DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Session 7: Data Journals

7-Data Journals & Editorialization of Open Data

Data Journals

Anne Baillot & **Marie Puren**, INRIA

What is a data journal?

Why is it called “journal” and to what extent is it different from a traditional journal? Why do we need data journals? What is advantage of publishing it in the traditional structure of a journal? => You can get some credit for it
This session will not be too technical but will try to reflect on what it means to have the opportunity to publish data papers, to construct data journals, what it means for the academic system in terms of recognition for digital research and for academic communication in general.
For the Australian National Data Service: Data journals are publications whose primary purpose is to expose datasets by providing the infrastructure and scholarly reward opportunities that will encourage researchers, funders and data centre managers to share research data outputs. Data journals have evolved from traditional journal model that describe datasets including supplementary material. Data journals have more in common than journals that publish articles or overlay papers that describe data but take the concept a few steps further.
As the primary purpose of data journals is to expose and share research data, this form of publishing may be of interest to researchers and data producers for whom data is a primary research output. It enables the author (or data producer) to focus on describing the data itself, rather than producing an extensive analysis of the data. Publishing a data paper may be regarded as best practice in data management as it includes an element of peer review of the dataset, it maximises opportunities for reuse of the dataset and it provides academic accreditation for data scientists as well as front-line researchers.
Data journals are nowadays well established and indexed, which is important for questions of credit, but until now data papers were mostly published in *mixed journals* - journals that have a separate section for data papers, in order to have journal articles and data papers altogether. The conclusion of the article “Data journals: a survey” in 2015 is that although there are platforms to publish data papers, they are still not open enough to foster data sharing and data reuse which is actually the point.

- “Scholarly publication of a searchable metadata document describing a particular online accessible data set, or a group of data sets, published in accordance to the standard academic practices.” Chavan & Peney, 2011 quoted in “Data journals: a survey”, 2015: http://onlinelibrary.wiley.com/doi/10.1002/asi.23358/abstract
- “This artefact is homologous with articles for traditional journals; it is expected to have an identifier and a content with title, authors, abstract, number of sections, and

references." "Data journals: a survey", 2015: http://onlinelibrary.wiley.com/doi/10.1002/asi.23358/abstract

My main thesis is that we tend to separate certification and evaluation from research itself, for different reasons like career pressure, the amount of scholarly publications, the development of questions specific to digital publication format and this leads to a deep lack of satisfaction from those who produce and disseminate scholarly knowledge. Since we won't be able to redesign the academic system in a quick and efficient way, we need to think of ways to improve the conditions which determine how we work and communicate the results of our work. This is the spirit in which this data journal model is being developed by DARIAH. This model is explicitly not purely research but at the interface of research and infrastructure - infrastructure is becoming more and more essential to the way we do research. And it is of crucial importance that researchers identify themselves with this kind of work at the interface of research and infrastructure.

Authorship

Do we still need peer-review? Data journals as a way of reconsidering our evaluation culture and our understanding of research

In this presentation, the idea is to give you a broader historical perspective on the question of authorship and try to identify systematically which aspects of peer review are misleading scholars and which aspects can be reappropriated in a more constructive way.
The core assumption of this presentation is that data can be a scholarly publication when they meet clear academic standards. This is one issue we encounter when dealing with inadequacy of our evaluation system is that it is author centered because it doesn't correspond to actual practices since scholarly work is hardly ever an individual endeavour. The concept of author, as it emerged in the 18th century, is mostly conceived to concentrate on one name, preferably a male name, all the authorship qualities. There are economical considerations behind this idea: big names are attractive and sell more than the mention of the actual contributors (copyist, lab experimenter, editor, publisher, etc.) will do. Also, copyright was conceived with this notion of single authorship which in turn encouraged single big name authorship practices. The opportunity to construct the publication system around a *dispatched authorship model* could have emerged with for example the European Republic of Letters.
If you look at the facts, there is probably no point in our publication history when the author who appears on the book cover was the sole producer of the content of their books. You can probably name isolated counter examples, but the general trend is that book production, especially literature and science production, is and has always been a collaborative phenomenon. We can even identify -with variable accuracy- the different spheres of influence (family, friends, lab assistants, publishers, etc.). We are aware that we have to decipher these modes of participation, but the knowledge of split text, book or scientific production remains some kind of hidden truth even if we know it. This awareness is not a major epistemological principle reflected at large in the humanities' understanding of authorship. The result is that literature history, and to a great extent also science and scholarship history, still live in the myth of the author, this great man.

Why is this a problem for *digital publications*?
Because part of the recognition we need has to do with split authorship or split producership. When it comes to digital publications, we are expecting something different, especially because the modes of cooperation don't obey the same hierarchy and rules than it is or was in the analog world. In digital publications, we don't want the publishers to appear separately anymore because we consider that design and funding is in the domain of scholarship, it doesn't have to be separated from the production of the work. Along with the designer, all intermediates (software designer, technician, etc.) also contribute to the final form of the publication that is offered to the reader. In digital publication, attribution and versioning are two key techniques which have always belonged to the core principles of IT archiving and publishing. The TEI has inscribed in the header the revision and version as a mandatory element for a good reason, and other elements such as institution and funding have a prominent place as well. It is the whole production context that is taken into account. The aim of such an inclusive understanding of text producing is not to make all of the instances involved accountable for the content in a legal sense, but to render the production context as extensively as possible. In other words, there are no technical challenges to the implementation of authorship distribution or split producership in the case of digital publications. There are, though, cultural issues: the change of mentalities that makes the bridge from traditional journal formats to data journals difficult to cross. Some data journals are consistently using micro attribution to address this issue. They name every participant to the production of a dataset by providing appropriate credits to each, by capturing their contribution. But this is not systematically implemented.

Two position papers: Reconsidering scholarly publications in the digital age

In two recently published texts by two working groups I am affiliated to, we listed the various possible authorship or contribution forms with the aim of showing the extension of this variety of functions in text production. We also insisted on the fact that digital publication can take a variety of forms (monographs, articles, edition, database, code, images, videos, etc.). It is not new, but it tends to show how narrow our understanding of a publication in the humanities has become in the course of the history.

The question of academic recognition is at the core of the debate in both papers:

- http://dhd-wp.hab.de/?q=content/empfehlungen_ag_digitales_publizieren
- https://www.merkur-zeitschrift.de/2016/10/24/siggenthesen

Additionally to the question of displaying various and complex authorship and contribution modes, there are two other aspects that make the implementation of data standards and any inherent certification even more difficult:

- Time machine problem: standards and evaluation criteria develop and change. This makes it difficult to attribute them for once and for all and to name them down in a manner that would be definitive.
- One of the most difficult thing to grasp for the traditional academic evaluation system is the fact that digital publication is hardly ever finished. Almost all of them are processual kinds of publications. Some hypotheses are only verified later and implemented in an update, new material is found, etc. There can be many reasons

why you would change a digital publication: emendations, enrichments, cross-checks, etc. The reactions that this processuality phenomenon provokes are not unanimous:

- Some see it as a chance to publish editorial material progressively, arguing that there is no need to wait 10 years to publish results; instead of that, you can enrich and improve progressively.
- Others find it hard to cope with the lack of liability inherent with this openness to change: what is the version of reference if you know that the publication is always going to change? How do you refer to that publication? Admittedly, tracking changes via log files and version history is not self-explaining: it has no equivalent in the print culture. So there is a question of scholarly culture and mentality that needs to be addressed specifically and that can't be changed at once.

=> The question of authorship is not just historical or secondary, it is really at the core of the whole academic system.

Peer Review conundrum

Pre-publication peer review was established at a point where it was not possible anymore to print everything. The analog production of all scholarly papers and books would have been too cost intensive. Nowadays, pre-publication peer review is considered on the one hand as the *best* way to evaluate good science, on the other hand as a system that has become *unreliable*. Peer review is taking more and more time as the number of scholars grows and as the concurrence increases in submissions. We have also heard that peer reviews are not really achieving their goal of generally contributing to opening up innovative research questions and answers. This question is regularly addressed in the Guardian Higher Education:

- https://www.theguardian.com/science/2016/sep/21/cut-throat-academia-leads-to-natural-selection-of-bad-science-claims-study
- https://www.theguardian.com/science/2011/sep/05/publish-perish-peer-review-science

Those of you who have received reviews after submitting a paper will know that quite a lot of the overall produced peer reviews consist in a reviewer being touchy because his or her work on the topic was not quoted. We have intrinsic problems with pre-publication peer review, especially because of its dominant position in the evaluation system and because it produces delays in the whole publication process without necessarily improving the quality of submitted papers. As editor for a journal, you have to wait a lot of time for the reviewers to accept to do the review, then you will have to wait for them to actually do the review and this is really delaying the publication of many journals; but on the other hand you know that reviewers have many other review requests pending.
As opposed to the *paper reality of the analog world*, there is no real room problem in the digital world. It doesn't matter if a paper has a predetermined amount of pages, because there is no need to calculate paper and binding cost. The argument is obsolete. Even if the digital production and maintenance of online publications is not at zero cost, institutional

repositories now exist for scholars and allow to make primary data and research accessible, readable, without any valid cost argument.
One model that counters this method is *post-publication review*. One advantage is that it is particularly relevant in the context of data journals as data or publications are being submitted and accepted for submission only if they already fulfill some basic editorial conditions of legibility and scholarship.
=> It means that you submit papers in a better quality if you know that they are consultable online before you submit them.
In this context, we still don't know what post-publication open peer review will bring in the long run, but it seems worth a try compared to the failure of pre-publication peer review we are now experiencing.
=> There is a clear gap between the reality of research, especially in the digital era, in terms of *temporality*, *contribution types*, *techniques available* to take all of these into account on the one hand, and the reality of the evaluation system on the other hand, which is slow, author-focused and in an authoritative position towards the research production.

Why data journals in the Arts and Humanities

This is precisely what we are trying to do with a workflow for data journals in the humanities, which is aiming at improving the recognition of the in-depth phenomena previously mentioned, especially in the case of digital scholarly editions. The initiative of the data journal as a structure comes from DARIAH-EU, it is supported by the French institution The Center for Direct Scientific Communication (CCSD) which hosts the episciences platform, and Inria. It is this infrastructure we are currently adapting in order to offer to the scholarly communities a data journal model in adequation with the reality of scholarship. This project started under the codename "*living sources*" (one example), because it is based on the core idea that digital resources are processual - they keep growing and need to be re-reviewed along time. The concept was first developed at the Max Planck Institute (MPIWG) and has been since then claimed by commercial platforms such as scienceopen. What matters is not only to emphasize the lively character of the process, but also the adequacy it wishes to generate, in the overall process of scholarship, between publication and evaluation. In this perspective, the role of the review is not to sort out the good from the bad for it to be published, nor is it to put a stamp on a digital publication. More importantly, the review is becoming an incentive to further development. The review is conceived as a dialogue with the digital resource, both of them working towards improvements.

Submission and review process

The envisioned process goes as follows: A scholar or a group of scholars submits a data paper and an OAI-PMH access to the corresponding metadata. This allows to gather the version of the data which will be reviewed. At that point, it is up to the editorial committee to decide whether technical and content review should be separated, whether this should be double-blind, single-blind, not blind at all or open and in which time frame they want to operate.
The publication can integrate a link to the review, which can be done in the form of a certification, but since there are scholarly contexts in which certifications can be a risky modus operandi, a simple link to the review seems at this point the most viable system.

The review can raise points that could be improved, and the resource's team could be offered to re-submit data when these points have been taken into account. It would then be possible to show clearly the progress achieved along time. Such an organ needs two driving forces:

- a motivated editorial board willing to define a review model and to gather a critical mass of reviewers
- a solid web interface

What DARIAH wants to offer is the technical background, so that the workflow is backed by a solid structure and team. We hope that scholarly communities will find this offer appealing enough to take advantage of the structure we are currently developing. The data journal sandbox is now opened, metadata have been imported from Ortolang and Nakala, the Deutsches Textarchiv and others trusted repositories should follow soon.

Data Journals on the episciences platform

- Episciences platform: episciences.org
- Our sandbox: datajournal.episciences.org
- Data journals: episciences.org/page/journals
- Other examples: https://www.cms.hu-berlin.de/de/dl/dataman/teilen/dokumentation/datajournal

The episciences platform has not been developed for data journals, it is an overlay journal platform, on top of a preprint archive or repository. An overlay journal is an open access electronic journal based on and composed of research articles that are submitted after being deposited in an open archive. The implementation has clearly been made easier by the French centralized repository structure HAL for the Arts and Humanities. An overlay structure requires submissions to be written and formatted properly before being submitted. It spares time in copyediting and formatting for the editorial team, but it requires that the authors take responsibility for their texts much more strongly than it is the case in traditional arts & humanities journals. Usually papers are submitted with linguistic problems, typos, but it doesn't matter because an editorial assistant will do it for you, but when you submit to a repository, it is your way of working that is becoming visible to the scientific community.

Why use the episciences platform for data journals?

In the context of the Journal of the Text Encoding Initiative, I have been working with Open Journal System (OJS), one of the major open access editorial workflow system. In comparison the editorial interface of episciences is incredibly flexible. It can be adapted for practically every editorial need, with a lot of functionalities. For example, as an editor, you have to send reminders to authors and reviewers. In episciences, you can completely automatise the whole process. In OJS, you have to do that by hand, OJS sends only one reminder and you can't change it.
The episciences platform has the advantage and the inconvenient that it relies on the quality of data repositories and requires a clear vision of the amount and type of relationships with the repositories that are envisioned. One of the very great advantages is

that it allows to certify or evaluate any kind of data: not only a research paper, it can also be video, software, a data set, etc.
The episciences platform is designed to harvest metadata via OAI-PMH, which is useful and necessary to gather the information needed for a data journal. Each scholarly community has to identify the resources or repositories relevant to their field, but some technical elements such as the OAI-PMH interface are necessary to exchange information on a reliable basis. This also means that the repository you will be working with has to have clear versioning strategies to allow to re-review data.
On episciences, nothing is kept on the platform itself, everything is harvestable and can be "called" via the metadata and the OAI-PMH interface, at any time as long as the repository offers such an interface. It is a great advantage compared to having data as "supplementary files" or to gather the data for the review as is currently most often practiced. It also allows to avoid proprietary archiving strategies of repositories. Episciences is built on top of open access repositories.
On the other hand, the layout question is left unsolved in the hands of the authors. It is only a minor issue when the scholarly communities are used to work with LaTeX, but arts and humanities scholars are used to editors taking care of the layout. And this is important because what makes a journal is also to have something nice to read in the end and not just an ugly word document in times new roman.

The dashboard offers different options depending on the role you have, but the general review process is:

- Submission
- Attribution of reviewers
- Reviewing process
- Final acceptation
- (Publication)

Hands-on session

Let's build a data journal in digital humanities within our episciences sandbox http://datajournal.episciences.org/

- Group 1: create a rating grid
- Group 2: define the form of peer review
- Group 3: write a rationale for the journal
- Group 4: find potential resources

Elements of guideline
=> Group 1

- Create your own rating grids by defining your evaluation criteria
- Examples (DH Commons):
    - http://dhcommons.org/journal/2016/women%E2%80%99s-print-history-project-1750-1836
    - http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heurecl%C3%A9-digital-heuristic
    - http://dhcommons.org/journal/review-guidelines

- Think about the level of “visibility” of each criterion. Why don’t we have access to one specific criterion? Or why do we have access to another? For instance: level of visibility of the review report? If a review is closed, what might be the consequences on the reviewer’s work?
- Quality of manuscript: writing, clarity, organization, adherence to template (of the journal)
- Criteria for assessing the effectiveness of the data paper content as a mean for accessing the data set(s)
- Data quality, criteria for assessing the methodologies leading to the production of the data set(s)
- Data reusability, criteria for assessing the actual reusability of the data set(s)
- Utility and contribution of data, criteria for assessing the potential of the data set(s) for the community

=> Group 2: Define the action scope of the different roles + type of peer review

=> Group 3: Write the rationale for the data journal (what is the journal about? what do we want to evaluate, with what aim?)

=> Group 4: Look for potential resources and defining which metadata fields are of use for the evaluation to work. Analyse the “quality” of the repositories in terms of metadata. Do the datasets comply the data paper criterias

- Make a list of data repositories. Ensure that data set(s) are usable for a data journal dedicated to Digital Humanities.
- Have a look at the conditions of use
- Do you easily find the information that you need? What do you have to do to obtain them? (email, form...)
- Metadata fields
    - Title
    - Authors
    - Abstract
    - Key words
    - References
- Potential resources - Repositories
    - Nakala
    - Ortolang
    - DTA
    - Registries of repositories: re3data and Open access directory (OAD)
- OAI-PMH
- Availability to provide data set access attributes => DOI or URL
- Competing interests: fundings declaration of any factor that might influence the data set (personal, financial)
- Coverage to provide data set “extent” attributes, including spatial and temporal coverage
- Format (format, encoding, language)
- Licence

- Microattribution: all the creators who contribute to the datasets
- Project: goal and funding
- Provenance: methodology and tools leading to the production of the dataset
- Quality (including data set limitations and anomalies)
- Reuse: information on the potential uses of the data set(s)

Conclusion

What is the benefit of opening a data journal for a scholarly community?
First of all, an editorial board wanting to engage in such an endeavour would benefit from the technical infrastructure and the ongoing reflections on workflow and assessment procedures. Then, a data journal by definition recognizes the value of data, something often still difficult to cope with in arts and humanities scholarly communities. This will contribute to the change in mentality this initiative wants to induce or at least contribute to. Beyond the certification, which might be considered as a first level of readability (for example for our colleagues and students that are too often unaware of quality criteria for digital resources), the second level is the reconciliation of the research process and the evaluation process. One part of the research process that will gain great recognition from this, namely data modelling. This is certainly one big mentality change but re-evaluating data modelling within the frame of data journals is something that could, in the end, also help people to understand better what DH are doing.

Contact

Anne Baillot was a trainee civil servant at the École Normale Supérieure in Paris between 1995 and 1999. She completed her PhD in 2002 in Paris. Since then, she has been living in Berlin where she worked as a post-doctoral researcher at various institutions. Between June 2010 and January 2016, she was junior research group leader at the Institute of German Literature of Humboldt University, funded by the DFG (German Research Foundation). As a junior group leader, Anne Baillot is the editor of Letters and Texts: Intellectual Berlin around 1800. Since 2013, Anne has been a member of the editorial board of fr.hypotheses and en.hypotheses, and since 2015, Anne has been a board member of the German DH association (DHd) and of the European Society for Textual Scholarship. She blogs about her research in English on http://digitalintellectuals.hypotheses.org/ and tweets as @AnneBaillot. Since February 2016, she has joined Laurent Romary's team and is working at the interface between research, infrastructure and cultural heritage institutions. She is Managing Editor for the Journal of the Text Encoding Initiative and is working towards developing new models for journals in the scholarly ecosystem. Her next book (to appear 2017) is dedicated to the relationships between writers and publishers between the late 18th and early 20th century in Germany.

Marie Puren is junior researcher in Digital Humanities at the French Institute for Research in Computer Science and Automation (INRIA) in Paris, members of the Alpage laboratory (INRIA – Paris Diderot University). As collaborators to the PARTHENOS H2020 project, she focuses her research on the development of standards for data management and research tools in Arts and Humanities, and she currently works on the creation of a Data Management Plan for this project. Marie Puren also contributes to the IPERION H2020 project, especially by upgrading its Data Management Plan. After being a lecturer and a

responsible for continuing education projects at the Ecole nationale des chartes, Marie Puren has been a visiting lecturer in Digital Humanities at the Paris Sciences et Lettres (PSL) Research University. Her main publications belong to fields including intellectual history of the XXth century, French studies and digital humanities. Marie Puren has been awarded a Ph.D. in History at the Ecole nationale des chartes – Sorbonne University. She holds Master's degrees in History and Political Science from the Institut d'Etudes Politiques de Paris, and in Digital Humanities from the Ecole nationale des chartes.

Anne Baillot & Marie Puren

Twitter: @AnneBaillot & @puren1406

Email: anne.baillot@gmail.com & marie.puren@inria.fr