Title	**Final Report and Recommendations (D11.1b)**
Work Package	WP11
Authors	Professor R. Sinnott (National e-Science Centre, Glasgow, UK)
Dissemination Level	PU (Public)

Summary/abstract

Part B of the tendered report entitled "Possibilities and Implications of Grid-enabling Social Science and Humanities Data Collections in the Context of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure".

Choices, Resources and Sustainability Issues Relating to a Future Council of European Social Science Data Archives (CESSDA) e-Infrastructure

Report Drafted as part of the CESSDA Preparatory Phase Project (CESSDA PPP)

Professor Richard O. Sinnott
National e-Science Centre
University of Glasgow
Kelvin Building
Glasgow, G12 8QQ
United Kingdom

Tel: 0141 330 8606
Fax: 0141 330 8625
Email: r.sinnott@nesc.gla.ac.uk

Document Control

Details

Document Title	Resource and Sustainability Issues relating to a Future Council of European Social Science Data Archives (CESSDA) e-Infrastructure
Funded By	University of Essex, UK Data Archive
Circulation	CESSDA PPP
Author	Prof. Richard O. Sinnott
Version	1.0
Date	30th May 2009
Authorised By	Hilary Beedham, Ken Miller, Kevin Schurer

Document History

Version	**Editor**	**Date**	**Description**
1.0	Richard Sinnott	30th May 2009	Version 1

Comments and feedback from CESSDA PPP members are welcomed on this draft including any areas where more information or clarification might be useful.

This work has been funded by the UK Data Archive, University of Essex.

Contents

Executive Summary

The existing Council of European Social Science Data Archives (CESSDA) is planning a major upgrade of its existing research infrastructure in order to ensure that European social science and humanities researchers have access to, and gain support for, data resources they require to conduct research of the highest quality, irrespective of the location of either researcher or data within the European research area. In addressing these concerns, the planned upgrade will develop CESSDA from the current situation in which the member organisations work with limited national resources, to create a common platform, sharing a common mission, with a stronger form of integration in which expertise is genuinely pooled, shared and applied in a co-ordinated pan-European experience. This will facilitate the delivery of a fully-integrated data archive infrastructure for social science and humanities researchers, allowing seamless access to as many data holdings across Europe as possible.

One way that such an infrastructure can be developed is through the Grid paradigm and associated middleware. A previous report [CR1] drafted for the CESSDA gave an overview and analysis of the possibilities and implications of Grid-enabling social science and humanities data collections in the context of the existing CESSDA research infrastructure to support a future *e-Infrastructure* (also referred to as cyber-Infrastructure) supporting seamless, secure access to distributed data collections. This report builds directly upon the previous report and focuses in particular upon the specific technological choices, and associated resource and sustainability issues relating to a future e-Infrastructure. In particular the report addresses the following key question: what are the effects of choosing a Grid or Cloud-based solution in terms of cost, effort, maintainability and future developments, when compared to traditional solutions like client server based infrastructures and/or non-Grid or Cloud-based service oriented architectures (SOA).

In addressing this question, the report discusses the pros and cons of three key choices for the implementation of a new CESSDA e-Infrastructure:

- CESSDA implements its own infrastructure based on existing client-server or SOA models;
- CESSDA joins an already ongoing initiative and shares resources with other disciplines using Grid technologies;
- CESSDA exploits cloud computing services.

These pros and cons are considered along with opportunities and threats that might be associated with a particular choice. In particular we consider the consequences of these choices in terms of initial and annual investments to CESSDA; the granularity of the potential charging and accounting models of these choices to CESSDA; the long term sustainability of these technologies and the impact on future funding and revenue streams associated with these choices to CESSDA.

The rest of this report is structured as follows. Section 1 begins with an overview of the European Grid context and summarises the efforts to establish a European-wide Grid-based e-Infrastructure and what this might mean in the context of the CESSDA. Section 2 explores cloud computing and identifies what this might mean to CESSDA and the ramifications and potential pitfalls of adopting a cloud based infrastructure for CESSDA. Section 3 focuses upon the advantages and disadvantages of CESSDA developing and supporting its own in-house non-Grid/cloud-based service-oriented architecture infrastructure. Section 4 looks at specific questions related to the impact and ramifications of the various possible approaches open to CESSDA on its future roadmap, both with regard to establishing and maintaining an e-Infrastructure for CESSDA and longer term impacts of these choices to CESSDA. Finally section 4 summarises the report and identifies various recommendations that should be followed for CESSDA to consider in its evolution.

This report is based upon both the experiences gained in development of numerous Grid-based e-Infrastructures at the National e-Science Centre (NeSC – www.nesc.ac.uk) at the University of Glasgow to support a multitude of researchers in different research domains exploiting a wide range of Grid middleware. NeSC Glasgow does not promote its own middleware and has no vested interest in any particular Grid, cloud or other middleware initiatives.

1 European Grid Initiatives and CESSDA

There are a range of national and international e-Infrastructure initiatives that could or should inform any strategic directions that the future CESSDA RI adopts. As identified in the first report [CR1] many of these efforts are not strictly aligned with the requirements of CESSDA RI since they primarily focus upon supporting communities requiring access to and use of high performance computing (HPC) facilities. However it is the case that the vision of the Grid and e-Research in general has been to provide research environments where access to and use of a variety of resources is made seamless – both HPC resources and data resources as typified by the CESSDA RI as well as other more specialised resources. With the future CESSDA RI supporting inter-disciplinary research communities including those not currently supported by or interacting with CESSDA, this model of seamless access to distributed resources more generally is essential.

In terms of major Grid infrastructures that exist today and proposals for the future there are numerous national initiatives. In the UK, the UK e-Science National Grid Service (NGS – www.ngs.ac.uk) has recently been awarded continuation funding by the Engineering and Physical Sciences Research Council (EPSRC – www.epsrc.ac.uk) and Joint Information Systems Committee (JISC – www.jisc.ac.uk); the German Federal Ministry of Education and Research has funded the second phase of development of the D-Grid (www.d-grid.de); the Baltic states (Belarus, Estonia, Latvia, Poland, Lithuania) have established joint e-Infrastructure efforts with Sweden and CERN in Switzerland through the BalticGrid (http://www.balticgrid.org/) initiative; the Scandinavian countries of Denmark, Finland, Norway and Sweden have established the NorduGrid (www.nordugrid.org) e-Infrastructure initiative and similar national Grid initiatives have taken place in the Netherlands (www.dutchgrid.nl), and Ireland (www.grid.ie). Internationally, the US TeraGrid (www.teragrid.org) and the Japanese Naregi Grid (www.naregi.org) represent two of the larger international efforts, with projects such as Grid Asia (www.gridatasia.net) exploring European, China and South Korea inter-Grid research efforts. In Europe, the predominant European-wide Grid efforts have largely focused around supporting communities such as the particle physicists through the Enabling Grids for e-Science (EGEE) project (www.eu-egee.org/) which is now nearing the end of its funding stream. EGEE-III is due to run to the end of 2009 and discussions are now taking place on the future European roadmap for e-Infrastructures.

To this end, the European Grid Initiative (EGI) Design Study [EGI] was started in September 2007 and funded under the European Commission's 7th Framework. The EGI Design Study represents an effort to link up the individual European nation's National Grid Initiatives (NGIs) into a sustainable pan-European distributed grid infrastructure which, it is proposed, will be application domain neutral. The goal of this design study is multi-fold. It aims to evaluate use cases for the applicability of a coordinated EGI effort; to identify processes and mechanisms for establishing an EGI; to define the structure of a corresponding coordinating body for the EGI, and ultimately to initiate the construction of the EGI organisation itself. In early 2009 it was decided that the EGI will be coordinated through Amsterdam, Netherlands. The EGI Design Study is itself currently supported by over 30 NGIs whose representatives sit on the EGI-Design Study policy board.

It is envisioned that the EGI organisation will commence operations in January 2010, so a European Grid infrastructure is in place before the end of EGEE-III. In June 2009, it is expected that the EGI policy board will sign an interim Memorandum of Understanding (MoU) framework prior to setting up the EGI organisation, to allow commencement of the EGI Council operations. This MoU will be associated with collection of funds to form a small central team and employ a Director of EGI. In the UK at least, both JISC and the Science and Technologies Facilities Council (STFC – www.stfc.ac.uk) have both indicated that they are prepared to be the legal entity which will sign the interim MoU and pay the first year's membership fee (€76,000). However it is still unclear who will pay this annual fee in the longer term. It is likely that the NGI for the UK will be the NGS. However, it is important to note that there is no absolute agreement that the EGI will happen, and if it does so that all NGIs will be involved and if so, to what extent.

Having said this, the potential relationship between CESSDA and the EGI is important to clarify. It is highly unlikely that any given organisation in the CESSDA, e.g. the UK Data Archives, would expect to formally engage in the EGI directly. Rather EGI is expected to be a federation of NGIs. This is a key point as it has both potential benefits and potential drawbacks that we outline in the following sections. In many ways the proposed EGI model is similar to the existing CESSDA collaborative model in its organisation structure with each member organisation/country responsible for their own national resources and providing federated access to remote international resources.

In terms of resource requirements, the EGI Design Study document currently foresees that the EGI will require significant resources to support the general operation of the infrastructure (17 FTEs); to deal with middleware interfaces and certification concerns (8 FTEs); to address application support and training (11 FTEs); to support external functions (4 FTEs) and management and administration (11 FTEs). The EGI Design Study document identifies that these 51 FTEs represent only a small fraction, equivalent to a few percent, of the total effort spent on Grid infrastructure in Europe today. Furthermore, the document also identifies that to run an NGI as part of EGI, it is estimated that between 2.5 and 30 FTEs are necessary to cover the basic regional and international tasks. The precise requirement depends on the size of the NGI, demands of the local user communities and on the commitment to take up international tasks. It argues that in countries with an operating Grid infrastructure, most of these resources already exist and are resourced accordingly.

In surveys and workshops that have been organised in the UK by EPSRC, researchers from multiple application domains and a variety of disciplines have said they would be keen to use the EGI if it would: enable collaboration; provide a common infrastructure; provide easy access to large scale computer resource power across European facilities; was cheap to use and had an adequate support system. However they also stated that they would be deterred from using the facility if: it was difficult to use or was unreliable; local resources were sufficient for their needs; security was not adequate; the financial and personal time costs were too high; it wasn't connected to local infrastructure and resources; did not allow interoperability with other grids; the infrastructure did not appear to be sustainable or if it became too bureaucratic. Many of these considerations map directly upon the demands of the CESSDA RI and the social science research community more generally.

In more detail, there are many pros and cons of establishing an EGI which are likely to directly impact upon the future direction of the CESSDA RI – should the CESSDA RI decide that it wishes to be part of an EGI. To this end a Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis of EGI was undertaken by EPSRC in the UK. This identified the following strengths, weaknesses, opportunities and threats. In each case we identify issues and points that are directly relevant to CESSDA.

We emphasise that the EGI is not yet confirmed. It represents a proposal that has been put forward and is currently being reviewed by NGIs and the funding bodies that fund such NGIs. It might well be the case that the EGI itself does not actually happen, or at least not in the form in which it is currently described. Nevertheless it represents arguably the most enhanced proposal for a future European-wide framework for a Grid infrastructure. As such, it is relevant to CESSDA and the advantages/disadvantages considered accordingly.

1.1 EGI Strengths and CESSDA

- *EU Support* – EGI Design Study is currently supported by over 30 NGIs whose representatives sit on the EGI Design Study Policy Board. EGI Design Study Policy Board will become the EGI Council when the EGI is created.
    - *There is thus a major European-wide support network that CESSDA RI might benefit from. However this support network is likely to be primarily to support HPC researchers. Nevertheless the proposed "application neutral"l emphasis of the EGI might enable wider data-oriented Grid support to be leveraged for domains like the social sciences.*

- *Status* - Europe will continue to host the largest multi-science grid.
    - *To a certain extent this is potentially less important for CESSDA and more a question of kudos on the international HPC supercomputing stage. However that said the scale could have importance if the wider e-Infrastructure was used for larger scale, wider inter-disciplinary collaborations, e.g. building large scale indexes of inter-*

disciplinary national and international data sets including the social sciences that subsequently allows rapid searching.

- *Coordination* - The EGI will ensure pan-European grid coordination by linking up the existing NGIs and actively supporting the set-up and initiation of new NGIs, and will aim at standardisation wherever reasonable.
    - *This has a direct impact upon the CESSDA and represents a potentially timely opportunity, especially with the future focus of CESSDA RI to engage with countries that are not yet part of the CESSDA as described in CESSDA PPP WP6 and WP7*

- *Efficiency* **-** The use of the EGI avoids each project having to create an international e-infrastructure with all other countries involved in each international project. If this n+n negotiation and e-infrastructure establishment were undertaken individually for each international project or community it would be considerably less efficient (and therefore more expensive at a national or European level) than using the EGI to support the interoperability and provide accounting information on the use of the e-infrastructure by each project.
    - *This is one of the primary goals of the CESSDA RI aims and objectives. Through the EGI it may thus be possible to establish this critical mass for CESSDA and leverage multi-national engagement in one fell swoop.*

- *Applicability* **-** The EGI hopes to provide a generic infrastructure which is application domain neutral. It intends to overcome the currently perceived barriers to outsiders to join the EGEE infrastructure as users or resource providers, and reduce the current overheads required by EGEE to join the e-infrastructure.
    - *At face value, this application neutrality is a highly desirable feature for the CESSDA RI since it does not preclude the social sciences. However some key aspects are worth noting on the current EGI Design Study document. Firstly, the study document states that the EGI will "not simply be a continuance of EGEE-III" and as a result be an infrastructure primarily for the Large Hadron Collider (LHC) particle physics community, i.e. an EGEE-IV. That said however, the technical description of work proposed specifically states that it proposes to adopt the EGEE-III infrastructure and technology as the starting point (Page 29, section 5.1). Furthermore, little mention is made of the impact that the LHC going live will have on the EGI. Up to now the EGEE infrastructure has been largely focused upon simulations of representative data. When the LHC is live, the data that is generated may put great demands upon both the EGEE infrastructure and the support personnel. This factor may well be key to the resource and impact on future choices of CESSDA when considered EGI involvement.*

1.2 *EGI Weaknesses and CESSDA*

- *Cost* **-** There are large cost implications to joining the EGI. Each country will have to a pay an annual membership fee (at least €76,000 for countries like the UK but decreasing for other countries down to 1,400€ for Moldova and Macedonia), and the estimated annual cost of an NGI is expected to be between €2 and €4M. It is noted that the cost of middleware (which is essential for the success of the EGI) is outside the EGI funding model and will require additional financial support by key stakeholders in EGI.
    - *The impact of cost to the CESSDA RI may be a moot point if it is the case that a NGI exists and is prepared to pay these costs. In the UK at least it is not clear who will pay these on-going yearly costs. If it happens, it is highly likely that the EPSRC and JISC would be the primary contributors however it might be the case that a research council-wide contribution might be sought. Thus if it was agreed that the CESSDA RI would be involved in the EGI, then the Economic and Social Sciences Research Council (ESRC – www.esrc.ac.uk) may be asked to contribute to either the yearly membership costs or the running and support costs. It is noted that it is highly likely*

that the more HPC oriented research disciplines and associated research councils would be expected to shoulder more of the costs. At the time of writing no cross council initiatives or discussions have taken place[1].

- *Sustainability* - None of the NGIs hold long-term funding commitments. In the UK at least, JISC and STFC are prepared to sign an interim MoU, but discussions over responsibility for long-term funding are yet to be held.
    - *This is a key factor for CESSDA RI as identified in the initial requirements of the tender document. The benefit of EGI in this respect is that long term sustainability issues will potentially not rest solely on CESSDA but can be shared across national initiatives. Potentially is underlined here since the sustainability of the EGI might not in itself support all aspects of sustainability that CESSDA requires, e.g. longer term data management, but instead focus upon sustainability of HPC-oriented e-Infrastructures ala EGEE.*

- *EU integration* - The EGI has not yet made arrangements with respect to providing a common Grid infrastructure aligned with similar on-going major efforts such as the Partnership for Advanced Computing in Europe (PRACE - www.prace-project.eu/), or the Distributed European Infrastructure for Supercomputing Applications (DEISA – www.deisa.eu). Although it is noted that the requirement for at least common authentication, authorization, accounting and data sharing exists and will be addressed later in 2009.
    - *In terms of CESSDA, such wider more HPC-oriented EU integration may not be a showstopper. However, having a common authentication, authorisation and data sharing infrastructure will address many of the key requirements of the future CESSDA RI as outlined in [CR1], thus leveraging these efforts would potentially be highly beneficial.*

1.3 EGI Opportunities and CESSDA

- *Research* - This is the major driving force behind the EGI. Researchers of one project should be enabled to work seamlessly together within a country and also across countries; this should encourage cross-disciplinary cooperation, the sharing of resources and data, and enable major research breakthroughs.
    - *This is at the heart of the future CESSDA RI. Leveraging EGI efforts from this respect would be highly beneficial for CESSDA since it would allow CESSDA to be involved in truly inter-disciplinary, international infrastructure and be directly involved in research involving social sciences and the wider non-social science research communities.*

- *Innovation* - The transfer of expertise to areas beyond science, e.g. e-Health, e-Government, e-Learning, and the use of e-Infrastructures as cost-efficient platforms for large-scale technological experimentation, e.g. the future Internet, are different dimensions that EGI could allow to be explored.
    - *For CESSDA, the EGI infrastructure itself could provide a platform for engaging in a wide range of social science related research. Thus exploring how infrastructures such as EGI change the dynamic of collaboration and the socio-economic aspects of data sharing. This is an indirect benefit however that might be of interest to some social scientists, but is not in itself the primary focus of the CESSDA RI efforts.*

- *Influence* - If a given country is a founding partner of the EGI, then they will have a seat on the EGI Council and therefore have more influence over the organisation and related services.

[1] I am on the EPSRC Strategic Advisory Team on HPC and e-Infrastructure and the EGI and its impact across the UK research communities has been discussed at meetings. EPSRC are still evaluating whether EGI is a good idea for the UK engineering and physical sciences. I note that EPSRC is distinct from the Science and Technology Facilities Council (STFC – www.stfc.ac.uk) who fund research such as particle physics and are heavily involved in EGEE work for example.

- *Similarly, if the social sciences are helping to drive forward the requirements of the NGI and EGI, then they would expect to have more influence in the strategic direction of the infrastructure and research that is supported. It is noted in the UK at least that the social sciences are actively driving forward many wider NGI requirements. This is exemplified both through major ESRC funding streams related to e-Social science, e.g. the National Centre for e-Social Science (NCeSS – www.ncess.ac.uk) and through major recent JISC funding streams in the area of social simulation such as the National e-Infrastructure for Social Simulation (NeISS – www.neiss.org.uk).*

1.4 EGI Threats and CESSDA

- *Sustainability* – Signing an interim MoU will raise expectations on long-term commitment. Disengagement could potentially reflect poorly on a given NGI. The issue of sustainability affects all of the current NGIs, and if for example a given NGI were to disengage, the costs of their contribution would have to be met by other partners and the EGI could thus be a weaker entity.
    - *This is an issue but less so if CESSDA were part of a wider NGI like the NGS in the UK for example. It should be noted that full sustainability of the EGI operations is expected to eventually be achieved using national funding only, helped by the expectation that effort to operate the EGI infrastructure can be gradually decreased thanks to streamlining and automation. However the EGI Design Study document also identifies that in the highly dynamic environment of distributed computing, EC funding for innovation has to continue — most logically on a project basis. It also identifies that for EGI sustainability, NGI co-funding is expected to total 20 M€/year with an equal matching EC contribution to cover the estimated costs of the EGI organisational tasks and international tasks in NGIs.*

- *Research Focus* - CERN is a lead partner in the EGI Design Study and the implementation of a Grid to replace EGEE-III is a huge driving force behind the EGI, although the EGI stakeholders are keen to stress that EGI is not simply a continuation of EGEE or any other infrastructure project. If a given country NGI does not engage at an early stage, there is a concern that their own national requirements will be taken over by the requirements of the particle physics community – should the EGI actually happen of course!
    - *This is perhaps the greatest concern related to the CESSDA RI and its potential relationship with the EGI. Will the requirements of the social sciences shape the EGI or will it be dictated to by larger and more established HPC-oriented communities? This question is currently difficult to answer with any degree of confidence. However looking at the UK NGI represented by the NGS, it is clear that they are committed to working with and supporting other less HPC-oriented domains like the social sciences. Up to now this has primarily been in relatively straightforward ways, e.g. supporting open source packages such as R on the NGS that can be used by social scientists and others. The larger challenge of engaging in more depth in application domains, e.g. in helping to support and maintain access to data and services for managing that data has not yet materialised nor (in my opinion) is it likely to since the NGS are not resourced for such activities. Rather this tends to happen through shorter term ESRC/JISC funded projects for example. Given this, the question of research focus and relationship between national social science research requirements of CESSDA organisations and their respective NGIs requires careful scoping and alignment. Thus there may well be strong links to e-Social Sciences in the UK with the NGS and major ESRC and JISC projects, but are other international communities and their e-Infrastructure/e-Research efforts similarly aligned and are future funding streams similarly supported? Looking at the wider international perspective, the relationship between CESSDA and EGI as a whole requires careful scoping and alignment. Thus existing members of CESSDA may not have any existing NGI or national resources to establish one for example?*

2 Cloud Computing and CESSDA

To understand the relationship and hence possibilities of cloud computing for a future CESSDA RI, it is necessary to clarify what is meant by cloud computing. Many regard access to remote software and/or services as cloud computing (often referred to as working in the cloud). Simple examples of this might be email accounts supported and managed by remote providers such as Microsoft hotmail or Google gmail. Such models have been refined and extended and exploit a range of web services and service oriented architectures to provide access to an increasingly wide range of distributed resources and capabilities. The ultimate vision of this Software as a Service (SaaS) model is that end users will no longer have to buy and install software on their own machine but will simply access and use resources that are offered remotely. Many major vendors including IBM, Microsoft are now offering such possibilities and are providing cloud-oriented development environments that allow exploitation of such capabilities, e.g. the Microsoft Azure platform (http://www.microsoft.com/azure), Google App Engine platform (http://code.google.com/appengine), Engine Yard's Vertebra platform (http://www.engineyard.com/vertebra) and the open source Eucalyptus platform (http://www.eucalyptus.com) to name but a few.

This is not the only model and/or interpretation of cloud computing however. Many providers such as Amazon now offer access to virtual resources, e.g. virtual servers running on server farms that Amazon manages. These can be used for a variety of purposes that are often application domain specific. This model is often referred to as Infrastructure (or Platform) as a Service (IaaS/PaaS). Other interpretations of this model include on-demand computing and utility computing. The Amazon Elastic Compute Cloud (EC2 - http://aws.amazon.com/ec2/) is one example of this where end users can purchase a variety of different virtual machines. These can be preconfigured depending upon the particular user requirements and the particular charging arrangements. The EC2 cloud for example offers a variety of virtual machines with different operating systems and database configurations that a user can pay for, e.g. a Windows 2008 server with SQLServer installed on a machine with 8Gb and 250Gb disk space. Alternatively the user can simply buy access to and use of virtual servers themselves and configure them as they see fit. This can be achieved through uploading and deployment of locally configured machine images.

Key to this model is the concept of virtualisation. Thus the end user does not buy a machine directly, but buys a virtual machine or more likely a set of virtual machines, that run on server cluster resources that Amazon manages. When a user providing a service requires more servers, e.g. to scale their systems, then they can simply acquire (buy) more virtual servers. There are a wide range of virtualisation technologies that exist today including VMware (http://www.vmware.com/), Xen (www.xen.org), and Microsoft's Hyper-V solution (http://www.microsoft.com/virtualization).

Other data-oriented models of cloud computing exist, e.g. the Amazon Simple Storage Service (S3 - http://aws.amazon.com/s3/). S3 focuses upon outsourcing the storage and management of data. Once again, users can add more resources or indeed return resources at any time when their data requirements change over time. The model is thus designed for on-demand resource utilisation that can scale accordingly. The business community, especially the small to medium business enterprise community, have been strong advocates of cloud computing and the leasing of infrastructure it provides.

There are many flavours of clouds that exist. Public clouds such as Amazon's EC2 service where resources are dynamically provisioned on a self-service basis to third party providers over the Internet via web services are the most common current model. Private clouds describe offerings that emulate cloud computing but running on private networks represent an alternative model. Hybrid combinations of these also exist, e.g. the Eucalptus system can be used for hybrid cloud-based solutions.

It is fair to say that cloud computing represents *the latest thing* with many academic research groups and commercial enterprises now involved in research into cloud based computing and offering cloud based systems and solutions. With this fluid state of affairs in mind, it is difficult to state with any great degree of certainty exactly what kind of impact cloud based approaches or solutions may have with regard to a future CESSDA RI. However currently we can identify the potential strengths, weaknesses, opportunities and threats offered by cloud computing to a future CESSDA RI.

2.1 Cloud Strengths and CESSDA

There are numerous potential benefits of cloud computing and related technologies in the context of the future CESSDA RI. We outline some of the key areas of clouds and identify their potential benefits to CESSDA.

- *Capital Expenditure* – clouds allow services providers and third party service providers to avoid capital expenditure investment on hardware, software and services. Instead cloud computing allows for a variety of other models, e.g. pay-as-you-go or subscription-oriented models of infrastructure usage.
    - *I believe that the existing CESSDA infrastructure is not limited by a lack of major hardware resources that are restricting growth and evolution of CESSDA. Unless a future CESSDA RI has far greater requirements on data management and importantly on data processing than currently exists, the running costs of managing a CESSDA data centre are not in themselves especially onerous. Rather it is a lack of integration of existing systems, harmonisation of data sets and security that are the primary issues that must be tackled and addressed by a future CESSDA RI*

- *Reduced Entry Barriers* – clouds can offer lower entry level access to potentially large-scale shared infrastructure, with low management overheads (in terms of managing large scale server infrastructures) and immediate access to a broad range of applications.
    - *The reduced entry barrier is not in itself a great benefit for a future CESSDA RI. As identified above, I believe that the CESSDA is not currently constrained by access to large scale infrastructures. However, the cloud model can allow for peak usage periods of CESSDA to be addressed. This might be achieved through replication of infrastructure, e.g. data services and data sets (subject to licensing and security considerations) on cloud infrastructures.*

- *Software Manageability* – clouds allow easier management and deployment of software on a larger scale for user communities. Thus rather than attempting to manage complex large scale deployments on heterogeneous infrastructures, through virtualisation, clouds allow virtual images to be securely uploaded for use by different user communities.
    - *Adopting the cloud based model of creating virtual images and uploading them and deploying them on remotely managed shared infrastructure is relatively easy to achieve. However this model has drawbacks for CESSDA. Firstly, the heterogeneity of the CESSDA middleware is not the major impediment to establishing a future CESSDA RI rather it is more important to have an infrastructure supporting secure access to distributed data resources. Secondly, interoperability of clouds is still very much an area being actively explored with little guarantees currently being offered with regards to inter-operability between cloud providers. Instead the onus is very much placed upon the third party providers who are using clouds to ensure that the services and resources that they make available through clouds are inter-operable with other resources, i.e. there is nothing in the clouds themselves that address this. Given this, the advantages of creating and uploading virtual images as a mechanism for addressing issues of software manageability complexity is impacted directly. It is worth noting that the Grid community themselves are currently actively pursuing issues related to Grid and cloud interoperability, e.g. the Open Grid Forum (OGF) has recently launched the Open Cloud Computing Interface Working Group (OCCI-WG – http://www.ogf.org/News/news.php?id=132) which aims to define an API for cloud infrastructure delivered on-demand.*

- *Cloud Focus and Novelty* – there is currently a huge amount of effort in cloud related technologies and systems, both in academia and in industry. Given this, many of the issues that clouds currently face are being explored on many fronts and open questions that need to be addressed by communities such as CESSDA can be fed in and used to drive the requirements of future cloud related research and technologies.

- *Whilst it is the case that a future CESSDA RI could benefit from "riding the current wave" of cloud interest and help to both lead and drive the requirements of cloud related research, the primary focus of CESSDA is not in itself technology-oriented. Rather it is delivery of stable services supporting the international social science community that is paramount. Nevertheless, it is acknowledged that such focus and effort is highly relevant since there is an active community exploring issues relating to cloud-based computing. This may not be so with particular flavours of middleware for example.*

2.2 Cloud Weaknesses

- *Vendor lock-in* - cloud computing can potentially limit the freedom of users/third party service providers by making them dependent upon particular cloud computing providers. Thus at present it is often only possible to use applications or services that the cloud provider is willing to offer. This can often give rise to difficulties when migrating applications between clouds for example. Furthermore, once a user has adopted a particular cloud provider and deployed a variety of applications and data sets, they are to a given extent, tied to whatever charging model and whatever fluctuations in charging that the cloud provider wishes to impose.
    - *This is a potential major concern for CESSDA and one that should be considered carefully before cloud models and cloud providers are considered. In the current business climate, there is no guarantee that any cloud provider will continue to exist and/or be able to offer their services at a guaranteed rate. Whilst it is unlikely that organisations like Google or Amazon will cease to exist entirely, it may well be the case that the business models that are offered by their existing cloud based solutions will continue in their current form. Indeed Richard Stallman (founder of the free software foundation) argues that cloud computing is simply a trap aimed at forcing more people to buy into locked, proprietary systems that would cost them more and more over time.*

- *Cloud Security* – some of the main concerns with cloud computing in their current incarnation are to do with security and the issues associated with data privacy and potential hosting of personal data on a third party resource.
    - *This is a major limitation for cloud based solutions in their current form for CESSDA. For more security-oriented data sets, e.g. social science data down to the individual level, it is often simply not tenable to host certain data sets on remote cloud-based service provider resources. Furthermore, existing cloud security models are still evolving. Different cloud providers will have their own individual approaches for establishing and managing security. Thus with Amazon for example, the individual purchasing resources on the EC2 cloud will download a public/private key pair which is used for future access to and use of the EC2 resources. This can be used to define firewall policies, activate certain services and deactivate other services or indeed set up particular groups that can access and use certain cloud resources. However this is primarily an Amazon based security solution. The access to individual data sets or supporting finer grained access control decisions on access to and use of particular services making certain data sets available described in [CR1] for example cannot easily be supported right now. Or put another way, this is left up to the individual user (service provider) making use of the cloud resources.*

- *Cloud Regulatory and Liability issues* – there are currently many open questions with regard to regulation and liability associated with cloud providers. Clouds can in principle exist on virtual resources scattered around the globe. As such they may become subject to complex geopolitical issues, e.g. the ethics of what data can be kept and where it can be kept may differ between countries and regions. Furthermore, cloud providers are especially wary of making commitments with regard to data loss or infrastructure failure more generally, indeed many cloud providers, e.g. Amazon EC2, explicitly state that it is up to the users of the cloud to ensure that they have addressed all regulatory concerns and that they themselves are liable for any issues associated with data loss for example.

   - *This is a major concern for CESSDA since there is in principle no guarantee that a cloud provider can/will make with regard to ensuring that the deployed services and/or data sets are not accessed and used by others. Rather this is pushed on to the user (service provider). Given this and the nature of many of the CESSDA RI data resources, this is a major limitation of clouds. However, if a given CESSDA organisation themselves wished to support a cloud based infrastructure, i.e. as opposed to accessing and using a cloud offered by an external third party cloud provider, then many of these issues would not arise.*

- *Cost Issues* – the immediate cost benefits of cloud computing can over the longer term, become a liability and cost more than owning and managing a local resource.
   - *Whilst the short term benefits of cloud computing can be a major factor in deciding to exploit clouds, longer term these costs can become excessive. As a concrete example of this albeit from a high performance computing consideration, the ScotGrid (www.scotgrid.ac.uk) infrastructure at the University of Glasgow was purchased at the end of 2006 for a price of £550k as a central compute facility for the campus and for use in a variety of collaborative e-Science projects. This facility has processed over 1.5 million jobs and clocked up over 6.7 million CPU hours with a typical utilisation of over 90%. ScotGrid runs close to 24 hours per day and 365 days per year excluding minor time outs due to system upgrades and individual server failures/adjustments. As a representative example of the cloud computing providers, Amazon's EC2 cloud currently offers similar specification servers to ScotGrid for approximately 30p per CPU hour, with data transfer rates into the Amazon EC2 cloud of 7p per GB and data transfer out of the Amazon EC2 cloud costing in the region of 10p per GB. The Amazon Elastic Block Store associated with EC2 costs approximately 8p per GB. Based upon these rates, if the CPU time on the ScotGrid infrastructure had been bought on the Amazon EC2 cloud, this would have resulted in an overall charge of £1.8M for CPU time alone! The data storage and transfer into/out of ScotGrid would have resulted in fees far greater than £2M. In short, longer term usage of infrastructure is far more cost effective when bought and managed locally when compared to existing pay on-demand cloud models. We note that we acknowledge the local system administration and electricity costs both for running and cooling ScotGrid also have to be factored in here, but this still equates to more than triple the actual cost of the ScotGrid infrastructure itself.*

- *Data management* – cloud computing offers a scalable and flexible model for data storage where additional resources can be requested and made available on-demand as data storage requirements vary. Clouds in themselves however do not address the many challenges of data management and the data lifecycle more generally.
   - *Major cloud providers such as the Amazon S3 service offer resources that can be used for data storage. However, whilst S3 allows for flexible arrangements on data storage the whole issue of data management is effectively pushed on to the user. Thus there is no direct facility for cloud computing that can be used to address issues of access to and use of data and meta-data directly. Rather, services which give access to data have to be defined by the cloud customer (service provider) for others (end users). We note that some cloud providers do provide certain capabilities, e.g. for data back-up, but the fundamental issues facing CESSDA with regard to data management including data access, variable naming and data harmonisation etc, still exist and clouds in themselves do nothing to address this.*

2.3 Cloud Opportunities and CESSDA

- *Cloud revolution* – whilst section 2.2 has identified the numerous open issues related to clouds, it is the case that despite this, there is huge interest in cloud-based systems. This interest also gives rise to various opportunities to CESSDA and access to potential current and future funding streams.
  - *Whilst the focus of CESSDA is not in itself on the exploitation of novel platforms for social scientists or explorations of next generation technologies like clouds in themselves, it is the case that this future technology focus can offer access to a wide variety of additional funding streams. Thus it is the case that novel research augmented with novel software systems often has a stronger chance of receiving funding than research based upon more traditional infrastructures. As discussed, there is a perception of clouds being the next wave and one that many communities are aligning themselves with (this includes many of the existing Grid communities).*

2.4 Cloud Threats and CESSDA

- *Cloud provider fluidity* – given the novelty of cloud based systems there is a current flux of commercial and open source cloud-based offerings. It is difficult to know right now which of these offerings and hence the providers will persist either as organisations themselves or in the shape and form of cloud offerings that they currently make available. One would expect that when more and more cloud providers exist, the competition and hence price for access to and use of cloud resources would be more competitive, i.e. decrease. However especially given the current economic climate, there is a very real danger that even major organisations like Amazon and Google will re-evaluate their cloud-oriented business models.
  - *With regard to CESSDA, it is difficult to predict whether the current incarnation of cloud service providers will persist, and if so whether this will be in the form that they current exist in.*

- *Cloud technology fluidity* – there are many forms of clouds that exist today and many associated technologies that are used to support clouds. These technologies are very much nascent however and subject to radical change and development. Given this state of affairs it is unsurprising that there is no real defined notion or standards for inter-operability between clouds and between cloud providers.
  - *This is one of the major concerns with regard to the future CESSDA RI and its adoption of cloud based solutions. The resources that are offered and the way in which they are offered are continually evolving and it is difficult to predict with any great certainty that any given cloud-based solution will exist in its current form in the coming years.*
  - *On a related point, it could be argued that many of these arguments can also be applied to Grid technologies with on-going development of standards by organisations such as the Open Grid Forum (OGF – www.ogf.org) and associated technologies as realised by middleware providers like Globus (www.globus.org) and the application of these technologies by many others including EGEE and potentially the EGI. There is no fixed core set of software that has been widely accepted by all e-Infrastructure providers and that seamlessly incorporates evolution and change of standards and software from providers such as Globus and EGEE. Rather many mainstream e-Infrastructure providers such as the UK e-Science National Grid Service (NGS) have adopted a pragmatic and conservative approach to software deployment and upgrades, e.g. they offer a software stack that has been tried and tested based upon past experience and one that does not support many of the more recent Grid solutions and standards.*

- *Perception of clouds* – due to their avoidance of clearly defined, fine-grained security models, clouds are perceived as a danger and potential threat to domains that can have more of a security focus such as the social sciences.
    - *This is a potential major threat to CESSDA. If a given data provider considered that the data that they were to make available through the CESSDA RI could potentially end up on a remote cloud, then the likelihood is that this data would never be released to CESSDA.*
    - *We note that there is a similar perception with regard to Grids and Grid security. As outlined in [CR1] whilst many domains are not overly concerned with authentication-driven security as typified with X509 credentials used to access accounts on HPC resources, this model of access to and use of Grid resources is not the only one. Rather finer-grained security models can be supported. However, by associated with the "Grid" the perception is that all Grids and e-Infrastructures more generally are a security risk and should be avoided at all costs.*

3 Own Infrastructure and CESSDA

The present CESSDA data portal allows access to social science data via the Nesstar (http://www.nesstar.com) web based statistical software package. This system is based upon use of Data Documentation Initiative (DDI - www.ddialliance.org) extensible mark-up language (XML) records; a Nesstar server and use of the CESSDA subject classification translated in to the member's native language.

Through this infrastructure, simple analysis such as cross tabulations of single datasets is supported, exploiting indexes that have been established using the Lucene (http://lucene.apache.org) technology for harvesting DDI information held in CESSDA member Nesstar servers. The only service that this infrastructure provides, over and above what is on offer at the individual member's web sites, is a centralised resource discovery tool based on multi-lingual controlled vocabularies.

Although the present infrastructure is a valuable tool for the simple analysis of single, primarily flat, rectangular datasets, any research wishing to harmonize and combine two or more datasets held across CESSDA has to perform these operations outside of the present data portal. Furthermore a key requirement for the data portal is to be able to accommodate different types of data, different metadata standards and different data analysis software, potentially using different or upcoming controlled vocabularies and different services. Fine-grained security is also required on access to many of the data sets that are distributed across member partner organisations, and ideally it should not be required that a researcher has to provide multiple authentications to multiple distributed resources. Thus the notion of single sign-on is a highly desirable feature.

Whilst Grid-based or cloud-based e-Infrastructures offer advantages and disadvantages as outlined previously, other possibilities also exist. In this section we identify the strengths, weaknesses, opportunities and threats associated with more traditional approaches based upon client-server oriented technologies and/or service-oriented architectures.

It is worth emphasising here that many Grid-based solutions would claim to support service-oriented architectures and/or be based around client/service-oriented architectures. The distinction between standards and technologies and associated nomenclature is often confusing. By service-oriented architectures here we mean traditional web service oriented approaches where services are accessible and can be invoked by clients through Simple Object Access Protocol (SOAP) [SOAP] messages. The services themselves that consume these SOAP messages can be implemented in a variety of ways using a variety of languages, e.g. Java, C++, C# using the Document Object Model (DOM) or Simple API over XML (SAX) etc. Clients are shielded from the heterogeneity of implementation through SOAP messaging and the abstractions that it offers.

This is one classification of service-oriented architecture but others are equally possible and valid, e.g. use of other messaging protocols for interacting with services available over the internet. One example of this might be services that are based on Representational State Transfer (REST) [REST] based approaches. In this model, services are accessed directly over http (Hypertext Transfer Protocol) [HTTP]. Furthermore, HTTP is itself a client-server protocol which provides mechanisms, e.g. GET, PUT, POST, DELETE etc, to access resources over the web. This model also has advantages in that the bloat of information sent in XML-based SOAP messages is greatly reduced. REST-based (RESTful) services can also be accessed and used by lightweight client applications, e.g. those that exploit Web 2.0 technologies.

3.1 Own Infrastructure Strengths and CESSDA

There are several advantages in adopting a traditional web service-oriented architecture for CESSDA which may impact upon the future CESSDA roadmap.

- *Variety and Flexibility* – there are many approaches to developing and supporting web services with rich integrated development environments. Many of these, e.g. those from the open source Apache Foundation (www.apache.org) offer a variety of approaches and technologies for developing, hosting and maintenance of web services. Alternatively commercial solutions and development environments also exist, e.g. the Microsoft .Net platform being one example. This variety means that CESSDA member organizations are able to select whichever solution best fits their needs and local expertise, e.g. if they are more comfortable with Java development and solutions such as Apache Tomcat servers for hosting services etc.

  - *Whilst this variety and flexibility means that CESSDA member organizations are able to adopt their own solutions based upon local expertise, care must be taken since arbitrary web service development does not in itself immediately guarantee an inter-operable service-oriented architecture. We note that at least some of the issues involved in supporting inter-operability between web services can be addressed. Thus, provided the web service interface specification is made available, e.g. through a Universal Description, Discovery and Integration (UDDI) registry of services, then at least in principle, clients can be built that can interact with these services. In principle is given here since many of the issues of inter-operability are not simply due to the naming of information given in the interface for example. Thus if a service has implemented its own form of web service security for example, then clients can only interact with these services if they too have the correct signatures and encryption information. In short, there is nothing implicit in web services themselves which will guarantee their inter-operability. This was described in some detail in Annex 3 of [CR1] which focused in particular upon the area of security of web service-based approaches.*

- *Stability of Technologies* – web services are, at least in principle, more stable than other offerings, e.g. cloud or Grid-based solutions. The core W3C and OASIS standards associated with web services such as the SOAP specification have been established for almost a decade now and used on a large scale to make a variety of content available across the internet.
  - *This is an advantage to CESSDA however despite the stability of many core standards it is still the case that a huge amount of work is still on-going by organizations such as IETF, W3C and OASIS on refinements to existing specifications and proposals for new web service and web service related specifications. New profiles for security, for interfacing with semantic web-based web services etc are all continuing. See http://www.phpfever.com/images/WS-Standards-2007-02-medium.jpg. Given the major on-going work in the web service standards domain there will continue to exist evolution of the standards and hence the technologies themselves.*

- *Focus on service/content delivery* – by adopting tried and tested technologies such as web services CESSDA can focus upon content and service delivery and not have to immerse themselves with non-production level research-oriented middleware technologies.
  - *This is a real advantage that should not be underestimated by CESSDA. It is the case that Grid and to a lesser extent cloud-based offerings have not yet been truly 'productionised'. Rather, there still remains a learning curve in adopting many Grid-based solutions. This learning curve impacts upon the developers and administrators of Grid infrastructures, and is particularly so when considering more complex Grid middleware offerings. Instead, many e-Infrastructure providers have adopted a simpler set of middleware/software stack, e.g. the UK e-Science National Grid Service which, whilst missing some advanced features, allows them to offer a production level service.*

- *CESSDA-driven infrastructure* – by CESSDA developing and supporting its own service-oriented architecture, it is able to dictate the strategic direction of its infrastructure developments and ensure that they meet the requirements of the social science research communities across Europe and not those of other HPC-oriented domains for example.
  - *If CESSDA were to join an existing initiative such as the EGI (if it happens), then there is a potential danger in CESSDA not being able to direct its own specific social science related infrastructure requirements. Rather this common e-Infrastructure may need to support a wide range of research disciplines with requirements, e.g.*

supporting high performance computing, not addressing or aligned with usability, security or data federation requirements of CESSDA and the social sciences. This would not happen if CESSDA were to continue with its own largely independent e-Infrastructure developments.

3.2 Own Infrastructure Weaknesses and CESSDA

There are several associated disadvantages if CESSDA decides upon adopting an entirely web service oriented approach.

- *Lesson Learned* – one of the primary drivers of the Grid, at least from the middleware perspective, was to overcome limitations of existing web services and traditional service-oriented architectures. Supporting finer-grained, VO-specific security and single sign-on across organisations in a heterogeneous environment is at the heart of Grid approaches and something that traditional web based service-oriented architectures do not easily address. Many of the challenges that web service-related standards such as WS-Policy, WS-Federation, WS-Trust (see Annex 3 of [CR1] are tackling have been explored in detail and software systems already exist that address them in the Grid domain.
    - *There is a great danger to CESSDA in that if they do adopt a traditional service-oriented architecture based approach, then they will end up having to tackle many of the issues that the Grid community have already faced and developed solutions for, e.g. supporting interfaces and protocols for secure, parallel data transfer, trust-related issues and certification authorities, single sign-on, delegation of authority, etc.*

- *Isolation* – there has been a large momentum for the last few years in Grid and more recently in cloud-related technologies. Whilst arguably not as great as it once was, this momentum is still driving much work and efforts across Europe such as the EGI and many national initiatives. The vision of the Grid in supporting seamless, inter-disciplinary research is made more difficult if different systems and solutions are proposed, e.g. attempting to compose services using Grid-based solutions and non-Grid based solutions. It is really only when inter-disciplinary research is undertaken (ideally on a shared infrastructure) that the issues in developing, supporting and managing inter-disciplinary research infrastructures are identified.
    - *With regard to CESSDA, one weakness of adopting a traditional service-oriented architecture may be from the perspective of not being "in the club". If the CESSDA RI is developed largely independently from other e-Infrastructure efforts such as the EGI, then there may be a potential sense of isolation and independence of CESSDA may arise. In this case it may not be as easy to leverage other national or international infrastructure resources or efforts more generally.*

- *Interoperability* – web services and traditional service-oriented architectures do not in themselves guarantee inter-operability or support single sign-on across inter-organisational collaborations. Interacting with services across multiple organisations is made much more difficult if different technologies are adopted. We note that we state "made more difficult" here deliberately since it is at least in some cases possible to develop particular clients that can interact with particular Grid services and traditional web services for example. However these clients and the work involved in developing them greatly increases the overall software development complexity for developers, especially if security and/or other non-common features across these services are required.
    - *Whilst web service technology and frameworks from organisations like Microsoft can be used to support a form of security-driven single sign-on, this does not work when other services or resources from other non-Microsoft providers are integrated. Thus it is possible to access and use .Net based web services with open source Java-based web service clients for example however the efforts in software development are increased considerably. Furthermore, if CESSDA develops its own traditional web*

service-based service-oriented architecture, then it will not be able to leverage national and/or international e-Infrastructure efforts and resources.

3.3 Own Infrastructure Opportunities and CESSDA

The new opportunities offered to CESSDA by supporting its own service oriented architecture are minimal. However it could be argued that some opportunities do exist.

- *Research-oriented focus* – by not adopting potentially complex Grid middleware systems, CESSDA can in turn focus on social science sources of funding being used more for social science research itself rather than on development and support of e-Infrastructure.
  - *Whether this is really an opportunity for CESSDA itself is debatable.*

3.4 Own Infrastructure Threats and CESSDA

If CESSDA decides to support its own service-oriented architecture then there are certain potential threats associated with this.

- *Missing opportunities* – by supporting its own service-oriented architecture-based e-Infrastructure, CESSDA is at risk of missing opportunities for inter-disciplinary sources of funding on a common infrastructure.
  - *This is a real threat to CESSDA. Future roadmaps for research infrastructures across Europe are currently being defined through efforts such as EGI. With these roadmaps it is highly likely that a variety of related funding streams will evolve to maximise the investments in a common e-Infrastructure across Europe. If CESSDA is not involved in using and driving these initiatives then there is a real possibility that it will miss out on future potential funding streams, including funding streams made possible by future inter-disciplinary research made possible by a common integrated European-wide e-Infrastructure.*

- *Technological Innovation* – it is the case that technical IT-related innovation will always continue. Mainstream efforts represented by Grid and cloud based systems, and upcoming efforts such as Web 2.0 technologies, will continue to push the boundaries of computing science research. Often this is achieved through interacting with a wide range of disciplines, e.g. the clinical sciences have helped shaped many Grid efforts including security protocols and standards for example. As such, it is essential for infrastructure providers such as CESSDA and the future CESSDA RI, that they are best informed by such developments. This can realistically only occur by being actively involved in these efforts.
  - *If CESSDA develops its own independent traditional service-oriented architecture based e-Infrastructure, then there is a potential danger that it will become isolated from technical innovations that other groups and infrastructure providers are developing and rolling out.*

4 CESSDA e-Infrastructure Considerations

Based upon the previous considerations this section addresses the various questions that were raised in the original tender document with regard to resource and sustainability issues relating to implementation of a future CESSDA e-Infrastructure. In particular the tender document wished this report to address key questions on possible scenarios for the future CESSDA RI infrastructure, namely should the CESSDA infrastructure be based on: implementing its own infrastructure based on existing client-server or SOA models; joining an already ongoing initiative and sharing resources with other disciplines using Grid technologies, or should it use cloud computing services.

In addition to any technological decisions, CESSDA requested that a detailed exploration and prediction of a variety of concerns were addressed. In particular, for each of the three choices given above, what would be the cost breakdown of:

- hardware and cost of ownership including acquisition, maintenance, replacement and support;
- software cost of ownership including software packages used, licenses required, installation and administration support, e.g. for security;
- personnel required for technical, support and coordination efforts;
- usage considerations including network resources, power consumption and general wear and replacement of equipment;
- education and training for developers, end users, control/management systems and administrators more generally;

CESSDA also wished to know it would be possible to charge individual institutes and/or individual users based on their usage, and where CESSDA might apply for funding for the maintenance and sustainability of the infrastructure?

Finally CESSDA wished to: have answers on the longevity of the chosen technology and possible upcoming threats and trends; have an overview of ongoing development initiatives on the middleware and cloud computing front, and whether CESSDA should join an already existing initiative and if so, what would be the likely longevity of the initiative itself and how long will its members support the infrastructure.

I note here that some of these questions are extremely (impossible?) difficult to answer with any degree of certainty due to the number of unknowns. Some of the most important of these are:

- Whether the EGI or a similar European-wide initiative will actually happen and if so will all envisaged NGIs get involved?
- What middleware will the future EGI (or similar infrastructure) actually support[2]?
- If the EGI happens then will the full complement of training and support staff be funded?
- If the EGI happens, will the social sciences be one of the supported application areas, and hence will EGI support groups be made available?
- How many new organizations would join a future CESSDA e-Infrastructure if it supported inter-operable access to inter-disciplinary research domains?
- How many new researchers from non-social science domains would want to access and use CESSDA services for large scale data analysis etc?
- Will existing cloud business models continue in their current form?

This list is not complete and the answers to the questions given could have a radical impact upon a future CESSDA e-Infrastructure. Despite these numerous unknowns, given knowledge of past technological investments and initiatives, and a current snapshot of European and international efforts it is possible to provide informed predictions – this is what I have attempted to do in the following sections. Bearing this in mind, I emphasize that the following sections should be regarded as informed speculations as opposed to being definitive.

[2]It is currently discussed in the EGI proposal that an integrated set of software derived from the NorduGrid (Advanced Resource Connector (ARC)) software stack; the EGEE (gLite) software stack and the DEISA Uniform Interface to Computing Resources (Unicore) stack would be made available, as part of a Unified Middleware Distribution software stack. However this is a proposal right now and this stack does not exist. Furthermore, the EGI is not resourced to integrate these different software solutions nor is it clear whether given NGIs will agree to any new set of middleware, e.g. based on requirements of their existing customer base.

4.1 Initial and Annual Cost for the CESSDA Consortium

The different models of Grid-based, cloud-based or development of an independent CESSDA e-Infrastructure offer different costing possibilities and implications to CESSDA. We treat each of these in turn.

4.1.1 Initial and Annual Cost for a Grid-based Infrastructure

There are a variety of possibilities and cost implications that might arise to CESSDA if it were to adopt a Grid-based e-Infrastructure. There are two possibilities here:

- CESSDA establishes its own Grid-based e-Infrastructure;
- CESSDA joins an existing initiative such as the EGI - if this happens!

The initial and annual cost of hardware to CESSDA if it established its own Grid-based e-Infrastructure from a hardware perspective would not in itself be hugely different than they would for a non-Grid based infrastructure since the specification of the servers and their associated pricing would be broadly similar. Put another way, there is no fundamental distinction between a server for a Grid service and a server hosting web services *per se*. As such, the cost implications would not be fundamentally different if the existing capabilities of the CESSDA RI were to stay the same. It is highly likely however that the whole point of CESSDA adopting a Grid-based approach and developing an e-Infrastructure are to support richer capabilities than are currently possible, and hence enhanced capabilities of a future CESSDA RI would be expected. This might be through enhanced international querying of data archives, larger scale statistical analysis of cross national data sets which may in turn require access to and usage of larger scale servers or HPC facilities, e.g. for complex joins or statistical analysis etc.

That said the actual cost of the hardware would of course depend greatly upon the architecture of the system design. As an example of a typical costing for a CESSDA member organization that has adopted a 3-tier architecture model comprising the user interface layer (portal); middle-layer where logic is defined and enforced including security on access to a range of Grid and Grid data services, and lower layer physical data management, the hardware infrastructure might comprise servers for:

- Hosting of portal based infrastructure which give access to portlets/clients which interact with local/remote Grid services which in turn interact with Grid-enabled data services;
    - *a typical server or PC can often be used for this purpose depending greatly upon the expected number of users and the complexity of the portlet functionality as well as the number of portlets themselves. For resilience and redundancy a given member organization would probably expect to have more than one portal in case of failures or for system upgrades. Furthermore, it might well be the case that each individual CESSDA organization has its own portal which gives access international and its own national CESSDA services and data resources.*

- Hosting of Grid services and their associated containers giving access to one or more data services;
    - *Hosting of Grid service environments including containers for hosting of Grid services does not in itself require large scale hardware infrastructure and a typical server can be used for this purpose. However, depending upon the CESSDA organization itself, it may well be the case that many different Grid services are expected to exist and be co-hosted on hardware infrastructure. Scalability and security requirements might mandate that different services are hosted on separate infrastructure. It is noted that access to particular CESSDA resources can be made by associating security policies (policy enforcement points) and enforcing access control decisions (policy decision points) based upon access to the services themselves, and secure and non-secure services can be hosted in the same containers, this is not advisable.*

- Hosting of one or more Grid-enabled data services. Depending upon the association of data sets with services there may well be a multitude of independent data services, i.e. on separate data servers, or a single service could be used to provide access a range of data resources (including upload and query functionality). Thus a given Grid-enabled data service could interact with multiple data resources in a federated manner and join returned data sets for example. The choice ultimately depends upon the trade-offs of managing more hardware and more services versus the scalability and flexibility of the system and its evolution. Thus a given service that interacts with multiple data resources directly would need to be extended when these data resources change and accommodate the changes in its service logic. When individual services are associated with individual data resources this change does not in itself demand other services are completely changed.

- *It is worth noting that it is possible that Grid services can be used to directly interact with data resources directly, i.e. without the need for Grid-enabled data services. In the case of relational databases for example this might be through Grid services directly using JDBC technology to connect to a given database for example. A better and more flexible model of access is through support of specific data services that Grid services can interact with. Once again though in terms of hardware, a Grid data server will be similar to a data server more generally. The precise specification of the server itself will ultimately depend on the data service itself, i.e. what kind and amount of returned data sets and subsequent joining/processing the service is required to support.*

- Hosting of an attribute authority (this could be a centralized or decentralized authority as described in [CR1]); a certificate management service (such as a MyProxy service) and individual resources for policy decision making;
  - *For services that require finer-grained security, it is necessary to establish attribute authorities which are used for delivering security credentials as outlined in [CR1]. Different possibilities exist here: a single centralized authority such as VOMS could be used across all of the CESSDA organizations for access control decisions (which would require a single typical server specification); a centralized authority used across individual national CESSDA member organizations (which would require a typical server specification per CESSDA member country); a complete federated collection of attribute authorities associated with individual CESSDA institutions. Ideally in the latter case, exploitation of existing national access management federations could be exploited for this purpose, e.g. the UK Access Management Federation. It is noted that for those countries yet to establish their own national federated access management infrastructures the costs to establish such an infrastructure are considerable. More precisely the costs are not simply related to equipment used for identity providers and WAYF servers (see [CR1]) but in the whole process of supporting security including issuance of certificates used for authentication and ensuring that sites adhere to policies on being part of the federation. It is noted that it is possible to establish virtual homes for individuals existing across CESSDA sites that do not have their own individual access management infrastructure, e.g. establish accounts at a recognized institution (IdP) where individuals can authenticate, however this model is deprecated and not scalable.*
  - *When an individual attempts to access a remote (protected) service, the service needs to be able to make a local authorization decision. Part of this process is establishing the identity of the individual themselves. In many situations (e.g. in the UK), this is achieved through X509 certificates being issued by a recognized Certificate Authority. In this case the distinguished name (DN) of the individual in possession of the certificate can be extracted and used by a local policy enforcement point to assess whether that individual has the appropriate privileges to access the given resource. To avoid users having to manage their own X509 certificates, many sites and countries more generally have established MyProxy services which manage these certificates on behalf of individuals. CESSDA as a whole or an individual CESSDA*

  member organization may well wish to support this model also and a suitable server would also be required.
  - *For each resource (service, data set) that requires protection at a given site, it is necessary to have a server (typically an LDAP server) and an authorization decision engine (such as PERMIS or XACAML as outlined in [CR1]) which is used for making the actual decision.*

- Hosting of data sets themselves on back-end data servers;
  - *The infrastructure required for hosting of CESSDA related data sets would not change if a Grid infrastructure was used – provided that the scope of data and user community remained the same. However, it might well be the case that inter-disciplinary research communities are generating new data that could/should be hosted as part of CESSDA, e.g. health and social science related data sets. Furthermore, if these data sets do scale with wider research communities, then it will be necessary to consider the scoping and additional tool support for processing, preparation and analysis of these data sets. Thus building of large scale data indexes and/or hosting of derived data sets on behalf of research communities for further analysis are all possible extensions and refinements to what CESSDA currently does.*

- Software and Licensing;
  - *The majority of Grid software itself that exists is based upon various forms of open source models, e.g. Gnu Public License models. As such, there is no direct cost in Grid middleware per se*[3]*. However, Grid technologies can be used to give access to licensed software itself. SDSS, STATA and SAS are some example of software that can be made available "on the Grid", i.e. a Grid service exists which allows a particular STATA routine to be run for some statistical analysis say. As identified in [CR1], whilst it is possible to define and enforce access control on use of this service to individuals with the appropriate privileges, e.g. those whose local identity providers assert that they have an individual, department or institutional license for STATA, this demands that the producers of the STATA software are satisfied with this arrangement. There is no widespread agreement yet on these kinds of practices. Indeed licensing is itself ad hoc with fixed server licenses, individual user licenses, floating licenses etc all existing and most license management software, e.g. FlexLM, designed to avoid any variations on license models. Case in point, MatLab is the package most requested on HPC resources, but it is not deployed on the UK e-Science NGS due to licensing issues and the concerns that once this is available on a free national resource then the MatLab customer base will use this and not renew their own individual/institutional licenses etc.*

In terms of initial and annual costs of CESSDA establishing its own Grid-based e-Infrastructure, many of the above issues are straightforward, e.g. purchasing servers for hosting portals, services etc. It is highly likely that, at least initially, existing CESSDA infrastructure could be used for this purpose and hence no cost would be incurred. This is not the case for those countries that are not yet part of CESSDA of course. Longer term costs for running servers and their maintenance would need to also be factored in. However, given that the specification of these servers is similar to existing servers used in CESSDA or indeed in any organization where web services or hosting of internet facing service infrastructure exists, then they are arguably best placed to assess how much they are paying for hardware and on-going maintenance contracts. My own personal feeling here is that it is more cost effective to purchase and host servers.

[3] There are of course commercial Grid-software systems that exist, but these tend to be specialised to particular application domains, e.g. IBM's Information Integrator for the life sciences, ABCD from Johnson & Johnson for drug discovery etc. Suffice to say that the vast majority of Grid middleware in use across Europe is based on open source models of access and usage and as such freely available.

The on-going maintenance contracts and warrantees for servers will vary with institutional arrangements on procurements, e.g. from preferential suppliers etc. Typically 3-year maintenance contracts and warrantees are the norm in the UK, and tend to be based on the lifetime of the research project they are used in as well as the general lifetime of servers themselves – or more often than not when the server specification is significantly lower than current market offerings.

I note that one key aspect of this that I have not factored in, is long term data storage, data management and curation more generally. The costs associated with this can be significant and far outweigh any initial costs for given servers. It is difficult to assess the costs of long term data management with any degree of accuracy. A general rule of thumb used by the BBC for example in the UK is that it costs £1m per petabyte of data for managed server storage, e.g. disc based storage with back-up and redundancy. These costs vary greatly depending on whether the actual content is accessed and used on a regular basis.

A further key factor is the costs that would be incurred for management and administration of the e-Infrastructure. This is difficult to gauge with any degree of accuracy. It can be assumed that systems administrators exist already at CESSDA institutions and are able to maintain server infrastructures hence the focus here is on additional costs of developing and maintaining Grid-oriented e-Infrastructures. To a large extent, this would depend greatly upon the data sets that CESSDA centers have available; the architecture upon which they are to be delivered, and the infrastructure that supports this delivery. Ideally, a given CESSDA member organization would have their own e-Infrastructure personnel charged with developing user interfaces (portals); for establishing and maintaining Grid and data services, and for defining and enforcing local security policies. Alternative models might be that the CESSDA as a whole has a small team of Grid-related experts responsible for developing e-Infrastructures across individual CESSDA organizations based upon a single one-stop shop portal giving access to a multitude of distributed data resources that this team establishes and supports. This has advantages in that expertise is centralized and can be used to ensure inter-operability across international resources. However, the obvious drawback is that local expertise is not available and CESSDA organizations are not able to develop and refine their own e-Infrastructure offerings. Many organizations would be unwilling or unable to support this model, e.g. for security considerations. In short there are a multitude of possibilities that might exist on the personnel required to establish and maintain a future Grid-enabled CESSDA RI. My own personnel feeling is that each site would require at least one or preferably more Grid-related experts to work on local e-Infrastructure efforts and integrating and delivery of national CESSDA resources into international CESSDA efforts.

Many of the above considerations on initial hardware specification apply broadly whether CESSDA decides to join an international Grid initiative or decides to support its own Grid-based e-Infrastructure. As identified in section 1.4, it is unlikely that any given NGI (and hence EGI) would be able to develop and support specific services accessing and using CESSDA resources themselves. They are neither resourced nor have the domain expertise to achieve this. Nevertheless CESSDA can leverage resources from Grid initiatives such as EGI. Examples of this are through:

- International trust agreements and recognition of certificate authorities;
    - *By joining in the EGI or a similar effort, CESSDA can exploit resources such as national certification authorities and international trust agreements between these authorities. Existing hardware infrastructure such as VOMS servers can be exploited directly without having to establish and support CESSDA individual authorities.*

- Access to larger scale infrastructure;
    - *By joining the EGI or a similar effort, CESSDA would have access to very large HPC resources for much greater data analysis and processing possibilities without incurring any direct expense themselves (at least for hardware procurement).*

- Software development and Grid expertise;
    - *By joining the EGI or a similar effort, CESSDA would have access to considerable expertise in Grid technology and know-how. It should be emphasized that many of*

these technologies are in a relatively non-production form and require software expertise in how best to deploy and manage them, and/or adapt them to the particular requirements of the CESSDA. This troubleshooting support should not be under-estimated. Furthermore, "if" the EGI happens, then considerable resources are already identified for support of application domains, however at present these appear to be largely HPC-oriented. It is noted that much of the expertise is generic, e.g. in terms of how to access and use a Grid resource or take care of a digital certificate, whereas it is likely that CESSDA would require more social-science oriented support and expertise.

- Personnel and Administration Costs;
  - *By joining the EGI or a similar effort, CESSDA could leverage a variety of personnel charged with administration and management of the EGI infrastructure and the NGIs it comprises. As noted in section 1, it is unlikely that an NGI such as the UK NGS could help with the detailed specification of the access and usage policies for CESSDA resources at given CESSDA member sites for example, however support in dealing with many of the technologies is available.*

- Training costs;
  - *By joining the EGI or a similar effort, CESSDA would have access to hands-on training materials and educators involved in all aspects of Grid technologies. This includes training materials, lecture materials and summer schools that have been established and run for many years, as well as hands on training courses where aspects of Grid can be brought out. These training materials have included in the UK, information targeted to administrators, developers, policy makers as well as end users themselves. It would be expected that CESSDA would develop its own training materials that complement these materials, e.g. informing communities on data resources that are available and how to access and use them etc.*

- Annual running costs;
  - *As noted previously the existing model put forward for EGI is that it would be national initiatives that would join and hence pay annual running costs. These costs vary across countries. It is also not clear who would pay these on-going costs in the UK (most likely EPSRC and JISC) but contributions from ESRC might be made if it was ensured that the social sciences and CESSDA were supported directly. This is all supposition right now though and definitive annual running costs to CESSDA cannot be given.*

4.1.2 Initial and Annual Cost for a Cloud-based CESSDA e-Infrastructure

As noted, one of the primary benefits of a cloud-oriented approach is that the initial costs of acquiring hardware can be removed entirely and pay on-demand models can be supported. The hardware and server specifications outlined in 4.1.1 could be offset entirely by outsourcing this to a third party cloud provider and hosting of virtual servers. There are many different charging models that currently exist for access to and use of cloud-based systems. Taking Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Solution (S3) as examples, the cloud access and charging models have varied greatly since they were initially offered, with the price of compute resources dropping significantly, but with other costs now arising, e.g. data transfer in/out and costs of data replication to Amazon S3 from EC2. Currently the Amazon pricing models are:

- $0.11 per CPU/hour for a 1.7Gb RAM virtual machine with 160Gb disk running Linux;
- $0.44 per CPU/hour for a 7.5GB RAM virtual machine with 850Gb disk running Linux;
- $0.88 per CPU/hour for a 15GB RAM virtual machine with 1690Gb disk running Linux;

With data transfer-in costs of $0.10 per Gb and data transfer out costs of:

- $0.17 per Gb first 10Tb per month;
- $0.13 per Gb next 40Tb per month;
- $0.11 per Gb next 100Tb per month;
- $0.10 per Gb next 150Tb per month.

With Amazon elastic block storage costs of:

- $0.11 per Gb-month of provisioned storage;
- $0.11 per 1-million I/O requests;

And Amazon elastic block storage snapshots to Amazon S3 of:

- $0.18 per Gb-month of data stored;
- $0.012 per 1000-Put requests when storing a snapshot;
- $0.012 per 1000-Get requests when retrieving a snapshot.

Other cloud providers have their own pricing arrangements. As identified in section 2.2 however, the longer term cost implications of cloud-based solutions can be considerably more expensive than actually acquiring and managing hardware directly – at least for high performance computing as demonstrated with the ScotGrid example. Whether this is the case for data-oriented analysis is unclear. How much data will CESSDA have to support when other national organizations get involved, and/or when other non-social science communities engage in inter-disciplinary collaborations with CESSDA? How often would data have to be backed up: nightly, fortnightly, monthly, annually? Long term data management and curation would potentially lead to exorbitant cloud fees.

As described in section 2, the use of cloud-based infrastructures does not in itself address any of the key software requirements facing CESSDA. Indeed there are many other factors, e.g. security, which might mean that cloud-based approaches are simply not tenable to CESSDA. The software that is required for CESSDA to operate an infrastructure will still have to be developed and supported, and cloud based infrastructures do not help in this regard. Indeed many issues are made more complex by the current incarnation of clouds, e.g. supporting single sign-on to distributed resources, where different cloud providers are generating their own public/private key pairs for users to access and use virtual servers. There is currently no trust or root certification authority behind these cloud providers, hence bridging based PKIs or other technical solutions would have to be engineered to support inter-operability.

4.1.3 Initial and Annual Cost for a CESSDA e-Infrastructure

In this model, the current and annual cost implications to CESSDA would largely remain unchanged from the description of many of the Grid-based infrastructure costs given in section 4.1.1. Specifically:

- server specifications of Grid vs non-Grid hardware systems are broadly similar;
- cost of ownership and general running costs of Grid vs non-Grid hardware including replacement of defective systems are likely to be the same (assuming no access to larger scale HPC resources is required);
- software costs are likely to be the same (since Grid middleware is primarily open source) and hence any license fees and support associated with specific software packages are similar;
- support and management costs associated with hardware are the same;
- networking resources, bandwidth and power consumption are likely to be the same (assuming that the scale of usage of CESSDA resources does not change drastically in making CESSDA resources available to potentially wider inter-disciplinary research communities);

Where differences do arise with initial and annual costs of a Grid-based e-Infrastructure compared to a traditional service-oriented architecture based infrastructures are in the expertise and personnel costs of maintaining a Grid-based e-Infrastructure versus supporting a more traditional infrastructure. Installation and support for particular Grid-middleware *may* require expertise beyond those of someone familiar with traditional web service-based service-oriented architectures. I emphasize *may* here, since many Grid-based systems are web service-based as discussed in section 3[4]. Furthermore it is not clear what middleware will be hosted on a future EGI if it happens. Or more precisely this uniform middleware distribution (UMD) that has been proposed for EGI does not yet exist. Rather three separate middleware stacks (gLITE, ARC and DEISA) exist largely independently. Thus it is not clear if this new software stack will be any easier or more difficult than previous versions for development, deployment and maintenance.[5]

Irrespective of the software stack that is used, it is clear that CESSDA needs to address security in an open, collaborative but secure manner. Traditional service-oriented architecture-based models have a variety of solutions for security with numerous standards and approaches put forward as outlined in Annex 3 of [CR1]. Personnel are required to ensure that end-end traditional service-oriented architecture-based security is realized. As shown in [CR1] the Grid-model of single sign-on with advanced authorization is supported already.

4.2 Is it possible to charge individual users and/or institutes based on their usage?

The whole area of resource usage and charging is one that many communities are wrestling with, and it is fair to say that at present there are many open issues both with the technologies and with the business and policy models associated with accounting and charging. In the UK for example, the research councils have moved to a model of funding projects based upon full economic costing, where access to and usage of <u>all</u> resources a project requires has to be charged to grant proposals directly. Thus a grant may well request monies for access to and usage of HPC facilities on campus. When collaborations take place where different sites have their own HPC facilities, then trade offs are possible, but this in turn raises issues, e.g. when HPC infrastructures with dissimilar server and/or storage specifications are used, what is the charging model that should be used? Should some sites be able to charge more for access and use of their resources than others? It is the case that for the most part, these issues remain open and ad hoc informal agreements between institutions have tended to be the norm in the UK. Furthermore, these agreements have by and large been based upon time on HPC facilities and not dealt with access to and use of distributed data sets as exemplified by CESSDA.

Nevertheless it is possible to identify how such charging models could be supported, at least from a technical perspective.

4.2.1 Grid Infrastructure Charging

As described in [CR1] the Internet2 Shibboleth technology can be used to access and use a variety of Grid-enabled resources. The typical default model for Shibboleth is based around a core set of *eduPerson* attributes that have been agreed across the UK Access Management Federation. Using information given in these attributes it is directly possible to determine from which site a request arose. Assuming that a given service provider has a charging model associated with it, then this information can then be used for associated charging to that institution for the resources that were used. It is noted that there a variety of charging models can be applied, e.g. a one time charge based upon the access itself, charging based on the time spent on accessing and using the resource etc. The Grid community has put forward a variety of specifications for collecting this accounting information, e.g. the resource usage service specification is one model that is supported across the NGS for example for monitoring time spent accessing and using HPC resources.

---

[4] It is difficult to pin this down since there are a multitude of possibilities and interpretations of Grid. Some Grid middleware, e.g. Globus uses its own Globus container; others, e.g. OMII can be deployed in a Tomcat environment, whilst others do not use any web service container at all.

[5] On a personal note I am wary of UMD as these software distributions are already extremely complex and are evolving in themselves, e.g. gLite continues to change its own software stack with new job management systems proposed and being rolled out, hence merging them to a uniform distribution could well become a yet more complex software stack.

Thus for CESSDA, if an international access management federation were established, then in principle it is possible to charge sites directly for access to and use of CESSDA resources across international boundaries. It is noted that the model presented here is the default model of access and use of Shibboleth enabled resources in the UK which is based upon institutional-level information only. That is, the *eduPerson* attributes have been specifically selected to not identify the individuals, but rather their institution and their role in the institution for example. This anonymous access and use of Shibboleth-enabled resources was deliberate and used to ensure user privacy.

As described in [CR1] however, it is possible to release other information in the signed SAML assertions from an identity provider to a given service provider. Thus it is possible to use the email address, specific roles and/or the distinguished name of the individual in the *eduPersonEntitlement* attribute for example. Indeed in the Grid-context such information is essential to enforce access control decisions as described in [CR1].

Provided consensus was agreed between the CESSDA member organizations on use of such additional information then it is directly possible to know which individuals are using which resources at given sites. Obviously this model only works if sites themselves have agreed what charging agreements are in place *a priori*, and that sites are monitoring and collecting such usage information. In short, *technically* it is quite possible to identify resource usage from the user level to the institutional level. However, many of the issues with charging are not simply technical. Trade offs between organizations and/or the issues of academic usage of resources are real concerns. Thus many academics will simply not access and use a resource if they have to pay for it, irrespective of the cost. As such many institutions in the UK have yet to enforce all policies on access and usage. I note that at Glasgow, we now allow all local (Glasgow) researchers free access to our HPC infrastructure despite initially having an institutional charging policy in place.

It is also worth noting that many models of charging are also based upon a virtual organization itself being charged. Thus a given project using the CESSDA resources could be charged based upon resources they have used. This is achieved through recognition of the VOMS organizations and accounting information being captured based upon individuals involved in that virtual organization for example.

4.2.2 Cloud Infrastructure Charging

Cloud infrastructures offer an immediate way to measure the amount of resources used on the cloud since the customer has to pay for this directly on an on-going basis. However, the purchaser of the cloud resources is typically dissimilar to the actual end users of the cloud resources that have been made available. Thus should a given CESSDA organization buy a virtual server from Amazon say, then they will pay for this server themselves, but the actual usage of the resources deployed on this server has to captured and accounted for. There is nothing implicit across different cloud providers that can be used for this purpose. Rather the usage and accounting information on access to this resource has to be captured by whoever has set up the virtual server.

4.2.3 Own Infrastructure Charging

CESSDA could in principal capture information on access to and usage of distributed CESSDA resources through capturing information in a variety of ways from different institutions/individuals. Traditional web monitoring tools such as GoogleAnalytics allow for recording individual information on web access and usage across the internet. For service-oriented architectures where services have username/password or similar authentication-oriented access, e.g. exploiting WS-security models, then it is directly possible to determine the identity of the individuals accessing the given resources and subsequently use this for accounting and charging.

Alternatively, if the services are made available through web service clients existing in a portal container or some other web accessible resources, then it is equally possible to capture individual/institutional information through access models such as Shibboleth as outlined in 4.2.1.

4.3 Where can CESSDA apply for funding for the maintenance and sustainability of the infrastructure?

There are a multitude of funding streams that could be used to continue the CESSDA e-Infrastructure. As identified in sections 1-3, if CESSDA does engage in European-wide e-Infrastructure initiatives such as the EGI then it is highly likely that there would be numerous possible further funding streams associated with this. The European frameworks and roadmaps for efforts such as ESFRI are being defined and there is a real and timely possibility for CESSDA to capitalize upon this.

With the Grid and associated e-Infrastructures it is equally possible to explore other funding streams crossing wider/global initiatives. The US cyber-infrastructure and efforts such as TeraGrid are supported by numerous funding efforts and streams. Indeed, the monies for scientific research in the US have grown significantly with the change of administration. The National Science Foundation and other national efforts will thus continue to be a key source for international research funding streams. Numerous international collaborations have taken place with countries such as Japan, China, Malaysia and Australia based on joint-funding initiatives based on exploitation of e-Infrastructures for collaborative research.

At a national level, different countries will continue to have their primary research funding streams. In the UK the JISC and ESRC would be the prime targets for continued support of international CESSDA e-Infrastructure efforts. If CESSDA is engaged in wider collaborations using resources such as the EGI, then it may well be the case that funding councils such as the EPSRC would be a further source of funding. It is noted also that many research councils in the UK and internationally are looking more and more at funding streams for inter-disciplinary research. As an example, the recently funded NeSC project, Scottish Health Informatics Platform for Research (www.scot-hip.ac.uk) was funded through a grant comprising monies from the Wellcome Trust, EPSRC, ESRC and the MRC in the UK. This builds upon work done in previous NeSC projects such as the MRC-funded VOTES project (www.nesc.ac.uk/hub/projects/votes). Given this, if it is shown that the CESSDA e-Infrastructure can be used to realize inter-disciplinary research collaborations, then in principle all research domains could be seen as potential funding streams, e.g. the geospatial sciences, the biological sciences, the clinical sciences etc.

Furthermore, as also identified in section 2, despite the numerous current deficiencies in cloud-based infrastructures, there is a huge push both from industry and the academic community in supporting research on all aspects of clouds and usage and exploitation of clouds. This push is likely to continue for some time to come and offer a variety of new funding streams that CESSDA may exploit.

If CESSDA decides to continue with its own e-Infrastructure, i.e. non-Grid or cloud-based, then it is likely that existing and known funding streams would continue to be sought. This includes upcoming EU funding streams such as ESFRI and funding streams from national countries, e.g. the JISC and ESRC in the UK. I would suggest that the future funding streams open to CESSDA would be diminished by this technological choice however.

4.4 What can be expected of the longevity of the chosen technology?

This is an extremely difficult question to answer in the current climate (ever?) hence the responses below can be regarded as informed opinions only.

4.4.1 Grid Technologies

In the past 8 years since the UK e-Science core program began, there have been a raft of standards, technologies and initiatives that have been put forward. Internationally the work in building large scale distributed systems has been going on for decades[6]. Recent moves to Web 2.0 technologies have been driven at least in part by the complexity of Grid middleware offerings. The question is, will the current incarnation of Grid middleware exist or will it be overtaken by newer simpler offerings, e.g. Web 2.0.

[6] My PhD was on open distributed systems and many of the ideas, challenges and solutions put forward have not evolved greatly.

My own personal feeling is that the Grid will continue in some form for a variety of reasons. The EU has invested huge amounts of money in projects such as EGEE and I find it inconceivable that it will simply stop pursuing these efforts. Indeed for the particle physicists without the EGI, it could be argued that the whole LHC experiment could be in jeopardy. If the EGI happens, it is thus highly likely that a significant Grid-oriented e-Infrastructure will persist for the next 5 years at least.

There are also a range of issues with recent technologies such as Web 2.0 where they have not addressed key challenges that the Grid community has been wrestling with for some time: dynamic virtual organizations, single sign-on, delegation of authority, advanced authorization and access to large scale heterogeneous resources such as HPC facilities being just some of the issues not yet adequately addressed.

Furthermore, it is the case that numerous countries continue to support national initiatives in the Grid-domain. As noted the EPSRC and JISC in the UK have recently agreed to continue to support the NGS and the OMII for example. Similar efforts continue internationally also, e.g. the German Grid efforts www.d-grid.de as one example.

Having said this, I also believe that at least some Grid efforts are starting to run out of steam. The major funding streams for Grid-related infrastructures and research in the UK are no longer at the scale that they once were[7]. The huge push to standardized Grid technologies and APIs as organized by efforts such as the Open Grid Forum are also starting to show signs of fatigue. Fewer people attend these events and the standards work itself is (in my opinion!) becoming more fragmented with fewer people involved in the specification of the standards being created.

There are also a variety of other approaches and solutions that people are pushing as the possible answers to data challenges facing numerous communities. The semantic web has been one area in particular that has seen a major effort to solve the many data discovery, data access and integration problems that exist in and across domains. However development of ontologies and tools to best exploit them remains a challenge requiring widespread agreements from domains that is often non-trivial to agree upon.

4.4.2 Cloud Technologies

Many of the cloud based technologies are not a new thing. Virtualization has been supported for a considerable time and numerous technologies for virtualization now exist.

My own personal opinion on clouds is that I remain largely skeptical. I see them primarily as a way for business to make money as opposed to them offering a solution that the research community requires addressing. Having said this, I am also not naïve. I know that where business and industry goes, then this is also where academia tends to follow. This was not unlike the major efforts of HP, IBM, Microsoft, Oracle, Sun amongst numerous others in pushing Grid technologies. All of these organizations had business–oriented reasons to push Grid-related efforts.

It is difficult to determine how long the burst of interest in clouds will continue. Many major players such as Amazon and Google are making concerted efforts in pushing the cloud paradigm. However as identified in section 2, aligning CESSDA entirely with clouds or a given cloud provider would be risky at best. Thus, there is no guarantee that a given provider will continue to offer the services that it currently makes available, or indeed whether the charging models on offer now will continue in the future.

4.4.3 Service-oriented Architecture Technologies

Should CESSDA decide to continue to develop and support its own non-Grid/cloud-based e-Infrastructure then there is little danger of the technologies becoming obsolete. Many of these technologies are driving the internet itself and have a vibrant open source community, e.g. the Apache foundation. Having said, this the next wave of technologies on the internet are also changing. Web 2.0 technologies, mash-ups and the increasing adoption of social networking sites and related technologies will continue (it seems?) to spread the access to and use of web-accessible information to a wider and wider community.

[7] This is not altogether surprising since the UK invested over £250m in the e-Science core program.

5. Conclusions

In this report I have provided an account of the strengths, weaknesses, opportunities and threats associated with Grids, clouds and traditional service-oriented architectures in the context of the future CESSDA research infrastructure. This report builds upon an earlier report to CESSDA [CR1] which gave an overview of the technologies and opportunities offered by Grid technologies for a future CESSDA e-Infrastructure.

In this report I have tried to be honest and objective and not biased with regard to any particular technologies. I have deliberately not tried to state that the Grid is the answer to all problems that CESSDA faces – it isn't and indeed many technologies I would suggest are completely orthogonal to the kinds of problems that CESSDA is facing. My main concern with regard to efforts such as EGI is that it is appearing to implicitly assume a starting point of EGEE and hence technologies like gLite. This technology has been established primarily for HPC-oriented domains and is especially complex.

That said I believe that the Grid and its use to develop e-Infrastructures can address many of the needs of the CESSDA RI including seamless access to federated data sets; single sign-on security models and exploitation of wider computational resources for larger scale analysis. However I fully recognise that it is the case that a multitude of choices and opportunities exist in this space right now. There are many unknowns – will the EGI actually happen? Will the social sciences and efforts such as CESSDA have any possibility to influence and direct EGI resources to the social sciences? It would be remiss of CESSDA to ignore EGI if it happens, but at the same time it is essential for CESSDA to not simply adopt a technology that is not solving the fundamental problems of CESSDA in tackling seamless access to and integration of heterogeneous distributed social science data sets.

Many of the issues facing CESSDA can only realistically be determined once a detailed design and implementation of the future CESSDA e-Infrastructure has been undertaken. Thus when determining how many personnel would be required to incorporate a given CESSDA member organisation resources into a European-wide e-Infrastructure this depends greatly upon the architecture and design of the infrastructure itself. Would all organisations require in-house Grid related personnel? In section 4, I proposed a hypothetical n-tier architecture where secure access to services and resources is achieved through user-oriented portals to tease out possible resource requirements. This will be greatly influenced by whether CESSDA wants a single portal for Europe giving access to international services and data sets; a federation of national portals (offering single sign on between them) or some other hybrid solution. All of these scenarios are possible from a technical perspective, but the precise specification is needed before any realistic personnel and hardware resource specification can be given.

To a great extent, the success of any given infrastructure and the longer term sustainability of this infrastructure stems not from the infrastructure itself or the technologies that underpin it, but in the user uptake of the solutions that it makes available. This is one of the reasons that the UK e-Science NGS has received continued funding. It is not the technologies that they have used, but the communities that they continue to support and indeed growing these communities and embracing other disciplines. To this regard, CESSDA is in charge of its own sustainability destiny and the technology is not the ultimate barrier to its success.

My final conclusion on this report is that it is not always necessarily a direct choice that exists on technologies, i.e. either: Grid, cloud or traditional service-oriented architecture technologies. It is possible to develop hybrid solutions composing web and Grid services for example. Although I note that combinations of these are made more difficult especially when dealing with security for example. Furthermore, many researchers groups are now looking at cloud technologies and doubtless inter-operability between cloud-based approaches and Grid/web services will be supported.

If I had to recommend anything to CESSDA then it would be to have pilot projects exploring realistic case studies applying given Grid, cloud and service-oriented architecture technologies and independently determining whether a technology is fit for CESSDA purpose.

References

[CR1] R.O. Sinnott, *Possibilities and Implications of Grid-enabling Social Science and Humanities Data Collections in the Context of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure*, First Report Drafted as part of the CESSDA Preparatory Phase Project (CESSDA PPP), February 2009.

[EGI] European Grid Infrastructure Design Study, EGI BLUEPRINT EU DELIVERABLE: D5.4 version 3.0, http://www.eu-egi.eu/blueprint.pdf, December 2008.

[SOAP] Simple Object Access Protocol (SOAP), Version 1.2 Part 1: Messaging Framework (Second Edition), W3C Rec. April 2007, http://www.w3.org/TR/2007/REC-soap12-part1-20070427

[REST] Representational State Transfer (REST), R.T. Fielding, *Architectural Styles and the Design of Network-based Software Architectures*, Doctoral dissertation, University of California, Irvine.

[HTTP] R. Fielding, J.Gettys, H. Frystyk, L. Masinter, and T. Berners-Lee, RFC 2616: Hypertext Transfer Protocol, HTTP/1.1, Jun. 1999, http://www.w3.org/Protocols/rfc2616/rfc2616.html.

Annex 1: CESSDA PPP Report Specification

This report was based upon the following outline specification of requirements of a future CESSDA cyber-infrastructure as given in the original tender document.

CESSDA needs to take into account resource-related and sustainability-related preconditions and consequences when deciding on the base-technology of their new cyber infrastructure. We want to know what effects choosing a grid based solution will have on cost, effort, maintainability and future developments, when compared to traditional solutions like client server based infrastructures and/or non-grid service oriented architectures. For this report, we assume there are three possible scenarios for the implementation of the new CESSDA cyber infrastructure:

1. implementing our own infrastructure based on existing client-server or SOA models,
2. joining an already ongoing initiative and sharing resources with other disciplines using Grid technologies,
3. using cloud computing services.

We are not interested in the technical or conceptual differences between these three scenarios. We only want to know the differences between them as far as the following variables are concerned:

- What will be the initial and annual cost for the CESSDA consortium?
  - _ Cost breakdown on:
  - _ Hardware/Cost of ownership (Acquisition, Maintenance, Replacement, Support)
  - _ Software/Cost of ownership
    - Which software packages are needed? (server containers, middleware, etc.)
    - Licenses, In-house/Custom implementation
    - Installation, Support
    - Administration (authorization, authentication)
  - _ Personnel
    - Technical (for maintenance, installation, etc.)
    - Support (for helpdesk, technical management, etc.)
    - Coordination (security-policy, general management)
  - _ Usage
    - Network resources/Bandwidth, Power consumption,
    - Wear and Replacement of defective systems
  - _ Education and Training
    - Developers (middleware, API's, etc.)
    - End-users (workflow applications, data-storage, etc.)
    - Control (management systems, monitoring, etc.)
    - Administration (usage, costs, reporting-tools, etc.)
- How can the CESSDA consortium cover the annual costs?
  - _ Is it possible to charge individual institutes based on their usage?
  - _ Is it possible to charge individual users based on their usage?
  - _ Where can CESSDA apply for funding for the maintenance and sustainability of the infrastructure?
- What can be expected of the longevity of the chosen technology?
  - _ Overview of ongoing development initiatives (e.g. Open Grid Forum, gLite (EGEE), Apache Axis, Amazon EC2/S3 etc.)
    - Current and long-term expected effort of initiatives
    - Current deliverables and status
    - Possible upcoming trends and threats
  - _ When joining an already existing initiative:
    - What can be expected of the longevity of the initiative itself?
    - Does the initiative have sustainability measures, policies and/or guidelines?
    - For how long will its members support the infrastructure?