Demonstrating Repository Trustworthiness through the Data Seal of Approval

by Stuart Macdonald[1], Ingrid Dillo[2], Sophia Lafferty-Hess[3], Lynn Woolfrey[4], Mary Vardigan[5]

Abstract
This paper is a summary of a panel session which consisted of five presentations given on trusted digital repository certification through the Data Seal of Approval (DSA) at IASSIST 2015 in Minneapolis. The paper begins with an overview of the DSA complemented by case studies illustrating how archives undertake the process of certification and concludes with future plans.

Keywords
DSA, Data Seal of Approval, trusted digital repository, certification, data stewardship, digital preservation

Introduction

The Data Seal of Approval: A fitting label for trustworthy data repositories

(Ingrid Dillo, DANS)

The Data Seal of Approval is a basic, transparent process for digital repositories to certify that they are sustainable and trustworthy. Assessments are conducted first internally by a repository and then reviewed by community peers. Assessments help data communities – producers, repositories, and consumers – increase compliance with an awareness of established standards.

Data Seal of Approval offers a basic, lightweight certification standard.

National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data.

If we want to share data, the long-term storage of those data in a trustworthy digital archive is a sine qua non. Data created and used by scientists should be managed, curated and archived in order to preserve the initial investment in collecting them. Researchers must be certain that research data provided by the archives for secondary use remain useful and meaningful, even in the long term.

The concept of sustainability is challenging and crosses several dimensions: organizational, technical, financial, legal, etc. Certification can be an important contribution for ensuring the reliability and durability of digital archives and, hence the possibilities for sharing data, over a long period of time.

The Data Seal of Approval offers a basic, lightweight certification standard. The DSA enables any organization, regardless of size or staff, to quickly self-assess how they are performing compared to data community standards.

The DSA, developed by DANS (Data Archiving and Networked Services) in the Netherlands, was first presented at the first African Digital Curation Conference in 2008. The DSA assessment criteria were initially developed for use in the Netherlands, but were soon found to be very useful in an international context as well. Thus, in 2009 the DSA was transferred to an international body, the DSA Board, which has since managed and further developed the guidelines and the peer review process.

The DSA aims to safeguard data, to ensure high quality and to guide reliable management of data for the future without requiring the implementation of new standards, regulations or heavy investments. The Data Seal of Approval:

- Gives researchers the assurance that their data will be stored in a reliable manner and can be reused;
- Provides funding bodies with the confidence that research data will remain available for reuse;
- Enables researchers to reliably assess the repositories holding the data they want to reuse; and
- Supports data repositories in the efficient archiving and distribution of data

There are 16 guidelines in the DSA: three focusing on the data producer, three on the data consumer, and ten on the data repository.

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.

2. The data producer provides the data in formats recommended by the data repository.

3. The data producer provides the data together with the metadata requested by the data repository.

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including: when applicable, regulations governing the protection of human subjects.

6. The data repository applies documented processes and procedures for managing data storage.

7. The data repository has a plan for long-term preservation of its digital assets.
8. Archiving takes place according to explicit work flows across the data life cycle.

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

10. The data repository enables the users to discover and use the data and reference them in a persistent way.

11. The data repository ensures the integrity of the digital objects and the metadata.

12. The data repository ensures the authenticity of the digital objects and the metadata.

13. The technical infrastructure explicitly supports the tasks and functions described in internationally-accepted archival standards like OAIS.

14. The data consumer complies with access regulations set by the data repository.

15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.

16. The data consumer respects the applicable licences of the data repository regarding the use of the data.

The DSA guidelines

The guidelines are based on five criteria: the data can be found on the Internet; the data are accessible (clear rights and licenses); the data are in a usable format; the data are reliable; and, the data are identified in a unique and persistent way so they can be referred to.

Obtaining the DSA involves two stages. First, the repository conducts a self-assessment, documenting and compiling evidence of compliance into an online tool. A community peer then evaluates this self-assessment by confirming and validating the evidence. The self-assessment, including all evidence, will only be published on the websites of the DSA and the applicant's data repository after the DSA has been awarded. Since approved applications, including any evidence and peer review comments, are publicly available on the DSA website, they can be used as references or samples. This openness fosters trust and accountability as the assessment is accessible by all stakeholders.

Today a total of 55 Seals, including 8 renewals, have been awarded, and some 45 digital archives are working on their DSA self-assessments. This steady growth shows that there is a clear demand for a less resource-intensive approach to certification of trustworthiness of digital archives.

The DSA is also a good springboard for repositories interested in completing the more rigorous and comprehensive assessments and certifications available such as the ISO16363 Audit and certification of trustworthy digital repositories. In addition, the DSA can be used as a roadmap and a planning tool for repositories that are just getting started.

Case Study 1 Odum Institute for Research in Social Science, University of North Carolina

Sophia Lafferty-Hess

Introduction

The H. W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill was founded in 1924 making it one of the oldest social science research institutes in the United States. The Odum Institute Data Archive was established in the late 1960s and today supports researchers' data management needs throughout the research lifecycle, including providing a trustworthy repository for the long-term preservation and dissemination of data. Because of its commitment to data stewardship, the Odum Institute Data Archive has been actively working to demonstrate its trustworthiness through transparent policies and procedures, self-assessments, and certifications.

As part of this initiative, the Odum Institute Data Archive successfully applied for, and was awarded, the Data Seal of Approval (DSA) in September 2013. The primary rationale for applying for the Seal was that it provides a transparent and public method to display to the broader community the Archive's commitment to following data curation standards and best practices through an external peer review process. The Data Seal of Approval also plays an important part in the Archive's overall strategic plan for self-assessment and ongoing improvement.

Data Seal of Approval Application Process

Successfully completing the application for the DSA involved a four-step process that included: education, assessment, documentation, and submission. The education step included building familiarity with the DSA guidelines and criteria, examining successful applications

published on the DSA website, and reviewing pertinent standards referenced in the guidelines (i.e., the OAIS Reference Model). The second step involved a comprehensive self-assessment of all current policies, procedures, and technical systems. During this step, policies and procedures were mapped to the relevant guidelines allowing the identification of documentation that needed to be updated or expanded to address the criteria in the guidelines. After completing this assessment, policy, procedural, and system documents were revised as needed and published on the Odum Institute website or internal wiki. Archive staff then drafted responses to the guidelines using this updated documentation. The final step was submitting the online application upon completion of a final comprehensive review by archive staff.

Reflections on Demonstrating Trustworthiness

Completing a successful application for the Data Seal of Approval allowed the Odum Institute Data Archive to not only demonstrate compliance with data stewardship best practices, but also provided an opportunity to reflect on the role of self-assessments and certifications in the broader context of their organization and the data curation field. Detailed below are some key reflections as well as the Odum Institute's plans for demonstrating trustworthiness into the future.

Demonstrating Trustworthiness is a Continuous Process: Demonstrating a repository's trustworthiness is not a simple task and requires commitment to continually striving towards transparency and improvement in procedures and systems. The DSA provides a useful first step for demonstrating trustworthiness through a relatively lightweight certification process.

Certifications Facilitate Structured Periods for Assessment: Often a repository can become so busy "doing data curation" that it requires conscious effort to assess whether procedures align with current best practices and standards within the field, which can be a moving target as technological and workflow developments continue to emerge. Certifications, such as the DSA, provide an ideal mechanism for repositories to slow down, take stock, modify procedures and systems, and update documentation in accordance with new developments.

Demonstrating Trustworthiness Benefits from a Supportive Community: The DSA community-driven structure benefits the entire data curation community by providing a low-cost method for data repositories to demonstrate their trustworthiness. Community members contribute by participating in the peer review process and these relatively nominal contributions, taken as a collective whole, make a significant impact on the data curation field and repository users who reap the benefits of trustworthy repositories.

Plans for Continuing to Demonstrate Trustworthiness: Since demonstrating trustworthiness is an ongoing process, the Odum Institute Data Archive plans to continue to strive for transparency and improvement through assessments and certifications. This plan includes renewing our DSA when updated guidelines are released and completing a self-audit using ISO 16363 in preparation for undergoing an external formal audit. The Odum Institute Data Archive will also support the broader DSA community through participation within the DSA General Assembly.

Case study 2 - Cornell institute for Social and Economic Research (CISER)

Stuart Macdonald

Introduction

The Cornell Institute for Social and Economic Research (CISER) was founded in 1981 and is home to one of the oldest, university-based social science data archives in the United States. Its mission is to anticipate and support the evolving computational and data needs of Cornell researchers throughout the entire data life cycle. The data archive houses an extensive collection of public and restricted-use numeric data files to support quantitative research in the social sciences with particular emphasis on studies that match the interests of Cornell researchers: demography, economics and labor, political and social behavior, family life, and health.

CISER Data Archive was awarded the Data Seal of Approval in July 2014.

DSA application process and approach

CISER have long been committed to long-term archiving and providing access to scholarly research data in a sustainable way and trustworthy manner. Formalisation of this commitment through the Data Seal of Approval self-assessment process commenced with a review of documentation and existing case studies (Archaeology Data Service and Finnish Social Science Data Archive) in November 2013.

This was followed by a scoping exercise primarily to gain familiarity with DSA regulations and compliance statements (detailing quality aspects with regard to creation, storage and reuse of data as it applies to data producer, consumer and archive or repository) as well as 16 guidelines underpinned by the following criteria that determine whether or not data may be qualified as being sustainably archived:

- data can be found on the Internet
- data are accessible
- data are available in a usable format
- data are reliable
- data can be referred to.

A cross-section of successful DSA applications were consulted that were identified as giving sufficient breadth of data archival practice in addition to providing discipline-specific guidance. Elements deemed relevant and pertinent to CISER Data Archive practices were collected and collated in a series of spreadsheets in association with relevant requirements from the Applicant Manual for each statement. Statements were then assigned to members of staff with particular expertise (storage, security, formatting, restricted data, metadata) and discussed in weekly meetings (1-2 hours) with separate meetings held to discuss in more depth individual assignments as required. Separate meetings were also held to update policies and craft new policy documents to underpin archival process and workflow where they didn't already exist. The application process from scoping to submission took approximately 12 person weeks (principally that of the Data Services Librarian plus colleagues).

Observations and lessons

To kick-start the application process 'quick wins' were identified and used to seed the submissions document such as referencing of existing policies, agreements, terms of use, guideline 0. Information gathering and evaluation was in part an iterative process with knowledge, workflow and procedure being 'scattered' across the organisation, existing inside people's heads, in technical documentation and legacy printed material (including policies), and in internal and external online links. As such, it was easy to underestimate the time required to assemble and craft new policies (such as Preservation and Storage, Security, Versioning, Data Collection), mission statement, and other public facing documentation as evidence to support the application. Proofreading, consistency of language, terminology and narrative also took time to rationalise, bearing in mind the 'different voices' of staff member experts.

Organisational and community benefits

A number of organisational and community benefits were gained from the Data Seal of Approval application process, namely:

- clarification and articulation of organisation's archival practices;
- promotion of trust and confidence between the three stakeholders in the data supply chain - producer, repository/archive and consumer are working to a common set of standards and principles;
- easier to conduct future systematic reviews of technical/human processes and procedures;
- better equipped to respond to necessary changes in data stewardship workflows as/when new compliant tools, technologies and standards emerge;
- identification of service gaps and areas for improvement or modernisation in archival process and procedure;
- raise the profile of the archive and preservation with Cornell senior managers;
- provide a holistic overview and perspective on the mechanics of a mature data archive for new archive staff;
- foundation for further institutional Trusted Digital Repository accreditation such as DIN 31644 (34 metrics) and ISO 16363 certificate or TDR Checklist (107 metrics);
- uncover areas of mutual interworking and interaction between archival colleagues for the purposes of streamlining operations;
- contribution to the social science data archiving community and the data stewardship profession by openly sharing processes, workflows and practice.

Summary

In summary, the Data Seal of Approval application process is beneficial as a learning and knowledge sharing experience for archival staff. It also provides the opportunity for an organisation to audit and enhance its archival operations. More importantly however, the Data Seal of Approval is a public pronouncement of an organisation's archival intent, to demonstrate reliable and trusted access to managed research data for the academic community both now and into the future.

Case Study 3 DataFirst, University of Cape Town

Lynn Woolfrey

Introduction

DataFirst's repository was awarded the Data Seal of Approval in 2014, and is the only African institution to achieve this certification to date. DataFirst is based at the University of Cape Town in South Africa, but our repository gives online access to African data for researchers around the world.

DataFirst's Data Service Model and the DSA Guidelines

Getting to the point of DSA certification was enabled by our twin strategies of, first, adhering to standards and, second, communicating regularly with our various stakeholders. The aim is to build confidence in our service as a trusted digital repository. Trust in our service leads to use of our resources for data-intensive research and quality research output, generating further usage and demand. It also encourages data deposits, as data producers gain confidence in our abilities to handle and share their data in a responsible manner.

DSA Guidelines Concerning the Data Life Cycle

DSA guidelines require that data curation should be carried out with a clear mission and according to documented and well-understood procedures. The requirement for repositories to understand and advertise their mission is stated as:

The data repository has an explicit mission in the area of digital archiving and promulgates it (Guideline 4)

DataFirst's mission is to support top quality research on South Africa and other African countries by providing researchers with access to African survey and administrative microdata. This is clearly stated on our website as a mission statement and in other information on the work we do.

Over 14 years of service provision, we have modelled the work of our data service. The model is based on the Open Archival Information System (OAIS) model for digital repositories. This model was originally designed for Space Data Systems. The OAIS has since become the standard for digital archives and is registered with the International Standards Organisation (ISO) as ISO 14721:2012.

The model in Figure 1 shows how we comply with DSA guidelines related to data curation processes undertaken by repositories. These include:

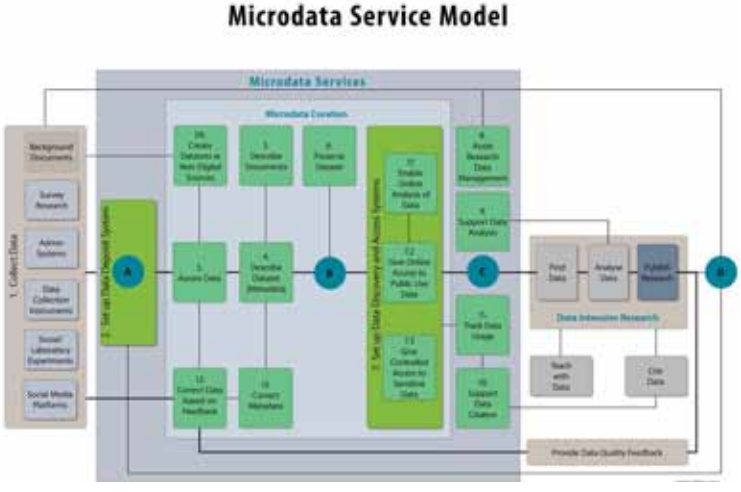

Figure 1. The DataFirst Microdata Service Model

- The data repository applies documented processes and procedures for managing data storage (Guideline 6)
- Archiving takes place according to explicit work flows across the data life-cycle (Guideline 8)
- The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS (Guideline 13)

Below we detail our compliance with the rest of the DSA guidelines, with reference to this model.

DSA Guidelines Related to Data Deposit

The data deposit stage is depicted as Stage 2 in the model. DSA guidelines related to this stage are:

- The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms (Guideline 1)
- The data producer provides the data in formats recommended by the data repository (Guideline 2)
- The data producer provides the data together with the metadata requested by the data repository (Guideline 3)

These guidelines ensure that data users are informed of the quality of the data in the repository. DataFirst provides feedback from data users to data depositors, to address anomalies in the data. The service provides data quality notes on each dataset to highlight issues. Data quality benefits can accrue from this type of independent assessment by academics. DataFirst supports this "virtuous cycle of data reuse" (as depicted in the model).

The DSA guidelines require depositors to provide data in formats that can be used to prepare a usable research dataset. Depositors can provide data in a number of formats. For example, administrative datasets deposited with the Service may be in inappropriate formats. However, DataFirst staff have developed the skills needed to convert these files into research-ready formats.

Another dimension of quality is interpretability, which is dependent on the availability of useful documentation to support sound data analysis (Statistics Canada quality guidelines 2014).

Guideline 3 requires depositors to provide adequate information to help repository managers create useful metadata. DataFirst communicates with depositors on an ongoing basis, to ensure this is the case. This interaction is also designed to ensure compliance with ethical norms. For example, depositors are required to confirm they have ownership of and permission to share the data.

DSA Guidelines Concerning Data Assurance

Data security needs to be assured by those sharing the data, to engender trust among depositors and users. DSA guidelines related to data assurance are:

- The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects (Guideline 5)
- The data repository ensures the authenticity of the digital objects and the metadata (Guideline 12)

These guidelines are fulfilled at the data assurance stage depicted in the model (Stage 3), and also during data preservation (Stage 6). Protection of confidential data is assured through disclosure control routines followed closely at the Service. We also encourage deposits of restricted-access data with our Secure Research Data Centre, which is a controlled environment at the university accessed by approved researchers.

DSA Guidelines on Data Preservation

DSA compliance requires long-term planning by data repositories. The guideline dealing with this is:

The data repository has a plan for long-term preservation of its digital assets (Guideline 7)

While nothing is permanent, DataFirst has operated a Data Service since 2001 and we have the support of our parent institution, the University of Cape Town, to ensure our continued existence and ability to preserve and disseminate data.

The integrity of preservation and dissemination of datasets needs to be protected, as DSA Guideline 11 states:

The data repository ensures the integrity of the digital objects and the metadata (11)

At DataFirst, all iterations of each dataset are stored on a secure server with password access. Checksums are used to ensure the preserved and shared datasets are not altered inadvertently.

DSA Guidelines on Data Discovery, Access and Citation

DSA guidelines dealing with data discovery and access are:

- The data repository assumes responsibility from the data producers for access and availability of the digital objects (Guideline 9)

- The data repository enables the users to discover and use the data and refer to them in a persistent way (Guideline 10)
- The data consumer complies with access regulations set by the data repository (Guideline 14).
- The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information (Guideline 15)
- The data consumer respects the applicable licences of the data repository regarding the use of the data (Guideline 16)

Data accessibility is an important component of data quality. This refers to how easy the data are to find and obtain (Statistics Canada quality guidelines, 2014). DataFirst's data portal enables researchers to discover and access data. Discovery is aided by detailed metadata prepared for each dataset. Metadata is created using Nesstar Publisher, which is free data markup software for the creation of XML compliant metadata. The Publisher software uses the Data Documentation Initiative (DDI) metadata standard. Public and licensed data can be downloaded from DataFirst's online data portal. The portal has been created using the National Data Archive (NADA) software, an open source data dissemination package developed by the World Bank.

Enabling researchers to use the data involves providing good metadata. It also involves answering user queries about the data. Users can contact us through our online support site or Facebook page. The support site is used mainly by early-career academics and postgraduates. Queries range from requests for help with data portal registration and downloads, to questions related to in-depth analysis of the data.

DSA guidelines concerned with codes of conduct and licensing relate to data security standards, as data licenses are a means of protecting data confidentiality. Researchers who download data from our site agree to a standard data usage license. This agreement commits them to preserving data confidentiality and citing data sources. Citing data in a standard manner assists other researchers to find data sources and assess or extend research based on the data. DataFirst provides a recommended citation for each of our datasets, based on the DataCite international data citation standard. On our website we also provide researchers with information on how to cite data in their publications.

Conclusion

This paper describes how we built the Service at DataFirst by complying with data curation standards and by working with depositors to ensure they follow data quality standards. Feedback from stakeholders is also essential to offering a good service. All stages of our data curation process have stakeholder communication dimensions built into them. Interactions with experts in the international data curation community also assists us with best practice. Government, academia and other Data Services are represented on our Board, to provide input to our work. We interact daily with researchers, and this enables us to understand their data needs. Finally, reviews like the Data Seal of Approval process enable us to judge the services we offer against international standards and community best practice.

Future Directions

(Mary Vardigan, ICPSR)

Sustainability

As mentioned above, the rapid uptake of the Data Seal of Approval around the world attests to the need for a basic, low-threshold certification process for repositories. As increasing numbers of repositories are applying for and being awarded the Seal, the DSA Community is growing in other ways as well. While the initiative started in the social sciences and humanities, we are now seeing repositories in the natural and physical sciences applying for the DSA, and the geographic spread is expanding also. These are all positive developments, but they raise the issue of sustainability: how do we ensure the future of the DSA initiative so that we can continue this forward momentum and provide more and more repositories with this validation of their trustworthiness?

The DSA regulations offer a clear path to a sustainable and community-driven organization through the mechanism of a General Assembly (GA). Any repository that has acquired the Seal is eligible to join the General Assembly. Membership in the GA naturally carries both rights and responsibilities. Each General Assembly Member commits to conduct a maximum of three peer reviews a year per DSA repository, thereby receiving voting rights in the GA, which elects the DSA Board and provides advice to the Board when needed. The idea is that any certified repository having earned the Seal itself should have enough expertise to participate in reviewing other repositories. This will enlarge and refresh the pool of reviewers and ultimately strengthen the organization.

As of this writing, the General Assembly had been convened and had elected a new Board. We are confident that this governance mechanism will go a long way toward assuring the stability of the DSA initiative. At the same time we are also evaluating different business models and approaching potential funders. While the DSA organization does not have a lot of overhead, it does need a minimal amount of funding to maintain the DSA assessment tool and website and manage and train the pool of peer reviewers. So far these activities have been undertaken by individuals contributing their time, but such in-kind contributions are not sustainable in the long term and must be bolstered by actual funding, even if minimal.

Common requirements for basic certification

Another interesting development relating to the future of the DSA is a project begun in 2014 to harmonize the guidelines of the DSA and the certification criteria of the World Data System, an interdisciplinary body of the International Council for Science (ICSU) created in 2008. The World Data System has built its own set of guidelines for trustworthy repositories, many of which overlap with the DSA criteria. Thus,

under the auspices of the Research Data Alliance Repository Certification Interest Group, an RDA Working Group was created to bring the requirements of the two certification catalogues together. The Case Statement of the Working Group also calls for the Group to develop common procedures and to create a shared testbed for assessment. The ultimate goal is to create a shared framework for certification that includes other standards as well, such as the nestorSEAL and ISO-16363/TRAC.

Goals of this DSA-WDS harmonization effort are to:

- Simplify the array of certification options
- Show the value of a certification procedure requiring relatively low investment
- Stimulate more certifications
- Foster greater trust in repositories
- Promote data sharing

Representatives from the DSA and the WDS have worked diligently to create harmonized criteria and as of this writing were poised to publish them to the RDA community. The harmonization work included constructing two-way mappings between the standards, analyzing the gaps and commonalities, and then crafting new language to bring the standards together. Certification procedures were also harmonized.

What will this project mean to the DSA going forward? The plan is that in the future both the DSA and the WDS will be using the common criteria and there will be greater collaboration and synergies across the organizations. This is still a work in progress, so stay tuned for more information as the project bears fruit in 2016.

With all of this ongoing activity, it is clear that many exciting changes lie ahead as the DSA expands into new territory. But even more exciting is the steady accumulation of repositories gaining the Seal, coming on board one by one, and together demonstrating the impact and importance of a growing federation of trusted repositories that are protecting valuable digital assets around the world.

Notes

1. Data Services Librarian, CISER, Cornell University. (University of Edinburgh) Email: stuart.macdonald@ed.ac.uk
2. Deputy Director, Data Archiving and Networked Services (DANS). Email: ingrid.dillo@dans.knaw.nl
3. Research Data Manager, Odum Institute for Research in Social Science, University of North Carolina. Email: slaffer@email.unc.edu
4. Data Services Manager, DataFirst, University of Cape Town. Email: lynn.woolfrey@uct.ac.za
5. Assistant Director, Inter-university Consortium for Political and Social Research (ICPSR). Email: vardigan@umich.edu
6. http://stardata.nrf.ac.za/nadicc/presentations/harmsen_henk.ppt
7. http://datasealofapproval.org/en/information/all-documentation/
8. ADS and the Data Seal of Approval – case study for the DCC - http://www.dcc.ac.uk/resources/case-studies/ads-dsa
9. Finnish Social Science Data Archive and the DSA: a case study - http://datasealofapproval.org/en/assessment/fsd-dsa-case-study/
10. http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm