IASSIST Quarterly

VOLUME 40 – Number 3– 2016

Demonstrating Repository Trustworthiness

The Role of Case Studies in Effective Data Sharing

Mitigating Survey Fraud and Human Error

Image Management as a Data Service

Online at: *iassistdata.org/iq*

COLOPHON

IASSIST Quarterly

The IASSIST Quarterly represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The Quarterly is normally published four times per year. Authors are encouraged to submit papers as word processing files (for further information see: *http://www.iassistdata.org/iq/instructions*-authors). Manuscripts should be sent to Editor: Karsten Boye Rasmussen.: Email: *kbr@sam.sdu.dk*

.Announcements of conferences, training sessions, or the like are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event

Editor

Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email: *kbr@sam.sdu.dk*

Deputy editor

Walter Piovesan
Simon Fraser University

Website editor

Michele Hayslett
University of North Carolina at Chapel Hill

In this issue

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137 © 1997 by IASSIST. All rights reserved.

Editor's Notes

Being international - and proud of it!

IASSIST is proud of being international. These days some us of find it important to emphasize how international collaboration has improved and made our lives more efficient. In the small but around-the-globe-reaching world of IASSIST, many national data archives have come into existence as well as continuing their development, through friendly international support and spreading of knowledge and good practices among IASSISTers. So let us cherish the 'International' in IASSIST. We are proud of the lead 'I' for 'International' in the IASSIST acronym and have no intention of changing that to 'N' for 'National'. It is also my impression that data archives all over the world simply don't have the facilities for storing 'alternative facts' as they are shy of all kinds of documentation.

Welcome to the third issue of Volume 40 of the IASSIST Quarterly (IQ 40:3, 2016). Four papers with authors from three continents are presented in this issue.

The paper 'Demonstrating Repository Trustworthiness through the Data Seal of Approval' is a summary of a panel session at the IASSIST 2015 conference in Minneapolis with panel members Stuart Macdonald, Ingrid Dillo, Sophia Lafferty-Hess, Lynn Woolfrey, and Mary Vardigan. The paper has an introduction from DANS in the Netherlands where the Data Seal of Approval (DSA) originated. Cases from the US and South Africa are presented and the future of the DSA including possible harmonization with other systems is discussed. DSA certifications are basically consumer guidance, clearly assisting all the involved parties. Depositors and funding bodies will be assured that data are reliably stored, researchers can reliably access the data repositories, and repositories are supported in their work of archiving and distribution of data.

The second article brings us to the actual use of data. From the UK Data Service, Rebecca Parsons and Scott Summers in 'The Role of Case Studies in Effective Data Sharing, Reuse and Impact' take us into positive narratives around secondary data. The background is that although the publishing of data is now recognised by funders, the authors find that 'showcasing' brings motivation for data sharing and reuse as well as improving the quality of data and documentation. The impact of case studies is all-sided and research, depositing data, and the brand recognition of the UK Data Service are among the areas investigated. The future is likely to include new case studies developed for use in teaching in schools, with easy linking to datasets, as well as for researchers being assisted to build their own portfolios. The appendix presents case studies on research and impact.

In the third article, we are situated in data creation. Muhammad F. Bhuiyan and Paula Lackie from Carleton College in Minnesota write on 'Mitigating Survey Fraud and Human Error: Lessons Learned from A Low Budget Village Census in Bangladesh'. As the 'fraud' term implies, they are looking into the problem of data creators being too creative, but more importantly they are investigating the essential area of data quality. The authors explain how selected technological assets like the use of geographic information systems (GIS) and audio-capturing smart pens improved data quality. The use of these tools is exemplified through many scenarios described in the paper. Furthermore, a procedure of daily monitoring and fast transcription lead to quick surveyor re-training and dismissal of others, thus minimising data errors. For those interested in false data and its detection, the introduction in particular has valuable references to literature.

In the last paper the difficult task of handling images is addressed in 'Image Management as a Data Service' by Berenica Vejvoda, K. Jane Burpee, and Paula Lackie. Vejvoda and Burpee work at McGill University in Montreal. You have already met Lackie from Carleton College in relation to the third paper above. The 'images' in the article are digital images, and the authors suggest that the knowledge of digital data services across the 'research data lifecycle' also benefits the management of digital images. Digital images are numerical data, and the article compares the data, metadata, and paradata of a survey respondent to the information on a digital image. Considerations from normal data concerning system formats and storage space also apply to management of images. In the last section the paper introduces copyright issues that are complicated, to say the least. Just as reuse of normal data can have ethical angles, it is even more apparent that images can have complicated issues of privacy and confidentiality.

Papers for the IASSIST Quarterly are always very welcome. We welcome input from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. When you are preparing a presentation, give a thought to turning your one-time presentation into a lasting contribution. We permit authors 'deep links' into the IQ as well as deposition of the paper in your local repository. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at *http://www.iassistdata.org*.

Authors are very welcome to take a look at the instructions and layout:
http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: *kbr@sam.sdu.dk*. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
January 2017
Editor

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

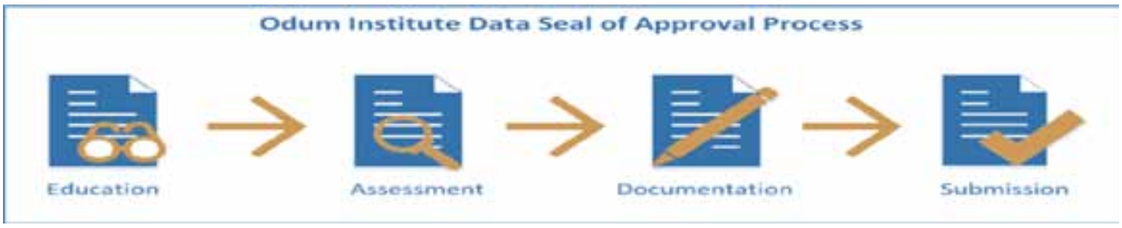
Introduction

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:

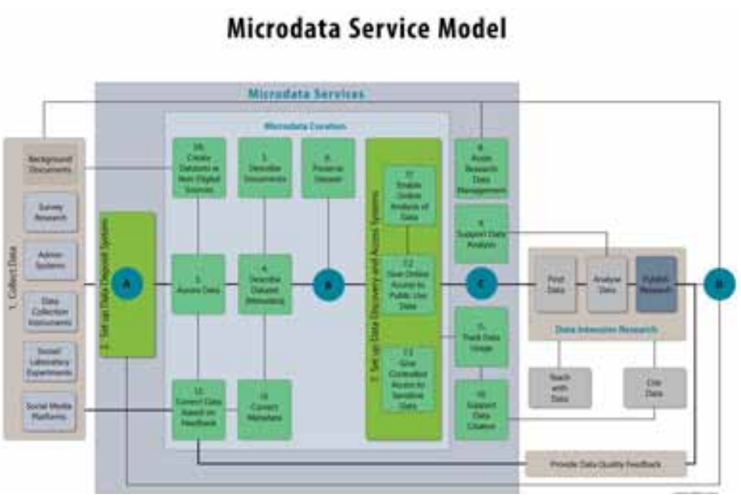


Figure 1. The DataFirst Microdata Service Model

- 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

The Role of Case Studies in Effective Data Sharing, Reuse and Impact

by Rebecca Parsons[1] and Scott Summers[2]

Abstract

The effectiveness and impact of social science research is under constant review. From the sharing, reuse and archiving of social science research data to the outcomes, reach and impact of research, social science professionals are under increasing pressure to realise the maximum potential of their data collections and their research findings. The UK Data Service is playing a key role in supporting researchers in this process and is using detailed and well-received case studies to provide them with guidance on the best practice for sharing and reusing data and also identifying and capturing the impact of research. The impact of research is now routinely considered when examining the 'success' of funded projects, but the reality of identifying and capturing impact can be a challenge. Publishing data in its own right is now recognised as being impactful to funders, yet exposing this narrative through 'showcasing' is under exploited. Such narratives can incentivise others to share data and also improve the quality of the data and documentation. This paper seeks to explore the role that case studies of research can play in this regard. The paper also examines the role that depositor and user case studies can play in enhancing the reuse of a showcased data collection. To achieve this, a variety of illustrative depositor and impact case studies are discussed, highlighting the role that these can have on research projects.

Keywords

Case Study, Impact, Data Sharing, Data Management.

Introduction

This paper explores the use of case studies in a variety of ways; First, it considers cases studies in regards to research and impact. Second, the role they can play in assisting other depositors with depositing data at the UK Data Service. Third, the role they can play in brand recognition and impact for the UK Data Service. Finally, the paper concludes by identifying areas where we can develop the use of case studies further in the future.

The UK Data Service has for a number of years been producing case studies on the data that we hold.[3] These have traditionally been interviews or short vignettes of research papers, linked to the catalogue record of the data, and were mainly intended to show other researchers how particular data had been used. These short case studies described the research question posed, the data used, the methodology applied to the research and the findings and publications produced by the researcher. Appendix 1 gives an example of this type of case study.

Case studies of this type are used primarily to give UK Data Service users accessing data information about research that had already used these data. They provide the basic 'narrative' for a piece of research, but do not look further at whether the research has been applied in practice, whether it has influenced policy or decision making, or whether any follow-up work has been completed.

Over the past few years' we have been looking in more detail at how the UK Data Service is used and have been starting to track the impact of research using data held in the collection. A small team was created in the UK Data Service Communications section, under the guidance of a Director for Communications and Impact, which was tasked to identify and track research impact. Identifying research impact is a time-consuming process and the team at the UK Data Service uses several methods to identify impactful research. The team follow the review documents produced for the national UK government and for local and regional government, policy reviews prepared by several leading UK think tanks and data use in the media. From these we can identify the research publications that have been used for evidence and, if it has come from our collection, track the data use back to the UK Data Service. As well as this backwards identification of impact, the team are also in contact with think tanks and individual researchers about work that they are currently undertaking, which may be impactful in the future. This allows us to follow individual projects as they are happening. We also encourage researchers to submit details of publications and the team can then search to see if they have been used in policy.

The focus on research impact is a relatively recent development in the Higher Education sector of the UK and has moved the definition of 'successful' research beyond simple metrics (for example a paper has been cited 'X' number of times, or a dataset has been downloaded 'X' times), to attempting to define how research has been used and what practical benefits it has brought. This change in measuring successful research has occurred at the same time as the expansion in providing open data, which anyone can download and use. For data which requires users to register with the UK Data Service, it is possible to track the metrics of data use, but for open data we do not record in detail how data are being used. There is now a greater emphasis on making data as open as practicable to allow greater use and reuse, but the focus of funders, research councils and the government is now not on the quantity of data being accessed, but on the quality of the research that comes from the data. Case studies allow us to record the quality of research.

We are still producing shorter case studies to document research, but we are supplementing these with case studies focussing on the impact that research has in the wider world, as well as the role that the UK Data Service plays in developing tools for data use and that share depositor stories for the benefit of other researchers. An example of an 'Impact' case study can be found in Appendix 2. The process of developing these case studies has been eye-opening, not just in seeing how widely the UK Data Service is used and the breadth of research that is coming from the data deposited with us and downloaded from us, but has also given us a greater understanding of the issues facing researchers in using and sharing data. Case studies have enabled us to spread the word about the work that we do, but also analyse how the Service is used and where we can improve our support for data users.

Research and impact

A key feature of case studies is that they allow us to demonstrate the reach and impact of the research that is carried out using data downloaded from, or deposited with, the UK Data Service. Within the UK there is a clear need for researchers to demonstrate the impact of their work, with 'impact' being a key indicator of success when universities are graded under the Research Excellence Framework (REF) – a national system for assessing the quality of research produced by UK institutions.[4] The REF uses case studies as a method of comparing the research output from universities and institutions who have to submit detailed case studies as part of their evaluation process. Universities were last graded by the REF in 2014 and are now familiar with the developing case studies to show impact. Case studies have become a familiar way of documenting the effect that research and projects have and most universities now employ staff specifically to promote research impact in preparation for the next REF in 2020.

Through talking with researchers that use the UK Data Service, we found that the expansion of these institutional impact case studies came with some issues for researchers. This is sometimes because the research has not yet produced any visible impact at the completion of the project, or because the researcher has moved on to new work and does not have the resources to identify the impact of their past work. This is particularly an issue for early career researchers, who may be building their portfolio of work and moving between academic institutions. UK universities are selective in which researchers are submitted to the REF for assessment, so researchers whose projects do not produce impact in a specific timescale can miss-out on institutional support for documenting the impact of their research. The UK Data Service identified this as one area where we could support researchers that use the Service, providing useful evidence for researchers to use in their own career development and, at the same time, boost the profile and use of the UK Data Service. Researchers can submit papers or projects that have used the Service to us and the impact team will follow the outputs of the research to see what impact there has been. The case study is then prepared giving a summary of the research, information about the data that was used and the impact that has come from it. The case study is then used by the UK Data Service to promote the use of the data and is shared with the researcher for them to use too. The development of impact specific case studies has shown the breadth of research that has been undertaken using data from the UK Data Service. Our case studies come from the fields of health,[5] social policy,[6] education,[7] technology use,[8] and business.[9]

Examples from our most recent case studies include, Dr Ivy Shiue who carried out a piece of research that compared health bio-markers in older adults with the temperature of their home and showed that living in a colder home has a detrimental effect on a person's health, a finding which was used in the UK government's Winter Planning Policy.[10] Dr Shiue has worked with the UK Data Service for several years and contacts the impact team when she publishes a piece of work that she thinks could be used by policy makers. Another project entitled 'Onward Migration,' which investigated the experiences of refugees across the UK, was used as evidence by the Scottish Refugee Council on the effect of forced relocation on refugee integration.[11] This research was identified through media coverage of the project that prompted the impact team to study the data used by the researchers and contact them about their work. As a final example, we even have evidence of energy use for English wine production, a piece of research which led to advice for English wine producers on reducing their energy costs and carbon footprint,[12] which was identified following the researcher submitting details of their research paper to the Service.

Assisting other depositors through illustrative case studies

Another use that the UK Data Service has identified for depositor case studies is that of providing assistance to other depositors. Put simply, depositor case studies allow us to point other depositors to illustrative examples of previous users that have deposited data with us and how they managed – and dealt with – any challenges they may have had when depositing data. This has been highly informative not just for potential depositors, but also for the team who engage in research data management training.

For example, one recent depositor case study helps to emphasise the message that we have been promoting for years at the UK Data Service, namely: the importance of planning for the sharing and managing of data early within the research project.[13] Dr Karon Gush's study, which investigated how couples managed their households during recessions did just that, enabling the data to be effectively deposited and shared with the UK Data Service.[14] Dr Gush and her team identified the potential challenges when it came to archiving and sharing the data early on in the project and sought to address them throughout the project so as to prevent future issues. For example, consent for sharing was gained from the participants when they were interviewed because the researchers had already identified that they wished to share the data and were aware that they would need 'fully informed consent' from the participants to do this. When it came to anonymising the data a careful balancing act had to be found so as to ensure that the data was as useful as it could be to future users, whilst still achieving the correct level of anonymity for participants. Dr Gush put it best in the case study interview when she said:

> "Anonymisation should be thought about at the beginning and should be seen as 'part and parcel' of the whole project… and the anonymisation process should be completed as you go along and not left until the end." [15]

This is an important message that we have always promoted. Now we can easily point depositors and users to the case study to highlight this.

Other depositor case studies have allowed us to explore and discuss a variety of other challenges; such as, the problems with archiving complex anthropological data ten years after the research is completed,[16] and how one can gain consent for data sharing and archiving retrospectively.[17]

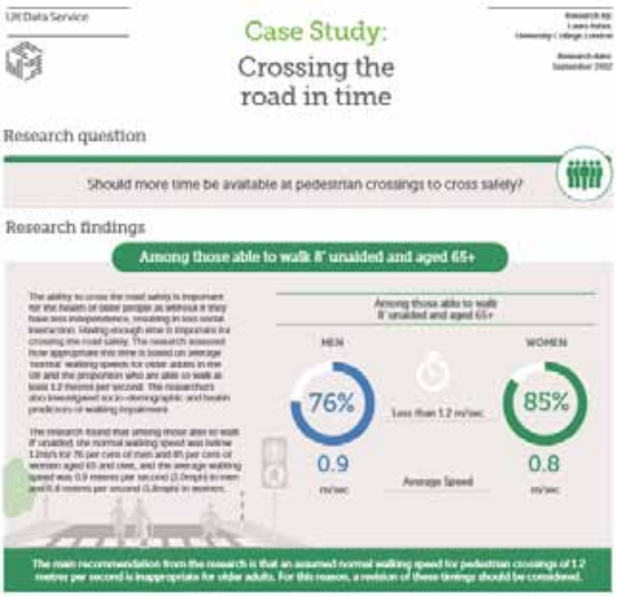
These case studies also help to illustrate the importance of adequate data management training, and planning being undertaken at the beginning of the depositor's research project. By doing so, it ensures that most challenges/issues related to the research, the data, sharing and archiving are identified early on. Thus, potential issues can be adequately addressed, which in turn leads to better quality sharable data. Alongside this, there is a saving of resources for the researcher (both in terms of finance and time).

Utilising depositor case studies in this way has a variety of core benefits for us, our depositors and, our users; Firstly, it allows us to actively encourage data deposits from users by showing how other data has been deposited and shared. Secondly, it provides illustrative examples of how other depositors have addressed potential data sharing issues in practice. Thirdly, it allows data depositors to build collaborations with users by providing further information on their research project and the steps that they went through to deposit their data. For example, this might be the anonymisation or consent procedure(s) they followed. Fourthly (and finally), it allows us as a Service to have impact in a variety of ways. Specifically, it allows us to help other researchers to deposit data with us, share their research and experiences, have additional impact with their research, and provides examples of impact for our funders.

Raising awareness of the UK Data Service

The final benefit to developing a range of case studies has been the collating of real-world evidence on the work of the UK Data Service. As with any publically funded organisation, the Service has a need to prove to our funders and supporters the usefulness of the work that we do. In many ways, 'data' and 'data archiving' can be rather vague terms, particularly for those who do not work in the archiving or social science areas, and explaining the breadth of the work of the UK Data Service to people in the world outside our niche becomes much easier if we can show real examples of the data we hold actually in use. The UK Data Service is funded by the Economic and Social Research Council (ESRC) and they have a remit to share the work that they fund with the general public.[18] The case studies that we have developed for our funders focus both on individual pieces of research using data from our collection and also specific studies on how the Service is being used. These more technical case studies look at the expertise and technology that the UK Data Service itself has and provide evidence for how the Service is used. An example of this would be our case study on InFuse – the interface that provides access to Census data.[19] We developed this case study to show how InFuse was collaboratively developed, who was currently using it and some examples of research that had made use of it. This case study is now used by the ESRC to illustrate the benefit of supporting data infrastructure like the UK Data Service.[20]

For use outside the data archiving community we worked with a data visualisation company to develop visual case studies, based on our impact and research case studies. These are an attempt to distil – on to one page – the research question, findings, methodology and policy implications of a piece of research, for example this case study looks at how long it takes elderly people to cross the road:



UK Data Service

Case Study:

Crossing the road in time

Research question

Should more time be available at pedestrian crossings to cross safely?

Research findings

Among those able to walk 8' unaided and aged 65+

The ability to cross the road safely is important for the health of older people as without it they have less independence, resulting in less social interaction. Having enough time is important for crossing the road safely. The research assessed how appropriate this time is based on average 'normal' walking speeds for older adults in the UK and the proportion who are able to walk at least 1.2 metres per second. The researchers also investigated socio-demographic and health predictors of walking impairment.

The research found that among those able to walk 8' unaided, the normal walking speed was below 1.2m/s for 76 per cent of men and 85 per cent of women aged 65 and over, and the average walking speed was 0.9 metres per second (2.0mph) in men and 0.8 metres per second (1.8mph) in women.

The main recommendation from the research is that an assumed normal walking speed for pedestrian crossings of 1.2 metres per second is inappropriate for older adults. For this reason, a revision of these timings should be considered.

This form of case study does have limitations: it is not suitable for very complex research questions and care needs to be taken that results or methodology are not over simplified by the visual representation. However, we have found these simple case studies useful when introducing the work of the Service, particularly for school students aged 16-18 years old and undergraduate students, who are just starting to use social science data for themselves and need an entry level summary of how some of the data is being used.

Conclusions and the future

In the future we would like to continue to engage with new UK Data Service users by developing case studies specifically for use in schools that can link to datasets suitable for teaching and are based on UK education curriculum. We also plan to develop case studies specifically focussing on early-career researchers, show-casing their work and helping them build a portfolio of impactful research. We will be continuing to add to our case study collection, with a particular focus on the innovative use of data.

In addition, we plan to further develop the use of depositor case studies to assist our other depositors and complement our current research data management training and guidance. In particular, we will be identifying depositors' collections which have 'acutely sensitive data' or are typically seen as 'data that cannot be shared easily', which we can point other users to. We will utilise these not only to illustrate how data can be effectively shared but also how these potential issues can be simply and effectively addressed in practice.

Appendix

1. Example Research Case Study

Health effects of industrial incinerators in England[21]

About the research: Research specific to waste incinerators has provided mixed evidence for the effects of proximity to incinerators on health. Older incinerators have been associated with increased incidence and mortality from selected cancers, while more recent reports show little association. Despite this, the effect of incinerator emissions remains a public health issue. This study assesses whether living close to industrial incinerators in England is associated with increased risk of cancer incidence and mortality.

The results show no evidence of elevated risk for those living in areas containing an incinerator compared to those living in matched areas without an incinerator. Moreover, within areas, there is little evidence of an increase in risk for those living in close proximity to an incinerator compared to those living further away.

About the data: This research draws on aggregate data from the 2001 Census for England. The data were key to the study for the identification of case circles around each incinerator, and for obtaining Lower Super Output Area level population counts, by five-year age groups and gender, for 2001.

Methodology: The researchers considered five regions with industrial incinerators in England, compared with five matched control regions, from 1998 to 2008. Spatial and temporal trends in annual health outcome data within each circular region are analysed. Initially, the researchers used a Poisson log-linear model including age-standardised expected count as an offset and covariates for case-control status, matched pair and deprivation, fitted to circle level data to investigate temporal trends. They later modelled data at a finer resolution – at the Lower Super Output Area (LSOA) level. A Poisson log-linear model was then used, as previously, with additional covariates for the effect of distance from the incinerator in case areas included as a multiplicative factor. Population data was used in the calculation of the age-standardised expected count.

2. Example Impact Case Study

Do higher energy prices affect international trade?[22]

About the research: Dr Misato Sato and Dr Antoine Dechezleprêtre at the Grantham Research Institute on Climate Change and the Environment at the London School of Economics and Political Science have been studying climate change policy and its effects on trade in a research project funded by the European Union Seventh Framework Programme under grant number no. 308481 (ENTRACTE Research Project), the Global Green Growth Institute, the Grantham Foundation and the Economic and Social Research Council (ESRC) through the Centre for Climate Change Economics and Policy.

Emissions trading policies, such as the EU emissions trading system (EU ETS), are regulations implemented by many countries and cities to reduce industrial greenhouse gas emissions cost-effectively. These regulations are a key tool for achieving emissions reduction targets. However, according to the European Commission on Climate Action they can result in carbon leakage and may affect the competitiveness of businesses. In fact, standard trade models suggest that policies that increase energy price put domestic firms at a disadvantage relative to foreign rivals facing lower energy prices. Producers of energy intensive products respond to higher energy prices by producing fewer energy-intensive goods which may lead to a decline in net exports and the partial relocation of production to a region with low energy prices.

In this study Dr Sato and Dr Dechezleprêtre explored whether, and to what degree, changes in relative energy prices might influence trade and competitiveness. This is the first study to analyse the effect that energy costs have on global trade using historical data on trade and energy prices. It analysed 62 business and industry sectors in 42 countries over a 15-year period, from 1996 to 2011, using data that covers 60% of global merchandise trade.

Findings showed that changes in relative energy prices have a statistically significant but very small impact on imports. On average, a 10% increase in the energy price difference between two country-sectors increases imports by 0.2%. The impact is larger for energy-intensive sectors but even within these, the effect is minor - changes in energy price differences across time explain less than 0.01% of the variation in trade flows. The authors calculated that a 30% increase in energy prices across Europe would cause exports to fall by only 0.5% and would increase imports by 0.07%. They concluded that "contrary to some claims, rises in energy prices do not have much effect on the global competitiveness of businesses. Even a sizeable difference in the price of energy relative to the rest of the world has only a very small impact on a country's imports and exports."

Methodology: The researchers estimated the short-term effects of energy price differences on bilateral trade at the sector level using a gravity model. Trade between countries is not only influenced by energy costs but many other factors such as transport costs, labour cost, exchange rates, tariffs, trade agreements, common language and common currencies. The study therefore brings together a variety of datasets to analyse the impact of relative energy prices on trade. The coverage and detailed disaggregation of the data used goes well beyond previous work, allowing the first global ex-post analysis of the relationship between trade and energy prices. Bilateral trade data were taken from the CEPII's BACI database which contains detailed bilateral import and export statistics from the UN Commodity Trade database. The researchers also used a unique and comprehensive dataset of industrial energy price indices at the country and sector levels covering 48 countries and 12 industry sectors for the period 1996 to 2011. The dataset was constructed in Sato et al., 2015 and uses data from the International Energy Agency World Energy Balances, the International Energy Agency Energy Prices and Taxes and the World Bank, as well as other sources. The procedures used to construct the dataset including the methodology developed to reduce missing data-points, are documented in the working paper, as are the full set of original data sources. This dataset is made publicly available for download here. Additionally, the researchers used data on GDP and population obtained from the International Monetary Fund's World Economic Outlook (2012) and data on wages obtained from the United Nations Industrial Development Organization (2011).

Findings for policy: This study finds unique evidence suggesting that concerns about the risks of carbon leakage may have been overplayed. Carbon pricing policies are intended to induce energy intensive industry to reduce carbon emissions by making it more expensive to pollute. However, if the carbon price is set too high, there is a risk that companies will respond by relocating production to regions with lax policies, rather than clean up their production. Whether this 'carbon leakage' occurs is an important issue in the debates around how to design emissions trading schemes, and whether or not leakage occurs is a question in much need of robust empirical evidence.

Impact of the research: The study's findings about the risks of carbon leakage have been included as research evidence in the 'Evaluation of the EU ETS Directive' report carried out by the European Commission in November 2015. The evaluation subsequently informed policy measures implemented by the Commission regarding the revision of the EU ETS Directive, as set out in the framework of measures of the conclusions of the European council in October 2014.

The study has also informed the following reports:

- The OECD Environment Working Paper No. 87, a review of literature on ex post empirical evaluations of the impacts of carbon prices on indicators of competitiveness.
- The OECD Economics Department Working Papers No. 1282, "Do environmental policies affect global value chains?"
- The New Climate Economy Working Paper 'Implementing Effective Carbon Pricing', which was written as a supporting document for the 2015 report of the Global Commission on the Economy and Climate, Seizing the Global Opportunity: Partnerships for Better Growth and a Better Climate
- Methods for evaluating the Performance of Emissions Trading Schemes, a Discussion Paper by Climate Strategies, prepared as part of the project "Evaluation of Emissions Trading Scheme Pilots in China" funded by Children's Investment Fund Foundation (CIFF) and executed by the Tsinghua University.

Notes

1. Rebecca Parsons, UK Data Service, UK Data Archive, University of Essex, Colchester, Essex, UK. rparsons@essex.ac.uk
2. Dr Scott Summers, UK Data Service, UK Data Archive, University of Essex, Colchester, Essex, UK. ssummers@essex.ac.uk
3. UK Data Service (2016), 'Case Studies' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-studies
4. Research Excellence Framework (2014), 'REF 2014' http://www.ref.ac.uk/
5. UK Data Service (2016), 'Nearly a third of Welsh adults struggling to cope with the pain of chronic health conditions' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=192
6. UK Data Service (2015), 'Children with psychological distress are more likely to become unemployed' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=168
7. UK Data Service (2015), 'Exploring the 'middle' in GCSE attainment' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=179

8. UK Data Service (2015), 'Screen-based media and well-being in adolescence' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=177
9. UK Data Service (2015), 'Investigating external and private benefits from investment in skills and training: UK innovators study' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=194
10. UK Data Service (2016), 'The impact of cold homes on health' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=195
11. UK Data Service (2016), 'Moving on? Dispersal policy, onward migration and integration of refugees in the UK' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=191
12. UK Data Service (2015), 'Energy use within the English wine production industry' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=176
13. UK Data Service (2016), 'Karon Gush – Depositor Stories' https://www.ukdataservice.ac.uk/deposit-data/stories/gush
14 ibid
15. ibid
16. UK Data Service (2016), 'Pat Caplan – Depositor Stories' https://www.ukdataservice.ac.uk/deposit-data/stories/caplan
17. UK Data Service (2016), 'Maggie Mort – Depositor Stories' https://www.ukdataservice.ac.uk/deposit-data/stories/mort
18. Economic and Social Research Council (2016), 'ESRC Shaping Society' http://www.esrc.ac.uk/
19. UK Data Service (2011), 'InFuse' http://infuse.ukdataservice.ac.uk/
20. Economic and Social Research Council (2016), 'Opening up census data for research' http://www.esrc.ac.uk/news-events-and-publications/impact-case-studies/opening-up-census-data-for-research/
21. UK Data Service (2014) 'Health effects of industrial incinerators in England' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=163
22. UK Data Service (2016), 'Do higher energy prices affect international trade?' https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=206

Mitigating Survey Fraud and Human Error:

Lessons Learned from A Low Budget Village Census in Bangladesh[1]

by Muhammad F. Bhuiyan[2] and Paula Lackie

Abstract

The paper suggests effective strategies for collecting high quality data in developing countries based on lessons learned from implementing a household level census of three villages in Bangladesh. In particular, we focus on low cost but effective techniques for reducing survey fraud (e.g. curbstoning) and human error (e.g. transcription errors) in conducting face-to-face questionnaire-based interviews by hired surveyors. We find the following strategies to greatly improve data quality: use of a geographic information system (GIS) and audio-capturing smart pens; daily monitoring; surveyor retraining; and swift firing of those showing consistent errors in judgment. Transcribing the data as soon as the surveys were completed helped locate and contain human errors, as well as fraudulent activities. The techniques suggested here are geared towards prevention of errors, rather than detecting fraud during post-survey validation.

Keywords

Curbstoning, Survey Fraud, Low Budget Survey, Developing Country, Data Quality

Introduction

"[i]n spite of early evidence of cheating in market research and other fields where professional surveyors are employed ..., the problem of surveyor cheating has largely been ignored in recent literature." – Harrison and Krauss (2002)

The process of mitigating data falsification is more difficult than it may seem at first. While the challenges of collecting reliable interview-based data are well known (Biemer and Stokes, 1989; Crespi, 1945), the literature on dealing with these challenges, particularly with regard to falsified data (Harrison and Krauss, 2002), is sparse. This lack of attention began to change in 2003 with the publication of the best practices list by AAPOR (2003). It succinctly describes the reality that "[e]ffective control of falsification is not the result of any single method, but of the combined aspects of the study-specific environment in which surveyors conduct their work." In this paper, we briefly outline lessons learned during the implementation of a census of three contiguous villages in Bangladesh (Bhuiyan and Szulga, 2013), denoted the Tangail Survey (TS). We focus on the strategies employed during the data collection process to mitigate human errors and data falsification. The research project had minimal funding so budgetary restrictions dictated many of the methods employed.

Roughly a dozen local graduate students were hired for the TS, which used a geographic information system (GIS) and smart pens to conduct face-to-face interviews (approximately 40 minutes each) at the household level. Our top priority was to gather a highly reliable and robust dataset on the villagers' subjective well-being (SWB) and their perceptions of relative economic position, by identifying households with international migrants, and collecting the geographic (latitude/ longitude) coordinates of household locations with a global positioning system (GPS). In agreement with the techniques suggested by Koczela et al. (2015), we provide evidence that the use of smart pens, a GIS mapping of the household location prior to conducting the interviews, and daily on-site monitoring of the hired surveyors, to be quite effective in catching survey fraud and reducing unintentional errors. The use of the mentioned technologies also made post-survey validation and catching transcription errors a relatively easy task.

Curbstoning, Falsified Data, and Cheating

The most common terms used to describe survey fraud are curbstoning (where face-to-face interview data are faked), partial falsification (where only a portion of the survey data are faked), and cheating (when the convenience of the surveyor takes precedence over the protocols of the survey). Blasius and Friedrichs (2012) provide a concise summary of the literature and describe faked interviews. They conclude that it is remarkably easy for surveyors to fabricate interviews in face-to-face surveys which may remain undetected when basic monitoring protocols are not followed. Basic protocols, while essential to achieving high quality data, do not necessarily guarantee this quality of data:

> [While] there has been an old and long discussion on the reasons why surveyors fake interviews (cf. Crespi, 1945, 1946; Bennet, 1948; Nelson and Kiecker, 1996), the most elementary reason has hardly been discussed. Falsifiers save a lot of time/earn more money if they (partly) fake their interviews. ... [and] the risk of getting detected is relatively low since control mechanisms as those proposed by AAPOR (2003) and Murphy et al. (2004) are relatively easy to bypass. ... Furthermore, a detailed introduction, good payment, no time pressure, and an interesting study do not necessarily guarantee well-done interviews.

Harrison and Krauss (2002), Waller (2012) and Koczela et al. (2015) focus on the motivations for surveyors to cheat. Waller (2012) provides the most comprehensive study of the motivations for and methods of falsifying data by the surveyors. In general, the literature focuses mostly on methods of detecting falsified survey data in pre-existing data (Bredl et al., 2011, 2012; Kuriakose and Robbins, 2015) and less so on the on-site prevention of the collection of falsified data. We contribute to the latter.

The Tangail Survey Process

The accuracy of survey data may be compromised due to a variety of reasons, ranging from human errors or misaligned incentives to technical problems before, during, and after the data gathering process. These errors are further intensified when the survey site is in a developing economy and the project has tight budgetary constraints. Both are true for the TS. To understand how the TS methodology was driven by the overarching goal of collecting robust accurate data, we provide a brief overview of the research agenda and the project workflow.

Research Objectives

The choice of topics covered in the TS is primarily determined by the principal investigators' (PIs) research interest - relative income, subjective well-being (SWB) and international migration. While the literature on relative consumption is vast, there is very little empirical work looking into the economic position of individuals relative to local reference groups. For instance, when talking about income relative to neighbors, most papers operationalize the definition of neighbors as all individuals who live within broad geographical regions such as villages (Fafchamps and Shilpi, 2008), areas within the same zip code (Knies et al., 2007), primary census units (Luttmer, 2005), states (Blanchflower and Oswald, 2004) or even countries (Easterlin, 1995). The PIs decided to collect data fine enough to be connected to more realistic definitions of local reference groups such as neighbors.

The TS gathers data on both objective and perceptive measures of relative economic position, as the literature does not provide a strong preference for either measure (Fafchamps and Shilpi, 2008; Ferrer-i-Carbonell, 2005; Luttmer, 2005; Mayraz et al., 2009; McBride, 2001; Senik, 2009). Although asking respondents about their perception of relative income compared to neighbors, siblings, colleagues, etc., is somewhat straightforward, data of this type are largely missing for developing countries. Hence, we decided to ask respondents about their perceptions of relative economic position directly. In terms of objective measures, we recognized that having geographic coordinates for every household in each village along with objective measures of their income, would be the best possible type of objective data on relative economic position that we could hope for.

One of the PIs research interests is understanding how local networks affect the choice of destination when it comes to international migration. For instance, if a potential migrant's neighbor is already an international migrant in the Middle East, is it more likely that they will put more weight on that region when choosing between multiple destinations. No data are available that can adequately address this question. After preliminary visits to Bangladesh and in consultation with local experts, villagers, government agencies, and development workers, the PIs chose three contiguous villages in the Tangail district of Bangladesh. These villages have a significant number of households with at least one international migrant and they are going to different regions of the world. The census nature of the survey along with the geographic coordinates of the household location, makes it possible to more closely study local neighborhood effects on migration decisions and subjective well-being.

From a policy standpoint, the TS offers valuable insights into the interaction of local neighborhood/community effects with international migration, conspicuous consumption and quality of life measures in rural Bangladesh. The paucity of data of this type, especially in the context of developing countries, makes this a very useful dataset for those interested in the aforementioned topics.

Project Workflow

In the pre-fieldwork and preparation phases, the PIs enhanced their local social network and improved their rapport with the hired surveyors (figure 1). Once the training and inter-coder reliability development stage commenced, the personal ties strengthened and surveyor specific

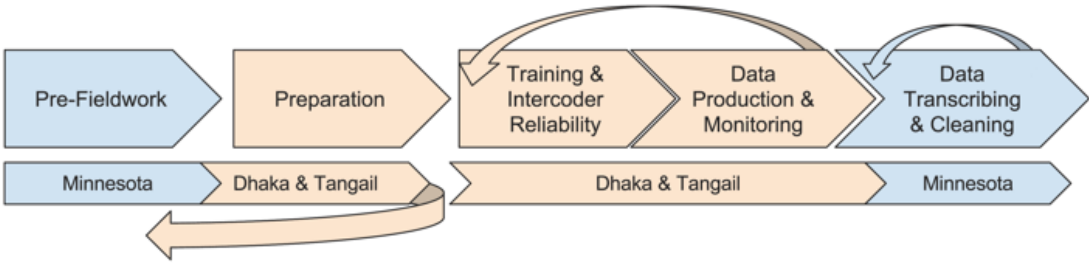

Pre-Fieldwork
Preparation
Training & Intercoder Reliability
Data Production & Monitoring
Data Transcribing & Cleaning
Minnesota
Dhaka & Tangail
Dhaka & Tangail
Minnesota

Figure 1: Overview of Data and Project Workflow

weaknesses with the delivery of certain survey questions became more apparent. During data production and monitoring, it was easy to verify suspicions and fire the surveyors who exhibited serious errors in judgement. The surveyors met to discuss questions of the survey, improve their understanding of how to deliver questions and interpret answers on a daily basis. This was both time consuming and exhausting, but an effective impediment to the would-be cheaters. (See appendix A for the overall project time-line and a daily schedule of survey activities.)

Due to financial and infrastructural constraints, we chose to (a) create a geographic information system of the survey area, and (b) use the inexpensive audio-capturing Livescribe™ smart pen technology. It not only allowed for collecting the data efficiently but also helped with monitoring and catching survey fraud and errors. It is worth noting that we were able to borrow the necessary GPS equipment from a local institution for free which kept our costs in check.

The Use of a GIS

Figure 2 provides a schematic example of the type of map employed for this project. The use of a GIS made sense on several levels. Mapping out the spatial location of households offered a nuanced understanding of local neighborhood effects. Additionally, the development of the GIS gave the supervisors, who were also the GIS mappers, experience inside each village. This familiarity came in handy during the survey process. The GIS also offered a simplified solution to:

- dividing the villages into specific areas that were then assigned to surveyors.
- assigning unique identifiers to the households.
- managing the paradata (which households were non-responders, unavailable, or chose to delay the survey), and tracking the overall survey progress.
- following up with the households to screen for fraudulent activities.
- randomly choosing households during post-survey validation efforts.

Developing the GIS itself produced some limited errors. A concrete example of this showed up during the TS when a son claimed he and his family ate separately while his father claimed the opposite. It was later found that the father and son had recently split and the father did not want to think of his son as living separately. These GIS errors became evident during the questionnaire survey phase and it was relatively simple to correct them. Yet another example of errors occurred when the responders would get confused about the definition of "household" and provide inaccurate answers. When the surveyors returned and found such households they were instructed to report them. The daily reporting session held every evening was a chance to discover these households and address the situation.

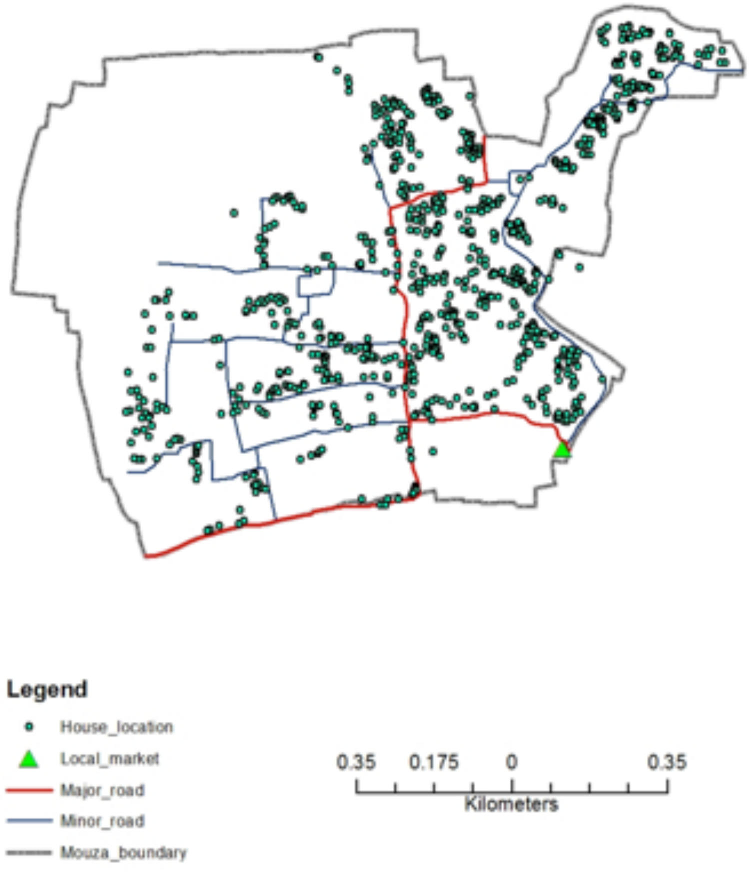


Figure 2: Example Household Location/GIS

The Use of Smart Pens

The basic feature of the smart pen is to have anything written on a special dot-paper recorded into the pen and synchronized with all audio that accompanies this writing. The resulting audio can be "played back" using Paper Replay™ with the pen and paper itself. It can also be replayed as an encapsulated Pencast™ using desktop software such as Adobe Flash®, PDF, PNG or M4A or within the Livescribe desktop software. A feature of this pen is that it captures only what the pen itself writes on special dot-paper and not what has been pre-printed (e.g. the survey form), on the paper (Lackie et al., 2014).

Once the survey process was underway, the benefits and problems of the smart pens became clear. At the end of each survey day the interviewing teams returned to the base (where there was more reliable access to electricity than in the villages). The supervisors immediately began processing the day's data: logging the survey forms, transferring the recorded "data" and audio from each pen to a computer, and backing all of that up again. Most importantly, the smart pen provided a rapid recap of every recorded survey. Each day the supervisors transcribed the paper surveys and listened to specific areas of the surveys that were suspect (transition points, questions that had been difficult to read or were simply skipped on the paper form). The

Pencast™ made it possible for the supervisors to immediately fast-forward the recording to a specific point in the survey where surveyors were facing challenges.

With this immediate and convenient method of double-checking the data, the surveyors and supervisors met daily to follow up on errors, specifically focusing on survey questions that were not being asked or recorded consistently across the surveyors. This daily iterative process continuously improved the quality of the data. It became clear to the surveyors that they had to conduct the surveys as instructed or risk getting caught and fired. The Echo™ smart pens:

- are significantly less expensive than any computer tablet or screen-based device.
- are small, easy to use, and not distracting in an interview.
- are robust enough to withstand intense heat and humidity.
- run on a single battery charge for the entire day (necessary as there was no way to recharge mid-day.)
- are capable of storing a full day of interview data with full audio within the pen.
- provides backup for every survey (i.e. paper forms and a digital document with full audio).

Lessons Learned and Effective Strategies

There are multiple ways of classifying errors that compromise data quality. Certain errors are intentional while others unintentional. An example of an intentional error is turning in fake data to avoid the effort of conducting a genuine survey. Reframing the survey question by modifying the language and mistakenly assuming this causes no bias is another type of unintentional error. From the PIs' perspective, the tools used to mitigate and repair these errors caused by unintentional mistakes, negligence, imprecision or fraud tend to be very similar. Where it is clear that the hired surveyor is intentionally conducting fraud, it is best to terminate their contract right away. However, in cases where it is not as obvious, the surveyor's ability to incorporate feedback and the magnitude of damage caused by their potentially unintentional errors, is the appropriate metric for deciding on termination decisions. As will become clear from the discussion below, the following four aspects of the TS significantly improved our ability to catch and prevent survey fraud and human errors:

- Having at our disposal the pre-survey GIS map of the study area.
- The use of Livescribe™ smart pens for audio-capturing the full interview.
- Checking the surveys daily (e.g. listening to the audio recording) for deviations from the survey protocols and debriefing the surveyors instantly.
- Sending supervisors to verify if certain households were surveyed properly when survey fraud was suspected.

Issues Mitigated by Requiring a Full Audio Transcript of the Interview

In this section, we provide scenarios of survey fraud and human errors along with techniques used to mitigate the errors during the implementation of TS. In particular, we focus on survey fraud and errors that were effectively mitigated using the smart pens. Scenarios 1- 3 are examples of curbstoning while scenarios 4-9 are examples of cheating or unintentional errors.

Scenario 1: The surveyor was unsure whether the respondent would cooperate once they were located, and so ventured into the market place and asked some random individual to complete the survey. As surveyors were able to start or pause the audio recording of the smart pen as they wanted, some thought they would be able to hide the fact that they were interviewing the wrong person.

Solution: The fact that the audio capturing was turned off strategically to avoid recording the name of the respondent raised red flags. Subsequently, supervisors were sent to these households to verify whether they were properly surveyed, if at all. Fraudulent surveyors were caught and fired.

Scenario 2: Surveyors claimed that the background noise from a weaving machine was too loud. Consequently, the interview could not be heard in the audio-capture.

Solution: The supervisors were aware that the audio-capturing features of these smart pens are robust to these types of noises. Thus, such claims also raised red flags and resulted in subsequent verification.

Scenario 3: Surveyors claimed that the pen was not working on a specific day.

Solution: Two approaches were used to deal with this. First, the surveyors were told that they could keep the smart pen at the end of the project, but only if it worked throughout the survey process. If their pen did not work, they would not be allowed to keep it. They valued the pen and consequently had proper incentives to protect the device. Second, when they reported a pen not working at the end of a particular day, a supervisor was sent out to verify if households were actually surveyed the day for which the audio could not be captured.

Scenario 4: Surveyors rephrased the questions incorrectly. For instance, replacing the phrase "life satisfaction" with "happy" in the question "How satisfied are you with your life?" Note that in the SWB literature they have very different meanings.

Solution: From the training sessions the supervisors and PI knew which questions would be most challenging and could quickly check the audio directly on the Paper Replay™ when the surveyors returned each evening. The surveyors who were caught not having followed their training were retrained. Repeat offenders faced the prospect of being fired.

Scenario 5:: Surveyors changed the language of the survey and assumed answers. For instance, when asking a question on the perception of relative economic position compared with neighbors, the surveyor might truncate the response from a five-point scale of "much worse", "worse", "same", "better" and "much better" to a three-point scale of "worse", "same" and "'better'. The surveyor then uses their own judgement to convert the answer to a five-point scale response.

Scenario 6: Surveyor unintentionally reverted to a pre-training way of speaking.
Solution (both 5 and 6): As a part of the training, the surveyors discussed their language use and developed a feeling for which questions were most likely to cause these problems. The immediate Paper Replay™ made it a simple matter to focus on the audio of how specific questions were asked and review them as the surveys were completed. In many cases, this was an unintentional result of fatigue and the tendency to revert back to their usual way of speaking. This was confirmed by listening to many of their surveys and hearing that they framed the questions properly during most interviews but slipped up on a few. Repeated listening made them more aware of their language use and they quickly learned not to do it.

Scenario 7: Socioeconomic differences played a role. The surveyors were educated and from the city, while the population surveyed are poor and rural. In the Bangladeshi context there is an implicit understanding of socioeconomic hierarchies which lead both sides to act in certain ways. For example: A surveyor harshly demanding answers to questions. "Why does it take you so long to answer this? answer quickly!" or showing anger or impatience with the respondent in any way. This lead to the respondents not thinking about their answers and answering quickly to get out of the circumstance.

Solution: The surveyors had to be trained about the deleterious effects of such behavior on the survey process and then to overcome these tendencies. They were regularly reminded that for the survey to be taken seriously, socioeconomic biases needed to be addressed. As the audio was being checked as they turned in their forms each night, this behavior was quickly detected.

Scenario 8:: Although trained not to, sometimes the surveyor would prompt the respondents with an answer in trying to explain the question.

Scenario 9:: When transcribing multiple fields of data, sometimes it does not show internal consistency. For instance, a family indicated as not having an international migrant, nonetheless, seems to be receiving a non-zero remittance from one of its household members. Another example involves transcribing the gender of a son of the household head to be a female. While it is obvious that there is a mistake here, it is not clear which field between the two contradictory ones is incorrect.

Solution (both 8 and 9): The audio playback provided the supervisor or PI with the necessary tools to repair the data entry.

Issues Mitigated by the GIS and Post-survey Validation
The pre-survey GIS data developed for the study area included information on the geographic coordinates of the household locations and the name of the household head. These data played a very important role in monitoring coverage, avoiding duplication of surveys, and dealing with erroneous transcription of household identifiers. It also made post survey validation a very quick and cost-effective process. Here are some problematic scenarios that the GIS helped to resolve:

Scenario A: Occasionally the surveyors miscommunicate which houses had already been surveyed and would duplicate efforts and skip other households altogether.

Solution: This was caught when the household identifier was matched between the GIS records and the survey data.

Scenario B: Transcription errors of household identification numbers were more difficult to catch. It resulted in certain households mistakenly connected with a different GIS location, such that two sets of data were then suspect.

Solution: Comparing the name of the household head in the GIS survey, with the questionnaire survey usually made it clear which survey had the incorrect household identification number.

Scenario C: Some cheating was caught by re-surveying 20% of the households. These were chosen at random and the households were asked three verifying questions about, (a) whether someone surveyed their household (b) the name of all individuals who lived in the household, and (c) whether the surveyor instructed the respondent to not cooperate or answer in a fraudulent manner.

Solution: While the basic questions are simple, the process of verification is still a weak link. Especially in the case of a violation of (c), the respondents may choose not to answer truthfully in the verification stage, out of fear that they would have to make the time to do the survey again.

Scenario D: Surveyors occasionally chose to skip some households, but claimed that the household said they did not want to take part in the survey. It did not happen very often legitimately, due to the good relationship they had developed with the villagers.

Solution: All households who refused to participate were followed up on.

Issues Mitigated by the GIS and Post-survey Validation

The tips from the few articles about mitigating surveyor fraud seemed to hold true for the TS (e.g. recording the interviews, providing random checks to confirm that the interview took place as expected, following up on protocol violations, and rigorous oversight with little isolation of surveyors). In addition, some new issues arose with this group. These hired surveyors were hand-picked, well-paid graduate university students who were gaining excellent field experience. Several had hopes of attending school in the US or elsewhere and needed letters of reference. Still, they required a great deal of attention and persistent following.

A few students did not like being monitored and in retrospect conducted much of the survey fraud. This group tried to unionize and extort a higher wage early in the process. They started rumors about the PI making money off their hard work and being involved in financial fraud. They realized that the PI was under a binding time constraint and so started to engage in extortionary behavior. The solution to these issues included persistence, openness, and being willing to fire and replace them very quickly. These surveyors were also the ones who worked to cheat whenever possible and exhibited a disdain that their "usual" methods of survey fulfillment (e.g. survey fraud tactics) would not work because of the rigorous and prompt data verification process. When they were fired, the rest of the survey team took notice and worked very well. A few proactive bad apples can seriously hamper the process and it is important to deal with them swiftly and transparently.

References

AAPOR (2003) Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection, and Repair of Its Effects, Survey Research, 2004 Newsletter from the Survey Research Laboratory, College of Urban Planning and Public Affairs, 35, 1, University of Illinois at Chicago.

Bhuiyan, M. F. (2012) Relative Consumption: A Model of Peers, Status, and Labor Supply, American Journal of Agricultural Economics.

Bhuiyan, M. and Szulga, R. (2013) The Tangail Survey: Household Level Census of Subjective Well- Being, Perceptions of Relative Economic Position, and International Migration: 2013 [Tangail, Bangladesh]., http://doi.org/10.3886/E61146V1.

Biemer, P. P. and Stokes, S. L. (1989) The Optimal Design of Quality Control Samples to Detect Interviewer Cheating, Journal of Official Statistics, 5, 23–39.

Blanchflower, D. G. and Oswald, A. J. (2004) Well-Being Over Time in Britain and the USA, Journal of Public Economics, 88, 1359–1386.

Blasius, J. and Friedrichs, J. (2012) Faked Interviews, in Methods, Theories, and Empirical Applications in the Social Sciences, Springer, pp. 49–56.

Bredl, S., Storfinger, N. and Menold, N. (2011) A Literature Review of Methods to Detect Fabricated Survey Data, Zentrum für internationale Entwicklungs und Umweltforschung 56, Courant Research Centre PEG.

Bredl, S., Winker, P. and Kötschau, K. (2012) A Statistical Approach to Detect Interviewer Falsification of Survey Data, Survey methodology, 38, 1–10.

Clark, A. E. and Oswald, A. J. (1996) Satisfaction and Comparison Income, Journal of Public Economics, 61, 359–381.

Cole, H. L., Mailath, G. J. and Postlewaite, A. (1998) Class Systems and The Enforcement of Social Norms, Journal of Public Economics, 70, 5–35.

Crespi, L. P. (1945) The Cheater Problem in Polling, Public Opinion Quarterly, 9, 431–445.

Deaton, A. and Stone, A. A. (2013) Two happiness puzzles, American Economic Review, 103, 591–597.

Diener, E., Suh, E. M. and Smith, H. (1999) Subjective Well-Being: Three Decades of Progress, Phsychological bulletin, 125, 276–303.

Duncan, O. D. (1975) Does Money Buy Satisfaction?, Social Indicators Research, 2, 267–274.

Dynan, K. E. and Ravina, E. (2007) Increasing Income Inequality, External Habits, and Self-Reported Happiness, American Economic Review, 97, 226–231.

Easterlin, R. A. (1995) Will Raising the Incomes of All Increase the Happiness of All?, Journal of Economic Behavior and Organization, 27, 35–47.

Easterlin, R. A. (2001) Income and Happiness: Towards an Unified Theory, Economic Journal, 111, 465–484.

Fafchamps, M. and Shilpi, F. (2008) Subjective Welfare, Isolation, and Relative Consumption, Journal of Development Economics, 86, 43–60.

Ferrer-i-Carbonell, A. (2005) Income and Well-Being: An Empirical Analysis of the Comparison Income Effect, Journal of Public Economics, 89, 997–1019.

Harrison, D. E. and Krauss, S. I. (2002) Interviewer Cheating: Implications for Research On Entrepreneurship in Africa, Journal of Developmental Entrepreneurship, 7, 319.

Knies, G., Burgess, S. and Propper, C. (2007) Keeping Up With the Schmidts: An Empirical Test of Relative Deprivation Theory in the Neighbourhood Context, Discussion Papers of DIW Berlin 697, DIW Berlin, German Institute for Economic Research.

Koczela, S., Furlong, C., McCarthy, J. and Mushtaq, A. (2015) Curbstoning and Beyond: Confronting Data Fabrication in Survey Research, Statistical Journal of the IAOS, pp. 1–10.

Kuriakose, N. and Robbins, M. (2015) Don't Get Duped: Fraud Through Duplication in Public Opinion Surveys, Statistical Journal of the IAOS (Forthcoming).

Lackie, P., Ketama, M., Loery, G., Mansour, C. B. and Strauss, B. (2014) Smartpens as Data Capture Devices: Survey Data from Handwriting on Paper to Automatic CSV File, Unpublished manuscript, Carleton College.

Luttmer, E. F. P. (2005) Neighbors as Negatives: Relative Earnings and Well-Being, Quarterly Journal of Economics, 120, 963–1002.

Mayraz, G., Schupp, J. and Wagner, G. G. (2009) Life Satisfaction and Relative Income: Perceptions and Evidence, CEP Discussion Papers, Centre for Economic Performance, LSE.

McBride, M. (2001) Relative-Income Effects on Subjective Well-Being in the Cross-Section, Journal of Economic Behavior and Organization, 45, 251–278.

McGuire, M. T., Fairbanks, L. A. and Raleigh, M. J. (1995) Life-history Strategies, Adaptation Variations, and Behavior-Physiologic Interventions: The Sociophysiology of Vervet Monkeys, New York: Oxford University Press.

Senik, C. (2009) Direct Evidence on Income Comparisons and Their Welfare Effects, Journal of Eco- nomic Behavior and Organization, 72, 408–424.

Solnick, S. J. and Hemenway, D. (1998) Is More Always Better?: A Survey on Positional Concerns, Journal of Economic Behavior and Organization, 37, 373–383.

Van De Stadt, H. A. K. and de Geer, S. V. (1985) The Impact of Changes in Income and Family Composition on Subjective Well-Being, Review of Economics and Statistics, 67, 179–187.

Waller, L. G. (2012) Interviewing the Surveyors: Factors Which Contribute to Questionnaire Falsification (curbstoning) Among Jamaican Field Surveyors, International Journal of Social Research Methodology, 16, 155–164.

Notes

[1] Acknowledgements: We are very grateful to the department of Economics at Carleton College for their valuable advice. The data gathering process has been a joint project of Radek Szulga (Economics, Lyon College) and Bhuiyan, and was funded by the Dean's Office and the Economics Department of Carleton College. Paula Lackie is the technology adviser for the project. In Bangladesh, we were helped by a dozen seniors of Dhaka University (DU) from various disciplines and Prof. A Q M Mahbub (Geography & Environment, DU). We are greatly indebted to them. The respondents in these villages took part in the surveys voluntarily and were very kind in donating their precious time to our cause. Last but not the least thanks to Yuping Huang, Iris Wang and Caroline Greenberg for their fantastic research assistance.

[2] Contact author Muhammad F. Bhuiyan is an Assistant Professor in the department of Economics, Carleton College, Northfield, MN 55044, USA. Email: fbhuiyan@carleton.edu. Corresponding author Paula Lackie is the Academic Technologist for Data, Information Technology Services, Carleton College, Northfield, MN 55044, USA. Email: plackie@carleton.edu

[3] The two PIs for this project are Muhammad F. Bhuiyan and Radek Szulga. Paula Lackie was the technical advisor for the project. The survey was funded by the Discretionary Fund of the Dean of the College Office (Carleton College, MN) and the department of Economics (Carleton College). Bhuiyan is currently an assistant professor of economics at Carleton College (MN, USA) while Szulga is an assistant professor of economics at Lyon College (AR, USA).

[4] See McGuire et al. (1995); Van De Stadt and de Geer (1985); Blanchflower and Oswald (2004); Clark and Oswald (1996); Dynan and Ravina (2007); Easterlin (2001); Duncan (1975); Cole et al. (1998); Diener et al. (1999); Deaton and Stone (2013); Luttmer (2005); J. Solnick and Hemenway (1998); Bhuiyan (2012); Mayraz et al. (2009).

[5] While the survey was conducted in Bengali, the answers were interpreted onto the survey form in English during the interview. Surveyors needed to learn to deliver the survey in a strict format and interpret the answers onto each form in a consistent way.

[6] LIVESCRIBE, PAPER REPLAY and NEVER MISS A WORD are trademarks of Livescribe Inc. All rights reserved. ©2014 Livescribe Inc. http://www.livescribe.com/en-us/faq/online_help/Maps/Connect_Desktop/r_ formats-for-sending-notes-and-audio.html

[7] For a discussion of downloadable dot paper: https://support.livescribe.com/entries/22263341-60013-Printing-free-Livescribe-3-or-Livescribe-wifi-smartpen-dot-paper

[8] Paper Replay™ is the process that the Livescribe™ smart pen uses to play back any audio recorded when specific text was written with the pen on special Livescribe™ dot paper. There is both a microphone and speaker built into the pen.

Image Management as a Data Service

by Berenica Vejvoda[1] K. Jane Burpee[2] Paula Lackie[3]

Abstract

Across all disciplines, researchers are creating or gaining access to an ever-growing body of digitized images. Since research data management includes the organization of 'all materials' intrinsic to a research project, a robust data management plan will include a path for images as well as data in the more traditional sense. While researchers across disciplines have a long history with the organization of numeric data, the inclusion of images as a resource set in research is only starting to take shape across the disciplines. This paper is intended for data librarians or academic support staff without expertise in image data management. The primary focus is to apply traditional data management practices to images and to discuss the challenges associated with managing image collections through the research data lifecycle.

Keywords

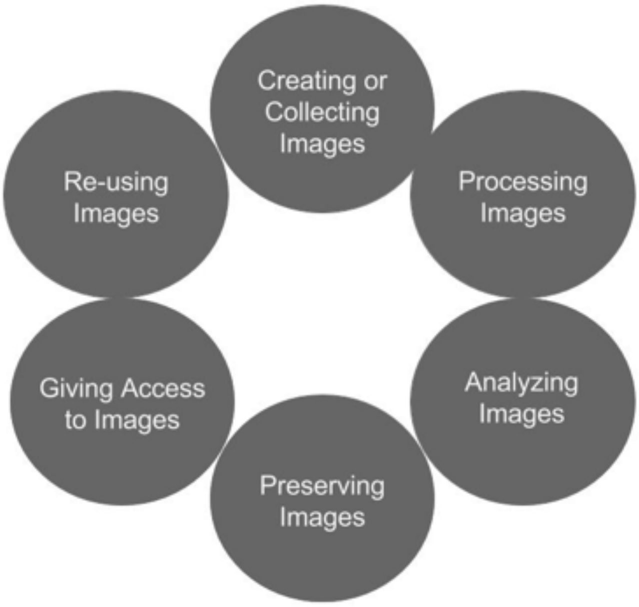
data management, images, research data lifecycle, preservation, sharing, mixed-methods research

Introduction

To move in small measure towards a greater understanding of image collections as a data management challenge, this article compares traditional numeric data management with organized image collections. By conceptualizing image collection management as a component of data service across the 'research data lifecycle' we hope to foster a better understanding of how data professionals can effectively transition their skills to include the management of images. This paper addresses images as data rather than as an object. The unique challenges associated with images (versus numeric data) will be highlighted through the points of the research data lifecycle which are most impactful for image management. Although the principles outlined here may apply to any digital object identified as an "image", this article will assume a format-based approach that includes any two-dimensional digital image format.

A Research Data Lifecycle Approach for Research-Related Image Collections

The following stages images in the data lifecycle (see diagram at right) will be discussed in the comparison with numeric data and research-related image collections: creating or collecting images; processing images; analyzing images; preserving images; giving access to and re-using images (adapted to images based on DCC, 2012 and MANTRA, 2014).



Adapted from, Create and Manage Data - Research Data Lifecycle. UK Data Archive, 2016. Retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle. Copyright 2002 -2016: University of Essex.

Creating and Collecting Images as Data

Managing data during the creation stage of the research data lifecycle can be challenging, most notably when the 'data' are in the form of images. Images can be collected in a number of different ways, e.g.: in-house or outsourced scanning or photography; digital creation from the outset; or purchased from vendors (Primary Research Group, 2013). Just like any data gathering process, for collection methods to be successful researchers will benefit if they make decisions before they begin the process of collecting and capturing images. Therefore, questions commonly asked when collecting numeric data offer a useful framework for effectively collecting and organizing image as a data-style resource (DCC, 2013).

- What type of data will be gathered?
- In what 'format' will the data appear? (will it change through the life cycle of the project?)
- What will be the expected 'volume' of data collected?

Types of Data ~ Types of Images

At the most fundamental level, digital images are numeric data. They are all ones and zeros stored on computer media. In practice however, images are often more like social science microdata, they can be both data as well as a datum at the same time. For example, a collection of images organized around a particular theme is comparable to a dataset and the individual image, a specific response source. When considering an image management strategy, simultaneously managing images with their metadata is like managing survey metadata, paradata, and data all at once. For example, both surveys and image collections may have metadata, paradata, and data important to the analysis or for reuse purposes.

For the purposes of data management planning, it might be helpful to think of these example parallels:

	Survey Respondent	Digital Image
Data	survey answers	Often simply the viewable image but it may also be the direct analytic content derived from that image
Metadata	e.g. age, education level, home address	e.g. camera make & model, image timestamp as set by the camera, aperture, shutter speed (EXIF data)
Paradata	e.g. respondents click-rate through a survey,	e.g. average image color, facial recognition material, image sequence in a set

working through choices in an image data management plan, comparing the process with data associated with a survey respondent can be a good place to start; both yield information elicited through data collection instruments, are inherently complex, and it is necessary to make choices regarding which aspects to focus on. Of course, the analogy is limited since unlike survey respondents, an image database may also contain the complete image - which may be flawlessly duplicated - unlike humans.

In addition to the need for understanding the complexity and structure of images as data, for images to be useful, to those who create them as well as for subsequent re-use, it is equally necessary to consider the format of image files as early as at the point of creation. Attention to format will ensure long term access and functionality.

Data Formats ~ Image Formats

Deciding on the best file format (i.e. the way information is encoded in a computer file) is understandably a question that applies to both numeric data and images. Both rely on applications or programs that will recognize the file format in order to access information within the file. According to a report published by a Digital Preservation Policy Working Group at Cornell (2001) file format for images consists of the bits that comprise the image and the header information on how to read and interpret the file. Similar to numeric data files, images can be stored in a wide variety of formats, including: bitmap (BMP), Joint Photographic Experts Group (JPEG), JPEG 2000, and Tagged Image File Format (TIFF). These standard image formats vary in relative file size, image quality and flexibility, and compatibility with software programs. Distinguishing between minimal requirements and recommended imaging requirements, the Cornell report gives preference to TIFF formats as a master image format since they do not compromise data, while JPEGs, a "lossy " format which compresses data, are included in the minimal criteria.

One research domain in particular that has championed the adoption of open standardized image data formats is the imaging community in the biological sciences, namely the JCB DataViewer initiative. Initiatives like the JCB Dataviewer align with the conventions of numeric data that recommend the adoption of open source formats in order to retain the best chance for future readability. If open source isn't an option then choosing formats that are in widespread use or agreed-on international standards will help achieve the objective of longevity and/or replicable research.

Volume: Counts and File Sizes

Related to formatting is the notion of volume. When data were first digitizable, disk memory was extremely limited. Researchers were resourceful in how they encoded and managed their data. As expectations for robust data analysis were fed by Moore's Law and a parallel rise in disk storage capacity enabled the rise in big data (e.g. moment-to-moment stock trade data or global social network data). Likewise, expectations for big data - in the form of images - has also risen. Like numeric data, a big concern for image data formats is related to a combination of storage space and computational power. Just like with numeric data; the numbers of image collections, the numbers of images in collections, and the size of individual images are all growing. Researchers should be aware of the trade-offs they are making when choosing either fewer images or images of lower quality than the original images as they were created.

Dealing with large numbers of lossless image files can quickly become unwieldy in terms of available storage space. Compressing large image files to smaller files is most easily produced using lossy compression and while the images may still provide adequate information for the immediate intended purpose, their longevity may suffer.

An advantage of the usual numeric data compression over image file compression is that they are lossless (e.g. .zip, .7z, and .gz.) On the other hand, choosing smaller image file formats (e.g. JPEGs) that are lossy over lossless image files (e.g. raw, TIFF) results in loss of image clarity; resolution, layers, and fidelity. As a result, the management of images becomes a more difficult decision when dealing with a large number of large files. In terms of image management, due to the usual lossy nature of compression, there is a clear preference for retaining uncompressed versions or for working to manage the balance of lossless compression against future format compatibility challenges. At the very least, it is recommended that researchers minimize the number of compression processes that need to be managed over the long term (Cornell, 2001).

The challenges associated with large image files are compounded by the sheer volume of images being produced across various disciplines and sectors. While the growth rate of images can be difficult to quantify and many claims appear as unsubstantiated hyperbole, the discourse surrounding the explosion of visual content agrees that it is undeniably large (Kane and Pear, 2016). Kane and Pear (2016) estimate that while 3.8 trillion photos were taken in all of human history until mid-2011, one trillion photos were taken in 2015 alone. In academia, specifically, the biological sciences, Moore, Allan, Burel, Loranger, MacDonald, Monk and Swedow (2008) noted eight years ago that 'most laboratories and imaging facilities do not have the means to store the volume of data generated by their microscopes in manageable and affordable way' (p. 557). Another testament to exponential growth in the biological sciences comes from the rapid progress in genome sequencing technology. In 2011 Gross noted that second-generation machines like Illumina's Genome Analyzer II create vast amounts of images and that the volume of these images was growing by five terabytes a day. The volume of images as data produced through medical imaging is also staggering. MarketandMarkets (2016) estimate that medical image archives are increasing by 20-40% annually, and they predicted that by 2012, there will be 1 billion medical images stored.

Processing Images as Data

Creating or Collecting Images
Processing Images
Analyzing Images
Preserving Images
Giving Access to Images
Re-using Images

The growth of digital images in size and number, the advent of powerful digital cameras and the willingness of libraries and archives to use them, has produced an overwhelming need for comprehensive image management software (Roy Rosenzweig Center for History and New Media, 2016). This need is documented across disciplines by researchers who struggle to manage collections of digital images. In response to this need, in 2015, the Andrew W. Mellon Foundation announced funding for a new project to develop Tropy an open-source software application that will help researchers collect and organize digital photographs, create metadata, and export photographs and associated metadata to other platforms (Centre for History and New Media, 2016).

Researchers in the sciences – particularly, in the medical and life sciences, are also expressing a need for image management systems. The Open Microscopy Environment (OME) Consortium has, for example, built a series of open source tools that assist researchers in managing large sets of complex images to support research in cell and developmental biology (Moore et. al., 2008). Researchers relying on medical imaging (e.g. CT, MRI, X-ray, NM, mammography, ultrasound, radiology) are also in need of image management systems to keep up with unprecedented growth. A case in point comes from the critical role of medical imaging, specifically image biomarkers, in clinical trials for Alzheimer's disease. Increased reliance on these medical images for study outcomes requires image management systems for effectively capturing, processing, analyzing, disseminating and archiving images (Jimenez-Maggiora, Thomas, Brewer, Bruschi, Hong, and Aisen, 2012). Jimenez-Maggiora et. al. (2012) note that these specialized systems are complex, inflexible and resource intensive.

Recommendations for managing traditional numeric data files at this stage of the research data lifecycle can provide a useful framework whether the data are "Big" or just awkward. Key considerations for organizing numeric data include data carpentry functions, such as: versioning, naming, and renaming (MANTRA, 2014). As with a numeric data file, an image file name needs to be carefully considered for consistency, logic and predictability so that users can effectively browse and retrieve image data and also to avoid confusion when multiple researchers are working and naming shared files. In other words, 'ThisImage.tiff' will be as problematic as 'ThisSASFile.sav'. And similarly, images will present with multiple files in various formats, multiple versions, across differing methodologies, etc.

A growing number of software tools exist to help organize images in a consistent and automated way through functions such as batch renaming. Renaming files may be especially useful for image metadata in instances where digital cameras automatically assign base filenames of sequential numbers. In the realm of numeric data file management, tools include: using the GREP command in UNIX or applications such as RenameIT . In addition, ImageMagick can perform various batch processing functions on image metadata. In addition to batching renaming, image management software can support image workflows by assisting with recording location, generating thumbnails and storing basic associated metadata that are embedded in image files (Jisc, 2016).

Analyzing Images as Data

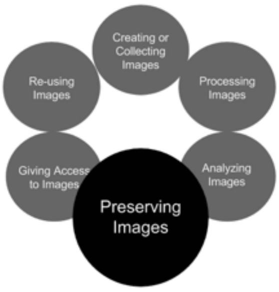
Creating or Collecting Images
Processing Images
Analyzing Images
Preserving Images
Giving Access to Images
Re-using Images

One of the defining features of data is that they are the raw material produced by primary research that is intended for analysis (Geraci, Humphrey, & Jacobs, 2012). Arguably, numeric research data are most often created for the intended purpose of analysis. In contrast, images may not always be intended for analysis at the point of creation. For example, many early digital library projects supported the creation of images to add to library collections, but in a large majority of cases such images were and continue to be used by researchers for making examples and illustrations versus serving as raw material for research. It is important to note that an image or collection of images may serve a variety of users and as such, they may also become data for analysis.

Recommended best practices for managing numeric data at the analysis stage of the research data lifecycle (MANTRA, 2014; DCC, 2012) include documenting analyses and file manipulations, managing versions of data files and deciding if analyzed data will be shared. With numeric datasets, it is fairly routine to document analysis and manipulation functions, usually in the form of programming code files. Similarly, documenting the manipulation and analysis of images helps researchers with their own image processing and analysis workflows (e.g., logging the numerous steps taken to geo-reference an image) and will also produce greater transparency of techniques, critical to successful replicability, which have in recent years emerged as potential sources of controversy across many disciplines. According to McCook (2016), companies such as Image Data Integrity (IDI) exist to help journal editors, publishers, funding agencies and institutions screen and verify whether image manipulation (e.g., blots and micrographs in biomedical materials) compromises the interpretation of the images for scientific purposes. In addition to documenting the process of analysis and manipulation, researchers working with numeric data also decide what form of data to ultimately share: i.e. raw, processed, analyzed, final.

The analysis stage of the research data life cycle presents unique challenges for those managing image data when using image analysis software tools. By way of example, ImageJ , which has existed for over 30 years, is a general-purpose, extensible scientific image-analysis program that is used to capture, display and enhance images in the biological sciences (Schneider, Rasband & Eliceiri, 2012). One key challenge noted by Schneider et. al. (2012) occurs when one uses the software to open and parse the countless variety of image file formats. In the case of proprietary file formats tied to specific software (e.g. with some microscopes), using such software can be especially problematic for reproducibility or any further sharing. It is preferable to have images from research processes be available independently of specialty equipment. ImageJ, for example, is able to connect directly to MatLab so that researchers are able to run statistical analyses as well as other tools such as Imaris which supports 3D and 4D image analysis (Schneider, et. al., 2012). In order to encapsulate diverse needs even within a particular domain w biological imaging, image analysis tools need to remain flexible and extensible.

Preserving Images as Data



One of the most critical aspects of data management, regardless of data type is the pairing of metadata to accurately and sufficiently describe datasets. It is important to note that like most other data management functions, metadata hold a central role across the entire span of the research data lifecycle. So while it is discussed in reference to preservation for the sake of structuring this discussion, it applies to other stages as well. Metadata for images is especially critical as a way to organize and search through growing libraries and repositories of images that are being produced by researchers and consumers alike.

As is the case with numeric data files, decisions need to be made about how much detail to record in image metadata records. The internationally accepted metadata standard for describing social science numeric data, DDI (Data Documentation Initiative) can be crudely parsed between study-level descriptions and variable-level metadata. This distinction can also apply to images. For example, low level descriptors such as 'title', 'creator', and 'size' are similar to DDI fields such as 'title', 'abstract', 'producer', 'distributor', and 'time period'.

For more detailed descriptions, DDI offers metadata fields at the variable-level – e.g. exact meaning of the datum (ICPSR, 2016). This level of description is created directly from formatted datasets. While the need for fuller descriptions of image content are also equally necessary, a key difference is the degree of subjectivity involved in producing more abstract, higher level meaning, such as feelings portrayed by a particular image, 'happy', 'sad' (Jisc, 2016). This challenge for describing images is discussed extensively by Eadie (2008) in his description of the development of the Jisc-funded Dublin Core Images Application Profile. Eadie (2008) aptly notes that unlike text-based materials that can be machine processed, images are not easily self-describing. This reasoning can also contrast with numeric data where machine-processing can quite easily produce meaningful variable and file-level metadata based on objective information embedded in formatted statistical data files. Digital images, on the other hand, have a more complex relationship with machine-assisted metadata-extraction; while pixel data and bit depth are objective data points, they provide little by way of meaningful contextualization of images as data (Eadie, 2008). So despite increased quality and quantity of camera sensor elements, it is neither practical nor meaningful to describe images to aid organization or querying based on millions of image pixels (Metadata Group, 2008).

In addition to subjectivity, images can be complex to describe because they often have relationships with other objects - which may even be embedded within them. So before you can even begin to say anything about an image you need to be very clear about which aspects of it, or its relationship with other objects, on which you are actually focusing (Jisc, 2016). For example, images can be found in slides, photographs, books, manuscripts, lectures, and presentations. There may also be interdependencies between images. For example, in the area of Geographic Information Systems (GIS), different layers of GIS data are superimposed to create a richer representation for spatial analysis.

Adding to the complexity of image description is that description consists of at least three different types (Eadie, 2008). First, there is technical information relating to the image. This is usually pretty straightforward as capturing technical metadata is usually automated (e.g., captured by digital cameras) and resides in the image itself. The main metadata container formats for images are: Exif (Exchangeable Image File Format) (for device properties), IPTC (International Press Telecommunication Council), IIM (Information Interchange Model) (workflow properties), and Adobe XMP (Extensible Metadata Platform). Each metadata container format has unique rules regarding how metadata properties are stored, ordered and encoded (Metadata Working Group, 2010). While technical metadata are fairly easily captured, how they are structured is considerably more complex. Even within the container format, metadata are stored, for example, according to various semantic groupings; within these groupings there can be numerous individual metadata properties. Perhaps the biggest issue concerning the structural complexity of technical metadata is that different applications and devices handle these technical specifications in different ways; hence, creating challenges for interoperability. Technical metadata also becomes more complicated for long term preservation as more metadata fields need to be added that are not normally captured by devices – for example, image format migration and versioning.

The second and third types of description for images as data relate to the content in an image; the application of abstract principles to the description of the image. Not only are content and abstraction difficult to describe in a standard way, but text-based descriptions will vary depending on the knowledge, culture, experience and point of view of the cataloguer (Jisc, 2016). Even more difficult is anticipating the needs of users in terms of what to describe in images. For example, in Figure 1, one researcher may be drawn to the couple while another, looking for depictions of leisure activities would find the hoop-rolling relevant. One way to deal with this challenge is to balance the time and resources available for describing images with the anticipation of what level of detail users of the image will require for effective discovery (Jisc, 2016). Given the inherent subjectivity and richness of images, however, rarely in large image libraries and archives are there sufficient resources for in-depth description.

In addition to the critical role of associated metadata, image preservation involves decisions about where and how to store them. As is the case for numeric data, depositing images in an archive or repository should facilitate discovery and preservation for the long-term. As mentioned earlier, some domains such as genomics produces such vast quantities of images (more than five terabytes a day; see Gross, 2011) that the practicality and cost of archiving these images for preservation purposes is often not feasible. Gross (2011) instead points to a motivation to alternatively invest in the development of real-time processing of the images 'to output only the base calls and the quality values' (p. R204). The distilled nature of the images for their originally intended purpose will limit future re-usability of these images but current technical and financial limitations mandate that some hard decisions are being made regarding precisely what to preserve.

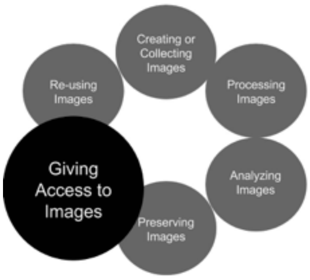
With the currently overwhelming volume of images as data we see additional kinds of re-use obstacles; image collections constitute "big data" in that they are larger than can currently be managed, adequate storage space is another issue, and even if storage were available, the transfer of such large files or sets of files is currently impractical. This should however be prefaced with a note that new infrastructure initiatives such as the pan-European project called ELIXIR (European Life Science Infrastructure for Biological Information) are looking for solutions that balance software compression with judicious data reduction. With ELIXIR, the ultimate aim is to bring compression down to 0.1 bits (0.01 bytes) for every base stored - which translates to a human genome taking up just 30MB of storage (Gross, 2011) (as opposed to the 1.5GB without the compression).

Not all image collections are so unwieldy. For the more common, manageable collections of images, it is possible to decide where to best archive image sets and their metadata. Because images as data are produced across a vast number of domains, repositories can range from individual solutions for photographers and other artists, to institutional photographic and slide collections, to archives and museum image repositories, and as institutional teaching and research archives.

As previously mentioned in reference to open/standardized data formats, JCB DataViewer was the first open repository in the life sciences that allowed for archiving and sharing of original image datasets to support published scientific articles (Linkert, Rueden, Allen, Burel, Moore, Patterson, Loranger, Moore, Neves, MacDonald, Tarkowska, Sticco, Hill, Rossner, Eliceiri and Swedlow, 2010). In addition to archiving the original binary image and associated metadata, additional information captured by acquisition software includes: acquisition settings, image size, and resolution.

While the imaging community in the life sciences already treats images as data and wherever possible has robust archiving solutions, this is not the case in all areas of research where digital images are produced. For these areas, the consideration for archiving numeric data can provide guidance for treating images as data in need of long-term preservation. Arguably, university repositories that have traditionally focused on archiving text-based resources may benefit from examining numeric data archiving practices as they can assist in archiving images as data.

As we know, preserving numeric data can be problematic due to the amount of data being generated (MANTRA, 2014). Given that image files are substantially larger, on average, this becomes a key consideration. Additionally, reliance on specific technologies for accessing anything digitized - becomes problematic since the technologies change quickly. Therefore, as with numeric data, it is essential to archive digital image data in a systematic way in order to minimize the chance of obsolescence or making images inaccessible over the long term.



Accessing and Sharing Images

The sharing and accessing phase of the data lifecycle is perhaps where those working with images become easily overwhelmed and/or frustrated when a discrepancy emerges between needs and search results (Chung & Yoon, 2011). As noted in other phases of the lifecycle this challenge is exacerbated by the explosive growth and availability of images. In order to allow for effective access to images, the growing image collections must be organized in ways that allow for efficient discovery, browsing, searching and retrieval (Rui, Huang & Chang, 1999).

According to Wang, Mohamad & Ismail (2010), an effective image retrieval system needs to be able to retrieve relevant images based on queries that conform as closely as possible to human perception. So unlike quantitative numeric data files, which are relatively straightforward to describe based on keywords that map to the represented measures, visual information is far more ambiguous and semantically rich (Wang & Ismail, 2010). Relying on traditional keyword querying systems of access will not be sufficient. Wang et. al. (2010) Note that image retrieval based on keyword querying, popular in the 1970s, relies on keywords used as descriptors to index an image. While the Jisc Digital Media Guide (2016) discusses the need for advanced search features (e.g. Boolean logic) to support relevant keyword image retrieval, Wang et. al. (2010) would argue that assigning keywords manually to images is not only time consuming but that keywords alone are inadequate and grossly inefficient to describe the rich content of images.

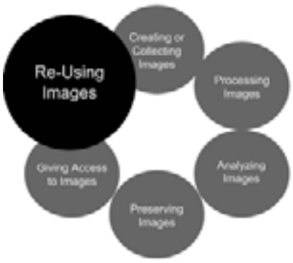
Another trend in image retrieval that supersedes text-based image retrieval (popular in the 1980s), is content-based image retrieval (CBIR). First used by IBM, this method of retrieval is based on extracting visual features from the image itself. While in theory CBIR systems can include the extraction of low or high level features, even with sophisticated algorithms that can combine multiple visual features, elements such as colour, texture, shape and spatial relationships do not come close to mirroring the richness of image content that users have in mind when searching for relevant images.

As a proposed solution, Wang et. al. (2010) discuss at length the most recent trend of developing semantic-based image retrieval systems that allow users to query image data using high-level concepts. In short, this retrieval method maps the automated process of extracting low level visual features from images with semantic descriptions also stored in an image database. It is hoped that this example of intelligent image retrieval, that can better represent the abstract concepts inherent to images, will help users discover and access relevant images.

While semantic-based image retrieval, in theory, allows users to better discover relevant images based on higher-level meaning, creating semantic descriptions still requires human intelligence which is time consuming and expensive. One innovative way to tackle this dilemma is to consider the need for metadata creation by users during the access and reuse phases of the research data lifecycle. Numerous web-based initiatives are testament to this type of metadata crowdsourcing (or 'social tagging'). ArtUk, an online site for art from every public collection in the UK, recently launched, ArtUK Tagger , which allows the public to add multiple tags to paintings. An algorithm then calculates which tags are likely most accurate and feeds these tags through to the Art UK website. Similarly, the Philadelphia Museum of Art encourages online visitors to tag objects in the online collection in order to improve access to works of art. Social tagging initiatives not only exist to provide better subject access to images but also to assist with quality control and processing functions. MicroPasts - for example, encourages users to help with location accuracy of artifact findspots and photographed scenes as well as the masking of photos intended for 3D modelling. Zooniverse is yet another platform that is designed to use volunteers to sort through and help classify excessive numbers of research images. "Our goal is to enable research that would not be possible, or practical, otherwise."

While crowdsourcing initiatives and semantic-based search functionality can improve access to relevant images, building an image database based on shared standards remains a challenge given the diversity of image collections, widely varying budgets and differing individual requirements of user groups (Bourne, 2005). Regarding the varying requirements of users, Chung and Yoon (2011) found that users were more likely to search based on abstract meaning when images were intended as objects but not when used as data. Another constraint that has similarly plagued the management of numeric data collections in university settings is that images (particularly slides) generate data that are very different from those handled by cataloguing systems created for books (Bourne, 2005).

Re-using Images



In addition to browsing, search and retrieval challenges associated with managing access to images, a discussion of the closely associated re-using phase of the data lifecycle is not complete without attention to image copyright as well as issues regarding confidentiality or data sensitivity. Generally speaking, factual numeric data in and of itself, represented in an obvious file structure, is not copyrightable in Canada as a work needs to be original for copyright to exist. This holds no matter how much work goes into collecting the data (Potvin, 2008). However, most numeric data need to be analyzed and processed and so the program code developed for these purposes is considered a 'literary work' (Potvin, 2008). Also, once data are formatted into, for example, a relational database, graph or dataset, it can be subject to copyright.

As previously mentioned, images, unlike traditional numeric data files, are not always created for the purpose of data analysis. As such, most literature discussing images and copyright reference images as artistic works that are copyrightable. Images considered to be artistic works in the UK, for example, include: blueprints, building plans, cartoons, charts, decorative graphics, diagrams, drawings, engravings, graphs, illustrations, logos, maps, moving images, paintings, photographs, sculptures and sketches. According to the Government of Canada images as artistic works include: patterns, art slides, maps, paintings, architectural drawings, plans, digital images, drawings, photographs, charts, and art prints.

Just as with copyrightable numeric datasets, it is necessary to clarify who has primary ownership of the datasets since copyright of a work comes into existence at the point of creation (i.e., authors/creators). If images are treated as data then similar ownership and rights issues apply when figuring out how the images will be managed and disseminated. This will mean assessing whose rights need to be considered, for example: funders, institutions, research participants, collaborators, publishers and the public (MANTRA, 2014).

Copyright is undeniably complicated and there are obvious, notable exceptions to the principle of creator as copyright holder. For example, U.S. federal copyright law denies copyright protection for works produced by the US federal government so NASA's images, for instance, are in the public domain, but individuals who create images based on data released by NASA can assert limited copyright because they have created derivative works or compilations.

Additionally, copyright surrounding images is context-specific. Images may simply exist as facts (equivalent to a numeric data file), in which case, they are not subject to copyright. Copyright may also not be an issue if copyright has expired or images are considered to reside in the public domain. Some image owners may also allow reuse for non-commercial purposes (i.e., education) but require attribution (e.g., Creative Commons Attribution). The educational sector may also find that images fall under fair use or fair dealing. In support of this, the Visual Resources Association (VRA) in the U.S. published a statement on the fair use of images for teaching, research and study (Wagner & Kohl, 2012). For teaching and study purposes, this statement covers preservation, use (both high-resolution and thumbnails), adaptations, sharing and reproduction. Also, if images are photographs, they are likely to be treated as original works and subject to normal copyright restrictions. Collections of images may be copyrightable if they exist as a database or as a result of researchers creating added value to images.

Ethical considerations, specifically privacy and confidentiality, also merit careful consideration when managing and sharing images. As with numeric datasets, researchers managing and disseminating images need to minimize the risk of disclosing confidential information and re-identifying study participants. One broad technique for safeguarding confidentiality of numeric data includes either collecting data without identifiable information or anonymizing data post-collection through de-identification processes. Best practices for handling sensitive image data may include the anonymization of facial and location identifiers in digital photographs. The actual techniques for doing so, however, may pose unique challenges for images. For example, in a 2016 email thread on the Jisc Research Data Management Listserv, the issue of anonymizing image data proved labour intensive and expensive for a use case where anonymization was required for a large collection of images. They couldn't find a freely available tool that could effectively bulk-blur identifying characteristics in the images. In another comment from the same thread, a researcher noted that obscuring faces by pixelating sections of a video image could greatly compromise the usefulness of data. Alternative strategies to anonymization noted by many researchers are to either gain consent to share, or to consider controlled access so that the usability of the images can remain unaltered. To maximize the effectiveness of these alternative strategies to anonymization, strong recommendations are made to consider and judge at an early stage the implications of depositing images with confidential information.

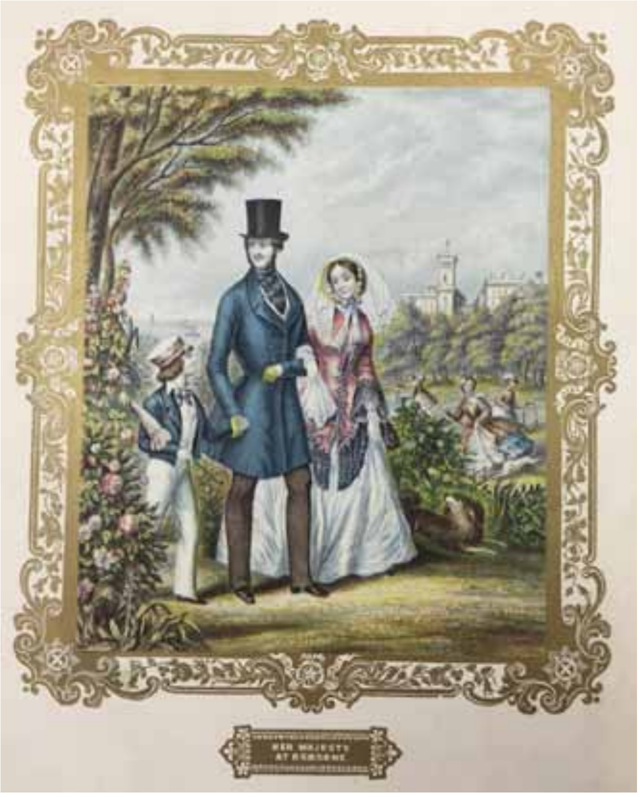


Figure 1

Source: LeBlond & Co. Her Majesty at Osborne. Regal series ca 1850. Print collection, Rare Books and Special Collections, McGill University.

Conclusion

The unprecedented growth of research image collections across disciplines, coupled with increasingly powerful instruments and devices for image capture, have created challenges and new opportunities for managing images across the research data lifecycle. This paper offers some preliminary recommendations for managing images as data by looking to established research data management practices for traditional numeric datasets.

Analogously, images, like survey respondents will provide information in the form of data. But like people, images are inherently richer than the discrete slices of data that are extracted by research instruments. Uniquely, images pose challenges in terms of size and volume, especially for storage and preservation. The creation of robust metadata is also complicated due to the subjectivity of image meaning and the difficulty in anticipating the search needs of users. Also, automating the process of metadata creation is difficult and manual description remains necessary. Some emerging solutions to these challenges are noted, including crowdsourcing initiatives for data processing and description as well as semantic-based retrieval systems.

References

Avondo, J. (2010) BioformatsConverter. (Available at cmpdartsvr1.cmp.uea.ac.uk/wiki/BanghamLab/index.php/BioformatsConverter)

Bourne, M. Image data. VRA Bulletin, 31(3), 26-29. (Available at http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=7&sid=106aa9eb-e435-4f6f-a71e-01e5c781e0f3%40sessionmgr106&hid=107)

Center for History and New Media (2016). RRCHNM to build software to help researchers organize digital photographs. (Available at https://chnm.gmu.edu/news/rrchnm-to-build-software-to-help-researchers-organize-digital-photographs)

Chung, E. and Yoon, J. (2011). Image needs in the context of image use: An exploratory study. Journal of Information Science, 37(2), 163-177 (Available at http://jis.sagepub.com/content/37/2/163.short)

Corti, L., Van den Eynden, V., Bishop, L., and Woollard, M. (2014). Managing and sharing research data: A guide to good practice. Los Angeles: Sage Publishing,

Humphrey, C. (2006) e-Science and the Life Cycle of Research. (Available at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc)

DDI Alliance (2013) Data Documentation Initiative. (Available at http://www.Ddialliance.org)

Data Curation Centre (DCC) (2012). (Available at http://www.dcc.ac.uk/resources/curation-lifecycle-model)

EDINA. (2014). MANTRA (Available at http://datalib.edina.ac.uk/mantra)

Eadie, M. (2008). Towards an application profile for images. ARIADNE: Web Magazine for Information Professionals. http://www.ariadne.ac.uk/issue55/eadie

Geraci, Humphrey, & Jacobs (2012). Data Basics: an introductory text. Unpublished. Local PDF file.

Gross, M. (2011). Riding the wave of biological data. Current Biology, 21(6), R204-R206. (Available at http://ac.els-cdn.com/S0960982211002818/1-s2.0-S0960982211002818-main.pdf?_tid=e46374de-026b-11e6-b386-00000aab0f26&acdnat=1460657489_27698d6d824b844bbc3b43f55c85558e)

ICPSR (2016) Data Management & Curation: Metadata (Available at https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/metadata.html)

Jimenez-Maggiora, G. A., Thomas, R. G., Brewer, J., Bruschi, S., Hong, P., and Aisen, P. S. ADCS Electronic Data Capture (EDC) - Integrated Multi-Modal Image Management for Clinical Trials in Alzheimer's disease. Neurosciences Department, University of California at San Diego: La Jolla, CA,(Available at https://www.researchgate.net/profile/Gustavo_Jimenez-Maggiora/publication/269038101_ADCS_Electronic_Data_Capture_(EDC)_-_Integrated_Multi-Modal_Image_Management_for_Clinical_Trials_in_Alzheimer's_Disease/links/548734e30cf268d28f071e7c.pdf)

Jisc Digital Media (2016) (Available at http://www.Jiscdigitalmedia.ac.uk)

Kane and Pear (2016) (Available at http://sloanreview.mit.edu/article/the-rise-of-visual-content-online)

Linkert, M., Rueden, C.T., Allan, C., Burel, J.M., Moore, W., Patterson, A., Loranger, B., Moore, J., Neves, C., MacDonald, D. and Tarkowska, A., (2010). Metadata matters: access to image data in the real world. The Journal of cell biology, 189(5), 777-782.

MarketsandMarkets. (2016). Rising volume of medical imaging data to increase the adoption of cloud computing in the healthcare sector. (Available at http://www.marketsandmarkets.com/ResearchInsight/north-america-healthcare-cloud-computing.asp)

McCook (2016) Retraction Watch. Don't trust an image in a scientific paper? Manipulation detective's company wants to help. Retraction Watch. (Available at http://retractionwatch.com/2016/02/24/dont-trust-an-image-a-new-company-can-help)

Metadata Working Group. Guidelines for handling image metadata. (2010). (Available at http://metadataworkinggroup.com/pdf/mwg_guidance.pdf)

Moore, Allan, Burel, Loranger, MacDonald, Monk and Swedow (2008). Open Tools for Storage and Management of Quantitative Image Data. Chapter 24 (Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.9978&rep=rep1&type=pdf)

Potvin, J. (2008). How is copyright relevant to source data and source code? Technology Innovation Management Review. (Available at http://timreview.ca/article/121)

Primary Research Group. (2013). Survey of Best Practices in Digital Image Management. New York: Primary Research Group.

Keeney, A. R. and Rieger, O. Y. (2001). Report of the Digital Preservation Policy Working Group on Establishing a Central Depository for Preserving Digital Image Collections. (Available at https://www.library.cornell.edu/preservation/IMLS/image_deposit_guidelines.pdf)

Research Data Canada, (2014) (Available at http://www.rdc-drc.ca)

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods, 9(7), 671-675. (Available at https://www.researchgate.net/profile/Kevin_Eliceiri/publication/228085958_NIH_Image_to_ImageJ_25_years_of_image_analysis/links/0fcfd4fee589e852eb000000.pdf)

Statistics Canada. (2015). Section 4: Data. (Available at http://www.statcan.gc.ca/eng/dli/guide/toc/3000276)

UKDA (2011) (Available at http://www.data-archive.ac.uk/media/2894/managingsharing.pdf)

Wagner, G. and Kohl, A. (2011). Visual Resources Association: Statement on the fair use of images for teaching, research and study. VRA Bulletin, 38(1), 1-18.

Wang, H. H., Mohamad, D., and Ismail, N. A. (2010). Toward semantic based image retrieval: review. Proc. SPIE 7546, Second International Conference on Digital Image Processing, 754626 (Available at doi:10.1117/12.853332).

Notes

1. Berenica Vejvoda (MISt, University of Toronto) is a data librarian at McGill University. Prior to McGill, Berenica worked as a data librarian at the University of California at San Diego and at the University of Toronto. berenica.vejvoda@mcgill.ca
2. K. Jane Burpee (MLIS, McGill University) is the Coordinator, Data Curation and Scholarly Communications at McGIll University. She has been active in the area of scholarly communication since 2000 when she became a scholarly communication librarian at the University of Guelph. She is a leading voice for open access and champions the transformation of scholarship. jane.burpee@mcgill.ca
3. Paula Lackie (MA, A.B.D., University of Southern California) is the Academic Technologist for Data at Carleton College. She is long-time research data advocate and social science and humanities technologist. plackie@carleton.edu
4. The Research Data Lifecycle (Humphrey, 2006; DCC, 2012; DDI Alliance, 2013, MANTRA, 2014).
5. Lossy compressions transform and simplify the media information in a way that gives much larger reductions in file size than lossless compressions. While the file becomes significantly smaller, quality of the image is compromised during the compression process (e.g. JPEG). Conversely, lossless compression results in no information loss, however, the image files are much larger (e.g. TIFFs) Jisc, 2016 (Available at http://www.Jiscdigitalmedia.ac.uk/infokit/file_formats/lossless-and-lossy-compression)
6. JCB DataViewer (https://datahub.io/dataset/jcb-dataviewer), launched in 2008, for archiving and sharing original image data in the life sciences, allows users to download original image data in an open, standardized data format and preserves the original image metadata (OME tagged image file format [TIFF]). Similarly, the Jisc-funded Data Management for Bio-Imaging project at the John Innes Centre developed BioformatsConverter software (Avondo, 2010) to batch convert bio images from a variety of proprietary microscopy image formats to the Open Microscopy Environment format, OME-TIFF. OME-TIFF, is an open file format that enables data sharing across platforms and maintains original image metadata in the file in XML format (UKDA, 2011).
7. Moore's Law refers to the long-standing pattern that computer processing power will double every two years.
8. Big data is currently an ill-defined term that at its root simply refers to extremely large data files that require greater than average computational power to manipulate and/or analyze.
9. Tropy: http://chnm.gmu.edu/news/rrchnm-to-build-software-to-help-researchers-organize-digital-photographs
10. RenameIT https://github.com/wernight/renameit
11. ImageMagick: http://www.imagemagick.org
12. A "thumbnail" is a very small version of the original image. Thumbnail versions are useful as a kind of wordless summary of the image.
13. ImageJ: https://imagej.nih.gov/ij/
14. ArtUK http://artuk.org/tagger
15. Micropasts Crowdsourcing: http://crowdsourced.micropasts.org
16. Zooniverse is a "platform for people-powered research": https://www.zooniverse.org The development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.