An investigation on the impact of service policy on energy usage in the Energy Data Centre

Catherine Jones (STFC)

Peter Holt (STFC)

Sarah James (STFC)

2021

Acknowledgements

We would like to thank Sam Pepler, Matt Pritchard and Andrew Harwood from the Centre for Environmental Data Analysis (CEDA) for their assistance on determining the share of JAMSIN usage for the EDC. Chris Dean, from UKRI/STFC Digital Infrastructure, supported us in measuring the usage of the machines we own and he operates. Jenny Mitcham and Paul Wheatley from the Digital Preservation Coalition (https://www.dpconline.org/) gave their time to sense check our plans and David Underdown, Hannah Merwood and Alex Green from the National Archives (https://www.nationalarchives.gov.uk/) for discussing DiAGRAM in more detail with us. Alan Ruddell from Energy Research Unit for sense checking our calculations.

This work was funded by an STFC Environmental Sustainability Concept Fund grant.

Contents

Figures

1. Introduction

The Energy Research Unit within the Technology Department provides the Energy Data Centre (EDC). The EDC is funded by the intra-research council Energy programme and is part of the wider UK Energy Research Centre (https://ukerc.ac.uk).

The UK Energy Research Centre aims to provide "*Independent whole systems research for a sustainable energy future*". UKERC is an independent research centre, with researchers based in 20 different institutions throughout the UK. The research addresses the challenges and opportunities presented by the transition to a net-zero energy system.

The EDC holds research data and information on publications and projects related to academic energy research. Its remit is to be the long-term preservation repository for this data. There is growing interest in quantifying the environmental impact of services which hold digital objects for the long-term and the EDC undertook a short project funded by the internal STFC Environmental Sustainability Concept Fund to assess our own environmental impact and establish changes to working practices to minimise this.

The policies and procedures for the service set the environment in which it operates, determining what is collected, how it is accessed, quality assurance procedures and how the service itself is maintained. This project aimed to review the policies in place and establish how the aspirations for the service may impact on the energy consumed.

One of the challenges of this project is that the relatively small size made seeing any distinct changes in the energy consumption difficult, but hopefully the approach will be of wider use.

It should be noted that the term "Data Centre" is used in this report to denote a specific domain focussed collection/service rather than a physical facility providing managed computing.

1.1 Aims and scope

The Energy Data Centre service is an in-house application which uses the PostgreSQL database system to hold the metadata and a web-based application to provide the discovery service. The in-house service run on physical machines owned by the EDC and research data is stored in the Centre for Environmental Data Analytics Archive but accessed through the EDC application.

In this short project we established the energy consumption of the equipment we are responsible for and identified variations from the baseline aligned to known routine tasks and ran some new preservation and validation tasks.

We:

- reviewed and updated our policies, procedures and preservation aspirations to see what impact this might have on power consumption
- estimated the impact of varying our policies on the load on the computing kit

- discussed how we can establish processes for identifying shares of larger common components.

The underpinning raw data and analysis and data management plan can be shared with others on request.

The following assumptions/scope restrictions were in place:

- The focus is on the current infrastructure and the policies & processes we are responsible for.
- Out of scope considerations which are not specific to the EDC:
    - While manufacture of computing equipment is acknowledged to be a significant part of the environmental impact of its lifespan.
    - The environmental impact of disposal of computing equipment.
    - The environmental impact of the staff who work in the EDC, so no consideration of office spaces, local computing or travel to work.
- While we will attempt to estimate the impact of shared resources, we will start with the research data stored in the CEDA Archive rather than other activities such as networking, central monitoring or the cooling of the machine rooms.
- While we develop the EDC system, we are not necessarily directly measuring the software development impact.

2. Background

There is much research, changes in practices and publications on the subject of energy efficiency within both data centres (the building) and computing based services. This section identifies some key topics.

2.1 Digital preservation

There has been growing interest in the environmental impact of digital archives. While all archives, physical or digital, use energy to help preserve their contents; digital items need energy to keep them viable. Policy changes such as open access to data and interest in reproducibility has led to an increase in storage of digital objects which are to be held for the long-term. For academic research, bodies such as UKRI have policies on research data [1] which expect publically funded data to be accessible for others, to be as *open as possible as closed as necessary.*

The article “Towards Environmentally Sustainable Digital Preservation” [2] caused the EDC to examine our sustainability. The authors discuss approaches to reduce the impact through technology by addressing efficiency measures such as energy saving settings, scheduling of jobs for off-peak electricity usage and finally through clean energy usage. They then discuss responses at an archive level by considering the appraisal process (what is collected and what file formats are used); permanence & acceptable loss (checksums, duplicate copies) and finally what is acceptable regarding availability (instant vs longer retrieval times). This theoretical piece has been followed up by a blog series [3] discussing some practical activities undertaken by the University of Houston Libraries. These archival level approaches are directed by the policy of the service.

There have been other activities discussed in this community, the 2009 Jisc funded Greening Information Management project [4] investigated the impact of changing information management practices on the process of greening ICT in higher education. It developed a framework (sadly no longer publically available) which had three stages: baselining (understanding the environment, looking for rationalisations), selecting options (such a de-deduplication, weeding, different storage technology etc) and finally assessment with a focus on new working practices.

The focus of “How to Improve the Sustainability of Digital Libraries and Information Services?” [5] Is on reducing the energy impact of a digital library by addressing the end users equipment which is used to interact with the online services, however it recognises that the service being used also has an impact.

Other UK based institutions thinking about this include the University of the Arts, London [6] which is considering environmental aspects of providing special digital collections. Matthew Addis from Arkivum [7] discusses in his DPC blog “Is digital preservation bad for the environment? Reflections on environmentally sustainable digital preservation in the cloud” some of the data centre improvements done by cloud services and gives some practical advice relating to the areas discussed in [2]

So the current practice in this area is investigating balancing the archival/repository good practice with the consequences to environmental impact of those policy decisions.

2.2 Data Centres

Minimising energy usage for data centres identify that keeping data live on spinning disk is an area where there might be possibility for reductions in energy consumption. The fact that articles are now written in the general computing press [8] shows that this is a mainstream topic.

In a 2016 report from Lawrence Berkeley National Laboratory [9] energy consumption for US data centres is considered and some scenarios for energy reduction included the consolidation of services onto fewer, high utilised machines as well as using hyperscale (i.e. very big) data centres. It estimated that the greatest improvements would come from bigger data centres but improvements in management and technology would also have a positive impact.

While there are existing methodologies for measuring energy use for IT hardware as discussed by Krumay & Brandtweiner [10] they suggest that those easiest to collect and which have an impact on financial return, such as energy consumption, were adopted by the experts they consulted. Demonstrating the fact that the ability to easily measure is an important factor is what is actually adopted in practice. Another article by Williams [11] discusses the environmental effects of information and communications technologies from the lifecycle assessments described in [9] to the wider whole systems changes in user behaviour.

Computer and chip manufacturers [12][13][14] are also interested in reducing the environmental impact of their activities and have environmental policies which aim to use renewable energy in manufacturing, reduce the energy usage of the computing kit produced and reduce waste disposal & water usage. Dell track the carbon footprint of the computers they build, and provide these publically on their website. IBM aims to use 90% renewable energy by 2030, from 75% in 2025.

It is recognised that large data centres, such as those supported by commercial cloud providers, can give economies of scale and the three major Cloud providers, Amazon [15], Microsoft Azure [16] and Google [17] all have environmental sustainability policies which have goals to use 100% renewable energy, reduce waste disposal & water usage and to run the data centres in the most efficient way. Additionally Microsoft is aiming to be carbon negative by 2030. Google has a dashboard to allow you to choose where your cloud machines are hosted by environmental impact.

It is clear that establishing and reducing the environmental impact is important for companies' reputations regardless of whether they manufacture computing equipment or provide computing services.

3. Policies and procedures

The Energy Data Centre aims to collect, disseminate, and preserve information for the whole system energy research community. This comprises of digital objects with metadata and metadata only records. The research data digital object storage and preservation has been outsourced by deposit into the CEDA Archive and the remaining digital objects are text-based PDFs which have not been traditionally treated as a preservation object.

Digital objects need active management to ensure the long-term integrity of the objects and associated usability. This includes processed to look for corrupt files, file formats going obsolete, retaining the metadata to know what the files you hold are, and recording what you may have done to the files to keep them usable (file format migration for example).

Service policies determine the content collected and the procedures undertaken to ensure the policy is enacted. There is a balance between activities to minimise the risk to the collection, such as number of copies, against their energy impact. This will be a different decision depending on the collection remit.

In 2020 and 2021 the EDC policies behind the existing procedures were reviewed and are summarised in the following sections. In 2021 two digital preservation focussed assessments were undertaken to establish a baseline for activities and risk appetite to continuously improve our digital preservation practice

3.1 EDC policy

There are two policies which determine how the system operates: the collection management and the preservation policy. Taking each of these in turn:

Collection management policy outlines the types of material collected by the EDC, any specific characteristics, such as uniqueness or additional value added by the EDC and retention policies. Currently all material is retained indefinitely. We have identified the parts of the collection where we are the primary repository and so need to ensure that we do not lose or corrupt this information.

We have not done an automated file format audit on our collection, but we know most of the content held locally is in common formats such as PDF, Excel or image files. Our research data collection is more varied.

A review of the metadata cataloguing policy has also identified that a routine semi-automated URL checking for external links process would enhance data quality.

The current policy therefore implicitly accepts that the content will grow year on year, which has implications for both the storage needed for the content and any routine activities.

The preservation policy sets out the preservation aims for the collection. The policy covers roles & responsibilities, ingest processes, infrastructure, and preservation actions.

As a result of formalising this policy and doing the assessments discussed in 3.1; there are some additional collection management activities that from a preservation

point of view should be done for our PDF collection. This is considered to be a low-risk collection as it is all in the same format and materials are available elsewhere, or we hold the original Word document that the PDF was created from in our internal STFC provided file storage.

Some of these would translate into new routine activities such as

- Checksum generation and checking for the PDF collection
- File format identification of the whole collection to identify potential risks related to formats.

We would also seek to do file format identification checks to ensure we understand the variety of files kept within the service to establish any rare or unusual format and establish a regular technology watch process to alert us to any changes in this area.

The current policy is suggesting that more routine activities need to be applied to the PDF collection, therefore this has an implication for energy use to run quality assurances processes on the content held.

For the research data, the checksum generation and checking is performed for the EDC by CEDA.

3.2 DPC Rapid Assessment Model

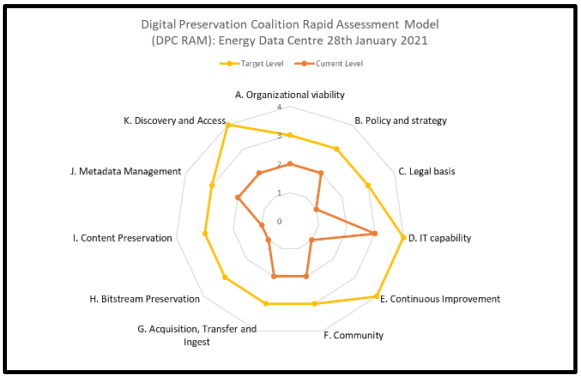
This tool from the Digital Preservation Coalition enables organisations to do an assessment of their digital management policy/practice and gives an opportunity to set the aspirations for improvements and to identify the activities to get there. The tool is available from https://www.dpconline.org/digipres/dpc-ram .

The diagram below shows the current assessed levels and aspirations. While the assessment has levels, it is not assumed that all organisations will want to attain level 4 in all areas and as can be shown from figure 1 the EDC has only identified 2 areas: IT capability and Continuous Improvement where we aspire to the highest level.

The areas which reflect the strengths of the service: consistent metadata, well managed infrastructure, resourcing and ingest procedures are operating at levels 2/3 and reflect the parts of the service which are more aligned to a dissemination service and the consistent staffing of the service to date.

The areas where improvements could be made are in specialist preservation areas. So, while we do not have concerns about losing objects we could improve on what information we hold to be able to understand and reuse them and what we record about preservation activities.

Figure 1 Visualisation of EDC's DPC Rapid Assessment Model



0. minimal awareness
1. awareness
2. basic
3. managed
4. optimised.

So, the impact of improving digital management policies, as discussed above, will have an impact on the number and frequency of routine activities run, for example more regular checks might be made for file corruption in the content held in the EDC.

3.3 DIAGRAM

This is a tool provided by The National Archives (https://nationalarchives.shinyapps.io/DiAGRAM/) designed to help archives assess the risk to the contents and gives an opportunity to visualise different scenarios. It differentiates between being able to open & use content (renderability) and understanding what you have, what rights you have on it and what you may have done to the content (intellectual control). This enables modelling of the impact of different policies and procedures on the risks to your own collection.

Figure 2 shows the differences in risks, from changing both the policy/internal management and the physical policies and procedures. It shows different risks to the PDF collection, which is more homogenous. So, as we adapt our procedures and enhance the information we hold about our material, the risk decreases. These changes are linked to the DPC RAM aspirations discussed in the previous section. There will always be risks to the material and our risk appetite will be different for material where we hold the prime copy against material where it is also held elsewhere.

Figure 2 Outputs from the DIAGRAM tool

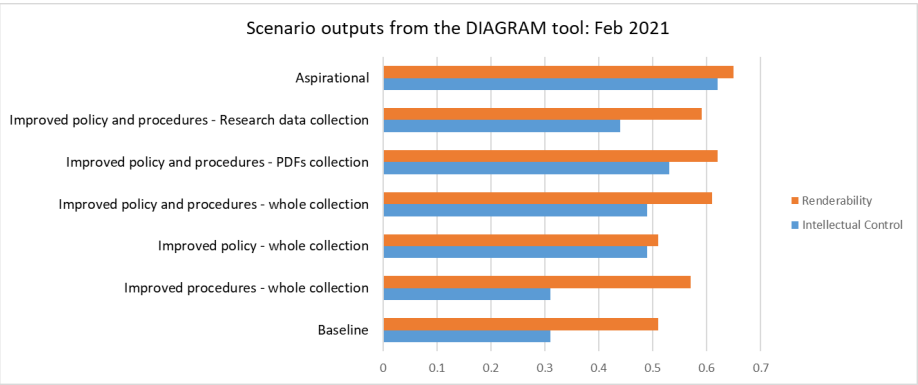
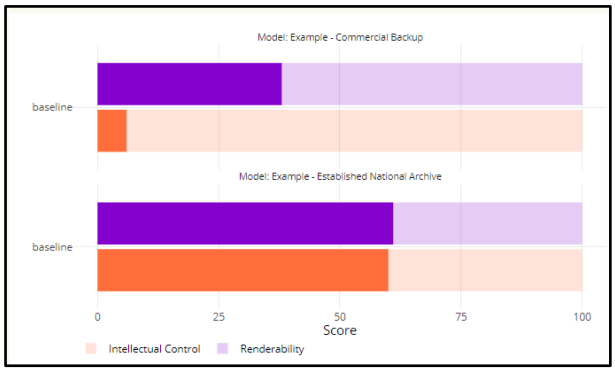


DiAGRAM uses a Bayesian network statistical method to generate the output where 1 means that for every 100 files you would be able to render them all and know all you need to know. This is not necessarily possible for all files in an

archival/preservation environment and figure 3 shows two reference models produced by the National Archives to help users put their own results in perspective. In the first example the content is just backed up on the Cloud, the second is a well-established archive with a large, varied collection which follows best practice. From these comparisons the EDC currently has some more risks that a well-established archive, but if we improved our policies & practices, including some of the activities discussed later in the report, our risk profile would reduce.

Figure 3 TNA DiAGRAM Reference models (April 2021)



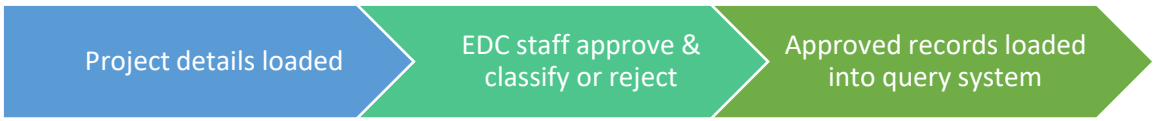
3.4 Review of procedures

In this section we consider our automated routine procedures. These can be split into two types of activity: (a) Ingest of content and (b) Content integrity and maintenance.

3.4.1 Ingest procedures

There are three types of content which are growing: metadata only records for projects; research data and our text-based PDF collection. Only the former is completely automated; for records with associated digital objects there are also manual loading processes, and the metadata is manually created. The automated loading process is outlined in the figure below:

Figure 4: Project data loading process



The first step in figure 4 is automated and scheduled to run once a week. The second step is a manual process where EDC expertise is used to approve and classify or reject projects. This is not done on a scheduled basis so that the approved grants load process (step 3), which moves approved grants into the live system runs every day regardless of whether there are grants to be processed.

Running a process regularly regardless of what processing is required is more straightforward than either kicking off a process when data is ready (more complicated) or running the routine process less frequently (data refresh rate slower). By investigating the energy consumption for these processes, we were able to recommend the most appropriate approach to this.

3.4.2 Content integrity and maintenance procedures

There are variety of jobs which are designed to ensure the data held about are valid.

i. Calculating changes to project values assigned to regions after changes for loads/approvals/manual amendments (weekly)
ii. Landscape text prepares text for searching (daily)
iii. Landscape sections reload (daily) updates the display fields
iv. Calculating changes to annual or financial year spend for projects after changes for loads/approvals/manual amendments (weekly)

These activities are responding to changes in the underlying data, and so although only run weekly, may not be needed if no new project information is added. See the discussion in section 3.4.1

Landscapes are a specific type of content held within the EDC.

3.5 Outcomes

As a result of reviewing our policies we have been more explicit about the purposes of the components of our collection and what activities we are doing or should improve to mitigate the risks of collection corruption or loss. We have reviewed the purpose of the routine activities and have identified some areas where to make the service easier to maintain, there may be processes that are consuming energy where other approaches may not.

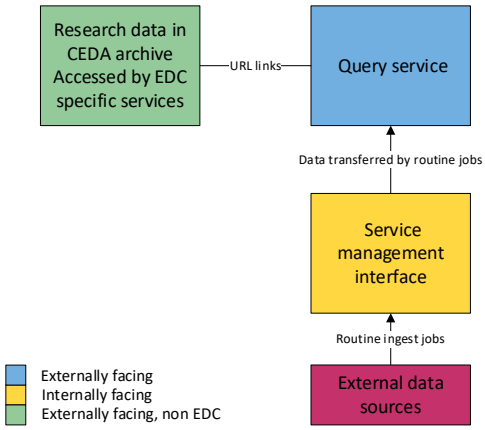
In summary:

i. Need to be clear about the purpose of the repository to ensure the right policy and associated processes are in place to ensure purpose can be achieved.
ii. Using external tools can make you reaffirm purpose and identify areas where practice can be improved.
iii. Long running services may accrete low level routine processes that aren't serving an active purpose: good to review these regularly.
iv. Need to ensure frequency of routine jobs reflect the frequency of the manual interventions.
v. Balance between risk, purpose and energy efficiency needs to be struck.

4. Infrastructure and energy consumption monitoring

The Energy Data Centre uses both EDC owned computing kit and a share of the CEDA Archive. All equipment is located in machine rooms, either RAL or DL. The EDC infrastructure was supported by Digital Infrastructure and is part of the standard system monitoring processes. The diagram below outlines the architecture.

Figure 5 Architecture of the Energy Data Centre



The EDC query service is set up to provide resilience through the use of a hot back-up machine at a different physical STFC site to the main production machine. These are kept in step with both the software and data. This choice reduces the risk of service outage and time to restore but increases the energy consumption profile as there is an idle duplicate running at all times. All the routine jobs run on the service management interface machine, whereas the query service machines has the public facing web application.

All full text documents are kept on the EDC server's storage and the research data is hosted by the CEDA Archive, and so eventually on JASMIN.

In the discussion there are some areas of the wider infrastructure which are considered out of scope:

- Any contributions from software development within CEDA to support EDC.
- Any contributions to data upload on CEDA

The approach has been to buy physical machines and run them in the production service while under warranty and then replace like for like. All the production service equipment is therefore changed at the same time. In part this is due to the funding source being long-term project based and partially due to the availability of alternatives and risk appetite in the past.

The current servers were purchased in 2019 and put into service in August 2020. As they were purchased with a long lifetime in mind, then currently the power of the machines is greater than the load observed.

All parts of the EDC are run in a well-managed environment.

4.1 Power consumption methodology

Before this project, power consumption was not routinely measured and we were guided by our technical support as to the possible methods of power monitoring.

We have tried two approaches to capturing the power data. Firstly by using Nagios (https://www.nagios.com/) and the Dell OpenManage Nagios plug-in (https://www.dell.com/support/kbdoc/en-uk/000178053/support-for-dell-emc-openmanage-plug-in-for-nagios-core). As the EDC machines are already monitored by the DI Nagios service to identify server issues, this seemed a good way to extend the information gathered. Using this we recorded the value in Watts for the W2_System_Board_Pwr_Consumption variable every 10 minutes along with a timestamp in a separate file. This is the method used to ascertain the basic energy consumption for the main servers. However, long-term we felt it would be better not to link our environmental measurements to a service designed for service monitoring.

The second approach was to use a Dell implementation of the Intelligent Platform Management Interface industry standard [18]. This uses a baseboard management controller to gather information from a variety of sensors on the motherboard. We then used this to retrieve power consumption of routine jobs during their running time at a variety of intervals from 1 second to 10 seconds interval.

Both these methods produced readings at specific Watt levels not a range as we might have expected. In routine circumstances these were 132W or 154W. Following further investigations, the Tango project's blog [19] explained that while the measurements done using IPMI are comprehensive, the figure reported is likely to be in specific steps as we had found due to the way it is generated and reported.. As for these measurements, there is quite a big jump between steps (22W). Work done by Kavanagh [20] considering the accurately of IPMI readings for energy model calibration, notes that the sensors used by IPMI are not as accurate as directly connected Watt meters but under-report power consumption and overall energy consumed. The main factors for inaccuracy are the latency of the arrival of the measurement from the sensor, the fact that it averages over a measurement window and the poll rate needs to be greater that the time taken for sensors to report values. In this paper to calibrate the data they adjust for the delays in averaging readings, which may be over 60 seconds, by removing the data from the first 60 seconds and including data for 60 seconds after the job has finished. This information on the IPMI averaging process means that while we have recorded data for very short jobs, we are not presenting it in this report as we believe the data to be too inaccurate to draw valid conclusions.

A value in Watts for the current power consumption is recorded by these methods, we have taken this value as the energy consumption of the machines that run the service.

So while we have produced graphs and will discuss the findings of this analysis, they are for indicative purposes only and the precise values are both tiny and inaccurate but they do demonstrate changes in energy consumption. The graphs show the energy for the duration of the job and we discuss what the difference is to adopting Kavanagh's recommendations. Going forward we will continue to refine our measurement processes based on the IPMI approach.

4.1 EDC owned baseline monitoring

The production query service machine and the production service management interface machines were monitored over an eight-week period in 2021. The average weekly energy consumption was 23.1kWh (Query) and 22.2kWh (service management). One of those weeks was over the Christmas period where we suspend all routine jobs and we have taken this weekly average as the baseline for the machine in steady state. The query interface is at a higher value, in part due to a period of exceptional load, but we have left this in the figures as unexpected events are part of running an external service. Using these measurements, we estimate that over a year the Query service consumes 1200kWh and Service management interface 1154 kWh.

We also monitored over the same period a piece of kit which had been decommissioned from the production service in 2020. This machine is older than the other machines, with a higher consumption according to the Dell published carbon footprint [21][22]. However during this time it wasn't running the application or PostgreSQL but was using an estimated 1104 kWh/year. If we installed the application, then it would give us an opportunity to establish what the energy consumption of the application itself is.

Comparing these energy consumption figures to the published figures for the EDC computing kit [21][22] which estimates energy demand measured using the standard Yearly TEC as 1480kWh, for the specification used by the production service machines, and 1760kWh for the decommissioned machine. Our measurements are slightly lower that the published Dell figures, but in the same range and may be due to IPMI under reporting energy consumption as discussed earlier.

4.2 Monitoring routine jobs

We monitored a selection of regularly running jobs to see if there were any observable changes in the overall energy consumption. The length of these jobs varied from 30 seconds to 26 minutes. For all but one of the jobs an increase in energy usage was observed during the time the job ran, subject to the caveats already discussed. Jobs taking less than one minute are not discussed in this report as the data is not reliable enough.

The schedule of routine jobs takes just over 2 hours (02:03:55) each week equating to four and half days per year. This means that the routine jobs currently take up 1.2% of the available time. While the scope of making appreciable differences to the energy consumption is limited, all changes may make a difference.

As many of the changes in energy consumption are very small, and the length of the job varies, the graphs in this section show the energy consumption of the job as a percentage of the energy needed for the service to run (using the Christmas baseline) over the known duration of the job. If a particular reading is 100% it means that no additional energy consumption was observed.

Adopting Kavanagh's suggestions for longer running jobs did not make a significant difference to the overall values.

4.2.1 Ingest

Figure 6 Energy consumption of weekly EPSRC project ingest, IPMI method

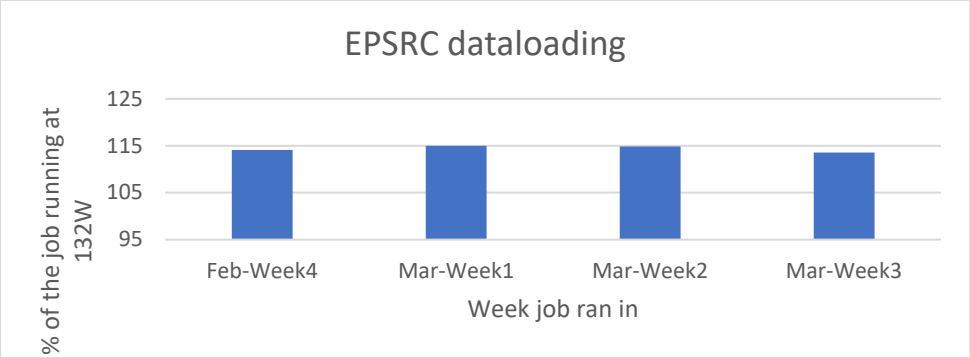
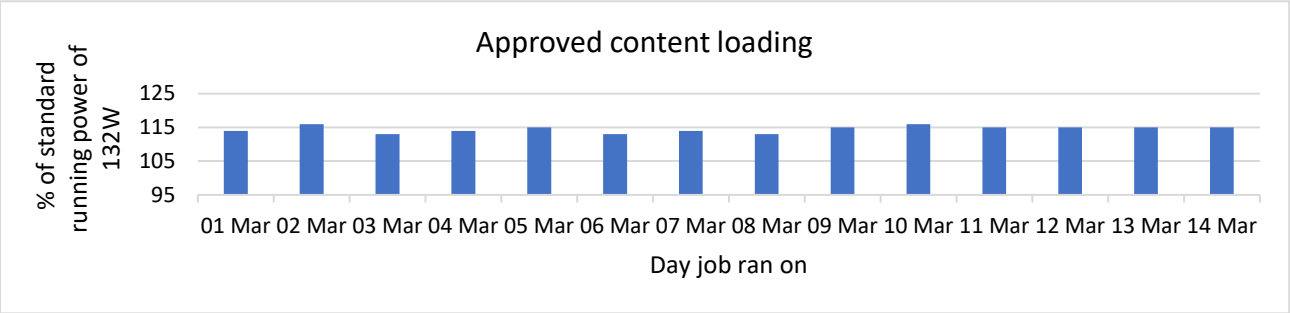


Figure 6 shows the power consumption for EPSRC loading, this processes a set of UKRI provided files and loads them into the database.

The job shows a rise in energy consumption while running which is to be expected as it performs file editing and database selects, inserts and deletes.

Figure 7 shows the energy taken for the processing once the content has been classified and approved as it is transferred from the processing SQL tables to the production tables. This job runs every day, regardless of whether there has been content approved. This is a short job, and so the conclusion to be drawn is that there is additional energy consumed and this appears to be very similar for each reading.

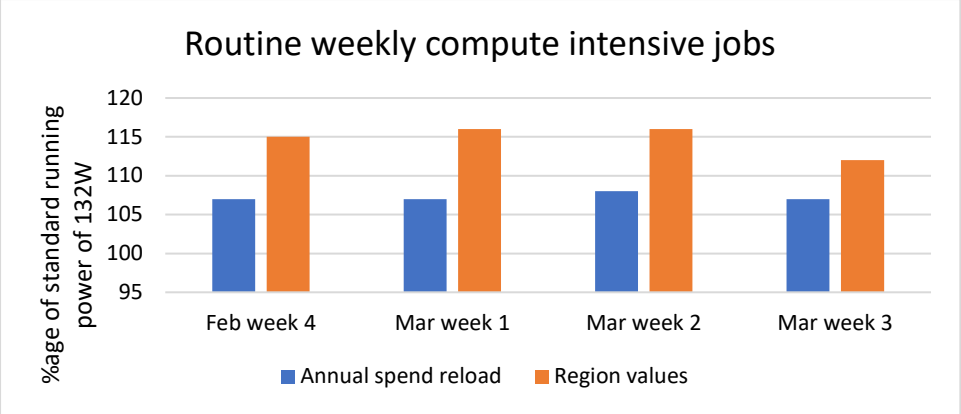
Figure 7 Approved content loading, IPMI method



4.2.2 Content integrity and maintenance

This section considers routine activities that ensure the quality of the content or are pre-computing results to make the interface work faster.

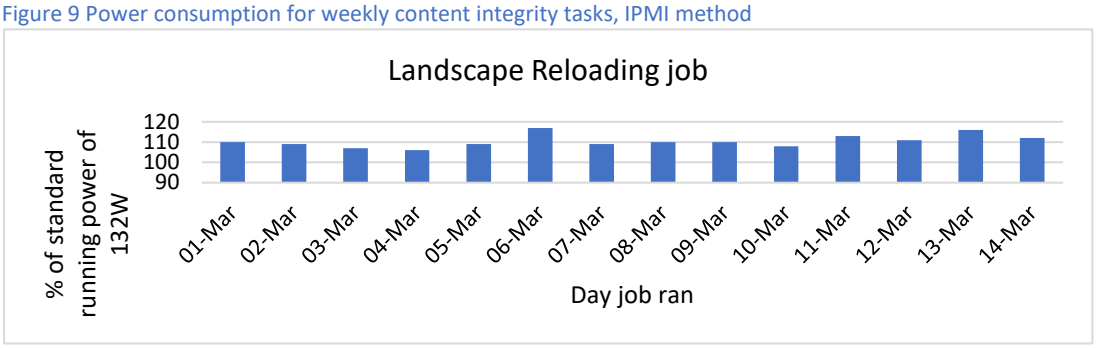
Figure 8 Routine weekly compute-intensive job



The two weekly jobs shown in figure 8, do compute intensive calculations on the value of grants within either calendar or financial years (running time 26 mins). Or value by region (running time 2.5 minutes). It should be noted that although energy consumption has been observed, the values computed will only change if new projects have been added to the catalogue.

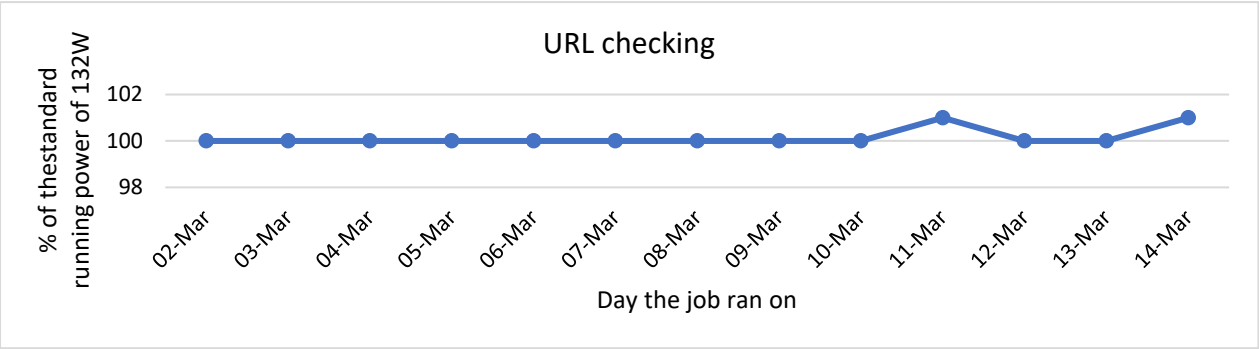
Figure 9 shows a job which update and modify content relating to our Landscapes material type, this reload job takes 11 minutes.

Figure 9 Power consumption for weekly content integrity tasks, IPMI method



As part of the project, we trialled a new job which checked that the URLs held in the metadata catalogue are still valid. This has been set to run daily to gather the measurements. The job looks up the URLs from a table within our PostgreSQL database system, establishes the status of URL and writes the output to a file. It runs for just under 5 minutes. This is not a compute intensive job as is demonstrated from figure 9. This shows that on most occasions, the power consumption is in the steady state while this job runs.

Figure 10 URL checking power consumption, IPMI method



From this experiment, we can improve the quality of the content of the EDC in this area without having a negative environmental impact on energy consumption. We now run this routine job on a different part of the system every month.

4.3 Analysis

It is shown by the analysis of the recorded measurements, that all but one of the routine processes that run routinely do increase the energy consumption in a detectable way; however, this increase is not a significant one over whole year. For ease of maintenance, there are several jobs which are set to run regularly but only have an impact if new content has been added to the system, and therefore changes to this processing approach have been identified.

There are also some jobs that were set up in the past with the expectation that content would change more frequently than it has turned out to do, these we will review and may turn off as part of the routine schedule.

Of the new activities identified in the preservation policy, one has been tested and due to the nature of the activity, it doesn't have a significant impact on the observed energy consumption. We will develop the other activities in the coming year, but we expect to be able to measure an impact. However all these preservation activities would be run on an infrequent basis.

4.4 CEDA infrastructure

The CEDA Archive infrastructure is used to store and preserve the research data held within the Energy Data Centre. This is part of the wider CEDA infrastructure which is hosted on JASMIN.

JASMIN monitors power consumption in four categories: storage; compute, virtual compute and other. The EDC service uses a fraction of storage and virtual compute as dedicated resources for EDC and a fraction of the services in the other categories. The data discussed in this section relates to 2019 as that was the complete year available at the point we agreed this process.

JASMIN is a unique computing resource, combining petabytes of storage and a variety of compute resource and a community cloud.; all with high-performance access to massive data resources. It has been operational since 2012 and in that time the underlying compute and storage devices have been replaced as part of the well managed resource. This can make measuring and comparing energy usage

across years more complicated as the underlying infrastructure may well not be the same. These complexities are not highlighted within the figures provided, but need to be taken into consideration over the longer period.

Areas of energy consumption, or share of energy consumption, which are not considered in this breakdown are: monitoring services, network costs, share of the data deposit function, any tape back-up costs and any energy costs associated with migration of storage technologies. For disks this happens every five years and will balance the energy in moving data against the energy consumption of newer disks.

The areas of EDC use of the CEDA Archive considered were: the virtual machines, VM, (Jasmin compute category) which are used for data browsing and user registration; a proportion of the storage (based on allocation not usage) and a proportion of the energy used for the storage audit which happens every six months.

It should be noted that for the power estimates for storage is based on a proportion of the total storage available on Jasmin for that period, 46PB. However the Data Archive is 10PB, hence in the calculations for audit contribution uses the Data Archive size.

The storage audit is a process run by the Data Archive to check for corrupt content through checking checksums. One of our aims for 2021 is to run a similar process for content held within the EDC and to see is there are any comparisons to be drawn.

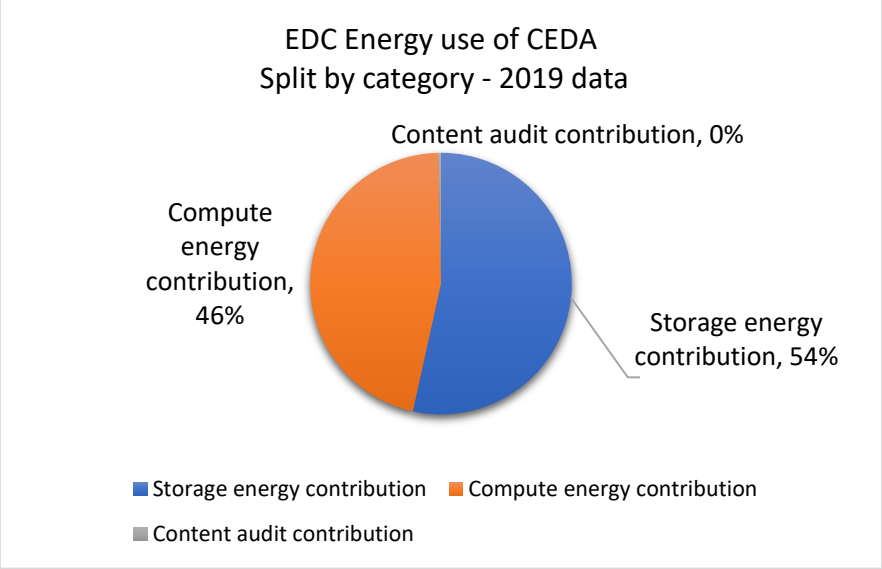


Figure 11 Split of energy use over the different categories in CEDA

From this it can be seen that the energy used for the virtual machines running the CEDA Archive part of the service currently use 46% of the total energy consumption, whereas storing the data takes 54%. However, it is likely that as the data rates increase the proportion of energy use for service delivery would drop as there would be no need to increase the number of virtual machines for registration and serving data.

We have been able to derive an energy consumption figure based on proportions of the total energy measured, this demonstrates the accepted theory that it is more

energy efficient to use a large Cloud or computing infrastructure than running your own as not very much energy was consumed.

It was a surprise to the team that the energy needed for the VMs was nearly as big as the storage. However, the calculation for the storage is more likely to be accurate, as this is a stable commodity. The compute energy consumption will be accurate, but the calculation uses the number of virtual machines which is more difficult to assess as some will have been used for the whole year and some may be more transient in nature. However, if the number used is an underestimate, this means that the total EDC share drops.

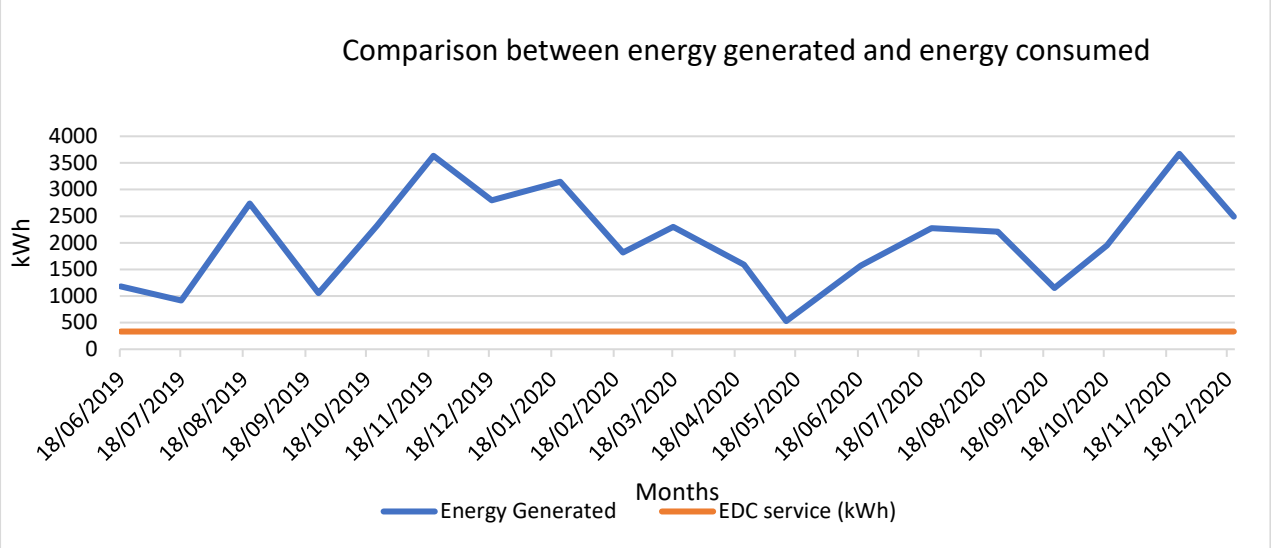
For this particular use case, using proportions of the whole was straightforward as the EDC had both dedicated VMs and storage allocation. We agreed not to consider other activities, such as monitoring or storage migration, but I think some of these might be harder to decide on which metric to use for the allocation algorithm, as use/energy consumption might not be linear with the metrics chosen.

4.5 Power generation

The Energy Research Unit operates a Britwind turbine (http://www.britwind.co.uk/), with a rated power of 12kW, and records the power generated from this. A monthly summary is publicly available from the ERU Meteorological data web site.

The figure below compares the record energy generated from the Britwind turbine against the energy consumed by the Energy Data Centre

Figure 12 Comparison between ERU energy generated and EDC energy consumed



Note that this is for illustrative purposes as the CEDA component figures of the service are for 2019 and the EDC owned figures are for 2021, and for the Britwind figures we are using the publicly accessible data and not the more accurate underpinning data and so is not directly comparable.

This illustration shows that the energy consumed by the Energy Data Centre each month is within the energy generated by the Energy Research Unit over the same

time period. Of course, in reality, the wind turbine doesn't always produce power and the EDC has a requirement for uninterrupted power, but it is shown for comparison purposes and show help put into context the relatively small amount of energy required to run the Energy Data Centre.

4.6 Greenhouse gas emissions conversion

With the data gathered from this exercise, we have used HM Government's 2020 Greenhouse Gas reporting: Conversion factors guidance [23] to estimate the yearly emissions for the service. Only considering the electricity used, then guidance on the *UK Electricity sheet*, provides a conversion factor of 0.23314 which is multiplied by the annual kWh consumed to calculate in kg of CO2e for the year.

The estimated annual kWh figure for all aspects of the EDC service is 3998 kWh. The overall figure calculated from our measurements is 932 kg CO2e per year.

5. Discussion

During our monitoring period, we had an unexpected event as our application didn't respond well to a routine STFC process. Whilst this means that there is a spike in the energy consumption, it also reminds us that unexpected events, from benign ones through to malicious ones such as Denial of Service attacks, will have an impact on the service's energy consumption so that important, but at first glance unrelated activities such as ensuring your service acts on any IT security vulnerabilities also has a contribution to make to your environmental footprint.

The timing of routine jobs can make an impact on energy costs/type of energy. So that running them outside of standard working hours may well mean that the electricity costs are reduced.

Some of the key decisions are now discussed.

5.1 Preservation content: energy consumption vs risk to content

The EDC aims to hold its content for the long-term and to provide access to it to our user community. This is the whole purpose of the EDC and so ensuring that content is accessible and uncorrupted is very important.

The research data is held within the CEDA Data Archive and thus is well-managed and in a large repository with the benefits of being part of a "hyperscale" data centre in environmental terms.

The other digital objects held, which are mostly PDFs, have not in the past been treated from a preservation perspective, although some good digital preservation practice is in place. The DiAGRAM tool demonstrates that if we amend our policies and procedures to increase preservation activities, we are reducing our risk of this content becoming inaccessible.

Some of these modifications are about changing what information we keep in the repository, but there are three activities that we would seek to introduce on a routine schedule: quarterly URL checks; six-monthly checksum validity checks and more infrequent file format profiling. We have established through this project that the URL checks process doesn't increase the energy consumption of the service. We expect that the checksum and file format profiling are likely to have a discernible impact as these are both more computationally intensive.

So, our approach to this is to increase our preservation activities to minimise risk to the collection, appreciating that some activities may be compute intensive and hence increase the energy consumption. We are hopeful in the short term that this increase will be balanced out by the changes in scheduling of some current routine jobs. In this case, minimising risk to the collection outweighs the energy consumption.

5.2 Preservation content: energy consumption vs content reputation

In 2017 we agreed with the Energy Technologies Institute (https://www.eti.co.uk/) to take their publicly accessible content as the Institute was closing. Once their website ceases operation in 2025, we will be the sole holders for the publicly accessible content and the importance of preserving this content has increased the reputational risk if the content was not available going forward.

So, for this part of the collection, it is important that we undertake effective content integrity checks even though most of the content is in standard file formats such as PDF as not following good digital preservation practice might lead to an event which damages our reputation as a reliable service and would then reduce the trust in the community for our long-term preservation remit.

5.3 Infrastructure risk appetite: energy consumption vs service reliability reputation

Our current EDC owned infrastructure has two characteristics: firstly that we are using physical machines and secondly that there is reliability built in by the use of hot back-ups.

It has been shown that using VM or cloud machines, that the overall energy consumption could be reduced, the CEDA VMs use far less energy that the recordings gathered from the EDC measurements. We note that we have not compared specifications here.

Buying physical machines can be considered to be a capital purchase, whereas using a cloud or a share of a VM cluster may be considered to be service, hence having a different financial impact, which is important to consider when the service is supported by grant income.

By having a hot back-up machine, it is more straight-forward and faster, to recover from a disaster as one can change the DNS so that the URL points to the hot back-up than having to re-install/reimage and retrieve the data from back-up, thus reduces the potential for reputational damage. It is also possible to take this approach as there are only a small number of computers in the infrastructure and there is the ability to afford to duplicate the main service.

There are plans, which are not fully realised, to put in place a hot back-up for the internal machine and data management interface. This project has established that the old machine UKERINT7 is less energy efficient from the information provided by Dell and that the CO2e impact is greater than the risk of the production internal machine being unavailable. This risk has not materialised in the last year, and while as the kit ages it may increase overall the risk to the external service by down-time for the internal service would be that new data would not be visible as quickly. So, as a result of this project, we will not be commissioning UKERINT7 and will consider our approach to service issues.

Over the longer-term, we will review our risk appetite to our infrastructure and consider whether physical machines are still the best approach, and if not what would provide a reliable and resilience service most energy efficiently.

5.4 Infrastructure kit specifications: energy consumption vs responsiveness

As discussed in section 4.1, the current physical kit is purchased with an expectation it will run the service for five years and so it has a high specification compared to the load anticipated at the point of purchase. This means that the service is very responsive at implementation stage and slowly degrades as the content increases and the machines age. While this report has considered the operations of the EDC, from a disposal environmental impact and whole life perspective, using machines for longer is better than replacing them quickly.

There are a variety of questions to be considered here:

How effective is that estimate of future load? Is energy being consumed because the specification is too high? Would using VM/Cloud machines were it possible to adjust the specification afterwards be more effective? These all address how the specification for the replacement kit is arrived at, whether it is driven for what can be bought for the budget allocated, or whether the computational requirements drive the purchase.

Would buying less high specified machines, which would be cheaper, and budgeting for changes sooner and hence buying more efficient kit be better? Would this lead to other environmental impacts? On the whole buying computing equipment more frequently is not a positive environmental approach as disposal is costly, and there is staff effort required for migration. However not having all the kit at the same age would reduce other risks associated with mass migrations.

Would in fact not buying our own kit and looking to use centrally provided VMs or Cloud machines provide a better environmental impact? This option is known to be more effective use of resources, but the VM cluster/Cloud needs to be responsive enough so that the shared aspect of the resources doesn't impact on the delivery of the service.

At this stage of the funding cycle, this is not top priority, but will form part of our resource planning.

5.5 Storage medium: energy consumption vs user experience

All our content is currently held on storage media which is instantly accessible. Using this type of media incurs a higher energy consumption that for other storage media such as tape. However, tape media does not provide for instant access to the content.

While the EDC uses infrastructure specialists for our equipment, we could choose to change our requirements for the infrastructure to include material to be held on tape. There are a variety of scenarios from everything on tape to older or less well used

digital content is moved from instant to delayed access, thus reducing the electricity consumption.

Policy decisions on this, are aligned those in library services, where less used stock may be put into storage. It can be a challenge to identify an algorithm which is effective to satisfy most user requirements without significant outliers.

Whether the impact on users would be to not use this content, or to be happy to wait, probably depends on the reason for using the resource, the retrieval time and how the service communicates delays to the user.

This is not currently on our roadmap for energy efficiency improvements, but we will keep a watching brief on storage developments in the services we use.

5.6 Application development: energy consumption vs resource required for adaptions

There is an increase in discussion in the wider world about how to make web applications more energy efficient, such as Greenwood [24] or Bergman [25]. One of the drivers of this is the development of mobile applications where the end user is much more aware of the energy use of the device as they are responsible for recharging it! There is also a growing movement around being a *Green Software Engineer* proposed by Hussain [26] which espouses eight principles to designing greener software and states that everyone has a contribution to make, and all contributions make an impact.

As with many software projects, it can be difficult to get the resources for existing working applications to make best practice adjustments. However, the EDC team are at the start of a new technical review and refresh project and we will aim to incorporate as much of the emerging best practice in this area.

While we have investigated the energy impact for specific routine scheduled jobs, we have not looked into other aspects of our application, and this will be on our list of activities for the coming year.

6. Questions for other service providers/developers to consider

By the very nature of this project, the results, discussions and conclusions are related to the specific circumstances of the EDC. This section suggests some areas and questions for colleagues elsewhere to consider.

- Purpose
  - Is the purpose of the system clear with a clear policy on what is added to it?
  - Are you clear about the balance of importance between environmental impact and facets of your service?
  - What are the service development aspirations? Have you considered the potential environmental impact?
- Policy
  - Is this implicit or explicit?
  - Does it cover environmental impact of decisions?
- Consider and document any risk factors which affect policy, such as:
  - Reputation
  - Uniqueness
  - Resilience & any operating requirements
- Resourcing
  - Do you have the skills and/or time to undertake a baselining activity?
- Hardware
  - What is your policy on where the application is hosted?
  - Are there any organisational/service policies which impact on these decisions?
- IT system/application
  - Are you able to measure the impact of the application?
  - If it is in-house have you investigated green software development?
  - If it is commercial/ open source can you discuss their approach to energy efficiency and measurements?
- System maintenance activities
  - When did you last review the routine system maintenance activities?
  - Has there been a major policy or IT change since the last review?
- Content
  - How fast is the content growing?
  - Can you differentiate between different parts of the collection?
  - Are there any parts of the collection which need higher levels of protection?
  - What are the requirements for accessing the content?
  - If you process the content to other file types/states of analysis, what is your policy about storing originals or intermediate stages?
  - What methods do you have in place for content validation? Are these still appropriate?

7. Conclusions and next steps

This exercise gave the EDC team an opportunity to review the policies of the service and identify which decisions we had made in the past have a greater impact on our environmental sustainability. As a result of this project, we changed our system maintenance jobs and save a small amount of energy.

When we started, we were not sure if we would be able to technically identify energy consumption from routine activities and we have demonstrated that this is possible, although as it is a relatively small-scale service, and with the constraints of the measuring tools, the outputs are imprecise and very small. However, we have learnt from this experience and future exercises will benefit from this. It has also demonstrated that the service does not consume a large amount of energy to operate.

As we embark on a technical review, this gives us an opportunity to modify the system to ensure that there is more of a link to doing compute intensive activities when we know there is new data to process and to continue to do these outside the working day. We are also going to concentrate on the energy efficiency of the user facing application, both through understanding the database schema & queries and through considering energy efficiency as part of the redesign.

We have identified a list of future activities:

7.1 Short-term and/or low effort

- Investigate the environmental impact of checksums checking for our text-based content. This will improve content validation but is likely to be compute intensive.
- Investigate what tools that PostgreSQL provides to analyse our SQL queries and processes, so that we can make the application more efficient.
- Review our risk appetite regarding the infrastructure.
- Establish what the energy consumption impact of running PostgreSQL (our database) and Apache (our webserver) is. The current baseline is for a fully operational system. Whilst the EDC won't work without these components, it would give a more rounded view of the system.

7.2 Longer term and/or significant effort

- Build the concept of an energy consumption dashboard into the Admin function of our system.
- Investigate and implement guidance available on building energy efficient web applications
- Consider whether it would be possible to experiment with virtual machines to see what the minimum specification to run the service without performance limitations would be and what the environmental difference would be.
- Consider the EDC approach to equipment procurement and disposal.
- Widen our environmental assessment to consider other parts of the landscape considered not in scope for this project, such as other underpinning

infrastructure, the impact of software development and the team's working environment.

As a service designed to support the Energy research community and the UKERC community in particular, our environmental footprint is a key factor in decisions going forward and this project has helped to set a framework for this.

8. References

[1] UKRI, 2021. Open Research [online] Available from:https://www.ukri.org/about-us/policies-standards-and-data/good-research-resource-hub/open-research/ (Accessed 14 April 2021).

[2] Pendergrass, KL., Sampson, W. , Walsh, T., and Alagna, L.. (2019), Toward Environmentally Sustainable Digital Preservation. The American Archivist 82 (1): 165–206 Available at http://nrs.harvard.edu/urn-3:HUL.InstRepos:40741399 (accessed 14/04/2021)

[3] Scott, B., 2021. Estimating energy use for digital preservation. Part I. [Blog] *BloggERS!* The blog of SAA's Electronic Records Section., Available at: https://saaers.wordpress.com/2020/10/06/estimating-energy-use-for-digital-preservation-part-i/ (Accessed 14 April 2021).

[4] McDonald, D, MacDonald, A & McCulloch, E (2009) Greening information management: final report. University of Strathclyde. Available from https://pure.strath.ac.uk/ws/portalfiles/portal/5855352/GIMFinalReport.pdf (accessed 14/04/2021)

[5] Chowdhury, G.G. (2016), How to improve the sustainability of digital libraries and information Services?. J Assn Inf Sci Tec, 67: 2379-2391. https://doi.org/10.1002/asi.23599

[6] Thurlow, E., 2020. It's not easy being green: Evaluating the impact of digital preservation. [Blog] *Digital Preservation Coalition blog*, Available at: https://www.dpconline.org/blog/wdpd/blog-elisabeth-thurlow-wdpd (Accessed 14 April 2021).

[7] Addis, M., 2020. Is digital preservation bad for the environment? Reflections on environmentally sustainable digital preservation in the cloud. [Blog] *Digital Preservation Coalition blog*, Available at: https://www.dpconline.org/blog/is-digital-preservation-bad-for-the-environment (Accessed 14 April 2021).

[8] Doige, F., 2021. Datacentre energy efficiency: Is the time now for a big switch-off?. *Computer Weekly*, [online] Available at: https://www .computerweekly.com/feature/Datacentre-energy-efficiency-Is-the-time-now-for-a-big-switch-off (Accessed 14 April 2021).

[9] Shehabi, A., Smith, S.J., Horner, N., Azevedo, I., Brown, R., Koomey, J., Masanet, E., Sartor, D., Herrlin, M., Lintner, W. 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775. Available at https://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf (Accessed 14 April 2021).

[10] Krumay, B and Brandtweiner, R (2016) Measuring The Environmental Impact Of ICT Hardware. International Journal of Sustainable Development and Planning, 11 (6). pp. 1064-1076. Available at: http://epub.wu.ac.at/5403/ (Accessed 14 April 2021).

[11] Williams, E (2011). Environmental effects of information and communications technologies. Nature 479, pp 354–358. https://www.nature.com/articles/nature10682

[12] Intel, 2020. *2019-2020 Corporate Responsibility at Intel*. [online] p.9. Available at: http://csrreportbuilder.intel.com/pdfbuilder/pdfs/CSR-2019-20-Full-Report.pdf#page=9 (Accessed 14 April 2021)

[13] Dell, 2020. *Global Environment Policy.* [online]. Available at https://www.dell.com/learn/uk/en/ukcorp1/corporate~corp-comm~en/documents~dell-global-environmental-policy.pdf (Accessed 14 April 2021)

[14] IBM.com. 2021. *IBM and the Environment - Environmental affairs policy*. [online] Available at: <https://www.ibm.com/ibm/environment/policy/> [Accessed 14 April 2021].

[15] Amazon, 2021. *Sustainability in the Cloud.* [online]. Available at https://sustainability.aboutamazon.com/environment/the-cloud?energyType=true / [Accessed 14 April 2021].

[16] Microsoft, 2021. *Corporate Social Responsibility.* [online]. Available at https://www.microsoft.com/en-us/corporate-responsibility/sustainability?activetab=pivot_1%3aprimaryr3 [Accessed 14 April 2021].

[17] GoogleCloud, 2021. *Cloud Sustainability.* [online]. Available at https://cloud.google.com/sustainability [Accessed 14 April 2021].

[18] Hargrave, J. An Introduction to the Intelligent Platform Management Interface, Dell (2004). Available from : https://www.dell.com/downloads/global/power/ps2q04-019.pdf). (Accessed 14 April 2021).

[19] Kavanagh, R, 2017. Quality of Monitoring Data with IPMI and RAPL. [Blog] *Tango blog*, Available at: http://www.tango-project.eu/content/quality-monitoring-data-ipmi-and-rapl (Accessed 14 April 2021).

[20] Kavanagh, RE orcid.org/0000-0002-9357-2459, Djemame, K orcid.org/0000-0001-5811-5263 and Armstrong, D (2017) Accuracy of Energy Model Calibration with IPMI. In: Cloud Computing (CLOUD), 2016 IEEE 9th International Conference on. 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), 27 Jun - 02 Jul 2016, San Francisco, United States. IEEE . ISBN 978-1-5090-2619-7. Available at: http://eprints.whiterose.ac.uk/102046/ (Accessed 14 April 2021).

[21] Dell, 2021 Carbon Footprint for the R440 [online] available at https://i.dell.com/sites/csdocuments/CorpComm_Docs/en/carbon-footprint-poweredge-r440.pdf (Accessed 16 April 2021)

[22] Dell, 2021 Carbon Footprint for the R430 [online] available at https://i.dell.com/sites/csdocuments/CorpComm_Docs/en/carbon-footprint-poweredge-r430.pdf (Accessed 16 April 2021)

[23] Department for Business, Energy & Industrial Strategy, 2020. *Greenhouse gas reporting: conversion factors 2020*. HMG. Available at https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2020 (Accessed 14 April 2021)

[24] Greenwood, T, 2019. 17 ways to make your website more energy efficient. [Blog] *WHOLEGRAINdigital blog*, Available at: https://www.wholegraindigital.com/blog/website-energy-efficiency/ (Accessed 14 April 2021).

[25] Bergman, S., 2020. How to measure the power consumption of your frontend application [Blog].. Microsoft Sustainable Software. Available at https://devblogs.microsoft.com/sustainable-software/how-to-measure-the-power-consumption-of-your-frontend-application/ (Accessed 16 April 2021)

[26] Hussain, A., 2021. Principles of Green Software Engineering • Principles of Green Software Engineering. [online] 🌱 Principles.Green. Available at: https://principles.green/ ([Accessed 16 April 2021).