Arkivum
Every bit archived

Arkivum Limited
R21 Langley Park Way
Chippenham
Wiltshire
SN15 1GE
UK
+44 1249 405060
info@arkivum.com
@Arkivum
arkivum.com

Estimating Research Data Volumes in UK HEI

Part Number	ARK/REPT/ALL/380
Version	1.0A
Date	15 Oct 2015
Status	Draft
Pages	12
Author	Matthew Addis (ORCID: 0000-0002-3837-2526)
DOI	10.6084/m9.figshare.1575831
License	CC-BY

Arkivum Limited

R21 Langley Park Way

Chippenham

Wiltshire

SN15 1GE

UK

Change Control

Date	Version	Author	Description
15 Oct 2015	1.0A	Matthew Addis	First Draft

Table of Contents

Title
Estimating Research Data Volumes in UK HEI

Part No
ARK/REPT/ALL/380

Version
1.0A

Date
15 Oct 2015

Status
Draft

3/12

Unclassified

1 Background

This report provides a very simple and quick approach to making a ‘back of the envelope’ estimate of research data volumes in UK Universities Higher Education Institutions (HEI)

The growing volume of research data in UK HEI is of interest to many parties. However, it remains a significant challenge to survey or calculate how much data is being created in research-intensive organisations such as Universities. Individual institutions may need this information when designing and provisioning Research Data Management (RDM) services, for example centralised storage. Funding bodies, for example Research Councils, may find this information useful given that data retention and access can be an eligible component of grant funding, or because the Research Council supports a national service for data deposit. Service providers, for example Arkivum, find this information useful because it helps understanding of the market and capacity planning for services. Driven by these needs,.

The analysis and report took just over one day to complete, the methodology is crude, the results have significant error bars, but nonetheless we hope that the approach is useful, especially for institutions who have no other means at their disposal for rapidly getting an estimate of how much research data they might be generating.

Title
Estimating Research Data Volumes in UK HEI

Part No
ARK/REPT/ALL/380

Version
1.0A

Date
15 Oct 2015

Status
Draft

2 Summary of findings

- We estimate that within UK Higher Education Institutions (HEI) there is on average approximately 5TB of research data for each researcher. See the summary tab of the spreadsheet 'DataVolumesPerResearcherUKHE' [12] for more details. A large University has 1000+ researchers and hence total research data volumes can be very substantial and are measured in PBs.
- On this basis, our conservative estimate of the total research data volume across English HEI is 450PB for 91,000 researchers in 156 institutions. See the 'Data Volumes' tab of the spreadsheet 'DataVolumesPerUKHEI' [13] for more details.
- It is over 3+ years since some of the underlying surveys were done on which our analysis is based. 'Data inflation' will have taken place since then which means actual data volumes may have doubled or more. There is also a bias towards under-representation of large datasets in our methodology. Therefore, it is entirely possible that there is actually over 1 Exabyte (EB) of research data within UK HEI.
- The vast majority of researchers generate or use relatively small volumes of data (a few hundred GB). It is only a small minority of researchers that generate a disproportionately large amount of the overall data volume. This data volume is often from research in the STEM disciplines, but not exclusively so. The effects of the 'large tail' of data has a significant skewing effect on the 'average data volume per researcher'.
- Research data types and data volumes are very diverse. Small research projects can generate large data volumes, e.g. social histories or archaeology. Likewise large research projects may generate small volumes of data, e.g. computational modelling. The type, volume or quality of research being done within an HEI is not always a good indicator of the data volumes that the HEI might generate. The simplest measure of research data volumes is, in our view, the number of staff doing research at the institution.
- Whilst there is clearly a large volume of research data in UK HEI, this is not to say that all this data has value or should be retained. Only a subset of the data will have real value, for example through reuse in further research projects, supporting the repeatability and verifiability of research outcomes, or in commercial exploitation. What data should be kept, why and for how long is the subject of a different analysis and is not considered in this report.

3 Challenges of estimating research data volumes in UK HEI

Estimating the volumes of research data in the UK University sector involves several factors and many uncertainties. For example:

- There is no national survey, league table or summary statistics for either the data types or the data volumes for research data that is created/held by UK HEI.
- HEI research data holdings are often distributed amongst schools, departments, groups and individuals. Several Universities have done surveys that show their holdings are very fragmented, for example residing on USB drives or local servers within individual research groups. This makes it difficult for HEI to audit let alone publicly report their research data holdings.
- Many Universities do not yet have centralised data storage services for all their researchers or groups. Sometimes those that do offer these services internally can see low levels of adoption. Therefore, the size of storage prevision within Universities is not a good indicator of data volumes.
- The scale of research done by individual Universities varies widely, e.g. between the larger research-intensive Russell Group Universities and the 'long tail' of smaller Universities, which can often be more focussed on teaching than research.
- Different types of research can generate wildly different data volumes. For example, even small-scale social science projects can generate large volumes of primary data such as audio or video recordings of people interviewed for oral histories. Large research projects can also generate relatively small amounts of data in some cases, for example High Performance Computing (HPC) projects doing numerical simulations can result in small output datasets. This means that the size of research projects or their level of research funding is not necessarily a good indicator of data volumes.
- Where Universities have done some form of internal survey on their holdings, the results are not always publicly available or if they are then they only apply to a part of the HEI and not the whole.

The exception to the above is a small number of Universities that have used the Data Asset Framework (DAF) [1] or similar approaches to surveying their researchers and their research data. These surveys contain a wealth of valuable and interesting information available and form the basis of our analysis.

4 Methodology

Published internal surveys from seven Universities (Bath, Exeter, Hertfordshire, Leeds, Lincoln, Nottingham and Sheffield) form the basis of our analysis.

Methodology

1. We used existing DAF surveys or similar to collect details of the research data generated per researcher/project at a range of institutions. This gives a small number of quantitative data points. There is a tab for each institution the DataVolumesPerResearcherUKHEI [12]. We have included the numbers from the published surveys and then added the averages.
2. We calculated the average number of TBs of research data per researcher by averaging across all institutions publishing survey results. This is in the summary tab of the DataVolumesPerResearcherUKHEI spreadsheet.
3. We used the table published by HESA (Higher Education Statistics Agency) [3] on the number of staff classified as Eligible [2] under the Research Excellence Framework (REF) [4] as a measure of the number of researchers in each HEI in England. The HESA data (2013/2014) is in the 'REF Template' tab of the DataVolumesPerUKHEI_V1 spreadsheet [13]
4. We multiply the number of researchers at each institution by the average amount of data per researcher to create an estimate of the total amount of research data at that institution. This is in the 'Data Volumes' tab that we have added to the 290183_REF_Contextual_table_1314 spreadsheet.

The above methodology is not intended to be rigorous. It gives a 'back of the envelope' estimation and should be considered no more than that given the relative paucity of statistics and information in this sector.

Notes and caveats

- The error bars in our analysis are likely to be very large because the lack of standardisation across the surveys and the low response rates in many cases. There are approx. 200,000 full or part time staff doing research within UK HEI [5]. The survey results used in this analysis cover less than 1% of these researchers.
- Each survey puts researcher data volumes into buckets, e.g. 1-50GB, 50-100GB, 100-500GB etc. To calculate the average size we use the mid point of each bucket. We then average across all the buckets weighted by the number of researchers that have data within the each bucket. The non-linear distribution of data set sizes means that this averaging process will only be a very rough approximation of the real data distribution.

- The data volumes are skewed by the small number of researchers who create very large research datasets. This has a significant impact on the average amount of data per researcher for the institution as a whole. Most of the surveys do not characterise this end of the data spectrum. Typically there is a 'catch all' category for large datasets, e.g. a bucket for all data that is >10TB. This means it is hard to know whether there are some research data sets that are significantly larger than this. Therefore, the only thing we could do for these 'open ended' buckets was to assume all data was at the small end of the scale, i.e. at the lower boundary of the bucket. In doing so the analysis will naturally underestimate the true volume of data.
- We used surveys from English HEI and we used the HESA table for REF eligible researchers. HEI don't just exist in England! A similar methodology can be used for the rest of the UK, for example using HESA statistics for staff in UK institutions [5]. However, it should be noted that due to the difference in the way staff are counted under HESA and REF, there would need to be an adjustment to HESA numbers to align with the REF eligibility criteria we have used. See ref [2] for more details.
- Research income for UK Universities is static, if not falling [6]. However, although research income might be flat, data volumes are rising, and are expected to rise. This is due to the falling cost of creating data. For example, the cost of Next Generation Sequencing fell by a factor of 1000 over 5 years [7] and this has driven an explosion in data volumes for NGS in research. This effect is true in many areas. For example, AV recordings such as oral histories in the humanities have followed the trend of audio -> low resolution video -> high definition video. This results in a 10 fold increase in data in under 5 years. This trend is likely to continue as technology advances. Therefore, a CAGR of 40% across many research types (data volumes double every two years) would not be unrealistic. Several Universities have their own estimates that are higher than this. The surveys used in the analysis date back to 2012 and haven't been adjusted for 'data inflation'. This could mean that the conclusions we have come under-estimate current data volumes by at least a factor of 2.
- There are several national services supported by UK Research Councils [11] that provide a place of deposit, safekeeping and on-going access for various types of research data (e.g. the UKDS [8] in social sciences, the ADS [9] in Archaeology, and the BADC [10] for atmospheric data to name but a few). The institutional surveys we used are not always clear about whether the data surveyed is destined for these services, will remain within an institution, or will be a mix of both. We take a conservative approach of not including any estimates of the data volumes in these external services for fear of double counting. Again this results in a bias towards under estimating UK research data volumes in HEI.

5 Example research data survey

DAF surveys collect a wide range of information. Some examples are shown below from the Nottingham survey. Nottingham provides a good case study because it is representative of many of the other surveys, it was completed relatively recently (2014), and it comes from a relatively large University and hence has a good number of responses so has statistics that are pretty much 'as good as it gets'.

The Nottingham survey shows that there are a very wide range of research data types, that datasets are frequently stored outside of centralised facilities e.g. on laptops or portable drives, and that datasets are in the main very small in size. The Nottingham survey also shows very clearly that there are a very small number of datasets (1.3%) that are over >100TB each and account for a significant amount of the total data by volume (19%).

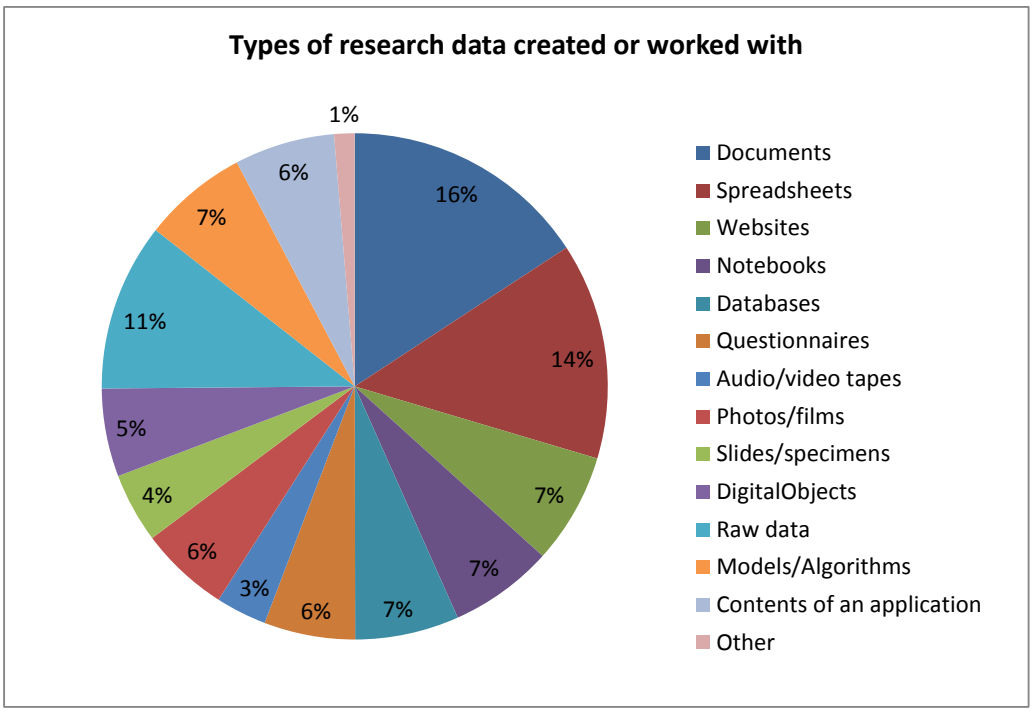

Figure 1 Types of research data. Reproduced from the ADMIRe survey at Nottingham. http://eprints.nottingham.ac.uk/1893/1/ADMIRe_Survey_Results_and_Analysis_2013.pdf

Title
Estimating Research Data Volumes in UK HEI

Part No
ARK/REPT/ALL/380

Version
1.0A

Date
15 Oct 2015

Status
Draft

9/12

Unclassified

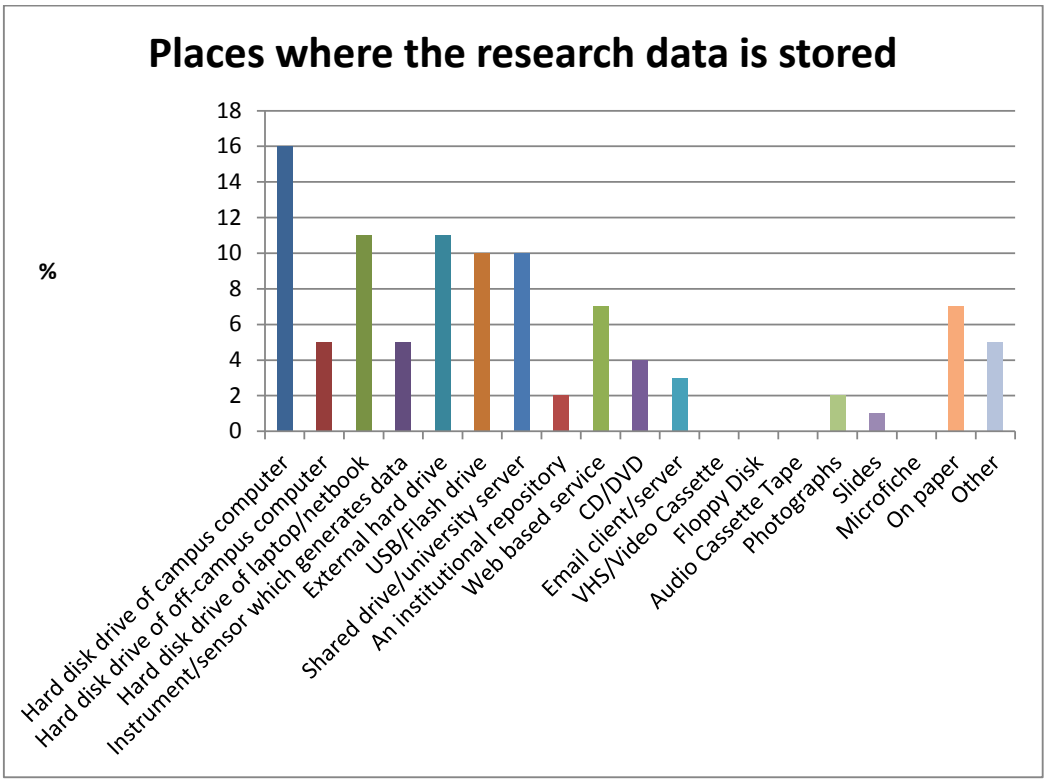

Figure 2 Research data storage locations. Reproduced from the ADMIRe survey at Nottingham.
http://eprints.nottingham.ac.uk/1893/1/ADMIRe_Survey_Results_and_Analysis_2013.pdf

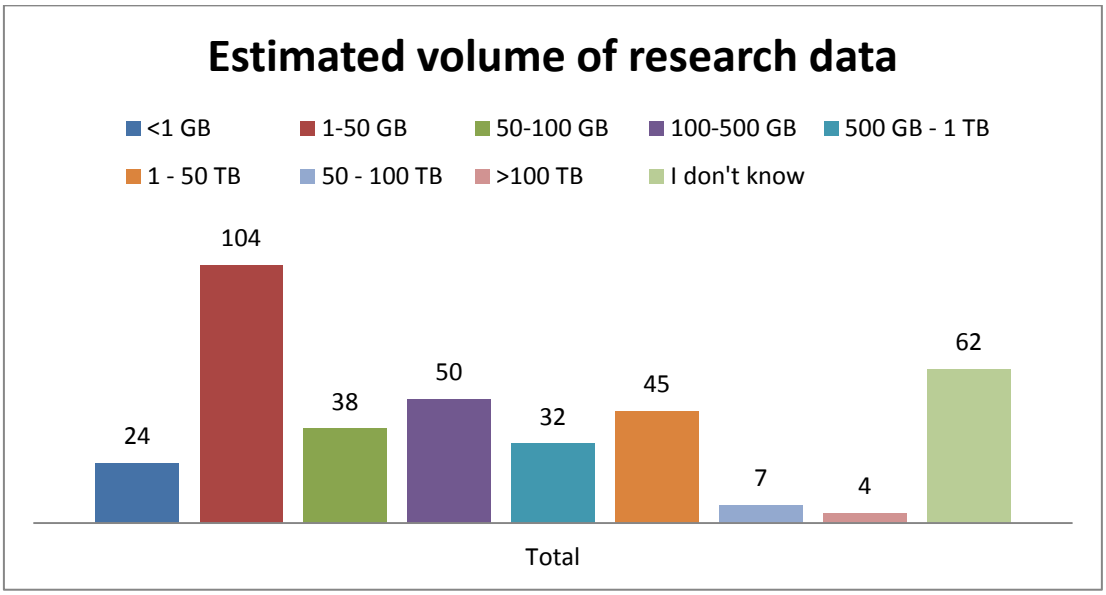

Figure 3 Volumes of research data. Reproduced from the ADMIRe survey at Nottingham.
http://eprints.nottingham.ac.uk/1893/1/ADMIRe_Survey_Results_and_Analysis_2013.pdf

Title
Estimating Research Data Volumes in UK HEI

Part No
ARK/REPT/ALL/380

Version
1.0A

Date
15 Oct 2015

Status
Draft

6 HEI research data surveys

Links to the research data surveys done by twelve UK institutions are provided below. Not all surveys include information on data volumes. Not all surveys adequately characterise data volumes, especially the number and size of large datasets. Therefore, we have excluded five of of the surveys from our analysis. The seven surveys left that we have used are from Bath, Exeter, Hertfordshire, Leeds, Lincoln, Nottingham and Sheffield.

Leeds	https://library.leeds.ac.uk/roadmap-project-outputs http://library.leeds.ac.uk/downloads/file/354/leeds_research_data_survey_results
Exeter	http://blogs.exeter.ac.uk/openexeterrdm/files/2012/04/survey-questions1.pdf https://ore.exeter.ac.uk/repository/handle/10036/3689 https://ore.exeter.ac.uk/repository/bitstream/handle/10036/3689/daf_report_public.pdf?sequence=1&isAllowed=y
Southampton	http://eprints.soton.ac.uk/195959/
Nottingham	http://eprints.nottingham.ac.uk/1893/ http://eprints.nottingham.ac.uk/1893/1/ADMIRe_Survey_Results_and_Analysis_2013.pdf
Newcastle	https://iridiummrd.files.wordpress.com/2012/06/iridium_arma_2012_low_res_v1_lw.pdf
Edinburgh	http://repository.jisc.ac.uk/283/1/edinburghDAFfinalreport_version2.pdf
Oxford	https://blogs.it.ox.ac.uk/damaro/2013/01/03/university-of-oxford-research-data-management-survey-2012-the-results/
Northampton	http://nectar.northampton.ac.uk/2736/1/Alexogiannopoulos20102736.pdf
Hertfordshire	http://research-data-toolkit.herts.ac.uk/document/rdtk-data-asset-survey-digest-july-2012/
Lincoln	http://orbital.blogs.lincoln.ac.uk/2012/04/30/data-assets-framework-survey-summary/
Essex	http://www.data-archive.ac.uk/media/391114/rdessex_staffsurveyreport.pdf
Bath	http://opus.bath.ac.uk/24960/1/DAF_report_May_2011.pdf
Sheffield	http://www.ijdc.net/index.php/ijdc/article/view/10.1.210/393

URLs retrieved 13 Oct 2015

7 References

URLs retrieved 13 Oct 2015

[1] http://www.data-audit.eu/

[2] https://www.hesa.ac.uk/pr/2672-ref2014

[3] https://www.hesa.ac.uk/

[4] http://www.ref.ac.uk/

[5] https://www.hesa.ac.uk/index.php?option=com_pubs&Itemid=&task=show_year&pubId=1717&versionId=27&yearId=313

[6] https://www.hesa.ac.uk/pr/3488-press-release-213

[7] http://www.genome.gov/sequencingcosts/

[8] http://ukdataservice.ac.uk/

[9] http://archaeologydataservice.ac.uk/

[10] http://badc.nerc.ac.uk/home/index.html

[11] http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

[12] 10.6084/m9.figshare.1577539

[13] 10.6084/m9.figshare.1577540

Title
Estimating Research Data Volumes in UK HEI

Part No
ARK/REPT/ALL/380

Version
1.0A

Date
15 Oct 2015

Status
Draft