The value of data

Barend Mons[1–4], Herman van Haagen[1], Christine Chichester[2,4], Peter-Bram 't Hoen[1,4], Johan T den Dunnen[1], Gertjan van Ommen[1,4], Erik van Mulligen[3,4], Bharat Singh[2,3], Rob Hooft[2,4], Marco Roos[1,2,4], Joel Hammond[5], Bruce Kiesel[5], Belinda Giardine[6], Jan Velterop[4,7], Paul Groth[4,8] & Erik Schultes[1,4]

Data citation and the derivation of semantic constructs directly from datasets have now both found their place in scientific communication. The social challenge facing us is to maintain the value of traditional narrative publications and their relationship to the datasets they report upon while at the same time developing appropriate metrics for citation of data and data constructs.

The chicken and the egg of scholarly communication

In data-intensive sciences, text is neither the only nor the most effective way to share scientific information. Aware of the paradox, we reintroduce the metaphor of the chicken and the egg to underscore our thesis that there is no meaningful information without data and conversely, data cannot be generated nor valued without prior knowledge. If we assume data to be the eggs, which need brooding (curation) to become chickens (articles), and we require the mating of complementary units of information to generate yet more fertile eggs, we have a reasonable frame of reference.

When datasets were sparse and only connected to the lab that produced them, we would brood every one of them, protect (patent) them and work on them in isolation in order to 'sell' them as chickens, usually in the form of a largely narrative article. Other scientists need to combine a minimum of two existing publications to generate new eggs and breed more chickens. However, chickens have become overabundant: more than 20 million articles exist in biomedicine alone. More recently, valuable aggregations of data were brought online (for example, data sets in GEO, curated databases such as SwissProt and locus-specific human gene variation databases (locus-specific databases such as the Leiden Open Variation Database LOVD). Now, data (eggs) have become a direct source of new *in silico* discoveries and a unit of scientific trade.

But the scientific market has no way to value eggs because the entire system is built upon judging and exchanging chickens for acknowledgement and credit (through citations and other measures of impact). On the other hand, for effective and evidence-based breeding, we need the eggs as well as information from the parent chickens to assess the value of the eggs. This is where a major challenge lies: in the long overdue adaptation in scholarly communication. The data-intensive science wave that has come over us calls for innovative ways of data sharing, stewardship and valuation. We must respect the connection between the articles and the data and value both appropriately.

A new market for data

We have all heard lamentations about datasets being difficult to find and the painfully slow increase in annotation and curation by the scientific community[1]. One bottleneck is the lack of a scientific reward system for depositing and curating data outside the mainstream of publishing conventional articles. But how do we implement such a system? In this issue, Giardine *et al.*[2] argue for "microattribution"[3], providing credit for contributors and curators of entries in databases of human gene variants. Here we comment on the technical as well as social challenges associated with broadening and sustaining that valuable approach.

Although we argue that the narrative form of scholarly communication will continue to be needed in the future, we also recognize that data-intensive sciences need computer-readable information[4]. It follows that *de novo* claims and the supporting data should be exchanged in machine-readable, unambiguous format. Ideally, this should be created at the same time the descriptive text is composed. Articles should tell us why we should believe the underlying data and the conclusions drawn, and they are perfectly suited for that task as they are.

A graphical analysis of the problem

We propose a new way to represent data, information and, in particular, assertions in the form of nanopublications[5]. A nanopublication is essentially the smallest unit of publication: a single assertion, associating two concepts by means of a predicate in machine-readable format with proper metadata on provenance and context[6]. Each concept in a nanopublication has an unambiguous,

[1]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [2]Netherlands Bioinformatics Center, Nijmegen, The Netherlands. [3]Department of Medical Informatics, Erasmus Medical Centre, Rotterdam, The Netherlands. [4]Concept Web Alliance, Nijmegen, The Netherlands. [5]Thomson Reuters, Philadelphia, Pennsylvania, USA. [6]Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania, USA. [7]Academic Concept Knowledge LTD., London, UK. [8]Free University, Amsterdam, The Netherlands. Correspondence should be addressed to B.M. (barend.mons@nbic.nl).

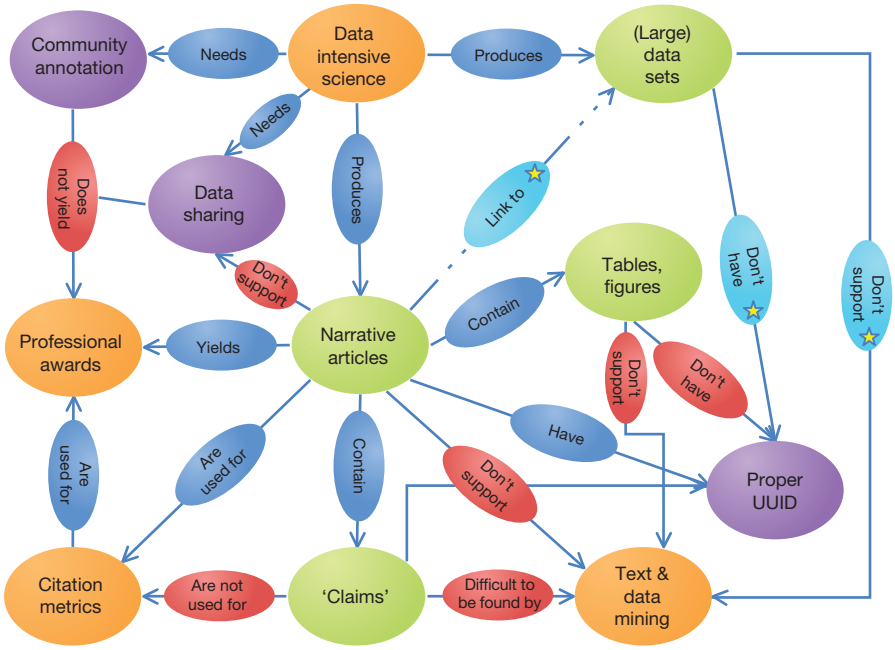

Figure 1 The current state of scholarly communication. Red ovals represent points where the current system is 'broken'. All predicates are represented as small ovals. An asterisk (yellow star in the figure) represents 'conditionality'. Illustrated concepts (large ovals) are discussed below in bold. **Data-intensive science** produces **large datasets**. It also produces **articles** in which the **narrative** of the experiments, methods and conclusions and some **figures and tables** are presented. The **datasets** that are too **large** to be part of the **article** itself are currently **'linked'** as **'supplementary data'**. Unless (as indicated by stars) there is good data stewardship by the journal (and journal-governed storage), the link to these data is a **weak point** in the system. Articles have a **proper ID** and can therefore be cited, and **citation metrics** are possible. However, datasets and databases as well as figures and tables usually (as indicated by stars) do not have such a stable, **unambiguous identifier**. Neither do terms, thus, retrospectively mapping terms in the narrative, in tables and in figures to unambiguous **concept identifiers** is extremely hard, leaving these research objects suboptimal for **text mining** and **data mining**. Therefore, it becomes difficult to 'recover' original **claims** from these research objects, and they cannot be properly **cited**. With **text-based articles** as the primary focus of scholarly communication, **data sharing** (that is, deposition and sharing) and **community annotation** will remain invisible for mainstream **citation metrics** and will not accrue the proper **credit** or **reward**.

non-semantic, stable unique and universal identification (UUID), to which different Uniform Resource Indentifiers (URIs) can be resolved[5]. Nanopublications support *in silico* knowledge discovery, tapping massive treasures of implicit information[7,8].

As an illustration of our vision, we represent the current state of scholarly communication in a graph of visualized nanopublications (**Fig. 1**). In the caption, we show that this picture can also be represented in narrative text, which is much more easily understood by people than the picture. However, making the picture did help structure the argument, making the text easier to write, and the picture is 'reasonable' for computers. However, it is hard to go the other way around, that is, to reconstruct the picture from the text.

We needed 217 words to describe the 21 assertions in **Figure 1**. Please note how we introduced near-synonyms, like 'scientific awards' for 'professional awards'. By using ambiguous terms and complicated sentence structure, we all contribute to 'knowledge burying'[4]. Above all these difficulties hovers another problem: much of what is worth mining is simply not findable or accessible with the current query methods and firewalls.

Imagine that we published **Figure 1** as a set of properly interconnected machine-readable nanopublications. This picture would then indeed be perfectly 'reasonable' in semantic computation engines such as the LarKC system[7]. Computer reasoning would most likely shift the narrative articles oval in **Figure 1** from its central position slightly off to the side and insert nanopublications in the central position (**Fig. 2**). This move would make all of the problematic 'red predicates' in **Figure 1** disappear by the virtues of the machine-readability of nanopublications. We think that it is worth looking at that suggestion.

Some argue that the rhetoric in articles is difficult to mine and to represent in the machine-readable format. Agreed, but frankly, why should we try? All nanopublications will be linked to their supporting article by its DOI. Many conceptually unique biological assertions are repeated time and again in texts and databases. Capturing the majority of them using a variety of trusted sources is one way to collect almost all relevant biomedical assertions ever made and to enrich them with a dynamic evidence factor based on frequency and conditionality[6].

When reasoning over the associations represented in such a computer readable set of assertions a scientist may have reason to check—even to doubt—any particular assertion, in the graph. Ideally, the list of underpinning articles and other data sources supporting that claim should be just a click away, enabled by the nanopublication provenance. The researcher can now judge the validity of the claim in question much better by reading the articles than by trying to judge a rhetoric argument that is painfully distorted into machine-readable format.

Practical implementation for knowledge discovery

A system very close to the one described in **Figure 2** will soon be put to the test in the recently launched Open PHACTS project of the Innovative Medicines Initiative to create an Open Pharmacological Space. This project will be based on representation technologies and tools currently used to expose a wide range of data in machine processable forms[9].

Giardine *et al.*[2] review a number of database entries relating human gene variants to hemoglobin phenotypes. The technical feasibility of microattribution for the submitted variant-phenotype associations has been the subject of a pilot study. We have collaborated to mine nanopublications from the article text as well as from the underlying databases and supplementary data. The pilot results illustrate the points made above. Using text mining and manual inspection, we recovered 698 nanopublications from the narrative covering all biomedical concepts in the paper. Only 13 of these directly assert genomic variation, for example, those of the composition [HGVS gene variant name][has][variant frequency].

Using simple unambiguous parsing routines, we have represented (in **Supplementary Tables 1** and **2**, respectively) two classes of nanopublications in the supplementary tables of Giardine *et al.*[2] of the form [HGVS gene variant name][has][variant frequency] and [HGVS gene variant name][has][OMIM allelic variant ID]. For those two classes, we found 1,855 instances of nanopublications now deposited at http://www.nanopub.org/

in XML format. These open access nanopublications are now provenance-linked to the article and the databases (and thus to the submitters, curators and authors). The provenance-linked nanopublications have the potential to increase the conventional citation rate of the article and database as well as provide the potential for microattribution. We estimated that the databases described in the article[2] currently contain around 40,000 nanopublications with meaning to geneticists and with citation potential.

Importantly, in this article, we deal with a particularly vulnerable subset of 'concepts', namely the so-called 'variants' in gene sequence. The HGVS nomenclature of such variants may increasingly be enforced by some prescient journals[10], but trying to find these variants and their synonyms in the broader literature is a notoriously difficult task that can only be done with some degree of success if one has access to the full text and, more importantly, all the supplementary data. In a related text-mining analysis, out of 4,940 different variants of 11 genes from the LOVD, only 16 variants could be identified in 10 million PubMed abstracts. Again, we see the tremendous advantage of data publication over text mining in exposing potential nanopublications. These results indicate that authors should construct and publish their data as nanopublications in tabular, ID-based databases as part of their submission and support these tables with narrative text.

The brooding question: can data curation duties be traded?

Only a minority of nanopublications in databases and datasets will ever make it into a narrative as an explicit textual assertion. Even if they do, they will be very difficult to recover retrospectively, for reasons related to access and the failings of mining technology, in confronting ambiguity and sentence construction. We estimated that describing the supplementary data of Giardine *et al.*[2] would require roughly 4 million words, with the result being a corpus hardly readable by machines.

On the other hand, a single LOVD website (http://www.dmd.nl/) consistently enjoyed more than 50 citations annually over the past three years. It is therefore reasonable to assume that proper formatting and exposure of the nanopublications contained in diverse sources such as locus-specific databases could allow these resources to be recognized for the important scientific contributions they actually are. Appropriate standards for proper measurement of these citable items seem to be the only remaining obstacle. So, let us agree to evolve these and to communicate more effectively.

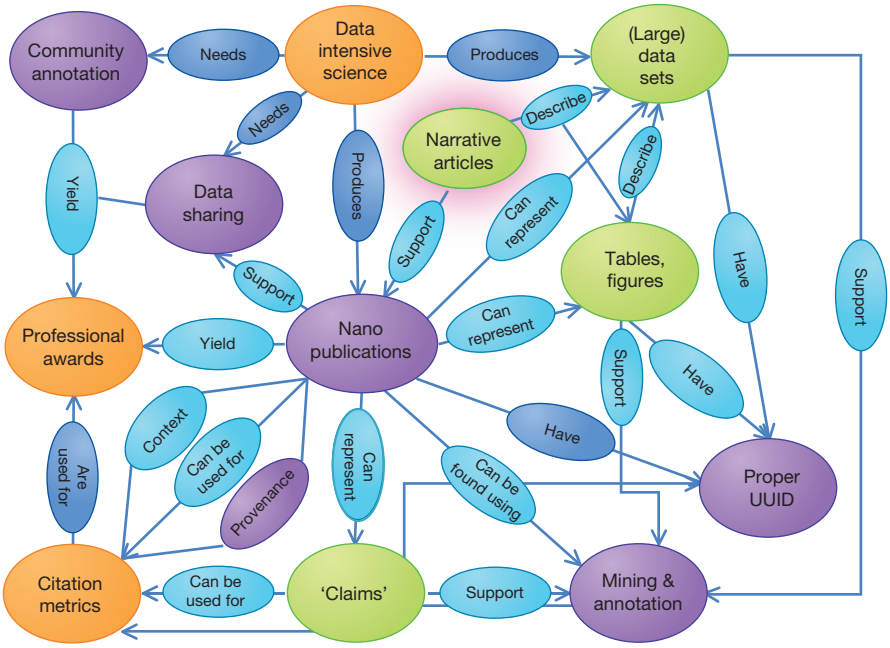

Figure 2 A proposal for the future of scholarly communication. The concepts and predicates are represented as in **Figure 1**. By placing machine-readable **nanopublications** at the core of communication and moving the **narrative** slightly in the graph, many problems may be solved. First, **datasets**, **tables** and the underlying data of **figures** (graphs), as well as their captions, can be represented as **nanopublications**. Because all concepts in a **nanopublication** graph, including those in the provenance and context parts, have a **proper UUID**, they automatically link through the **provenance** to their **source** and their position in that source. The narrative article now becomes supplementary to the data: it provides the detailed description of the **context** as well as the **rhetoric** building the argument as to why the data are valid and the claims correct. Good news for **article** publishers is that each **nanopublication** in a conventional **article**, a supplementary **dataset** (for instance, for two co-expressing genes) or a related **database** (for instance, for a variant-phenotype association) is now intrinsically **findable**, **citable** and **hard-linked** (through the article **UUID** in the provenance) to the **underpinning article**. Therefore, **nanopublications** will have the potential to increase the hit rate of an **article** and will promote proper citation of the **article** in which the 'first' **claim** was made. As **claims** are now individually **citable**, **citation metrics** are possible, and **community curation** and **annotation** can be traced back to the **contributing scientists**. Now, proper **scientific reward** is possible by **precise attribution**, and in the case of the simplest multiples, **nanoattribution**.

URLs. Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/; SwissProt at ExPASy Proteomics Server, http://expasy.org/sprot/; Leiden Open Variation Database (LOVD), http://www.lovd.nl/2.0/; The IMI Open PHACTS project, http://www.openphacts.org; PubMed, http://www.ncbi.nlm.nih.gov/pubmed/; HbVar, http://globin.cse.psu.edu/hbvar/; OMIM, http://www.ncbi.nlm.nih.gov/omim.

AUTHOR CONTRIBUTIONS

B.M., J.T.d.D., G.v.O., J.H., B.K. and P.G. conceived of the experiment and supervised the research. H.v.H., C.C., P.-B.t.H., E.v.M., B.S. and E.S. performed the experiments. R.H., B.G., M.R. and J.V. commented on the experiments and the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

1. Mons, B. *et al. Genome Biol.* **9**, R89 (2008).
2. Giardine, B. *et al. Nat. Genet.* **43**, 295–301 (2011).
3. Editorial. *Nat. Genet.* **39**, 423–423 (2007).
4. Mons, B. *BMC Bioinformatics* **6**, 142 (2005).
5. Mons, B. & Velterop, J. *Workshop on Semantic Web Applications in Scientific Discourse* (*SWASD 2009*) (Washington, DC, USA, 2009).
6. Groth, P., Gibson, A. & Velterop, J. *Information Services & Use* **30**, 51–56 (2010).
7. Van Haagen, H. *et al. PLoS One* **4**, e7894 (2009).
8. Van Haagen, H. *et al. Proteomics.* **11**, 843–853 (2011).
9. Bizer, C., Heath, T. & Berners-Lee, T. *Int. J. Semant. Web Inf. Syst.* (2009).
10. den Dunnen, J.T. & Antonarakis, S. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.* **15**, 7–12 (2000). Erratum in *Hum. Mutat.* **20**, 403 (2002).