiple different digital resources from content within the same analogue collection for example), or the digitisation of related content being undertaken by different partners. This data highlights the deeply fragmented nature of digital resources and the general lack of consistency in infrastructure (usually resulting in subsequent inconsistencies in design).

These relationships are considered in more detail in **Section 2.2.6 Relationships between collections**.

## 2.2.1 Subject Coverage

As the largest source of links to digital content with over 120,000 links, the DRAI Project’s coverage of Intute is the best measure of the limitations in the subject coverage of Phase 1. As mentioned above, the DRAI Project could not aggregate Intute in its entirety, which, as aggregation was performed sequentially under subject headings significantly skews the data in favour of those categories which were wholly aggregated. The table below shows which sections of Intute were completely aggregated into the database (as of 19/10/07). Many of the Intute links are duplicated from other sources, such as AHDS. Therefore the value ‘not aggregated’ below does not mean that the bulk of resources within a category were not input, only that they were not aggregated *from this source*.

The table below shows that two categories (containing nearly 55,000 records) were fully aggregated: Arts and Humanities and Social Sciences. Science, Engineering and Technology and Health and Life Sciences were not completed. Therefore the data on subject coverage is certainly skewed towards the Arts, Humanities, and Social Sciences.

### Intute categories aggregated into the Digital Repositories and Archive Inventory

Intute Section	Heading	Coverage
Science, Engineering and Technology (33,658 records)	All	Approx 20% overall, not accounting for repetitions
	Astronomy	100%
	Chemistry	100%
	Computing	Not aggregated
	Earth Sciences	Not aggregated
	Engineering	Not aggregated
	Environment	Not aggregated
	General Sciences	Not aggregated

	Geography	Not aggregated
	Mathematics	Not aggregated
	Physics	Not aggregated
Arts & Humanities (22,125 records)	All	100%
Social Sciences (32,575 records)	All	100%
Health and Life Sciences (31,155 records)	All	Approx 20% overall, not accounting for repetitions
	Medicine	Approx 75% (a – t inclusive)
	Nursing, Midwifery and Allied Health	Not aggregated
	Veterinary	Not aggregated
	Bioresearch	Not aggregated
	Natural History	Not aggregated
	Agriculture, Food, and Forestry	Not aggregated
	BioethicsWeb	Not aggregated
	MedHist	Not aggregated
	Psci-com	Not aggregated

Figure 2 provides an overview of the subject coverage of the information collected and, of course, reflects the bias in the data collection. However, there are still conclusions that can be drawn from this data; where the disciplines have been fully aggregated from Intute and are therefore as close to possible a comprehensive catalogue of the material available to UK researchers, the proportional popularity of digital resource will not change significantly in relation to one another. Analysis will therefore be restricted to the Arts and Humanities and Social Sciences data. It must be noted that the proportional of these subjects will drop overall as more collections from the Science categories are added to the data during DRAI Phase 2.

It is interesting to note that ‘Historical and philosophical studies’ is by far the most common HESA subject category represented, followed by ‘Creative Arts and Design’.<sup>11</sup> The latest HESA statistics available indicate that 6.7% of the students in the UK are engaged in Creative Arts and Design and 4.3% are engaged in Historical and Philosophical studies.<sup>12</sup> Conversely, ‘Business & administrative studies’ which is studied by around 13% of UK students has only a 2% share of the digital collections. From this we can see that the provision of digital resources does not correlate with the popularity of subjects being studied at universities.

One reason for this disparity could be that digital collections are not created to satisfy a particular need in the HE community, but their direction is instead defined by the interests of their creators, and their availability shaped by the institutions that manage the resources and make them available. It is surely no coincidence that History and Creative Arts resources are proportionally so high when four AHDS subject centres (Archaeology, History, Performing Arts, and Visual Arts) curate and make available large numbers of collections within these domains,<sup>13</sup> and that the provision of

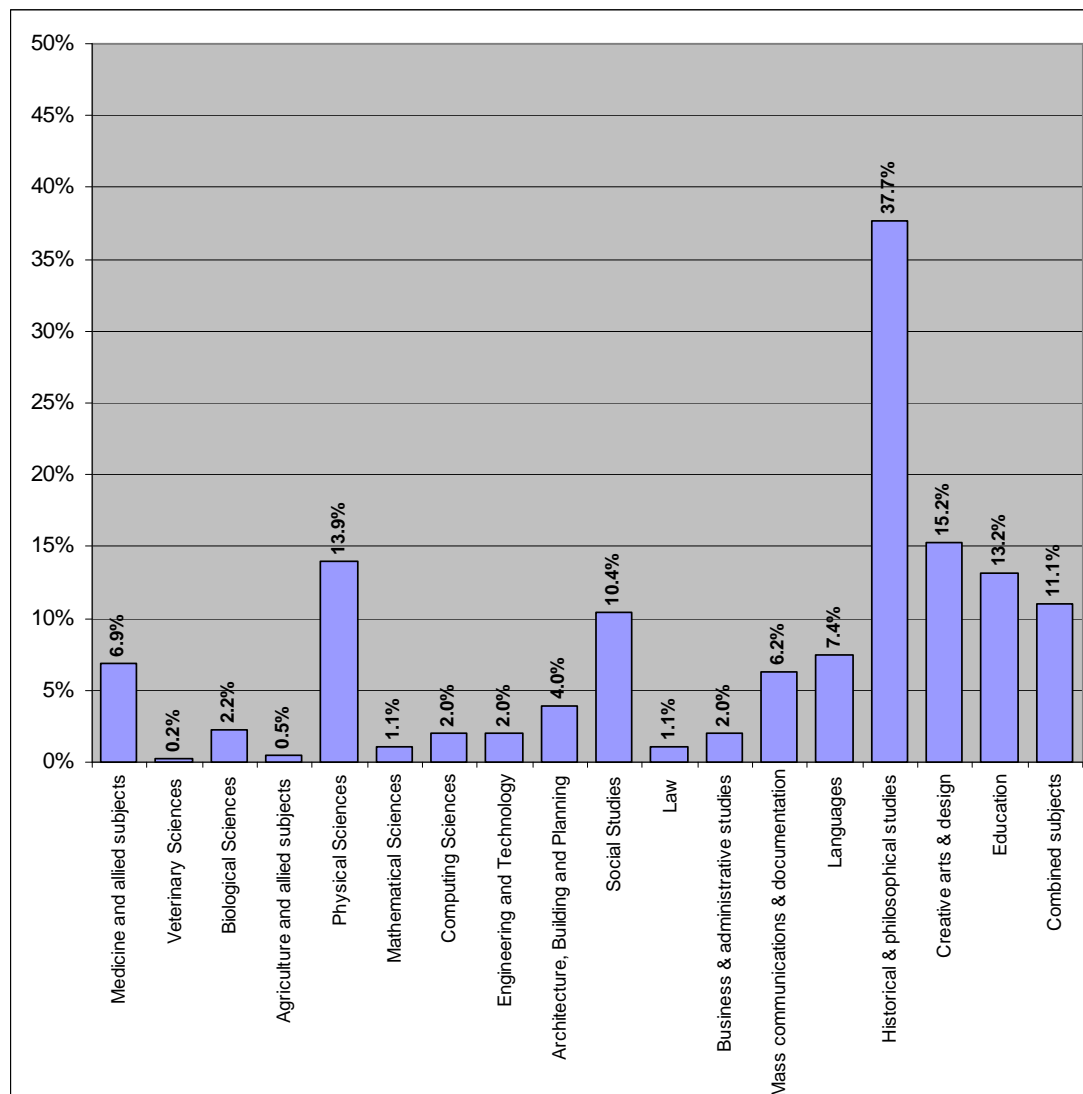
<sup>11</sup> A full listing of HESA subject categories is at Appendix 1.

<sup>12</sup> <http://www.hesa.ac.uk> (2005/6 reporting year). Data on the subject areas of university staff are not available, however it is a reasonable assumption that the student figures are accurate of the popularity of subjects across the higher education environment.

<sup>13</sup> ‘Languages’ (the subject area of AHDS Literature, Language, Linguistics) is also overrepresented when compared to course popularity.

resources for ‘Mass communication’ so exceed the student proportions when there is a large, well-funded, and technically expert organisation such as the British Universities Film and Video Council<sup>14</sup> which creates and preserves collections.

**Figure 2 - subject relevance of collections by HESA category**



More details about the variety of collection owners is available in **Section 2.3 Collection Owners**, however, it should be noted that the creators of digital content are extremely varied, encompassing heritage organisations, public and private memory institutions, commerce, individuals, and individually funded research or access projects, as well as educational and research centres and that meeting user demand in teaching is often not the primary goal of digital resource creation.

The data does appear to challenge the common perception of the arts and humanities as lagging behind in the use of digital content. The fact that many collection creators are heritage organisations could suggest that the arts and humanities, whilst slow to grasp the e-prints open access agenda, instead prefer to embrace the digitisation and dissemination of primary sources, most particularly image collections and textual sources.

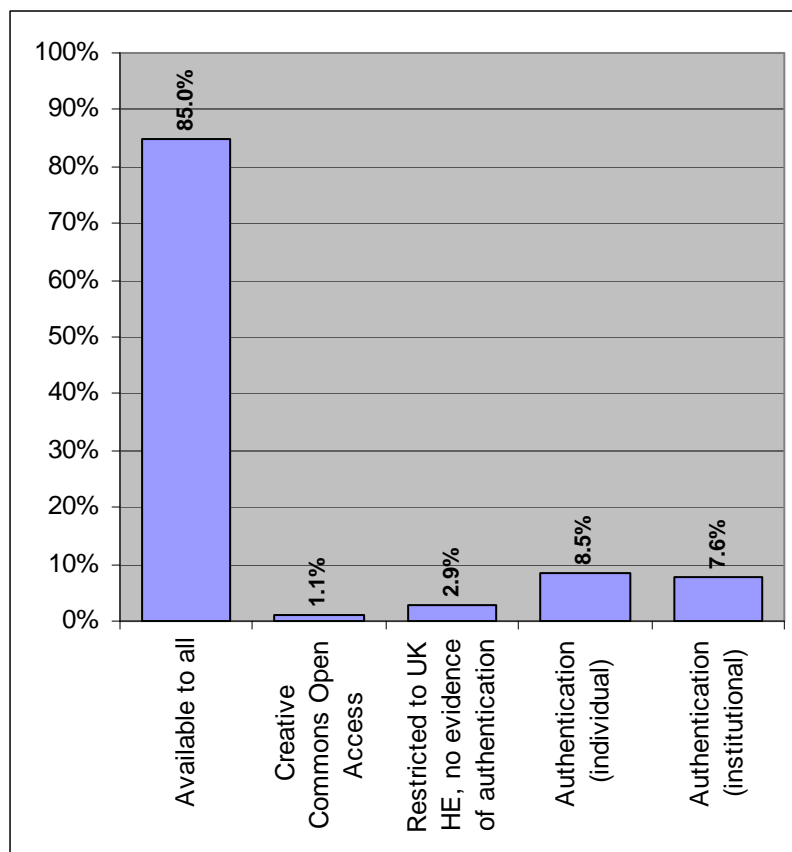
<sup>14</sup> <http://www.bufvc.ac.uk/>

Although it is difficult to comment on the Science and Health subject areas, the results may also be a reflection of the lack of accessibility of medical and science data due to ethical and confidentiality concerns. Different research practices within subject domains may also affect the subject coverage of collections, for example, if data is shared within communities of researchers rather than being made publicly available, it would not be represented within the DRAI data. It is to be hoped that the RIN study currently underway to investigate trends and practices in data publication<sup>15</sup> may throw light on data sharing and publication behaviours.

## 2.2.2 Access: methods and controls

As Figure 3 shows, the vast majority (85%) of digital resources are freely available to all users without restriction. 7.6% require some sort of authentication or free subscription at an institutional level, and 8.5% require individuals to register for access.

Figure 3 - Allowed access to collections



It is important to note that the DRAI definition of “freely available” considers the cost *at the point of use*, therefore some collections which require a paid-for subscription at an institutional level nevertheless offer free access to individuals from within that organisation.

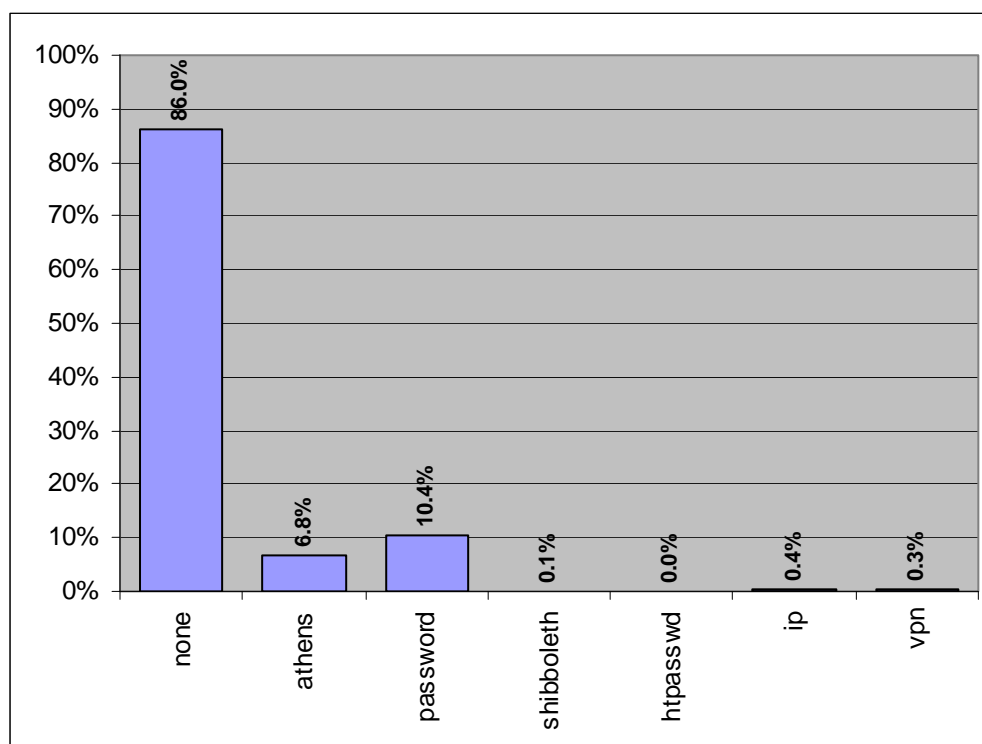
<sup>15</sup> <http://www.rin.ac.uk/data-publication>

The categories shown as part of this graph are not mutually exclusive and there is overlap with some collections allowing both access through a university or library subscription and individual registrations. Access opportunities cannot always be thought of as simple divisions, and can have varied complexity. For example, SCRAN<sup>16</sup> offers limited access to its considerable collections for free, with restrictions on the way content is used. It also offers tiered-cost institutional subscriptions which allow free (at the point of use) access to users, with additional rights over content, as well as individual subscriptions. Therefore this particular collection has three modes of access. In general however, collections tended to have a single mode of access.

The very low figure shown for Creative Commons Open Access (which was anticipated as being a useful separation from 'Available to all') shows a limited engagement with this form of access, however it is speculated that this figure may be lower than in actuality due to the subject coverage of collections thus far aggregated and the lack of explicit information about access on collection Websites.

Figure 4 shows what form of authentication is required for collections, reflecting their modes of access. Categories are based on the options available in JISC's IESR schema.<sup>17</sup>

**Figure 4 - Authentication for access control**



This data proved difficult and time-consuming to discover, as access to some resources was not transparent (i.e. the distinction between there being no

<sup>16</sup> <http://www.scran.ac.uk>

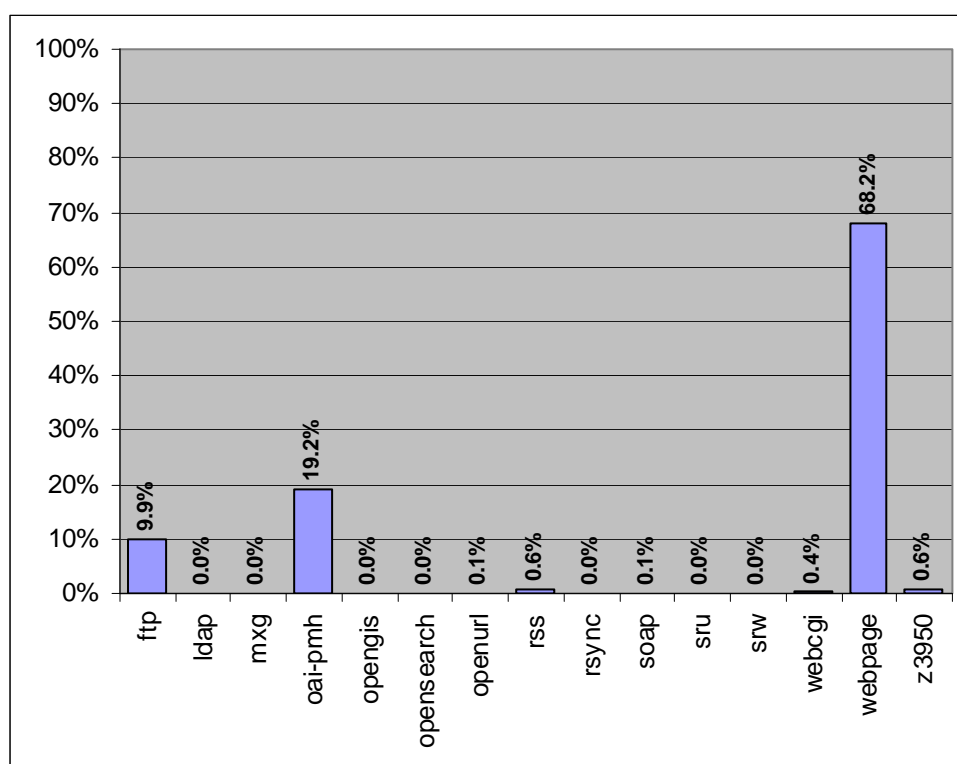
<sup>17</sup> <http://iesr.ac.uk/profile/vocabs/index.html/#AuthList>

authentication and the Athens authentication being invisible was not always apparent).

Nevertheless, somewhat surprisingly, Athens authentication was identified in only 6.8% of records. Again, this reflects a paucity of technical information available through collection interfaces, as 2.9% of records advertised themselves to be restricted to UK HE use but information about how this restriction was enforced could not be found. It is likely that at least some of these collections rely on Athens, however this supposition could not be verified.

Other forms of authentication were either not prevalent (Shibboleth for example is a relatively new solution and is not yet widely used), or were not able to be identified from publicly available information.

**Figure 5 - Access methods**



Apart from Web pages and downloads, technical information about the access method to each collection was even more difficult to identify, leading to a large number of NULL fields in the Access Method property. The categories are as defined in the JISC IESR schema.<sup>18</sup> Web page here refers to where the content itself is actually delivered on a Web page (not where the Web page is simply a conduit for ftp downloads for example).

The data shows that nearly one fifth of collections are OAI-PMH compliant. This data includes all the collections in the AHDS Collections Repository. However, due to lack of publicly accessible information, this category is likely to be under-represented. Although all collections delivered through software such as DSpace can support OAI-

<sup>18</sup> <http://iesr.ac.uk/profile/vocabs/index.html/#AccMthdList>

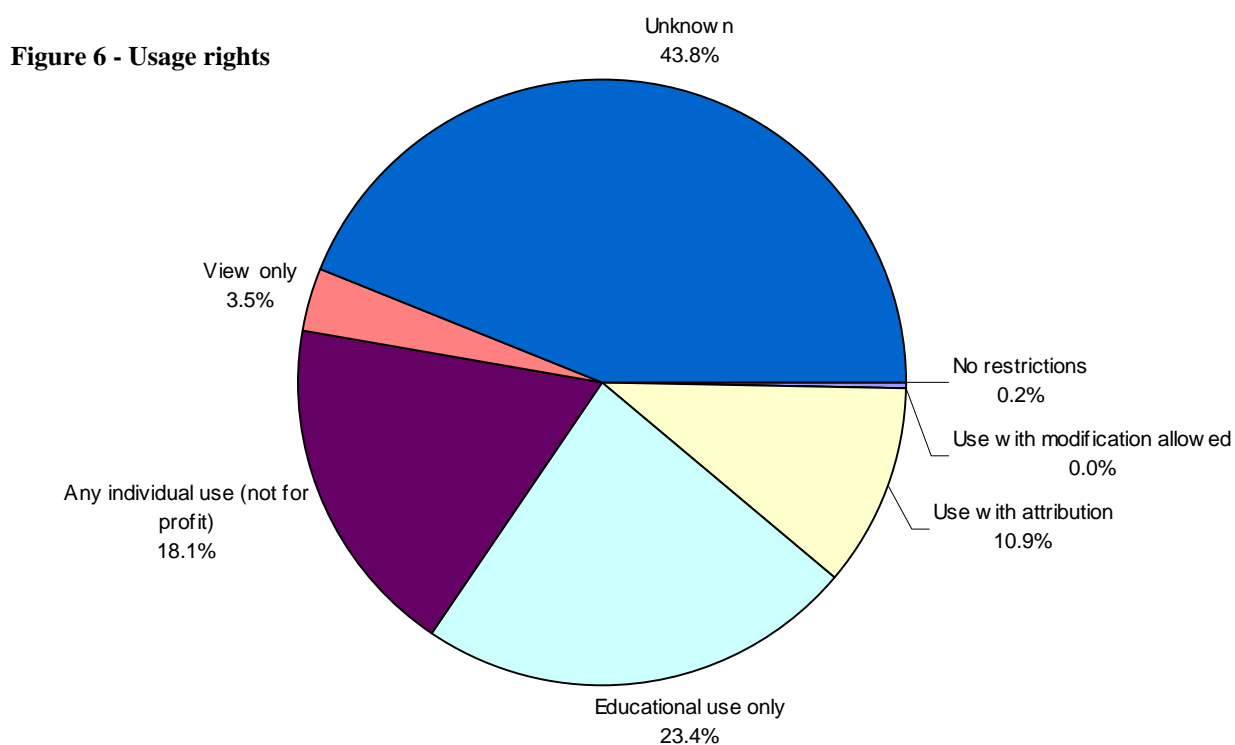
PMH, the user can choose to turn this functionality on or off<sup>19</sup> so it is not possible to make blanket assumptions about compliancy.

In Phase 2 of the DRAI project, the records will be enhanced property by property so it will be possible to use sources such as the Open Archives Register and OAIster<sup>20</sup> to clarify remaining null fields.

### 2.2.3 Use Rights

The following chart shows the information collected on use rights for the digital content within collections. Where it existed, a link was captured for each collection's usage policy, and information was classified as follows: view only (passive use only allowed); Educational use only; any individual not-for-profit use; use with attribution/agreement; any use (modifications allowed); no restrictions on re-use. It is worrying that usage information was not readily available for nearly half of collections (43.8%). This identifies a lack of awareness within the community of the importance of stating use rights clearly, and runs the risk of users assuming that any content freely available on the Web can be used without restriction.

Over 41% of collections explicitly stated their restriction to not-for-profit uses, and it is expected that many of the collections for which information was not identified will also fall into this category. Only 4 collections (0.2%) advertised their content as free to use without restriction.



There was clear evidence showing that collections which have explicit rights assigned

<sup>19</sup> <http://www.dspace.org/faqs/index.html#oai>. The delivery software was captured in the Inventory, so it is possible to see how many resources are delivered through DSpace for example.

<sup>20</sup> <http://www.openarchives.org/Register/BrowseSites>, <http://www.oaister.org/>

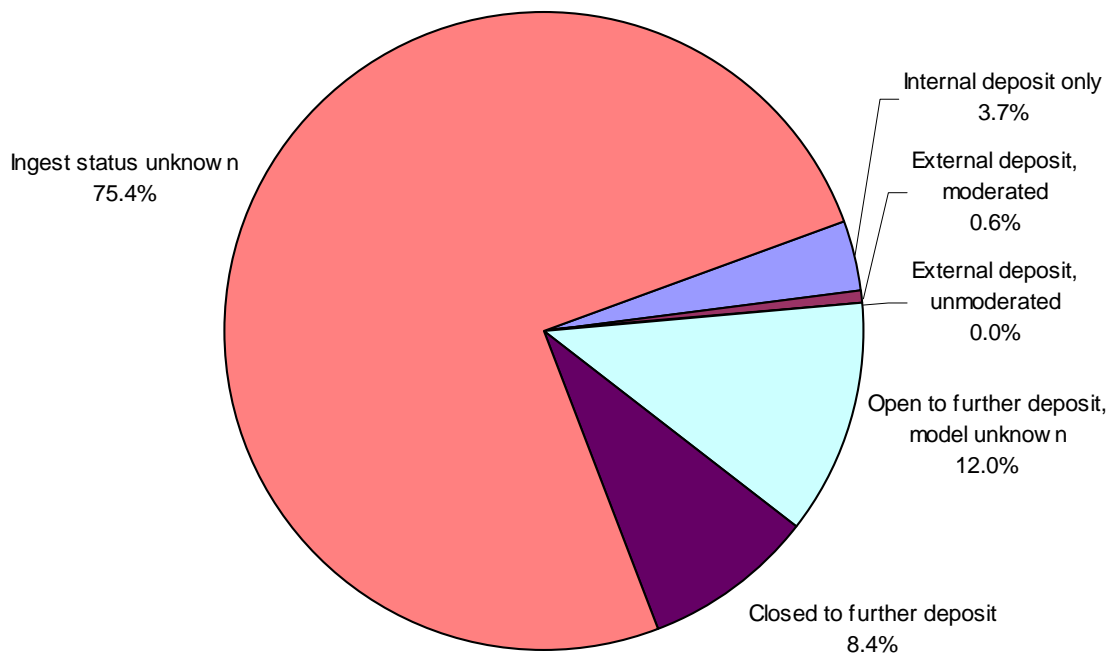
to them via Web information or usage policies tend to come from managed repositories whereas information is more scarce from isolated or 'standalone' resources managed individually. Usage policies are more important to organisations managing more than one resource, and in general, the more expert the organisation, the more likely it is to both clearly define use rights and to make this information easy for users to find.

### 2.2.4 Ingest Policies

The majority of collections did not have information available regarding their policies on ingest of new data. Of those that did provide information (which tended to be held in managed repositories) it is notable that the majority (16.3%) are open to further deposit of records or items, resulting in a large number of constantly growing resources. This demonstrates the dynamic nature of so many digital resources and clearly has implications for curation of the data.

8.4% of collections stated that they were closed to further deposit. There are two major tendencies in this group: firstly that the collection has completed its mission (for example comprehensive digitisation of a particular analogue source); or secondly that funding for the resource has ended so there are no staff to add data and the collection is considered closed (complete or otherwise). It is surprising that, in an environment where so many digital collections are funded by one-off grants rather than ongoing funding, there seem to be more collections that are open to new deposits than closed/completed resources. There is perhaps a bias in the information known, it seems more likely that a collection which is still growing would advertise this fact than a static resource, however this is not verifiable.

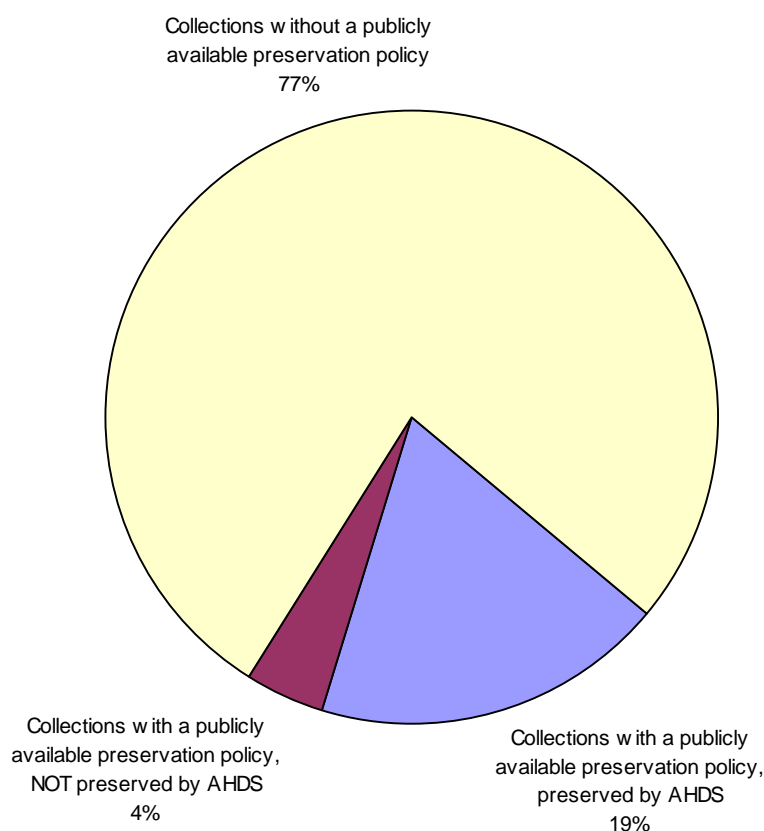
Figure 7 - Ingest of new data to collections



## 2.2.5 Preservation

Of the 1,924 collections aggregated, only 439 (23%) had evidence of a preservation policy, and of these the vast majority were from within the AHDS Collections Repository. This situation is extremely worrying and demonstrates just how fragile digital content is when not part of a preservation infrastructure. Whilst a preservation policy not being easily identified from the Website does not prove that a collection does not have one, it does provide an indication of the level of importance attached to digital preservation by that collection or owner. Over three-quarters of collections aggregated risk loss of intellectual content and investment due to lacking an infrastructure for long-term preservation and access. With the loss of funding to the AHDS, a further 19% of collections will come under threat within the next 5 years.

**Figure 8 - Preservation policies**



## 2.2.6 Relationships between collections

The DRAI Project has successfully catalogued every major 'parent' level repository identified in the UK and has comprehensive coverage of the sub-collections of many repositories. However several major repositories, such as the Oxford Text Archive,<sup>21</sup> contain many hundreds of individual collections or resources which it was not possible to comprehensively aggregate into the catalogue during the timespan of the project. Parent-level repositories or portals of which 100% of references were checked and aggregated if appropriate are:

- AHDS Archaeology

<sup>21</sup> <http://ota.ahds.ac.uk/>

- Archaeology Data Service
- AHDS Literature, Language, Linguistics
- AHDS Performing Arts
- AHDS Visual Arts<sup>22</sup>
- BUFVC
- All institutional repositories referenced by Opendoar
- Intute: Arts & Humanities
- Intute: Social Sciences

Many sub-collections which are part of larger ‘parent’ repositories have also been aggregated into the inventory, however the DRAI Project can only guarantee the completeness of the above repositories.

Of the total 1,984 collections aggregated, 332 (17%) act as ‘parents’ to sub collections. This figure was reached by examining which collections hold ‘Collection’ as an item type (see **Figure 1 – Item types**). The relationships between collections are, however, complex. Collections exist in multiple levels of hierarchy, some ‘sub’ collections themselves acting as parents to further resources. It is not uncommon for the hierarchy to have three or more sub-levels. However, it must be acknowledged that what constitutes a parent collection is a matter of judgement, and judgements about content and granularity are often based on the way the collections are managed and presented on Websites as well as any inherent relationships between them. For example BBC Radio 4 - Science : A Twist of Life, is clearly part of BBC Radio 4 - Science Archive, which in turn is part of BBC Radio 4 - Audio Programmes<sup>23</sup> and the siblings of these collections are connected by content and owner. Conversely, most of the collections in the AHDS Collections Repository are not natural siblings, their connection is simply that they are preserved by the same organisation.

Furthermore, collections do not exist in a simple tree structure. Related to the fact that collections can have multiple owners (see **Section 2.3 Collection Owners** below), the same digital resource can have multiple instances, and therefore be part of several different repositories. For example, a research centre based at a university could create a resource, offer it for access through their own Web page, deposit it with their institutional repository, and also with the AHDS. This collection then could have three parent ‘archives’ or ‘repositories’: the research centre’s cumulative digital output perhaps delivered online through a search on their homepage, via the institutional repository, and also through the AHDS Collection Repository’s cross-search feature. This leads to different contextual relationships for individual collections as well as multiple ‘parents’. An example from the aggregated data is Cecilia<sup>24</sup> which (although delivered through its own Website) is part of IAML (UK) as well as being preserved as part of AHDS Performing Arts’ collections. The DRAI Project could not hope to capture these complex relationships and when completing the Is Part of property (which describes a single relationship to another collection) a

---

<sup>22</sup> AHDS History contains 657 collections which it was unfortunately not possible to complete cataloguing on an individual basis – the aggregators estimate that around 1/3 of the individual collections were input into the database.

<sup>23</sup> <http://www.bbc.co.uk/radio4/science/atwisttolife.shtml>,  
[http://www.bbc.co.uk/radio4/science/archive\\_index.shtml](http://www.bbc.co.uk/radio4/science/archive_index.shtml),  
<http://www.bbc.co.uk/radio/podcasts/directory/station/radio4/>,

<sup>24</sup> <http://www.cecilia-uk.org/>

judgement was made as to which was the primary parent. Add to this complexity the fact that the resource may be referenced in many different online hubs, portals, catalogues, and other information sources (for example Opendoar, Intute, and JISC's IESR). Although these 'sources' were only included in the DRA Inventory if the catalogue itself could be considered to have significant research value, the distinction to many users may not be clear and they may mistake a collection being *referenced* by another collection for it actually being *part of* that collection.

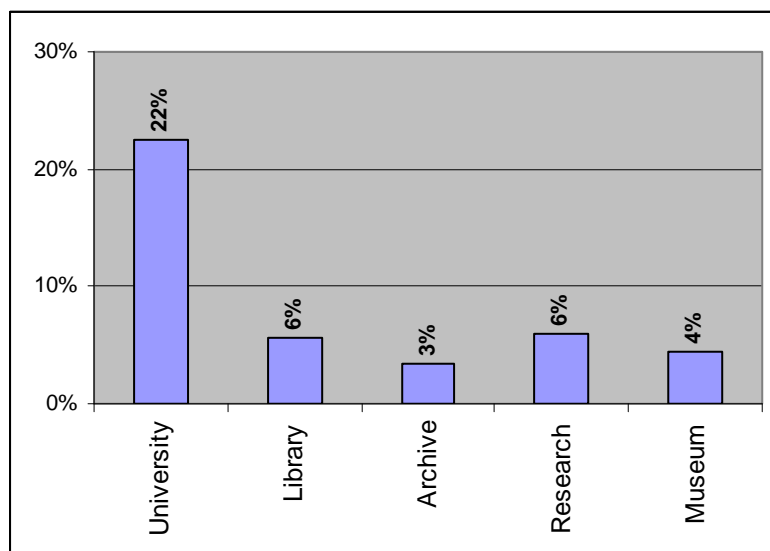
It is worth reiterating the highly fragmented nature of individual resources mentioned in **Section 2.2 Collections Data**, above. The relationships between collections are highly complex and multi-faceted and rarely is all the data from any large analogue collection presented digitally as a single, discrete resource. There is a sense from an overview of all data gathered that some resources are seen as 'incomplete' (for example represent just one section of a large archive). This presents a frustration in terms of information retrieval for the data, as well as inconsistencies in the design of related digital resources posing an obstacle to swift, accurate research.

A fuller understanding of the relationships between collections (for example, seeing all the children of one particular parent) could be achieved by detailed querying of the DRAI Project XML data but was not within the scope of DRAI Phase 1.

### 2.3 Collection Owners

Perhaps one of the most surprising discoveries of the DRAI Project is the number and variety of collection owners. There are 1,051 different owners of 1,924 collections. This demonstrates just how fragmented digital resource ownership is, especially given the fact that many national organisations can claim ownership of at least ten collections each.

**Figure 9 - Collection owners**



Collection owners include: universities and their individual departments, archives, galleries, charities, colleges, companies, government departments, libraries, councils, museums, research institutes, heritage organisations, national parks, professional

bodies, national corporations, individuals, foundations, voluntary agencies,<sup>25</sup> and political parties!

Often the collection is effectively independent and the owner is the same as the collection<sup>26</sup> whilst other collections have multiple owners, as a result of collaboration in either the creation of the resource or its management and curation. What is surprising is the proportion of owners who are not educational or memory institutions. As Figure 9 shows, although universities own at least one-fifth of collections<sup>27</sup> other memory institutions do not have the large proportion of overall ownership expected, or if they do have overall ownership of collections, this is largely invisible to users. This reiterates the fractured nature of digital collections and suggests that it extends to their administration as well as content.

The relationships between owners of digital collections can be as complex as the context of the collection itself. For example, one ownership model could be: A collaborates with B to create a collection from the analogue holdings of C with technical development by D. The collection is then hosted at A but also given to E for long-term preservation where migration is performed on the data to prevent obsolescence. The ownership rights over the data held within this collection are extremely complex and a contest of ownership would be difficult and time-consuming to resolve, almost certainly to the detriment of public access to the content.

### 3 Conclusions and Recommendations

#### 3.1.1 Number of digital collections within the scope of DRAI

Based on the results of the inventory and taking into account the Intute sections hitherto uncovered, there are likely to be well over 3,000 digital resources which are based at UK organisations, of use to higher education, and (potentially) available to researchers.<sup>28</sup> This was a higher number than expected, given the very short timescale of the DRAI Project. (Phase 1). There are also likely to be a large number of small collections which do not appear in information sources and were therefore not aggregated into the inventory.

**Recommendation:** that the data gathering exercise is completed, allowing as comprehensive a vision as possible of digital resource provision in the UK. As of December 2007, JISC has committed further funding towards this task and a subsequent report (available approx. July 2008) will provide a more

---

<sup>25</sup> E.g. Alcohol Concern: <http://www.alcoholconcern.org.uk/servlets/home>

<sup>26</sup> For example, the British War Memorial Project: <http://www.britishwargraves.org.uk/>

<sup>27</sup> Please note, this data does not include organisations without 'university'/'library' etc. in the title, therefore if a collection is owned by an organisation such as the Bakhtin Centre (<http://www.shef.ac.uk/bakhtin/contact.html>) it is not included in this count, even though it is part of the University of Sheffield. The DRAI project captured a 'type' for each owner which allows for a more accurate analysis of owners if required. Note, these categories are not mutually exclusive, to allow for the correct inclusion of dual-type owners such as University Archives.

<sup>28</sup> Figures for the sections of Intute covered 100% indicate that 1/40 of the links provided qualified for inclusion under the terms of the DRAI Project. In that approximately 80% of the remaining two sections was not aggregated, and their total number of records is around 65,000, a reasonable estimate for digital resources yet to be discovered is 1,325, assuming negligible variance in repetition and relevance between sections.

complete picture of the results initiated above. It would be valuable to enhance this data with newer collections and those which are not referenced by a portal or hub. This task goes beyond aggregation and would require a more active approach, such as contacting individual university libraries and departments with a request for information about all of their collections, then cross-referencing with the data already aggregated.

### 3.1.2 Availability of information about collections

Whilst cataloguing information is often present, most collections do not make many details about their technical infrastructure or curation environment readily available to the public. Although OAI-PMH compliancy is promising, the opportunities for information discovery offered by this standard do not seem to have fulfilled their potential.

**Recommendation:** Having technical and curation information readily available in a structured form would be invaluable to researchers, particularly those undertaking surveys similar to DRAI, as well as the projects and institutions who are actually responsible for the digital content in question. It would also reduce the number of individual queries directed at content owners about the technical details of their collections. It is recommended that exemplars of structured data on this subject (such as the JISC IESR) work with funders and major content owners in order to produce widely accepted templates for the presentation of technical information, which can then be published on organisations' or collections' Websites.

Additionally, it is recommended that structures are put in place to allow communities to make more use of the information that is already provided through OAI-PMH. For example, an aggregation tool aimed at a specific community of practice would both increase use of the resources referenced and encourage new collections to expose their metadata through OAI-PMH.

### 3.1.3 Acknowledging the complexity of the field

Digital resources in the UK and their relationships and management structures are extremely complex. There are multiple modes of creation, presentation, curation, management, access and delivery. Due largely to funding structures, digitisation activity is fragmentary and ad hoc, producing large numbers of small collections which often incompletely represent a corresponding analogue collection and certainly provide a somewhat fractured picture of the subject area. The relationships between collections are highly complex and multifaceted, some collections cover the same data but are not linked together; others imply complete coverage without providing it. The scope and range of digital collections is often difficult to unravel for researchers. Many collections are 'hidden' behind search interfaces and within Web pages. This situation leads to difficulties in discovering the existence of data, and inconsistencies between collections present a danger in interpreting the content accurately.

**Recommendation:** Further research into user needs in discovering digital collections would allow national organisations to come up with an appropriate strategy for facilitating use of the existing resources. Resource discovery and use would be increased by separate collections being aggregated logically based on their content (as opposed to geographically and administratively as suggested by the Institutional Repository model) and the creation of more

parent-level, subject-based collaborations such as Open Emblem Portal<sup>29</sup> or specific community-driven aggregations of OAI-PMH data as mentioned above is recommended. It is speculated that an ideal level of granularity will be somewhere between the ‘catch-all’ strategy of Intute (which has excellent coverage but suffers from low precision in searching) and the fragmented nature of collections held in individual departments of educational or memory institutions which rely on Google to refer users to their content. Ideally, logical collections of resources will go beyond simply aggregating links to individual projects and also be committed to data preservation and open access, acting as a managing repository for the data under their responsibility (and an indispensable resource for their research communities). It is also recommended that the information from DRAI is incorporated into the JISC IESR as this will allow searching within a very comprehensive inventory of digital collections at a higher level of accuracy than can be achieved through link hubs.

### 3.1.4 Administrative and management relationships

Digital collections are owned by a large number and variety of organisations and a surprisingly small proportion of these owners are educational or memory institutions. Relationships between organisations can be extremely complex and ownership over content is often blurred.

**Recommendation:** It is recommended that research is undertaken to investigate ways of helping institutions to clarify and make more explicit ownership over collaboratively produced or managed content. One outcome could be the production of freely available, customisable ‘licence’ templates which organisations could use to agree on ownership of content.

### 3.1.5 Digital resource provision for HE

UK researchers may not be the primary target of many digital resources with creation being driven by the agenda of the creators and accessibility often dependent on specific organisations, each with a different scope. As such, digital collection provision does not correlate particularly with its potential use in UK HE.

**Recommendation:** The recommendation here is that more training in the discovery and use of digital resources is necessary to allow individual teachers or researchers to make the most of the digital resources that do exist for their subject interest.

### 3.1.6 Usage rights

Almost half of collections do not have clear information on usage rights. In general there is a lack of publicly available information in an emerging field, and a definite lack of awareness as to the importance of this information amongst the increasing numbers of people creating and managing digital collections.

**Recommendation:** JISC has a role increasing awareness of open access and Creative Commons, as well as generic information about rights over digital assets and collections.

---

<sup>29</sup> <http://media.library.uiuc.edu/projects/oebp/>

### 3.1.7 Preservation environment

Only 4% of collections had freely available evidence of a preservation policy (excluding all the collections preserved until March 31<sup>st</sup> 2008 by the AHDS). Most 'standalone' collections lack policies to regulate their use, and ensure their sustainability for the future. This demonstrates that the preservation infrastructure is extremely weak and is a critical situation for the immense intellectual and financial value of the UK's digital assets. The UK is seeing an explosion in research data collections made available outside the managed repository framework which creates huge potential problems for the future.

**Recommendation:** Content creators, owners, and managers need to acknowledge the importance of sustainability planning before, during, and after the project. It is recommended that more outreach and training about sustainability and best practice in digital resource creation and curation is provided. They also require technical and financial support in order to develop and act on preservation policies. Sustainability costings should be written into grant applications with satisfactory plans for long-term preservation being a condition of the award. Organisations such as JISC could work with research councils and other funders to develop methods for providing this support, and formal collaborations between subject and technical/curation specialists.

## **Appendix 1: HESA Subject Categories**

Agriculture & related subjects  
Architecture, building & planning  
Biological sciences  
Business & administrative studies  
Combined subjects  
Computer sciences  
Creative arts & design  
Education  
Engineering and technology  
Historical and philosophical studies  
Languages  
Law  
Mass communications & documentation  
Mathematical sciences  
Physical Sciences  
Social Studies  
Subjects allied to medicine  
Veterinary science