THE ONGOING CHALLENGES OF CITING THE RESULTS OF SCHOLARLY RESEARCH

BY MAUREEN C. KELLY, PUBLISHING CONSULTANT

"Scholarly communication has built up an important tradition of citation. It reflects the fact that in all areas of research [...] we progress by building on the past. And we acknowledge our debt to the past by citation to it. By doing so, we assure that our sources can be checked, verified, validated. But that implies that the material so referenced, so cited must be available for checking, verifying, validating. What happens if the source data [is electronic] and has been erased, or worse yet, altered since it was last used? The entire structure of scholarly progress would collapse."
—Dr. Robert Hayes (1992).

Content Changed, Standards Followed

Dr. Hayes made an important observation. But we need to look very closely at the transition from print journals to electronic records of research in order to fully understand the challenges we've faced in the past few decades, and the ones that lie ahead. And we need to keep in mind that the goal is not necessarily about creating traditional "citations." Rather, it is about being able to reliably identify, locate, and access prior research records.

We have a long tradition of communicating research results neatly wrapped in a journal article and packaged in a journal issue. The practice dates back to 1665, when the Royal Society first published Philosophical Transactions. That publication "pioneered the concepts of scientific priority and peer review which, together with archiving and dissemination, provide the model for almost 30,000 scientific journals today" (Royal Society, 2016). And for most of those 350 years, journals were distributed in printed formats.

Over the past half century, technology has driven significant changes in that paradigm. In the early stages of change, e-journals were offered as complements to print journals. Libraries often subscribed to both forms, and the digital version was basically a replica of the print one: a PDF

version of the printed journal pages. The content remained stable and citable. Publishers' production systems and library delivery systems adapted to meet those changes, but the content was still locked within traditional journals and articles. There was an "official," fixed presentation of the research results.

Changes in this process were initially driven by pressure to speed up the publication cycle. Journals began to publish e-first articles, which created challenges for traditional citation metadata because the page numbers were often unavailable when the e-first version was released. Sometimes the e-version was later updated to include the pagination. Questions arose as to which was the version of record. But still the content remained discoverable and citable.

Over the last ten years we have seen a more substantive shift in publishing and library practices as e-journals have largely replaced print journals in libraries and in the economics of scholarly publishing. Over that time, we have also seen a refocusing away from the journal and journal issue as the container for scholarly content. Now, the focus is on electronic databases of articles where the journal and issue information are used simply as supporting metadata. We also see cases in which the electronic version of a journal issue contains more information than its print counterpart.

Fully electronic versions of scholarly content, with available XML and HTML, offer significant advantages, but they also bring on new challenges. Content is packaged in large databases and is remotely accessible. Search engines have replaced abstracting and indexing services as the tools for discovery. Different versions of articles may be available from preprint servers and institutional repositories. Google Scholar lets us search across multiple databases and often provides landing pages for content that is behind a firewall. Gone are the days when we would go the library and scan the shelves or ask the librarian to locate an article for us—now, scholarly information discovery and access have become a do-it-yourself enterprise for researchers.

As scholarly content moved to electronic formats, supporting standards followed. We still use style guides to prepare our citations when submitting articles for publication, but we now have the advantage of bibliographic data formats that are more actionable in an electronic environment.

In the late 1960s the ISBN came on the scene, followed by the ISSN in the early 1970s. The bar-coded versions of these standards have become important to inventory and purchasing systems. (The ISSN gained traction after the post office made it a requirement for second-class mailing permits.) But the major standards initiative for streamlining

CONTINUED »

content management in an electronic environment has been the DOI (Digital Object Identifier) System. It complements rather than replaces standard citations by providing a concise code for each digital object, rather than a user-friendly description of the resource. The latest implementation of the DOI is presented as a URL (Uniform Resource Locator) structure. In spite of this change in format, it remains a reliable identifier of the digital object rather than its location on the Internet, which can be unstable. The URL resolves to an underlying registration system rather than the open Internet. So it remains a surrogate for 'the cite' rather than 'a site.' The work of the DOI System has been complemented by work in the areas of library link resolvers, most-appropriate-copy resolution, smart landing pages, and similar developments. The DOI has been incredibly successful but it remains a work in progress as it continues to adapt to new content types and to new discovery and delivery environments.

Content Continues to Change; Standards Follow

We have begun to see changes in the output of scholarly research that impact what constitutes a publishable knowledge object. These changes go beyond whether the journal article is in print or electronic format. New types of research results are causing us to reevaluate whether a traditional journal article is a sufficient container for distributing scholarly knowledge.

The DOI has been incredibly successful but it remains a work in progress as it continues to adapt to new content types and to new discovery and delivery environments.

Scholarly research methods have changed since 1665, and new types of research results have become critical to the research process. It has become clear that the traditional journal article model is insufficient. In 1997, I gave a talk at the ICSTI meeting followed by an article in ICSTI Forum (Kelly, 1997).

In that article, I argued that "neither print journal articles or books nor their electronic equivalents are sufficient to the task ahead." Rather, "we must look beyond the current, text-centric paradigm." It was becoming apparent as early as 1997 that we needed new channels for sharing research results in a way that was more functional and reusable than a journal article presentation method could accommodate. Research results were becoming increasingly complex and much value was being lost by reducing them to static text and tables.

GenBank was the first major data initiative to break out of the journal article publishing paradigm. It began at Los Alamos and transitioned to the National Center for Biotechnology Information (NCBI). In the 1980s, as genetic sequences became a significant research output, journals struggled to publish these strings of letters in text form. Yes, they really did print pages and pages of A-T-G-C combinations. Imagine the challenge of copyediting such strings. But more important, imagine the loss of value that resulted from reducing this knowledge into simple text strings. As GenBank became established, publishers joined the move by declining to publish articles until the sequences had been deposited in GenBank. Little by little, GenBank has grown in sophistication and functionality to become a cornerstone of genetic research, regularly facilitating important new discoveries.

Many other communities, such as astronomy and geology, now rely on databases to facilitate collaboration and discovery. Publishing an article in a traditional journal remains important for giving recognition and creating career opportunities for researchers. But researchers increasingly need access to the underlying data; they want to cite it and incorporate it into new research.

This will not be an easy transition for scholarly communications, as there are many challenges associated with publishing research data. There are valid concerns regarding the challenges of peer reviewing data and risks of data piracy. Sufficient metadata is needed to provide the context of data collection and to support discovery and reuse. The list of challenges is long and valid, and these issues must be addressed. Researchers need and deserve recognition for their work, and they need a publication and citation environment that will work in this changing research environment.

Publishing and Citing Data

Given that scholarly research is all about collecting and analyzing data, we need to develop rigorous practices for sharing and citing data collections. Botanical gardens, with their vast stores of pressed plant leaves, are an early acknowledgement of the value in preserving data collections. As electronic datasets grow in size and complexity, the challenges for making them citable also grow.

Datasets are more complex than text, and since they are not static (and risk failing Dr. Hayes's test for reliable citation), we need provisions for versioning. But for datasets to be citable requires more than just the development of a data-citation format. We need an infrastructure that supports storage, curation, distribution, discovery, and access control for these new knowledge assets.

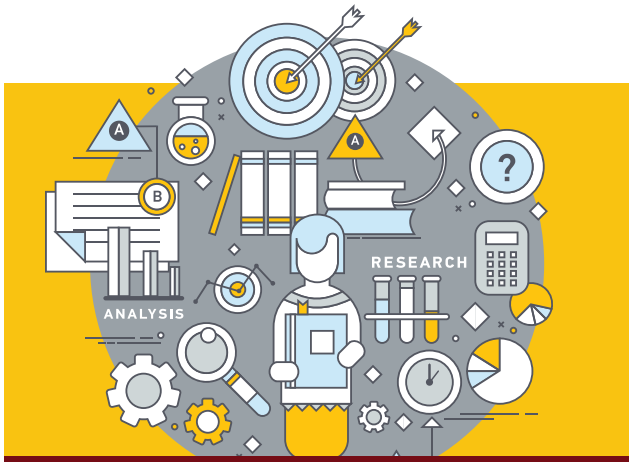
University libraries have served that role for journal articles through the print and electronic publishing eras. The value they bring to a university and its researchers is well understood. Yes, there are issues about journal pricing, and yes, there are drivers for open-access publishing. But the costs of this infrastructure have been considered a legitimate part of a university's role in serving researchers and students.

There is no comparable infrastructure available for archiving, curating, and providing access (and rights management) to our growing collections of datasets. The government, of course, plays a critical role in supporting data collections such as GenBank, but government data repositories are not sufficient.

Each research subject area has different requirements and practices. Consider, for example, the differences between genetic and astronomical data. Consider also the differences between experimental data and field data. Experiments can be repeated and the results validated; field data cannot be collected again under identical circumstances. Marcia McNutt has recently published a useful discussion on this topic including the issue of citation standards for data (McNutt, 2016).

While there is currently no business model for constructing and maintaining the necessary infrastructure for data collections, scholars haven't waited for official solutions: instead they saw the need and an opportunity to advance knowledge, and have moved forward. Infrastructure and standards must follow. At present, we have a variety of solutions in play and a variety of data-citation practices under development.

In 2008, NISO conducted a "Thought Leader Meeting on Research Data" at which the topic of data citation was discussed. Among the observations made was the need to accommodate the differing requirements found across disciplines and fields of research. One of the group's recommendations was for NISO to work collaboratively



Given that scholarly research is all about collecting and analyzing data, we need to develop rigorous practices for sharing and citing data collections.

with other organizations to develop guidelines for data citation. In 2009 (rev. 2010), Toby Green prepared a white paper for OECD expressing concerns about long term discoverability and accessibility of datasets: "We Need Publishing Standards for Datasets and Data Tables" (Green, 2009). Style guides now include provisions for citing electronic and web resources. MLA even has guidelines for citing a Twitter post. They call for typical citation metadata, but with special accommodations such as using the archive where the dataset is housed in lieu of the publisher. When a DOI is unavailable, they recommend use of a persistent URL.

Since the DOI System was designed to support traditional publishers, it is not automatically suited to accommodating large numbers of researchers directly. In 2009, DataCite.org was established by the British Library, the Technical Information Center of Denmark, TU Delft Library, the National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI), California Digital Library, Purdue University, and the German National Library of Science and Technology. The group assists members in using the DataCite Service for minting DOIs and registering associated metadata. DataCite also offers a metadata schema (Starr, 2011) and recommends that a new DOI be issued when changes are made to the dataset or new versions are created.

In September 2013, CODATA and ICSTI authored a report with the clever title: "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data" (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). Its abstract echoes and updates the observations of Dr. Hayes: "The use of

"As technological factors, such as faster processors, better storage, and increased bandwidth, have enabled the much greated production and capture of data, the creation of standards to manage these data had not kept pace."

published digital data, like the use of digitally published literature, depends upon the ability to identify, authenticate, locate, access, and interpret them. Data citations provide necessary support for these functions…." The report also stresses that "As technological factors, such as faster processors, better storage, and increased bandwidth, have enabled the much greater production and capture of data, the creation of standards to manage these data has not kept pace." The report offers a set of "guiding principles" as well as challenges to implementation.

Publishers are also working to facilitate data publishing and citation. Nature.com produces and hosts Scientific Data, an open-access, peer-reviewed journal for descriptions of scientifically valuable datasets in the natural sciences. The primary article-type is a "Data Descriptor" that is designed to make data more discoverable, interpretable, and reusable. It does not store the data but rather relies upon public, community-recognized repositories. (See http://www.nature.com/sdata/publish/for-authors#aims-scope)

Thomson Reuters offers a Data Citation Index. The Data Citation Index captures all available metadata for the data repositories it indexes. Since the metadata in those repositories can vary in format and detail, Thomson Reuters is working to establish a more consistent, descriptive data-citation format. (See http://wokinfo.com//products_tools/multidisciplinary/dci/repositories/)

Elsevier also has initiatives for supporting research data. The company offers its own data repository via Mendeley. Each dataset is given its own DOI and is archived through Data Archiving and Networked Services (DANS). Elsevier offers an Open Data pilot initiative where research data can be made openly available on ScienceDirect under a CC-BY license. It also has a Data Link tool that supports data discovery and includes a database search engine, an automatic data-citation generator, a data article writing tool for the *Genomics Data* journal, and a data visualization tool. (See https://www.elsevier.com/about/open-science/research-data)

Biomed Central is working with DataCite to address concerns raised about their OpenData policy and the legal (copyright) status of data published in their Open Access journals. (See https://www.biomedcentral.com/about/policies/open-data).

Dataverse.org at Harvard is an open source web application to share, preserve, cite, explore, and analyze research data. Harvard's Institute for Quantitative Social Science (IQSS), the creator of the application, is working on a set of guidelines for tiered access. The levels of access include Open; Guestbook; Required Acceptance of Terms of Use; and Restricted Access, which requires a specific access request. Dataverse's statement of best practices (Data Science at The Institute for Quantitative and Social Science, 2015) is very useful in laying out the many factors that come into play when using a dataset.

Force11 has developed a Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014) that is endorsed by many commercial and scholarly publishers including Elsevier, Nature Publishing Group, AGU, AIP, APS, PLIO and, of course, NISO. These principles describe what is needed for a data citation to be functional:

1. ***Importance:***
   Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

2. ***Credit and Attribution:***
   Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

3. ***Evidence:***
   In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

4. ***Unique Identification:***
   A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

5. ***Access:***
   Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials as are necessary for both humans and machines to make informed use of the referenced data.

6. ***Persistence:***
   Unique identifiers, and metadata describing the data, and its disposition, should persist – even beyond the lifespan of the data they describe.

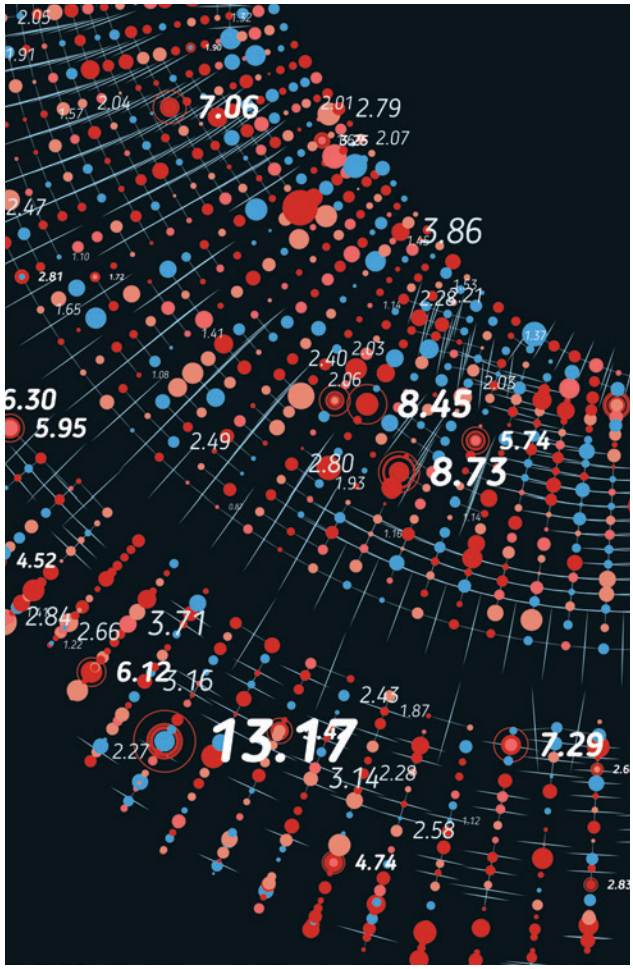
7. ***Specificity and Verifiability:***
   Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version, and/or granular portion of data retrieved subsequently is the same as was originally cited.

8. ***Interoperability and flexibility:***
   Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

As the volume of data continues to grow, agreed-upon standards and practices, adapted for the needs of individual areas of research, become ever more important in ensuring that data can be cited and discovered. It is reassuring to see such a strong community-wide effort toward developing and promoting reliable data-citation practices, but much work is required in the area of credentialing and supporting the data repositories that are needed to maintain, curate, and provide access to datasets.

Publishing and Citing Software

Datasets, for all their challenges, are not the last hurdle for sustaining our ability to cite the knowledge objects produced by today's researchers. We have come a long way since it was sufficient for researchers to tuck their data into Excel spreadsheets. Datasets are now more complex and require more sophisticated tools to analyze and extract value.



Datasets, for all their challenges, are not the last hurdle for sustaining our ability to cite the knowledge objects produced by today's researchers. We have come a long way since it was sufficient for researchers to tuck their data into Excel spreadsheets.

In my 1997 ICSTI paper, I anticipated that new forms of research output would include "computational models and simulations along with other collections of functional information." As research becomes more data intensive, software is developed to process the data; that software is an integral part of making data sets functional. Elsevier estimates that 38 percent of researchers now spend at least one day per week creating software to analyze the data they have collected.

It has become important for software to be treated as a valid part of the scholarly record. When custom software is the means by which the data is processed and conclusions are drawn, like data, it needs to be published in a functional form. It cannot be usefully reduced to text in a journal article any more than genetic sequences could usefully be published as printed strings of A-T-G-C combinations.

Versioning is a critical issue for publishing and citing software. And it is tricky to accomplish. While a version of software can be cited, there is a possibility that the software includes a call-out to a code library that may have been changed.

To this end, the software community has been active in developing software repositories and version control tools. GitHub, one of the most widely used, supports private repositories and free, open-source accounts. Another approach to making software reusable and citable has been the development and use of "reproducible reports" (Visser, 2014). The motivation for this approach was offered by Donald E. Knuth back in 1984, when he said, "Let us change our traditional attitude to the construction of programs. Instead of imagining that our main task is to instruct a *computer* what to do, let us concentrate rather on explaining to *human beings* what we want a computer to do" (Knuth, 1984).

Reproducible reporting provides a way for researchers to package together all the components of their work, including the workflow, data, and code into a shareable–and potentially citable–package.

Reproducible reporting provides a way for researchers to package together all the components of their work, including the workflow, data, and code into a sharable—and potentially citable—package. Among the better known tools that have been developed to aid researchers in preparing reproducible reports are Galaxy, Jupyter Notebook (formerly IPython), and knitr, a dynamic report generator for R.

We are also seeing commercial publishers providing support for making software citable. Articles published in *Nature Methods* increasingly support supplementary software files, most of which include source code. Nature encourages the use of code repositories such as GitHub prior to submission of an article. Using these repositories expedites the peer-review process and avoids the necessity for reviewers to test the code on their own computers.

Elsevier has started *Original Software Publications* to describe significant software and/or code. The software will be peer-reviewed and considered "one body of work" for citation and indexing purposes. The software/code itself will be deposited on the journal's GitHub, and Elsevier states that, "all software and code published is, and will remain, fully owned by their developers."

Loose Ends and New Forms of Content

Technology has made it possible for scholarly content to be distributed in new and less formal ways. In the past we talked about the "Invisible College" and grey literature. Now we talk about scholarly collaboration networks (SCNs) and Social Sharing Networks (SSNs) such as Mendeley (now owned by Elsevier) with five million members, Academia.edu with 30 million monthly users, and ResearchGate with six million members—levels of usage that make these important channels for scholarly communication. And to the extent that publishing means making information public, these networks represent a new form of publishing. The topic, particular concerns about sharing of journal articles on these networks, has been discussed on the Scholarly Kitchen site (Meadows, 2015). It is difficult to conceive of how such communication can be made citable, but it is an issue that warrants creative consideration.

E-print servers such as arXiv (at Cornell), bioRxiv (at Cold Spring Harbor) and SSRN (Social Science Research Network, recently acquired by Elsevier) are well-regarded

content distribution channels in their fields. Like scholarly collaboration networks, they provide for rapid communication of research results. While they do not offer the final, citable form of the paper, they are widely used and important components of the scholarly communication process.

New, informal channels continue to pop up. Consider @scholarlycomm on Twitter. It's part of Columbia's scholarly communication program and explores new ways to share, curate, and preserve new knowledge. There are also Twitter handles from SSP (@ScholarlyPub) and Harvard (@oscharvard), among others. While they do not constitute formal communication, they have become important channels for sharing scholarly information. These new, informal channels do not lend themselves to traditional citation, but they have become an integral part of how today's scholars communicate. And as mentioned earlier, they are sufficiently important that style guides now provide instructions on citing this type of content.

This is just a sampling of how scholars continue to work and communicate in new ways. Some of these assets and communications warrant citation in bibliographies (sometimes with URLs though not DOIs). But once they have been cited, however informally, we are faced with Dr. Hayes's concern about how they will be made available over time for "checking, verifying, validating."

My concerns for the future of citation are not limited to the standards and structure of citations for new types of content. We need to think seriously about how scholarly research findings will be archived, curated, and made accessible for future use and reference. The role of libraries as archives of scholarship is changing, as are publishing business models. Scholarly societies, once such important publishers of scholarly research, are being overtaken by large commercial publishers that have the resources to invest in new functionality needed to deal with dataset and software citation. We can hope that the government will continue to fill some of this role. But that will not be sufficient. Commercial publishers will certainly play a significant role. But will open source publishing enterprises like PLoS have the resources to preserve their collection? What role will libraries play in this new paradigm?

Archiving is not easy; curating is not easy. It requires a long commitment and the resources to sustain that commitment. The problem exists in many fields beyond publishing from old movies to bacterial cultures. But for us, it is a problem of the sustainability of the citations we create. Creating citations, and creating standards for citations will not be enough if they all resolve to dead links.

REFERENCES

CODATA-ICSTI Task Group on Data Citation Standards and Practices. "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data." *Data Science Journal* 12 (September 2013).
http://doi.org/10.2481/dsj.OSOM13-043

Data Citation Synthesis Group. "Joint Declaration of Data Citation Principles." Martone M. (ed.) San Diego CA: FORCE11. (2014).
https://www.force11.org/group/joint-declaration-data-citation-principles-final

Data Science at The Institute for Quantitative and Social Science. "Harvard Dataverse General Terms of Use." (2015).
http://best-practices.dataverse.org/harvard-policies/harvard-terms-of-use.html

Elsevier. "Original Software Publications." (2016).
https://www.elsevier.com/books-and-journals/content-innovation/original-software-publications

Green, T. "We Need Publishing Standards for Datasets and Data Tables." OECD Publishing White Papers (2009). doi:10.1787/787355886123
http://dx.doi.org/10.1787/787355886123

Hayes, Robert M. "The Needs of Science and Technology." *Science and Technology Libraries* 12, no. 4 (1992): 3-33.

Kelly, Maureen C. "The Role of A&I Services in Facilitating Access to the E-Archive of Science." ICSTI Forum: The Quarterly Newsletter of the International Scientific and Technical Information, no. 26 (November 1997).
http://www.informedstrategies.com/wp-content/uploads/2015/10/Facilitating_Access_to_the_eArchive_of_Science_Nov_97_MCKelly_ICSTI_.pdf

Knuth, Donald E. *Literate Programming. Center for the Study of Language and Information. (1984).*

McNutt, M. "Liberating field science samples and data." *Science 351*, Issue 6277 (4 March 2016): 1024-1026. DOI: 10.1126/science.aad7048

http://science.sciencemag.org/content/351/6277/1024 /

Meadows, Alice. "Article Sharing on Scholarly Collaboration Networks – An Interview with Fred Dylla about STM's Draft Guidelines and Consultation." Scholarly Kitchen. (February 24, 2015).
https://scholarlykitchen.sspnet.org/2015/02/24/article-sharing-on-scholarly-collaboration-networks-an-interview-with-fred-dylla-about-stms-draft-guidelines-and-consultation

The Royal Society. "350 Years of Scientific Publishing." (2016).
https://royalsociety.org/journals/publishing-activities/publishing350/.

Starr, Joan. "isCitedBy: A Metadata Scheme for DataCite." D-Lib Magazine 17, no. 1/2. (January/February 2011).
http://www.dlib.org/dlib/january11/starr/01starr.html

Visser, Ingmar. "Why Reproducible Reporting?" Open Science Collaboration. (October 2014).
http://osc.centerforopenscience.org/2014/10/30/reproducible-reporting/