# Best practices in Publishing of Data

December 2012

The accessibility and management of digital data, especially “big data[1]” is rapidly gaining increased attention from key stakeholders in scholarly publishing and there is a new focus on making data sets accessible and citable in open repositories. New organizations such as DataCite[2] have recently formed to establish easier access to research data, increase acceptance of research data as legitimate contributions to the scholarly record, and to support data archiving that enables results to be verified and re-purposed. New forms of publication called “Data Publications” are also being developed that publish descriptive articles about datasets intended to make them citable, interpretable and re-usable[3]. STM welcomes these developments and fully supports efforts to improve the discovery, interpretation, citation, re-use, curation, and preservation of digital data.

Because legal and rights issues are inevitably raised with publication and re-use of any material, even data (see Appendix A: Rights and Data), STM believes that it can help promote a supportive rights environment for this new kind of publication by issuing the following best practice guidelines for the scholarly publication of digital data:

1) For data contained and explained within manuscripts that describe, analyse and interpret a research project, STM recommends that publishers should seek only a non-exclusive license to reference and link to such data.

2) For data contained and explained in supplemental material that accompanies manuscripts which describe, analyse and interpret a research project, STM also recommends only a non-exclusive linking or citation license.

3) For raw data and datasets that underlie, but do not accompany, manuscripts submitted for publication STM recommends that publishers, at most, use a non-exclusive license to reference and link to those data.

4) For manuscripts that describe datasets, or parts of datasets held in data centres or repositories, i.e. data publications, STM recommends that publishers use a non-exclusive license to reference and link to those data that are the subject of the data publication – and not the entire body of data itself.

[1] i.e. extremely large and complex collections of data

[2] http://www.datacite.org/

[3] Opportunities for Data Exchange (ODE) 2011 Report on Integration of Data and Publications – see http://www.alliancepermanentaccess.org/index.php/2011/10/24/ode-report-on-integration-of-data-and-publications-published/

## Appendix A:
## Rights and Data

Publishers need to be aware of the characteristics of the data which they encounter in the course of their activities because those characteristics affect the rights that subsist in this material and consequently the rights publishers need to secure in order to respect the intellectual property rights of others. The following overview highlights key issues for publisher consideration:

Data are often intensively factual and numeric and big data projects conducted by multiple institutions can be highly modular. There are few "authors" of such projects who are involved in all aspects of the project or its results. Authorship itself can be attenuated, with greater reliance on software and computing. Many data projects are also ongoing, with data and data interpretations being added on a regular basis. The sharing of data is increasingly a funder requirement, although there are still areas of research with strong concerns about confidentiality, often in applied research.

Nonetheless the designer of a research project, and possibly the research funder or employing institution, may have some rights in the results of a research project that generates data. Copyright protection, however, probably does not yet attach to that "raw" data. This is because copyright protection does not arise until information is structured, formatted, and excerpted to provide an expressive object that illustrates a key research idea or point of view; or structured into a database in a jurisdiction that recognizes either database rights or copyright compilation rights. As most publishers know, copyright likely attaches to manuscript elements like charts or illustrations because of their expressive nature even if the essential factual elements of those materials are not in and of themselves copyrightable.

Sometimes the concepts of curation and stewardship are conflated with the concept of some form of "ownership", and this can in turn cause confusion about intellectual property rights. Researchers who generate data might well have rights in such data, but those rights might equally be attenuated and undeveloped. Some curators or host sites may in fact have no intellectual property rights in the hosted data. Further, it is often the case that when data is posted online that it will be posted under a license to users or with the assumption by the "owner" that the data will be used without significant restriction – perhaps through the use of a Creative Commons license[4].

Other intellectual property (IP) rights, including trade secret or patent protection, may reside in data or sets of data and researchers or others who wish to obtain such IP rights must be careful about the timing of such disclosures. Public disclosure will also eliminate the possibility of protecting technology through trade secret laws which require the continued maintenance of confidentiality. In certain types of research, information about medical patients or other research subjects may be included in such data projects, and the rights of such patients in terms of informed consent and data protection/privacy will need to be protected. The disclosure of sensitive personal information should always be stringently guarded against. Many research funders, national research organizations and research institutions also have strict policies with respect to research subject identification and informed consent, which must be

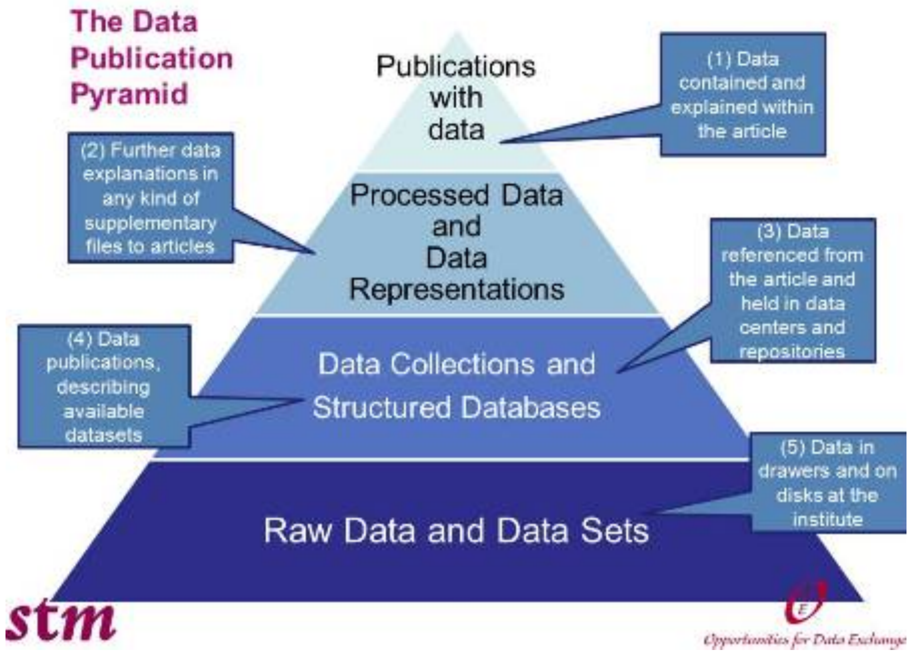
[4] see http://creativecommons.org/

complied with.  Finally, policies concerning tests and experiments will also have legal and compliance dimensions for researchers engaged in data-generating research.

In 2006 STM issued a statement on a statement on *Databases, data sets, and data accessibility*[5] regarding rights distinctions between raw data, data sets, and databases and we continue to endorse the principles expressed in that document. Also, in June of this year, STM, in collaboration with DataCite, issued a statement on the *Linkability and Citability of Research Data*[6] designed to achieve commonly shared principles for best practices between publishers, researchers and data repositories regarding linking between data and publications and citation practices for data sets. Both statements are available from the STM website – see http://www.stm-assoc.org/.

Research data comes in many different manifestation forms. Publications have always contained data, usually in a very condensed, processed and summarised way via graphs, tables and illustrations. At the other end of the spectrum is raw data and original data sets which too often remain unaccessible on people's computers, hard disks or in drawers. Many authors add their underlying research data in supplements to journal articles. In disciplines with community supported data archives (examples are Genbank, World Protein Database and Pangaea) researchers can deposit their data in a safe and relKiable way and publishers can ensure persistent links between the data and related publications. (See the full report available at: http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf )

For background on the classification of data and how data inter-relates with published science, the following Data Publication Pyramid has found to be instructive:


The Data Publication Pyramid
Publications with data
(1) Data contained and explained within the article
Processed Data and Data Representations
(2) Further data explanations in any kind of supplementary files to articles
(3) Data referenced from the article and held in data centers and repositories
Data Collections and Structured Databases
(4) Data publications, describing available datasets
(5) Data in drawers and on disks at the institute
Raw Data and Data Sets
stm
Opportunities for Data Exchange

[5] see  http://www.stm-assoc.org/2006_06_01_STM_ALPSP_Data_Statement.pdf
[6] see http://www.stm-assoc.org/2012_06_14_STM_DataCite_Joint_Statement.pdf