



JISC Final Report

Table of Contents

Acknowledgements.....	1
1. Executive Summary	2
2. Background	2
3. Aims and Objectives	3
4. Methodology.....	3
5. Implementation.....	4
Implementation details	4
Problems and lessons learned	7
6. Outputs and Results.....	8
7. Outcomes.....	8
Project achievements.....	8
Impact of the project	11
8. Conclusions.....	12
9. Implications	12
10. Recommendations	12
11. References.....	12
12. Appendices	13
Appendix 1. Catalogue schema showing mapping to IESR and DPE names.....	13
Appendix 2. Django Interface	16
Appendix 3. Quality assurance and data checking.....	17
Appendix 4. Database to XML transformation	18
Appendix 5. Coverage of sources.....	20

Acknowledgements

The Digital Repositories and Archives Inventory (DRAI) project was funded by JISC as part of the Digital Repositories Programme 2007-8.

The project was undertaken by the following partners:

- Arts and Humanities Data Service (AHDS), King's College, London¹ (Sheila Anderson: Principal Investigator)
- AHDS Performing Arts, HATII, University of Glasgow² (Daisy Abbott: Project Manager; Stephen Armstrong, Juline Baird, Isabel Fernandes: Aggregators)
- Digital Preservation Europe (DPE), HATII, University of Glasgow³ (Brian Aitken, Graeme Cannon: Technical Developers)

The DRAI project also wishes to acknowledge significant technical and interface support from Adam Rusbridge (DCC, LOCKSS Technical Support Service, HATII).⁴

¹ <http://ahds.ac.uk/>

² <http://www.ahds.ac.uk/performingarts/index.htm>, <http://www.hatii.arts.gla.ac.uk/>

³ <http://www.digitalpreservationeurope.eu/>

⁴ <http://www.dcc.ac.uk/lockss/>

1. Executive Summary

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.⁵ 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.

The overall approach was focussed on interoperability with the JISC IESR and in future the DRAI data can be incorporated with the minimum of effort. The DRAI project aggregated and classified nearly 2000 records from a variety of critical existing sources and reports, including checking over 60,000 of Intute’s links to digital resources. A wide variety of information on the access to and preservation of these digital collections was recorded into a database and XML document. This allowed the production of a report, *Digital Repositories and Archives Inventory: Conclusions and Recommendations*, detailing the project’s conclusions regarding the current preservation environment and recommendations for future issues relating to the preservation of the digital resources identified. 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.

2. Background

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.

JISC’s Information Environment Service Registry⁶ aims to make it easier for other applications to discover and use materials which will help their users’ learning, teaching and research and has been developed with portals and other applications in mind, so that portal developers can access up-to-date information about available resources, without having to maintain this information themselves. Interoperability between different sources of information about digital content is a key factor in supporting the JISC mission. Additionally, other reports such as the AHDS Performing Arts Scoping Study identify the need for a “one stop shop” for information discovery across different digital collections as a primary user need.⁷ The catalogue of resources created during this project complement the material already in the IESR, and provide specific information about the preservation

⁵ <http://www.ahds.ac.uk/performingarts/pubs/scoping-study-2006.pdf>

⁶ <http://iesr.ac.uk/>

⁷ <http://www.ahds.ac.uk/performingarts/pubs/scoping-study-2006.pdf>

of each collection (information which is not part of the scope of previous portals such as Intute).⁸ This information is crucial to understand the current preservation environment in the UK and builds on previous work by the AHDS and JISC (amongst others) in building strategies for digital preservation.

3. Aims and Objectives

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.

These aims formed the core activities of the DRAI project. More information on the fulfilment of these aims can be found in Section 6, Outputs and Results.

4. Methodology

The invitation to tender and initial discussions with JISC set the definition of what constitutes “repositories and archives of digital content” as very wide. Therefore 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 more narrow definition of ‘repository’ or ‘archive’ onto content which it does not necessarily describe.

Catalogue design

It was decided to refine the work being done by Digital Preservation Europe for their Registry of Repositories⁹ by combining the schema with the JISC Information Service Registry metadata schema, into which a great deal of effort on defining the properties of digital collections and archives has already been put.¹⁰ This was to build upon previous efforts and to ensure maximum possible re-use for the data generated during the DRAI project.

The demand for a single portal listing sources of educational digital resources was the principal user need identified by respondents to the AHDS Performing Arts Scoping Study.¹¹ Whilst making the catalogue publicly accessible is not within the scope of this project, there is a clearly articulated demand within the research and teaching community for such a resource to be accessible online, therefore a primary design feature of the catalogue was to allow it to be easily adapted for this potential future use case (for example, providing the information to allow the repositories to be searched by subject area). It also became apparent during the initial weeks of the project that consistency across classification terms is extremely important to allow meaningful analysis of the data. Therefore the approach taken was to develop a database with an interface which allows aggregators to use drop down boxes of controlled values. This approach not only enforces consistent data entry but significantly speeds up the process. The MySQL database data was delivered at the end of the project in conjunction with the data transformed into XML.

⁸ <http://www.intute.ac.uk/>

⁹ <http://www.digitalpreservationeurope.eu/repositories/>

¹⁰ <http://iesr.ac.uk/guidelines/content/>

¹¹ <http://www.ahds.ac.uk/performingarts/pubs/scoping-study-2006.pdf>

During consultation, it was also agreed to capture information about the preservation environment of each digital resource to allow conclusions to be drawn for the final report.

Data aggregation

Due to the wide scope of what is considered to be a digital 'repository' and the anticipated scale of the task, the approach of the DRAI project was to maximise the time available for aggregation activity. Therefore, three aggregators were appointed to work 14 weeks each, giving a total of 42 aggregation weeks over the 17 week duration of the project. Capturing data on as many resources as possible was considered to be a more important deliverable than completing a full catalogue record for each, so aggregators were instructed to fill out only the Required and Highly Desired properties on their 'first pass', with the potential to enhance the record at a later time if possible.

In some instances, repositories contained many hundreds of individually catalogued collections of research or teaching data. As each can be considered to be a standalone resource, and given the wide scope of what defined a digital repository, the DRAI project attempted to classify all individual collections as supplementary information.

As one of the key aims of both JISC and the project partners is to support the long term preservation and curation of digital data, the classification model and cataloguing activity will include information to facilitate understanding of the preservation environment of these repositories.

Documentation

In order to make this research repeatable, documentation was produced throughout the project, including:

- Project plan, budget, progress report, and final report, based on JISC templates.
- Notes or transcripts of internal communication between JISC and the project team
- Guidelines for aggregating and classifying data
- Notes of classification decisions made by data aggregation staff
- Documentation of quality assurance performed, along with any resulting actions necessary
- Acknowledgement of any assumptions made during data analysis and the statistical approach used.
- Report on the tools and methodologies used, to be included in the final report, including any lessons learned.

Together these materials will form an open, coherent overview of how the project was conceived and executed, allowing it to be repeated in the future and making explicit any factors affecting results.

Reporting and recommendations

The DRAI project team felt that it was crucial to reflect on the data produced within the timescale of the project in order to consider the current picture of digital resource provision for Higher Education in the UK, identifying relative strengths and weaknesses across subject areas. The opportunities for discovery, access, and use of this e-content, and its preservation and sustainability environment have been summarised as part of this report (see conclusions and recommendations, below) and are presented in an additional deliverable, the short report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*. This deliverable allowed the AHDS to draw on their significant expertise to provide recommendations for a high-level long-term curation strategy for the repositories identified.

5. Implementation

Implementation details

Catalogue design

During liaison with JISC, it was decided to design the catalogue to match JISC's IESR as closely as possible within the scope of the project, whilst accounting for the extra information required under the terms of this project. This schema replicates almost all of the DPE Registry of Repositories properties (although under different names) so the catalogue would be compatible with both with the minimum of effort. Properties were given the same names as the corresponding property in the IESR where possible, and in every case use the same controlled value list. The schema defined properties as

Required, Highly Desirable, or Supplementary (in the latter case, information was be entered if discovered, but was not actively sought). The catalogue included a Boolean value indicating whether an entry qualifies for inclusion in the catalogue based on JISC's requirements, however, in implementation it was decided that to record even basic information for resources which did not fit the criteria would have been too time consuming, so this value is universally 'Yes' if the resource record was entered. The Is Part Of property in the catalogue allowed aggregators to represent multiple levels of 'parent/child' relationships, indicating for each collection the host repository (or overall collection) within which it sits. A tabular definition of the schema, along with IESR mapping issues, is included at Appendix 1.

The project team at HATII designed a MySQL database and interface for data entry into the catalogue, ensuring that the data could be output as an XML catalogue (writing to XML using a PHP script) once the data aggregation stage was complete. To minimise effort in building the interface, the technical developer used the Django Web Framework¹² as a means of saving programming time (which would be required for a bespoke PHP interface). Django is a high-level Python-based Web framework which uses customisable templates for creating database interfaces, and provides automatic tools such as an admin interface and various simple filters. This allowed the project team to build a customised database interface quickly (it took a technical developer roughly one week to learn how Django works and to create the interface) and aggregators to enter data without needing any technical skills, freeing up time to concentrate on classification decisions and desk research. Screenshots of the Django admin interface and the main table for collection records to be entered is shown at Appendix 2.

Populating the catalogue

Collation of information was achieved through a detailed review of previous research done in this area and by aggregating resources referenced by online portals and reviews on digital collections.

Aggregation from previous research, reviews and reports was carried out successfully and repositories and collections mentioned in the following were included in the database:

- Guide to Digital Resources for the Humanities,
- JISC Digital Repositories Review,
- AHDS Performing Arts Scoping Study,
- Peer Review and Evaluation of Digital Resources for the Arts and Humanities Research Project,
- RePAH Report,
- LAIRAH Report,
- Report of the East of England Digital Preservation Regional Pilot Project
- Using Digital Resources in Teaching,
- Learning and Research in the Visual Arts,
- eLearning Review
- AHRC ICT Programme review of ICT outputs¹³

Those collections which fit the JISC criteria were entered from the following major online portals, memory institutions, subject specific organisations, centres of excellence, consortia, funders and research councils, other JISC projects and initiatives:

Online portals and specific organisations or centres of excellence

- Opendoar,

¹² <http://www.djangoproject.com/>

¹³ <http://users.ox.ac.uk/~ctitext2/resguide2000/contents.shtml>,
http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf,
<http://www.ahds.ac.uk/performingarts/pubs/scoping-study-2006.pdf>,
<http://www.history.ac.uk/digit/peer/>, <http://repah.dmu.ac.uk/>, <http://www.ucl.ac.uk/slais/claire-warwick/publications/LAIRAHreport.pdf>, <http://www.data-archive.ac.uk/news/publications/darp2006.pdf>,
http://vads.ahds.ac.uk/guides/using_guide/contents.html,
<http://www.ukoln.ac.uk/repositories/digirep/index/SharingeLearningMaterialsSynthesis>

- Intute,
- AHDS,
- Community Led image Collections (CLIC),
- TechExtra
- SHERPA DRIVER project
- Higher Education Academy subject networks.
- EDINA,
- MIMAS,
- UK Data Archive,
- JISC Collections
- British Universities Film and Video Council,
- ARKive,
- National Theatre,
- British History Online,
- Clinical Medicine NetPrints,
- Cogprints (Cognitive Sciences e-prints),
- Cultural Studies e-Archive,
- e-Crystals,
- ESRC Social Science's repository,
- GRADE,¹⁴
- and individual HE departments and institutes

Memory institutions

- National Archives' National Digital Archive of Datasets (NDAD),
- British Library¹⁵

Funders and research councils

- British Academy Portal
- AHRC,
- EPSRC,
- and Heritage Lottery Fund¹⁶

The 14-week data aggregation period included a familiarisation session in using the database, the entry of controlled value lists, and an initial 3 hour quality assurance session which included all three aggregators and the project manager where duplicate records were deliberately created in order to cross-compare between aggregators and clarify understanding of properties. In total, 1924 Collection records were entered into the inventory, resulting in an average of 10.2 records per aggregator per working day.¹⁷

In order to maximise the comprehensiveness of the catalogue, the data aggregation was intended to be performed in three phases:

1. A 'first pass', identifying all immediately available information from parent-level repositories, and aiming for comprehensive subject coverage.

¹⁴ <http://www.opendoar.org>, <http://www.intute.ac.uk>, <http://ahds.ac.uk/collections/index.htm>, <http://www.oucs.ox.ac.uk/ltg/projects/clic/matrix.html>, <http://www.techxtra.ac.uk/>, <http://www.sherpa.ac.uk/projects/driver.html>, <http://www.heacademy.ac.uk/SubjectNetwork.htm>, <http://edina.ed.ac.uk/>, <http://www.mimas.ac.uk/>, <http://www.data-archive.ac.uk/>, http://www.jisc.ac.uk/whatwedo/services/services_jiscollections.aspx, <http://www.bufvc.ac.uk/>, <http://www.arkive.org/>, <http://www.nationaltheatre.org.uk/archive>, <http://www.british-history.ac.uk/>, <http://clinmed.netprints.org/>, <http://cogprints.org/>, <http://www.culturemachine.net/csearch>, <http://ecrystals.chem.soton.ac.uk/>, <http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/AdvancedSearchPage3.aspx>, <http://edina.ac.uk/projects/grade/doc.html>

¹⁵ <http://www.ndad.ulcc.ac.uk/>, <http://www.bl.uk/>

¹⁶ <http://www.britac.ac.uk/portal/misc/areas.html>, <http://www.ahrc.ac.uk/>, <http://www.epsrc.ac.uk/default.htm>, <http://www.hlf.org.uk/English/>

¹⁷ This average takes into account the days when aggregators were off sick or on holiday, therefore the total number of working days is 188 (14 weeks x 3 aggregators - time off + 5 days from extra aggregator).

2. A detailed investigation, discovering as much detail as possible about each repository and adding data for 'child' collections.
3. Filling gaps in data by contacting collection administrators to tie in with the JISC survey and checking the coverage of the data by investigating any subject gap which have become apparent.

However, as the aggregation activity progressed, it became apparent that the first two tasks would take up the bulk of the time allowed and that it would not be possible to add a third level of data enhancement.

No major edits were made to the data model or the database in once the data aggregation had begun, although it was necessary to set some of the fields to default values, both in order to save time for the aggregators and to ensure accuracy. (For example Fulfils JISC Criteria was set to default to Yes, once it was determined that it was too time-consuming to create records which did not fit the criteria).

It was also decided that it would be useful to include a notes field within each record. This was for two main reasons: firstly, became apparent that the complexities of some records made them difficult to classify (for example complex relationships in the Owner or IsPartOf properties), therefore the notes property allowed ambiguous classification decisions to be recorded in a way that was implicitly attached to the record, instead of on a separate document. Secondly, there were a few cases where it was impossible to record required information in the field set aside for it, for example, where a collection did not provide an email address as a means to contact them but instead provided a Web form, which would not validate as an email address. The free text of the notes field allowed aggregators to add or clarify complex information relating to each record. The aggregators also identified some discrepancies between information listed publicly (for example through an aggregator such as Opendoar) and that listed on the actual repository/collection website. The aggregators used research and judgement to determine the value most likely to be correct and where discrepancies were found they have been noted in the notes property.

Documentation

Documentation was prepared on an ongoing basis and includes all of the items mentioned in Section 4, above. A table of quality assurance activities, problems, and resulting actions is at Appendix 3. Documentation of classification decisions made by aggregators was included in the catalogue in the Notes field, as described above. The aggregators were included in HATII's departmental policy of producing weekly reports outlining progress, therefore their reports are held in the HATII administration database.

Reporting

The reporting part the project was led by Sheila Anderson (AHDS Executive). It was decided to produce a separate report reflecting on the results of the aggregation and classification activity. This report is made available as a separate deliverable: *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

Problems and lessons learned

Of the risks identified in the project plan, two had an effect on the implementation of the DRAI project.

The first was that the scope set at the start of the project was extremely wide leading to an underestimation of the number of resources to be included, and an underestimation of the time needed to both filter out the resources which did not fulfil the criteria, and to discover the full range of information for each collection. It was acknowledged at the beginning of the project that the wide definition being applied would lead to a very large number of resources for inclusion. This issue, in itself, would not have been a significant risk. However, the range of information sought for each collection was also large, and 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 projects specifically based on the details of digital access or preservation, such as AHDS or Opendoar) make this kind of information readily available. As part of planning for this risk, it was agreed that a minimum of properties would be deemed to be 'Required'. The DRAI catalogue has a complete record all required fields, however in some cases not all of the 'Highly Desirable' properties could be completed. It is difficult to see how

this problem could have been avoided given the short timescale of the project, however, it may have speeded up initial data entry to demarcate more clearly in the interface which properties were 'Supplementary' i.e. only necessary if the information was immediately at hand. More details on this issue are provided in Section 6 below.

The second risk which affected the aggregation task was an unusually high level of staff illness. This could not be foreseen but led to a loss of aggregating time. An extra worker was reassigned from within the project team in the last two aggregation weeks to mitigate this problem.

6. Outputs and Results

The DRAI project has resulted in a highly adaptable catalogue of digital resources, delivered in XML and as an SQL file. There were a total of 1924 different collections entered in the catalogue, belonging to a total of 1022 owners.

The knowledge gained from this inventory is presented in a short report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

The DRAI project has also facilitated the gaining and sharing of knowledge amongst the project team, particularly in the areas of digital content access methods, preservation policies, and cataloguing solutions which can be implemented in a minimum of time.

7. Outcomes

This project was very short, therefore evaluations of the methodology and procedures was ongoing and is presented here.

Project achievements

The aims and objectives of the DRAI project are listed below, with comments on to what extent the aim was achieved and any further lessons learned whilst working towards each objective.

To gain an understanding of digital resources available to UK Higher Education.

The production of the inventory has a series of tangible benefits for the digital preservation community and creators and users of digital resources. The catalogue could, with the minimum of future effort, be made available as a portal of references to digital resources with a specific and relevant focus for HE researchers. It may also be incorporated into the JISC IESR and DPE Registry of Repositories to maximise its impact. The inventory's focus was also clearly on technical strategies in resource delivery (such as details of how the collection is accessed and preserved). This information has allowed the project team to produce a report discussing the context of digital preservation for UK collections and to make recommendations for the future. For more details on the understanding gained by the cataloguing activity, see the report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

The fact that the inventory has also been delivered as a MySQL database means that a great deal more in depth analysis will be possible in the future by performing specific queries on the data.

To produce a comprehensive XML catalogue of digital resources freely available for educational use in the UK.

The DRAI project gathered information on 1924 digital resources available to the UK HE community and allows various different ways to filter the records. Some of the properties discovered related to access, for example, was whether the resource was free to access for educational use, required any user verification, or was delivered under some sort of payment model. This allows a more detailed view of current digital resources provision than simple classification into free/not free for educational use.

The aggregation task was achieved using a MySQL database which was also delivered as a supplement to the XML catalogue. The XML file was itself produced using a script to transform data from the database into XML. The XML template is attached at Appendix 4. The PHP script which converted the data was also delivered as part of the project.

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. The primary measure of comprehensiveness of coverage of the inventory is by comparison to the subject categories within the Intute portal. This portal has a very wide scope and includes links to Web pages for analogue collections, individual commercial organisations, university departments, information for tourists, blogs, message boards, and other portals, as well as links to Web resources providing actual digital content or catalogues of content. With nearly 120,000 links, Intute is perhaps the most comprehensive of existing sources and is therefore relatively likely to reference repositories, archives, or collections suitable for inclusion in the Digital Repositories and Archive Inventory. Unfortunately, the very general nature of the links provided by Intute meant that extracting links to bona fide *collections of digital content* as defined by the DRAI project (as opposed to simple Web links) was extremely time-consuming even using the advanced search tools provided by the site. To provide a comparison, initial quality assurance identified that there is a significant difference in the time taken to input data that is not immediately apparent (e.g. assessing an Intute link for inclusion, following the link, desk research to re-assess the source for inclusion, research to discover information from 'About' pages, entering information into inventory) and the time taken to aggregate data from an existing quality source (e.g. Opendoar or AHDS). For example, by the time one aggregator had input 75 records using Intute as a source, the aggregators working primarily on AHDS and Opendoar references had input around 135 records each. Evaluating Intute links took the majority of the aggregation time and it was not possible to complete all subject sub-categories within the 14 weeks. A detailed list of which sections were and were not covered is included at Appendix 5 and the potential skewing of subject results has been acknowledged in the conclusions of the project.

The Intute categories contain the following numbers of links:¹⁸

Science, Engineering and Technology	33,658
Arts and Humanities	22,125
Social Sciences	32,575
Health and Life Sciences	31,155

Of these categories, the 54,700 records in the Arts and Humanities and Social Sciences categories were fully checked and aggregated where appropriate into the database. Approximately 20% of the remaining two categories was checked. See Appendix 4 for more detail about Intute coverage.

With the exception of Intute, all other sources noted in Section 5: Populating the catalogue were successfully aggregated into the inventory.

To examine the overall provision of digital resources across subject areas and formats.

The results of this outcome are presented in the report deliverable of the DRAI project, *Digital Repositories and Archives Inventory: Conclusions and Recommendations*. It was acknowledged within the analysis that results will have been affected by the fact that not all Intute subject categories were aggregated.

To discover and analyse the preservation environment for these digital resources.

Preservation information was discovered and aggregated as a 'Highly Desirable' property for all records. The results of this outcome are presented in the report deliverable of the DRAI project, *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

Part of the context for digital resources includes their modes of access and, as can be seen from the schema, it was intended to capture information about several aspects of collections ingest, access and preservation. It was discovered during the course of the DRAI project, that whilst some information is readily available (particularly from sources with an emphasis on preservation and access themselves) it is relatively difficult to discover some technical details about collections from the

¹⁸ As of 22/10/07

information available on their public Websites. Properties which proved very time-consuming and difficult to discover include:

- Delivery Software
- Access Control
- Access Method
- Metadata Used
- and perhaps surprisingly, Temporal

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 still has value despite the large number of null fields in many of the properties as, 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.

To report on the catalogue, providing conclusions and recommendations.

The results of this outcome are presented in the report deliverable of the DRAI project, *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

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.

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,¹⁹ 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
- Archaeology Data Service
- AHDS Literature, Language, Linguistics
- AHDS Performing Arts
- AHDS Visual Arts²⁰
- 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.

As described above, all UK-based collections were included in the catalogue which met the project's wide definition of what constituted a digital 'archive' and were categorised by various other properties, such as modes of access. This means that there is no secondary catalogue, instead the main catalogued can be intelligently queried to separate out those collection which are, for example, accessed through a paid subscription model. A very few collections or repositories were included that do not meet the definition (usually due to being hosted outside the UK but available within the UK). These can be filtered using the Fulfils JISC Criteria for Inclusion property.

¹⁹ <http://ota.ahds.ac.uk/>

²⁰ 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.

It should be noted that where a collection may theoretically available for use *in the future* (for example, collections which are not currently accessible but are held for 'Preservation Only' by the AHDS) they have been included in the inventory.

To fully document the project in order that it can be easily repeated in the future.

Documentation of the project is included in the JISC template reports, appendices, and separate files. The most useful documentation is likely to be the table showing mapping issues to the JISC IESR at Appendix 1.

Impact of the project

As stated above, the DRAI project captured a large amount of technical information alongside content information for 1924 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 Digital Repositories and Archives Inventory will be of significant use to other resource aggregators such as the JISC IESR which is a reliable source of information that other applications, such as portals, can freely access through machine-to-machine protocols, to enhance information discovery and retrieval. Once incorporated into the DPE Registry of Repositories, the data gathered will have an impact on an international level (and hopefully encourage other partner countries to provide similar information to the registry).

Specific lessons learned throughout the 17 week duration of the project are detailed above. Major lessons learned which will be of use to other similar projects can be summarised as follows:

1. It is extremely time-consuming to discover this level of detail about collection content and infrastructure without direct contact with collection owners, administrators, and technical staff. JISC has other surveys in place to gain information through direct contact and this was outside the scope of the DRAI project.
2. An aggregator can catalogue around 10 of records to a basic level of detail (i.e. filling in some but not all of the Highly Desirable properties in the DRAI schema) per working day. Bearing in mind that it is possible to catalogue around 80 resources per day if working from a high-quality structured source and only filling in the Required and obvious fields,²¹ the time taken to perform desk research on collection information should not be underestimated. To provide full detail for all records was not possible. A trade off was necessary between completeness of the survey and completeness of the individual records, with a comprehensive survey taking first priority.
3. Structured, technical sources are far quicker and more accurate to aggregate than very large, non-specific sources such as Intute. Intute coverage indicated that the aggregators checked around 66,000 links from Intute alone, so a very high level of time-consuming 'filtering' is required, even for a project with a relatively wide scope of what constitutes a digital resource.
4. Based on the results of the inventory and taking into account the Intute sections hitherto uncovered, there are likely to be over 3,000 digital resources which are based at UK organisations, of use to higher education, and (potentially) available to researchers.²² This was a higher number than expected, given the very short timescale of the DRAI project.

More information about the impact of the results of the inventory data can be found in the project report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

²¹ Figures from one aggregator who focussed purely on adding basic information for AHDS collections.

²² 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.

8. Conclusions

Conclusions are presented in the separate report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

9. Implications

Implications are presented in the separate report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

10. Recommendations

Recommendations are presented in the separate report *Digital Repositories and Archives Inventory: Conclusions and Recommendations*.

11. References

All reports, reviews, and online sources are listed and referenced in Section 5, above.

12. Appendices

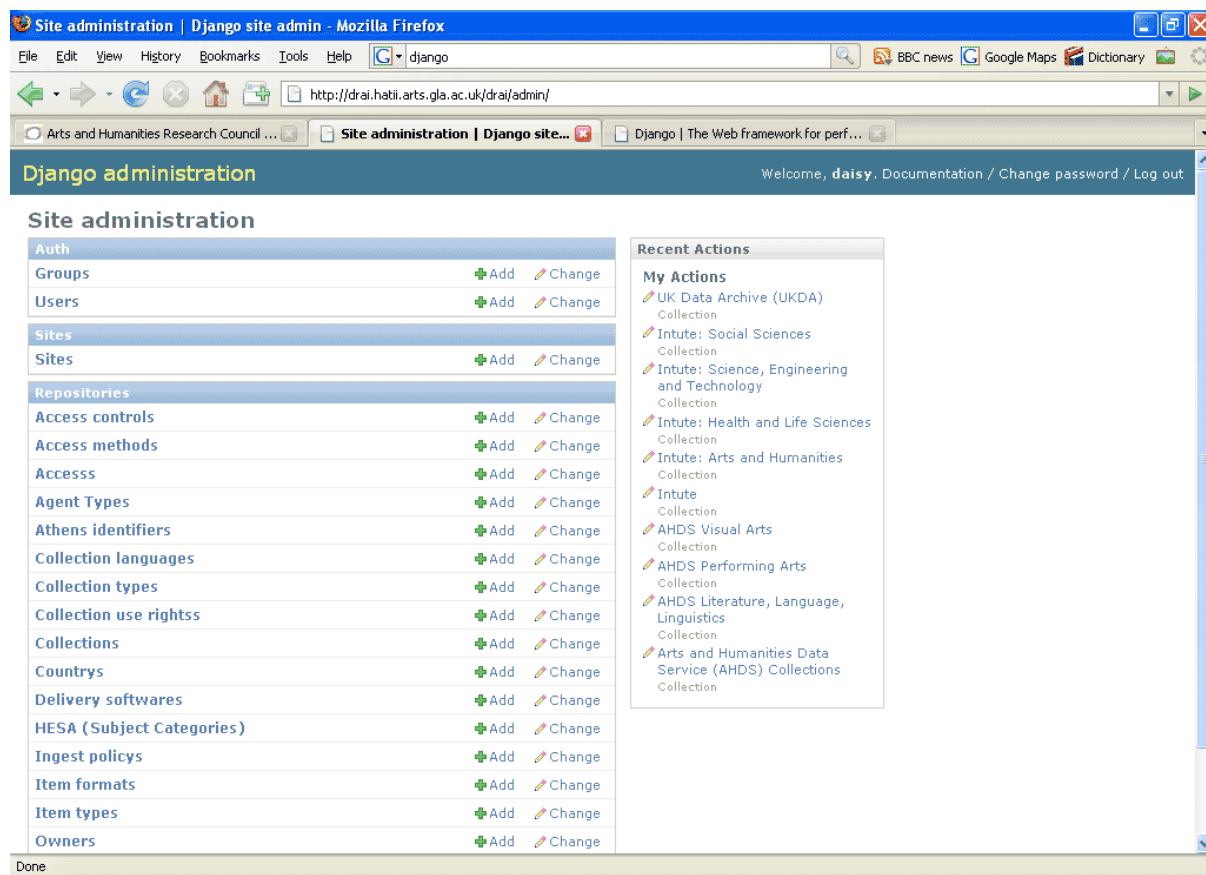
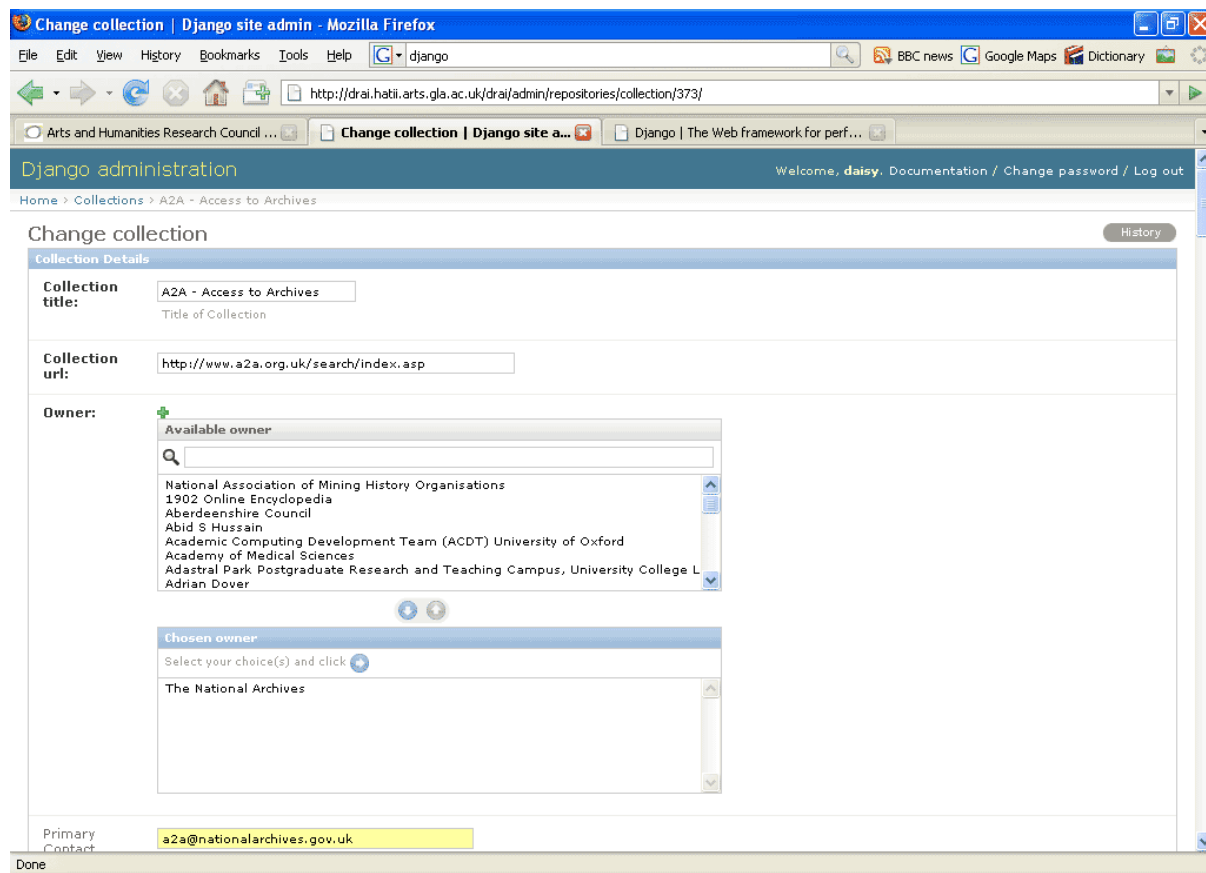
Appendix 1. Catalogue schema showing mapping to IESR and DPE names

DRAI			JISC IESR				DPE Registry of Repositories
Name	Type	Multiplicity	Name	Type	Multiplicity	Mapping Issues	Name
Collection: Title	String	Single	Collection: Title	String	Single	-	
Collection: Alternative Title	String	Multiple	Alternative Title	String	Multiple	-	
-			HasService	IESR internal identifier (relationship to Service: Title)	Multiple	Contradictory information in the IESR guidelines about whether this is Required or Optional. If required will need to be input.	
Owner (Agent)	String (relationship to Agent)	Multiple	Owner (Agent)	IESR internal identifier (relationship to Agent: Title)	Multiple	Contradictory information in the IESR guidelines about whether this is Required or Optional. Should be easily replicable by the association with the DRAI Owner (Agent) property.	
Collection URL	String	Single	-	-	-	-	
Collection: Contact	String	Single	-	-	-	-	
Collection: Description	String	Single	Collection: Description	String	Single	-	
Is Part Of	String (relationship to Collection: Title)	Multiple	Is Part Of	String (URL)	Multiple	Should be relatively easy to assign URL instead of Collection:Title	
Collection Type	C.V.	Multiple	Collection Type	C.V.	Multiple	-	
Item Type	C.V.	Multiple	Item Type	C.V.	Multiple	-	
Item Format	C.V.	Multiple	Item Format	C.V.	Multiple	-	
Collection: Language	C.V.	Multiple	Collection: Language	C.V.	Multiple	-	
Size	String	Single	Size	String	Single	-	
Maturity	String	Single	-	-	-	-	
Temporal	Date	Multiple	Temporal	Date	Multiple	-	
Contents Date Range	Date	Multiple	Contents Date Range	Date	Multiple	-	
HESA subject category	C.V.	Multiple	Subject: (JACS)	C.V.	Multiple	Controlled list is a String in DRAI, number in IESR. Should be relatively trivial to map.	
Subject (Dewey)	C.V.	Multiple	Subject (Dewey)	C.V.	Multiple	-	
-	-	-	Subject (all others)	Various	Multiple	-	
Access	String (imposed CV list)	Single	Access	String (imposed CV list)	Single	It is technically possible for this property to be a multiple value due to the interface. May need to be checked.	
Metadata	String	Single	-	-	-	-	

used							
Packaging standards	String	Single	-	-	-	-	
Collection: Use Rights	String	Single	Collection: Use Rights	String	Single	-	
Collection: Use Rights URL	String	Single	Collection: Use Rights URL	String	Single	-	
Ingest policy	String	Single	-	-	-	-	
Ingest policy URL	String	Single	-	-	-	-	
Legal Mandate to Preserve?	Boolean	Single	-	-	-	-	
Long Term Preservation Policy?	Boolean	Single	-	-	-	-	
Preservation Policy	String	Single	-	-	-	-	
-	-	-	Collection: Logo	URL	Single	-	
-	-	-	Uses Controlled List	C.V.	Multiple	-	
-	-	-	Educational level	C.V.	Multiple	-	
			Has Association	URL (relationship)	Multiple	-	
			IsReference dBy	URL (relationship)	Multiple	-	
Service Properties							
Delivery Software	String	Single	Service: Title	String	Single	Required in IESR but not in DRAI	Software
-	-	-	Administrator (Agent)	String (relationship to Agent:Title)	Multiple	Contradictory information in IESR on whether required or optional. If required will have to be input.	
-	-	-	Serves (Collection)	Internal identifier (relationship to Collection: Title)	Multiple	-	
Access Control	C.V.	Single	Access Control	C.V.	Single	Required in IESR but not in DRAI. Uses same CV list. It is technically possible for this property to be a multiple value due to the interface. May need to be checked.	
Access Method	C.V.	Multiple	Access Method	C.V.	Multiple	Required in IESR but not in DRAI. Uses same CV list.	
-	-	-	Service Function	C.V.	Multiple	-	
-	-	-	Locator	URL	Single	Required in IESR but not in DRAI. Will have to be input	
-	-	-	Interface	Varies	Varies	Sometimes required in IESR but not in DRAI. Will have to be input	
-	-	-	Domain Available	String	Multiple	-	
-	-	-	Service: Description	String	Single	-	
-	-	-	Service: Language	C.V.	Multiple	-	
-	-	-	Mediator (Shibboleth)	URL	Multiple	-	
-	-	-	Service: Use Rights	String	Single	-	
-	-	-	Service: Use Rights URL	URL	Single	-	
-	-	-	Standards	C.V.	Multiple	-	

-	-	-	level				
-	-	-	Service Help	URL	Multiple	-	
-	-	-	Shibboleth Info	URL	Multiple	-	
-	-	-	Service SLA	URL	Multiple	-	
-	-	-	Button Image	URL	Single	-	
-	-	-	Link Text	String	Single	-	
Agent Properties							
Agent	String	Single	Agent: Title	String	Single	Replicated by the Owner (Agent) relationship	Managing Institution
Agent Type	String (imposed CV list)	Multiple	-	-	-	-	
-	-	-	Owns (Collection)	IESR Internal Identifier (relationship to Collection:Title)	Single	-	
-	-	-	Administers (Service)	IESR Internal Identifier (relationship to Service: Title)	Single	-	
-	-	-	Contact	String	Single	-	
-	-	-	Agent: Description	String	Single	-	
Address	String	Single	Address	String	Single	-	
Postcode	String	Single	Postcode	String	Single	-	
Country	C.V.	Single	Country	C.V.	Single	-	
-	-	-	Telephone	String	Single	-	
URL	String	Single	URL	String	Single	-	
Athens Institution Identifier	C.V.	Single	Athens Institution Identifier	C.V.	Single	-	
-	-	-	Agent: Logo	URL	Single	-	
Administrative Metadata							
Creator	C.V.	Single	Creator	C.V.	Single	-	
Date and time	Automatically generated	Single	-	-	-	-	
-	-	-	Contributor	C.V.	Multiple	-	
Fulfils JISC criteria for inclusion?*	Boolean	Single	-	-	-	-	
Reason for exclusion	String (imposed CV list)	Single	-	-	-	-	
Admin Notes	String	Single	-	-	-	-	

Appendix 2. Django Interface



Appendix 3. Quality assurance and data checking

Checks performed	Date	Actions
Checking understanding of terms	16/07	Working with aggregators, added help text to drai interface.
Consistency of classification checks	16/07	Tested all aggregators against same record, discussed and advised any differences.
Consistency checks	06/08	None
Consistency and accuracy tests	10/09	Cleared up a few typos, enforced using “[title], The” to prevent duplication of records and ease finding.
Comprehensiveness test (using opendoar as test case)	10/09	None
Progress checks	10/09	Sought advice from JISC, decision to continue aggregating from Intute.
Request from aggregators – problem identified in that the list of controlled value filetypes at http://www.iana.org/assignments/media-types/ does not appear to have entries for either mp3 or aiff filetypes.	10/09	Difficult to deal with this issue meaningfully. Mp3 was entered as audio/mpeg instead and aiff was generally ignored as no value seemed to accurately represent this filetype.
Progress check	01/10	Reassigned work amongst aggregating team to share workload equally.
First batch of filtering tests: accuracy of entries	01/10	Some fields from the very first records entered (about 12 records) were identified as having lost data when changes were made to the db structure. This was corrected.
Testing filtering system itself	01/10	Need to add a means to identify null values in certain properties. IN PROGRESS.
Initial analysis	10/10	Concern about lack of representation of AHDS records and the fact that AHDS Archaeology data (already entered) is skewing the records, particularly re preservation policies and subject focus. Assigned extra worker to fill in all required fields for remaining AHDS collections
Accuracy checking	10/10	Removed some unwanted whitespace.
Final accuracy and comprehensiveness of parent collection checking performed.	17/10	None.

Appendix 4. Database to XML transformation

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
- <!--
```

```
template to show the data source for each element in the XML version of the DRAI database
```

```
-->
```

```
:- <drai>
```

```
:- <collection>
```

```
<collection_title>repositories_collection.collection_title</collection_title>
```

```
:- <collection_alternate_titles>
```

```
<collection_alternate_title>repositories_collectionalternativetitle.collection_alternative_title</collection_alternate_title>
```

```
</collection_alternate_titles>
```

```
:- <owners>
```

```
:- <owner>
```

```
<agent>repositories_owner.agent</agent>
```

```
:- <agent_types>
```

```
<agent_type>repositories_agent_type.type</agent_type>
```

```
<agent_type>repositories_agent_type.type</agent_type>
```

```
</agent_types>
```

```
<address>repositories_owner.address</address>
```

```
<postcode>repositories_owner.postcode</postcode>
```

```
<country>repositories_country.country</country>
```

```
<url>repositories_owner.url</url>
```

```
<athens_institution_identifier>repositories_athensidentifier.athens_identifier</athens_institution_i  
dentifier>
```

```
</owner>
```

```
</owners>
```

```
<collection_url>repositories_collection.collection_url</collection_url>
```

```
<collection_contact>repositories_collection.collection_contact</collection_contact>
```

```
<collection_description>repositories_collection.collection_description</collection_description>
```

```
:- <is_part_of>
```

```
<part_of>repositories_collection.collection_title</part_of>
```

```
</is_part_of>
```

```
<collection_type>repositories_collectiontype.collection_type</collection_type>
```

```
:- <item_types>
```

```
<item_type>repositories_itemtype.item_type</item_type>
```

```
</item_types>
```

```
:- <item_formats>
```

```
<item_format>repositories_itemformat.item_format</item_format>
```

```
</item_formats>
```

```
:- <collection_languages>
```

```
<collection_language>repositories_collectionlanguage.language</collection_language>
```

```
</collection_languages>
```

```
<size>repositories_collection.size</size>
```

```
<maturity>repositories_collection.maturity</maturity>
```

```
<contents_date_range>repositories_contentdaterange.content_date_range</contents_date_rang  
e>
```

```
<temporal>repositories_temporal.temporal</temporal>
```

```
:- <hesa_subject_categories>
```

```
<hesa_subject_category>repositories_hesa_subject_category.hesa</hesa_subject_category>
```

```
</hesa_subject_categories>
```

```
:- <subjects_dewey>
```

```
<subject_dewey>repositories_subject_dewey.classification</subject_dewey>
```

```
</subjects_dewey>
<access>repositories_access.access</access>
<delivery_software>repositories_deliverysoftware.software</delivery_software>
- <access_controls>
  <access_control>repositories_accesscontrol.access_control</access_control>
</access_controls>
- <access_methods>
  <access_method>repositories_accessmethod.access_method</access_method>
</access_methods>
<metadata_used>repositories_collection.metadata_used</metadata_used>
<packaging_standards>repositories_collection.packaging_standards</packaging_standards>
- <collection_use_rights>
  <use_rights>repositories_collectionuserights.collection_use_rights</use_rights>
</collection_use_rights>

<collection_use_rights_url>repositories_collection.collection_use_rights_url</collection_use_rights_url>
- <ingest_policies>
  <ingest_policy>repositories_ingestpolicy.ingest_policy</ingest_policy>
</ingest_policies>
  <ingest_policy_url>repositories_collection.ingest_policy_url</ingest_policy_url>
  <legal_mandate>repositories_collection.legal_mandate_to_preserve</legal_mandate>

<long_term_preservation_policy>repositories_collection.long_term_preservation_policy</long_term_preservation_policy>
  <preservation_policy>repositories_collection.preservation_policy</preservation_policy>
- <admin_metadata>
  <creator_id>auth_user.creator_id</creator_id>
  <date>repositories_collection.date_created</date>

<fulfils_jisc_inclusion_criteria>repositories_collection.fulfils_jisc_inclusion_criteria</fulfils_jisc_inclusion_criteria>

<reason_for_exclusion>repositories_reasonforexclusion.exclusion_reason</reason_for_exclusion>
- <admin_documentation>repositories_collection.exclusion_notes</admin_documentation>
</admin_metadata>
</collection>
</drai>
```

Appendix 5. Coverage of sources

All sources mentioned in Section 5 were catalogued comprehensively with the exception of Intute. The table below shows which sections of Intute were completely aggregated into the database as of 19/10/07. It is important to note the very high level of repetition of resources between categories, especially where hard to specifically define by Intute's subheadings, especially in the Health and Life Sciences section where many of the resources are listed under the MeSH headings and included in several categories. There are also several subheadings in different sections which could cause duplication of links between categories, such as Philosophy of Science appearing in both the Science, Engineering and Technology and the Arts and Humanities category, and Agriculture having subcategories under both Science and Health and Life Sciences. The aggregators estimated that there is, on average, around 10% repetition of links between categories, however, the estimate from the Intute staff is < 5%.²³ Additionally, 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, resulting in an estimated 20% of the nearly 65,000 records being aggregated. Therefore the subject coverage of the results is likely to be skewed towards the Arts, Humanities and Social Sciences.

Intute categories not aggregated into the Digital Repositories and Archive Inventory

Intute Section	Heading	Coverage
Science, Engineering and Technology (33,658 records)	All	Approx 20%, 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%, 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

²³ Phone call with Lisa Charnock, 22/10/07