# A Perspective on Archiving the Scholarly Web

Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos, NM, USA
http://orcid.org/0000-0002-0715-6126/
herbertv@lanl.gov – @hvdsomp

Andrew Treloar
Australian National Data Service
Melbourne, VIC, Australia
http://orcid.org/0000-0002-8911-3081/
andrew.treloar@ands.org.au – @atreloar

## ABSTRACT
As the scholarly communication system evolves to become natively web-based and starts supporting the communication of a wide variety of objects, the manner in which its essential functions – registration, certification, awareness, archiving - are fulfilled co-evolves. This paper focuses on the nature of the archival function based on a perspective of the developing future scholarly communication infrastructure.

## General Terms
Infrastructure, preservation strategies and workflows, theory of digital preservation

## Keywords
Scholarly communication, web preservation

## 1. THE FUTURE ACCORDING TO THE PAST
The 2004 paper "Rethinking Scholarly Communication" [3] observed significant trends in the way scholarly communication was then evolving as a result of the gradual, yet steady, transition towards a digital and network-based endeavour.

Based on these observations, the paper revisited the perspective of what constitutes a unit of communication, moving beyond journal publications, and including a wide variety of objects such as datasets, simulations, software as well as compound aggregations of such objects linked together using appropriate relationships.

The paper also pointed at the possibility of a profound reconfiguration of the scholarly communication system enabled by the networked technologies. It did so guided by the theoretical perspective developed by [2] of the essential functions that must be fulfilled by any system of scholarly communication, irrespective of its implementation:

- Registration: Allows claims of precedence for a scholarly finding
- Certification: Establishes validity of claim
- Awareness: Allows actors in the system to remain aware of new claims
- Archiving: Preserves the scholarly record over time

In the system of journals, these functions were vertically integrated in the journal-centric ecosystem: a journal took care of registering claims by accepting manuscripts, of certification through the peer-review process it coordinated, of awareness through its availability in libraries, and of (distributed) archiving by means of its long-term presence on library shelves, worldwide.

However, as soon as the Web made it possible to communicate digital information across a global network, signs of a future in which the functions of scholarly communication would no longer be fulfilled in a vertically integrated manner became apparent:

- The preprint movement, led by arXiv.org, then still at the Los Alamos National Laboratory, demonstrated the value of allowing manuscripts to be submitted (registration) and discovered (awareness) with out applying a certification process to them.
- As soon as journals went digital and were baptized e-journals, their preservation was no longer the sole concern of libraries. Quite to the contrary, publishers and special-purpose organizations such as Portico started fulfilling the archival function, thereby disconnecting it from the tight connection it had for centuries with the awareness function.

The 2004 paper drew these indicators to their logical conclusion by pointing at the future possibility of a web-based scholarly system in which the essential functions are fulfilled in discreet, disaggregated, and distributed manners, and in which a variety of networked pathways interconnect the autonomous hubs that fulfil these functions. Inspired by preprints, and motivated by a desire to increase the speed of discovery, the paper further made a plea in support of early registration – decoupled from certification - of the brave new objects of scholarly communication.

## 2. THE FUTURE IS NOW
Ten years later, indicators of both the changing nature of the objects of scholarly communication and of the disaggregated fulfilment of the essential functions of a scholarly communication system are abundant. To quote William Gibson, "The future is already here – it's just not very evenly distributed"[1]. Although indicators exist across scholarly disciplines, the life sciences provide the most compelling and complete range of examples, and so will be used to illustrate the ideas presented in this paper:

- Registration: A wide variety of life science objects are being registered in various systems. BioRxiv[2] is a preprint service modelled on arXiv.org. The RCSB Protein Data Bank (PDB)[3] enables the registration of experimentally determined structures of proteins, nucleic acids, and complex assemblies. While autonomous, it does have a tight binding to the journal publication and hence certification process – submissions about such structures will not be accepted without an assigned PDB identifier. WikiPathways[4] provides a

[1] http://www.npr.org/templates/story/story.php?storyId=1067220

[2] http://biorxiv.org

[3] http://pdb.org/pdb/home/home.do

[4] http://wikipathways.org/index.php/WikiPathways

iPres 2014 conference proceedings will be made available under a Creative Commons license.
With the exception of any logos, emblems, trademarks or other nominated third-party images/text, this work is available for re-use under a Creative Commons Attribution 3.0 unported license. Authorship of this work must be attributed. View a copy of this licence.

collaborative platform for the registration and curation of biological pathways; because it is based on wiki technology, a version history of all pathways is maintained. NeuroLex[5] is a platform for managing neuroscience terminology; it supports versioning to accommodate terminology evolution over time. NanoPublications[6] are targeted at machine consumption and convey a set of discrete scientific assertions and their provenance expressed as RDF, and are typically obtained by mining journal publications. MyExperiment[7] allows for the registration of scientific workflows.

- Certification: A number of systems exist to enable the community to certify the validity of findings in a manner that is disconnected from the journal's peer-review process. PubMed Commons[8] and PubPeer[9] both allow for post-publication commentary on methods or results. MyExperiment supports certification through social network indicators such as views, downloads, favourites. And, machines are starting to play a role in certification, as exemplified in Project FeederWatch[10] where software detects possible errors in bird species observation/identification data and passes potential errors on to humans for resolution.
- Awareness: Examples of support for the awareness function for novel objects include myExperiment's workflow search engine, and the RSS alerting mechanism used by eLabNotebook[11] to keep researchers informed about experiments as they are conducted.
- Archiving: The archiving function for journals is fulfilled by dedicated services such as Portico[12] and CLOCKSS[13]. For novel objects, dedicated archives exist depending on the content type. For example, the PDB enables the archiving of experimentally determined structures of proteins, nucleic acids. Genbank[14] maintains an annotated collection of all publicly available DNA sequences. While neither of these systems has the long-term commitment of a national archive, they do provide an implied level of ongoing availability.

# 3. CHARACTERISING THE FUTURE

While [3] anticipated some of the characteristics of a future communication system that have meanwhile emerged, many further characteristics of its ongoing evolution can be observed at this point. These observations pertain both to scholarly communication as such and to the objects that are being communicated; they are summarized in Figure 1 and Figure 2, respectively. In both figures, the left hand side reflects the status of a process or property in the system of journals, whereas the right hand side reflects its status in a future, emerging system, which this paper refers to as the web of objects.

The major observation depicted in Figure 1 is the transition of the research process itself from being hidden in the system of journals towards being visible in the web of objects. Indeed, the increased use of commodity networked technologies such as on demand cloud computing infrastructure and collaboration/sharing platforms for a variety of objects including software and workflows, make sharing objects that are created during that process not only possible but also attractive. MyExperiment, GitHub, Dropbox, networked lab notebooks, scientific wikis and blogs stand out as obvious examples of this.

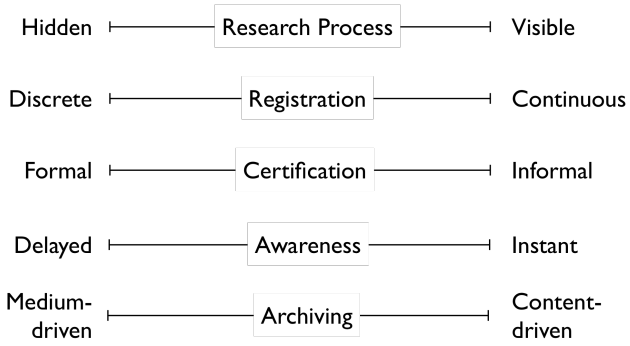


**Figure 1: Scholarly Communication Evolution**

Figure 1 also considers the evolving nature of the essential functions of a scholarly communication system [2]:

Registration is on a continuum that was characterized by discrete submissions of manuscript to continuous registration of a wide variety of objects, enabled by the aforementioned, networked commodity platforms that are used during the research process.

Certification in the system of journals is conducted in a formal peer-review process, but an evolution towards the inclusion of informal certification approaches, for example, based on indicators extracted from social network interactions is apparent.

Awareness in the system of journals was often delayed by years, in part as the result of lengthy peer-review processes but also due to its – originally – paper based nature. Within the system of journals, a trend towards faster communication can be observed, made possible by electronic distribution but also through an increased focus by certain journals on rapid turn-around peer-review. In the web of objects, awareness is already instantaneous: as soon as an object has a URI, notification technologies (Twitter, RSS, Dropbox alerts, etc.) make immediate discovery possible.

Archiving in the paper-based journal system was characterised by the medium that was being archived. Journals were printed on paper and libraries archived paper irrespective of the content that was printed on it. As the scholarly record evolves to include a variety of objects, including digital journals, an evolution towards content-driven archiving becomes apparent. Indeed, the expertise and infrastructure required to digitally preserve collections of PDF files, discipline-specific datasets, scientific blogs, etc. is significantly different and calls for archival specialization driven by content: Portico archives journal articles, GenBank archives genome sequences, web archives archive web pages, etc.

Figure 2 observes the changing nature of the objects that are communicated in the scholarly communication system, confirming the evolution from fixed to varying, from atomic to compound, from uniform to diverse, and from standalone to inter-related or networked that was anticipated in [3]. In addition, it observes the evolution from journal articles that exhibit a clear sense of fixity towards dynamic objects that (at least during part

[5] http://neurolex.org/

[6] http://nanopub.org/wordpress/

[7] http://mypexperiment.org

[8] http://www.ncbi.nlm.nih.gov/pubmedcommons/

[9] http://pubpeer.com

[10] http://feederwatch.org

[11] http://ourexperiment.org

[12] http://portico.org

[13] http://www.clockss.org/clockss/Home

[14] http://www.ncbi.nlm.nih.gov/genbank/

of their visible life cycle) are continuously changing (for example as they are being collaboratively edited on the aforementioned commodity platforms). The ongoing evolution from restricted to unconstrained access to scholarly objects catalysed by the Open Access and Open Science movements is also depicted.

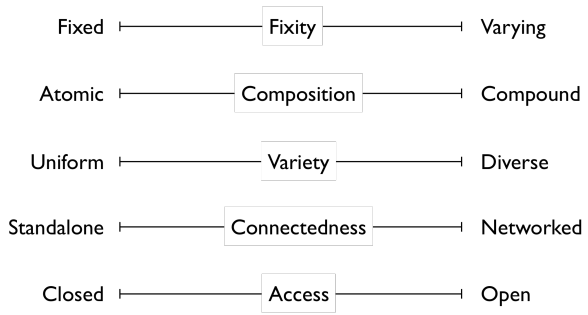


**Figure 2: Communicated Object Evolution**

# 4. ARCHIVING THE FUTURE

Several of the aforementioned indicators of the evolution of the scholarly communication system and the communicated objects have a significant impact on the way in which the archival function of a future system can and will be fulfilled. For example:

- In a closed access system, content has to be transported from its original custodian to the designated archive through restricted back office processes. An open system allows for both organized and accidental archiving by means of the open Web, and puts no constraints on the number or kinds of parties that can hold archived copies.
- The suggested evolution from medium-driven to content-driven archives yields an ecosystem of specialized, distributed archives and calls for appropriate levels of cross-archive interoperability in order to support seamless, uniform access to archived objects.

The remainder of this section zooms in on two important areas in which this evolution impacts the archival function of scholarly communication: the increased visibility of the research process and the dynamic, inter-related nature of communicated objects.

## 4.1 Recording is not archiving

The increased visibility of the research process is, among others, enabled by the adoption of commodity web platforms to record and expose the process. The use of GitHub for the purpose of scientific software development serves as an excellent example of a class of such platforms that share a number of characteristics.

These platforms were not designed with a focus on scholarly use cases, but nevertheless excel at the way in which they fulfil several of the functions of a scholarly communication system. Registration is supported not just by allowing submissions, but also by accurate time stamping and elaborate versioning support. Certification is achieved through a range of reputation-based features such as collaboration, commentary, activity indicators, and likes. Awareness is fulfilled in ways that directly result from the mere presence of these platforms on the open web. This yields discoverability of objects submitted to these platforms through common search engines and the possibility to advertise them on social platforms.

Although these platforms have numerous features that are highly attractive from the perspective of scholarly use cases, it must be observed that they do not fulfil the scholarly archiving function even though their capability to record objects with fine versioning granularity might give the impression they do. Indications that these platforms are excellent recorders of the scholarly process but do not have the long-term commitment to preservation that is expected for objects that are part of the scholarly record can be found in their legal terms and conditions.

Staying with the GitHub example, here are some excerpts from the terms of service[15] that make it explicit that GitHub is not in the archive or persistence business:

> GitHub reserves the right at any time and from time to time to modify or discontinue, temporarily or permanently, the Service (or any part thereof) with or without notice. (E.1)
>
> GitHub does not warrant that (i) the service will meet your specific requirements, (ii) the service will be uninterrupted, timely, secure, or error-free, (iii) the results that may be obtained from the use of the service will be accurate or reliable, (iv) the quality of any products, services, information, or other material purchased or obtained by you through the service will meet your expectations, and (v) any errors in the Service will be corrected. (D.4)

To further clarify the suggested difference between recording and archiving, Table 1 lists some distinguishing characteristics.

**Table 1: Recording vs. Archiving**

| Recording | Archiving |
|---|---|
| Short-term | Longer-term |
| No guarantees provided | Attempt to provide guarantees |
| Write many/read many | Write once/Read many |
| Scholarly process | Scholarly record |

It follows that, as these platforms are increasingly embraced for scholarly use, an appropriate archival function must be overlaid on them to guarantee the long-term integrity of the web based scholarly record. The awareness of this need is growing, as illustrated by the recent announcement[16] of a bridge between GitHub and CERN's Zenodo research output sharing platform, which aims at enabling citation and preservation of code. But, in order to deal with the wide variety of web platforms that is and will be used for scholarship, solutions that connect two distinct environments will not suffice. A systemic solution for the transfer of scholarly objects from the recording platforms into archival environments is required.

## 4.2 Archiving can not be atomic

The dynamic, compound, and inter-related nature of scholarly communication objects yields significant challenges for the fulfilment of the archival function. In order to illustrate this, consider a comparison between the print era of the system of journals and the web of objects towards which the scholarly communication system evolves.

When published, a journal article references other articles, published in the same or other journals. Both the referencing and the referenced articles are preserved in library stacks, worldwide. In order to revisit the article and its context of referenced articles some time after publication, it suffices to visit the library stacks and pull the appropriate journal issues. Since the content was

[15] http://help.github.com/articles/github-terms-of-service

[16] http://home.web.cern.ch/about/updates/2014/03/tool-developed-cern-makes-software-citation-easier

printed on paper and fixed, the combination of the article and its surrounding context of referenced articles remains the same as it was on the day of the article's publication. Gathering all articles may require some library hopping, but the original information bundle can accurately be recreated.

Reconsider this scenario for the web of objects. Starting conservatively from the same point of a journal article, rather than another scholarly object such as software, still serves as a sufficient illustration. The web-based article not only references other articles but also links to a variety of other objects that reside on the web, such as software, data, project web sites, scientific blogs, etc. Recreating the information bundle made up of the article and its surrounding context some time after publication in this scenario is far less trivial as a result of the dynamic nature of the web, the malleability of content inherent to digital media, and the dynamic nature of scholarly objects especially the ones created in the course of the research process. Indeed, even the links in the article are subject to reference rot, a term coined in the Hiberlink project[17] to refer to the combination of link rot, also known as 404 Not Found, and content drift, the evolution of a web resource's content away from what it was at the moment it was linked, possibly up to a point that it becomes unrepresentative of the content intended by the link. And, while referenced articles themselves may still be frozen in time, they are increasingly embedded in web environments with dynamic content such as commentary, metrics, etc.

The combination of these considerations aptly illustrates the archival challenges that result from the core characteristics of the new, web-native objects of scholarly communication. It also illustrates that the atomic perspective that underlies journal archiving is inappropriate for archiving in the era of the web of objects. Journals can be archived one by one, independent of each other. The fixed nature of their content and of their references guarantees that each article's information context can be recreated by visiting journal archives. The web of objects calls for another archival paradigm that inherently takes the interlinked and dynamic nature of the new scholarly objects into account.

Web archiving can serve as inspiration with this regard, especially since all objects of scholarly communication reside on the web and link to other web resources, both traditional articles and novel objects. When archiving web pages, web archives will not just archive a page's HTML but also embedded resources such as images and linked resources. As such, the information bundles that web archives collect are not dissimilar from the interlinked compound scholarly objects. And, web archives allow revisiting pages as well as their linked context as they existed at some time in the past. The Memento protocol [4] even supports including multiple web archives as well as versioning management systems in the recreation of the past. This is not dissimilar from the need to revisit a scholarly object and its linked context (see Figure 3).

Although the web archiving paradigm seems appropriate for the task of archiving the web of objects, the current practice is not sufficient to achieve accurate recreations of dynamic interlinked objects. This is aptly illustrated by Figure 5 drawn from a paper that explores temporal incoherence of pages in web archives [1]. The figure shows a page recreated by the Internet Archive. Although the page is the weather report for the city of Varina in Iowa on October 9th 2004, it doesn't take too much imagination to find similarities with a scholarly object, for example, by its inclusion of graphs, data points, data visualizations. The figure critically reveals that the page has been recreated by means of archived web resources with archival dates that range between 20 days prior and 9 months after October 9th 2004, a result of web crawling strategies and, likely, archive de-duplication processes. The recreated page is temporally incoherent and actually never existed in the way the web archive recreates it.

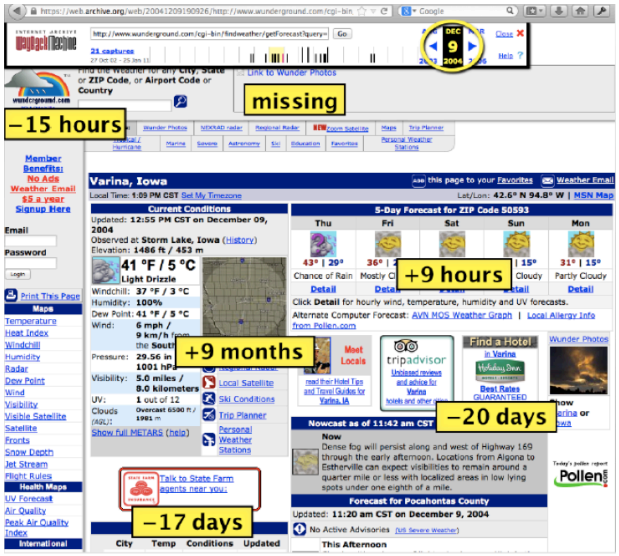


**Figure 3: Temporal incoherence of an archived web page**

On-demand web archives such as archive.is[18] do not exhibit this deficiency as they collect a resource and its embedded resources at the very moment a user requests it. Although this type of web archive typically does not collect linked web pages, the snapshot approach is more aligned with the requirements for archiving the web of objects. Still, the suggested trend towards content-driven archiving means that constituent or linked resources of a scholarly object can not be archived in a single place, but rather in specialized, distributed archives. This requires some sort of orchestration driven, among others, by content type of the archival process. As is the case with regular web archives, the Memento protocol could serve as the interoperable glue to recreate specific states of objects from snapshots available in multiple archives.

# 5. THE FUTURE OF ARCHIVING

As described, the emerging web of objects has fundamentally different characteristics than its predecessor, the system of journals. Several of those characteristics, especially the interlinked, dynamic, and heterogeneous nature of the objects suggest the need for a different archiving paradigm. The web archiving paradigm can provide inspiration as it is based on the understanding that, on the web, resources are interlinked and their interpretation critically depends on their network context.

## 5.1 Infrastructure considerations

Figure 4 provides a high-level view of a future scholarly infrastructure based on the above discussion, and inspired by [5]. Contributing to the planning and building of such an infrastructure is the subject of ongoing work by the authors and their colleagues.

A researcher conducts some of her research on private infrastructure, personal computing facilities. Since the objects created in this environment reside in local namespaces, their inclusion in a web-wide archival solution is hindered. As a result,

[17] http://hiberlink.org/

[18] http://archive.is

from a scholarly system perspective, objects in private infrastructures are ephemeral, and cannot be considered parts of the scholarly record.

A variety of incentives lead the researcher to move her objects from the private infrastructure to the web-based recording infrastructure. These include sharing with self across computing platforms, sharing with a team of collaborators, or even complying with a requirement from a funding agency. As described, recording platforms have attractive features when it comes to fulfilling the registration, certification and awareness functions of a scholarly communication system, but archival platforms they are not. However, because the recording platforms are embedded in the web, objects now reside in a global namespace and are network accessible. Hence, they are within reach of web-scale processes aimed at selectively moving objects from the recording infrastructure into the archival infrastructure, and hence into the permanent scholarly record.

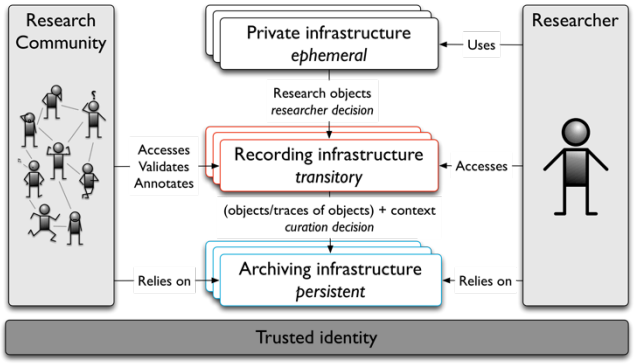


**Figure 4: High-level view of a future scholarly infrastructure**

Core aspects of these processes include the ability to snapshot the state of interlinked objects at specific moments in their lifecycle, to transfer these snapshots from a variety of recording platforms to appropriate distributed, content-driven archives, and curatorial policies aimed at deciding what should be archived when.

Underpinning the entire infrastructure is a trust component that provides assurances regarding identity and authorizations.

## 5.2 Curatorial considerations

Assuming the existence of web-scale processes that are able to transfer objects from their operational state in the recording infrastructure to an archival state in the archiving infrastructure, significant questions of a curatorial nature remain. Indeed, in order for the archival infrastructure to stand a chance at sustainability, significant curatorial filters will be required.

A first consideration pertains to what the archival object should be, or, to use the above terminology, what the nature of a snapshot is. For certain objects this may be a copy of the actual object, for others metadata that describes the state, or provenance information that can be used to recreate the state.

A second consideration pertains to the inputs that trigger the transition from operational to archival state. A variety of options present themselves with this regard. In the conservative scenario of Figure 4, the submission of a manuscript or the publication of a paper may launch a process aimed at collecting snapshots of all linked resources. Network-derived metrics, such as altmetrics[19] that measure impact of scholarly objects by means of their presence in the social network flow, or their use, could be used to guide a decision. Decisions could be on-demand, initiated by a researcher as a means to preserve what she considers an important state of one of her own objects, or to safeguard the state of a colleague's object before starting to build on it. It might also be worthwhile to introduce a level of randomness in the decision making to increase the chances of capturing objects that might be serendipitously interesting in the future.

A third consideration is around how the archiving decision is made. Given the vast number of objects that will reside in the recording infrastructure, largely automated decision making driven by heuristics like the aforementioned ones seems essential.

The implications of these considerations are being worked through with archival specialists at DANS[20].

# 6. CONCLUSIONS

This paper has provided a perspective of a future scholarly communication system, called the web of objects, and has focused on the impact of that system on the fulfilment of the archival function. A core observation was the increased use of web-based recording platforms that excel at registration, certification, and awareness but provide no guarantees regarding archiving. Hence, the introduction of web-scale processes aimed at transferring objects from recording platforms to appropriate archives, subject to curatorial filters was proposed.

Archival infrastructure that underpins research communication needs to be trustable and hence sustainable for the long term. Sustainability, in light of the heterogeneity and number of objects requires a distributed approach. A distributed archival approach to present the web-based scholarly record in a uniform, interconnected manner, requires interoperability and thus standards.

# 7. ACKNOWLEDGMENTS

The authors would like to acknowledge DANS for bringing them together to work on the ideas presented here.

# 8. REFERENCES

[1] Ainsworth, S., Nelson, M. L., and Van de Sompel, H. 2014. A Framework for Evaluation of Composite Memento Temporal Coherence. arXiv:1402:0928

[2] Roosendaal, H. E. and Geurts, P. A. Th. M 1997. Forces and functions in scientific communication: an analysis of their interplay. In *Proceedings of CRISP 97*. http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html

[3] Van de Sompel, H. et al., "Rethinking Scholarly Communication: Building the System that Scholars Deserve", *DLib* (September, 2004). doi:10.1045/september2004-vandesompel

[4] Van de Sompel, H., Nelson, M. L., Sanderson, R. 2013. RFC7089 HTTP Framework for time-based access to resource-states – Memento. http://tools.ietf.org/rfc/rfc7089.txt

[5] Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". *Proceedings of VALA 2008*, Melbourne, (February 2008). http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

[19] http://altmetrics.org/

[20] http://dans.knaw.nl/