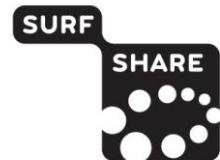


This is the translation of chapter 13 in the book 'Toegang tot Onderzoeksdata', which appeared in July 2011 as one of the products of the SURFshare programme. In this Dutch book various aspects of providing access to research data in the Dutch landscape are discussed.



It is essential for research data to be linked to publications

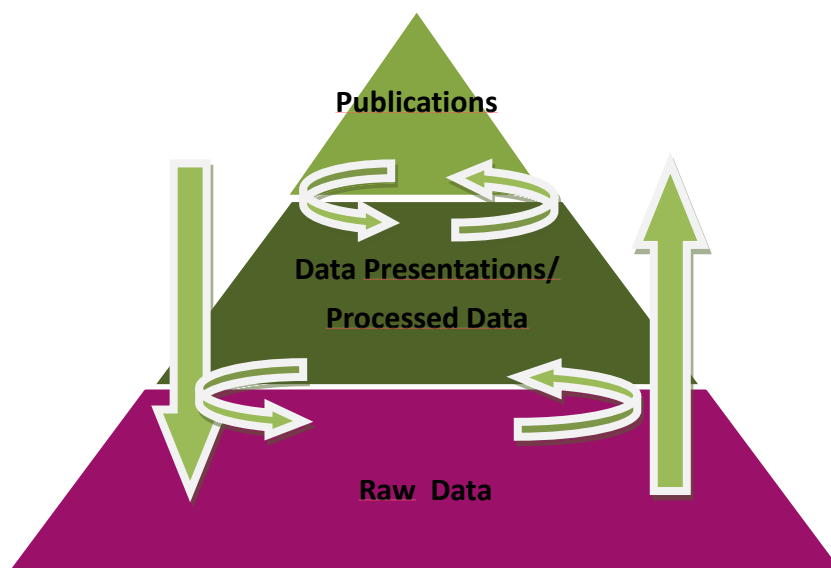
Interview with Eefke Smit on the role of publishers with respect to research data – Monday April 4, 2011 by Lilian van der Vaart

Eefke Smit's interest in research and high-tech developments started early in her working life, when she was a journalist for the Dutch newspaper NRC Handelsblad. She is now the Director of Standards and Technology at STM and coordinates the activities of the STM Future Lab Committee. Next to her STM work, Eefke works as an independent consultant in new business development, e-publishing and innovations. She brings to her work an extensive professional background in print and e-publishing, and was responsible for the development of e.g. Scopus™, Scirus™ and ScienceDirect™.



Eefke Smit (PhotoA)

"Where is the information we have lost in data?" This is the question publishers think they can help answer when it comes to their role and value-added in handling the 'data deluge'. "It is crucial that research data and publications 'stay together'", Eefke says: "data are relatively meaningless, or in any case hard to interpret, without the context of the publication, and the publication is enriched by the data underlying its claims and conclusions."



She points to the fact that many researchers are still reluctant to share their data with others¹ and that in many cases this is for fear of mis-interpretation or mis-use. This fear can be substantially reduced by providing the data in the context of the article. This way, the author also gets the credit for his data. Furthermore, data need meta-data to be accessible and re-usable, and “what better meta-data than the publication?” Paraphrasing a statement by Andrew Treloar² during a recent JISC workshop on Managing Research Data, Eefke says data are now often

- 1) Unavailable
- 2) Unfindable
- 3) Uninterpretable
- 4) Unreusable

The publishers’ role and added value with respect to data can be, to help address the first three of these issues: by providing the publication linked to its underlying data, these become more easily available, findable and interpretable. And perhaps there’s even a role to play in facilitating reuse, though this issue is more fuzzy. This involvement is a logical first step that fits in well with many publishers’ mission to concentrate further on content enrichment in an environment that will be increasingly interlinked and open.

“Having said this, there’s also a substantial dilemma for publishers to deal with”, Eefke continues. “Though linking data with the publication is a necessary improvement, no one is served by actually storing them together; that only leads to fragmentation at the data level; and to significant problems for publishers and their journal editors.” Experience of publishers who offer authors the option to submit data along with their article, or even require it, points towards problems with e.g. dealing with large varieties of formats, hosting a rapidly increasing volume of data on publishers’ websites, and handling the peer review of the data. Especially this latter problem has even led the **Journal of Neuroscience** to change their editorial policy and allow authors to include a link to their supplemental materials hosted on another site, but no longer accept those materials along with the article and host them on the Journal site.

In the increasingly dynamic and open environment, certification, trustworthiness and discoverability may become even more important. The archives may have a role with respect to the ‘provenance’ of data, publishers in the ‘discoverability’ – e.g. personification/filtering applications. Peer review of data may have to focus more on the methods & techniques used in data collection and analysis, on integrity of the data, in other words be ore process-oriented. It could become a service, to be delivered (at a price) or not.

How about copyright?

The Brussels Declaration concerns raw data, the base level in the data pyramid. The top of the pyramid is the publishers’ traditional focus area. This now shifts to the second level. Data presentations often form part of the publication and are copyrightable; selections of processed data could be. Whether publishers will ask for copyright transfer is an undecided matter yet. At present almost 60% still don’t, according to the PARSE Insight Report, but almost 50% apply the same access and use terms as for publications.

¹ The PARSE.Insight Report 2010 shows that though 60% of researchers would like to use data from others, 40% have real problems sharing their own. 25% of researchers are currently making their data publicly available. *PARSE.Insight. “Insight into digital preservation of research output in Europe.”*(June 2010) and Survey report (December 2009). http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf and http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

² Director of Technology for the Australian National Data Service

A better approach, Eefke feels, may be to follow the emerging example in a number of fields³, where large, central subject-oriented data archives are being developed; journal policies require authors to store their data there and provide stable links from their article to their stored datasets. These are usually community-endorsed archives, with their own established quality control mechanisms and policies with respect to e.g. persistent identifiers and preservation, so that trust and sustainability issues are – at least partly – resolved. Such data archive policies are essential conditions for this linkable data in certified repositories. Only in this way can seamless access and persistent linking between the article and the data be ensured.

STM is involved in a number of EU-funded projects on data, such as ODE and APARSEN, to participate in the development of standards and best practices in cooperation with other players across the information chain.

Consequently, such an approach requires close cooperation between data archives and publishers. This is not common practice yet in many disciplines and will require tackling practical as well as cultural issues. Furthermore, it cannot be done by individual organisations, neither on the publisher side nor on the data archive or research institutions' side. Data are too complex, varied and dynamic to be dealt with at a low aggregation level, whether that be the individual researcher, his institutional repository, journal or publisher, or even at the national level. An integrated approach to the handling of data can be very valuable, if it provides scale, volume, options for searching and combining throughout the collection, etc. This requires a normalisation and standardisation effort which can only be achieved at sufficient aggregation level: international and discipline oriented. Compartmentalization would form an important obstacle to the 'openness' of data, which is endorsed by publishers in their Brussels Declaration on STM Publishing⁴.

³ E.g. in sub-disciplines of the biological sciences, geo-sciences, earth and environmental sciences and chemistry.

⁴ Brussels Declaration on STM Publishing, by the international scientific, technical and medical (STM) publishing community, 1 November 2007, http://www.stm-assoc.org/public_affairs_brussels_declaration.php