

## 1.1. Culture of data-sharing

Overview: *Initiatives pursued by the funders and publishers of research to promote data-sharing have moved the agenda forward, however, cultural and technical barriers remain. Attitudes and abilities vary across specialities within biomedical research and though obliging data sharing has had some success, incentives and expertise are still lacking in many areas.*

### 1.1.1. Current Situation

The last ten years have seen consistent and targeted promotion of data sharing in research by organisations such as the Research Councils, the National Cancer Research Institute (NCRI), research charities and other funders and publishers of research. The aims are clear: “Ensuring data are made widely available to the research community accelerates the pace of discovery and enhances the efficiency of the research enterprise.”<sup>1</sup>

The assumption is that data sharing is a ‘good thing’ and that connectivity between data will enable greater research potential. There has been a lot of activity in areas such as access and governance, initiatives providing portals and developing standards and there is evidence that “in many research fields – from genetics and molecular biology to the social sciences –data sharing is ingrained in how researchers work”<sup>2</sup> and that, with regard to Systems Biology at least, “data sharing, despite some anomalies, is the prevailing ethic in this community.”<sup>3</sup>

The picture is mixed across the life-sciences, however, as noted in this recent call for contributions to a thematic series on data standardisation, sharing and publication by the online Journal BM Research Notes: “different disciplines have embraced the possibilities of data sharing and open data to differing extents, and it can take the leadership of a small number of individuals to develop and promote their standard to secure widespread adoption, and enable interoperability of scientific data... In other cases a standard of data collection and preparation might be well known amongst circles of experts but perhaps unknown to researchers in different or even related fields. But with few journals considering data-driven articles and apparent inconsistencies in incentives and rewards for data publication, the availability of definitive and freely-available examples of re-usable, standardized data across the life sciences is patchy at best.”<sup>4</sup>

Funders and journals are addressing this issue, promoting data sharing by various means including policies obliging researchers to make their output publicly available. As noted in regard to Systems Biology, however, “Where funder policies do not reach, there is a mix of results. Some researchers make great efforts to share data while others may retain their findings or publish in a form that means that although data are available, they are not readily accessible”<sup>5</sup> or ‘protecting by pdf’ as the practice is known within the community.

The situation is least advanced in Social and Public Health Sciences: “There are many datasets produced by individual researchers or small project teams that could have long term viability if they were offered to an appropriate data centre, but this tends not to be the natural course of things. The sharing of datasets from small scale research projects appears to be relatively uncommon at present.”<sup>6</sup> Other analysts concur: “By contrast, this culture has yet to be widely embraced by the public health research community.”<sup>7</sup>

---

<sup>1</sup> Walport M, Brest P. *Sharing research data to improve public health*. The Lancet, Early Online Publication, 10 January 2011

<sup>2</sup> Ibid.

<sup>3</sup> *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* - Report commissioned by the Research Information Network (RIN) in association with the Joint Information Systems Committee and the National Environment Research Council (NERC) – published June 2008. This report covered six discrete research areas, two of which were Social and Public Health Sciences and Genomics and two interdisciplinary areas, one of which was Systems Biology.

<sup>4</sup> A call for BMC Research Notes contributions promoting best practice in data standardization, sharing and publication; <http://www.biomedcentral.com/1756-0500/3/235/>

<sup>5</sup> Ibid.

<sup>6</sup> Ibid.

<sup>7</sup> Walport M, Brest P. *Sharing research data to improve public health*. The Lancet, Early Online Publication, 10 January 2011

Whilst the battle for cultural change appears to be far advanced in some areas and at least engaged in others, attitudes are not the only barrier: “Problems of reuse centre around... technical issues... – the variety of formats, the non-standardisation of formats, the need for proprietary software and so forth.”<sup>8</sup>

The picture is not uniform across all fields within biomedical research - in some areas data sensitivity and cultural barriers remain more challenging to address than technical and ethical issues. Others, genomics for example, appear comparatively mature, with both cultural and technical issues well in hand.

### 1.1.2. Funders policies

With regard to cancer, the NCRI Data Sharing Policy served as the first step in encouraging wider access to data generated in research by encouraging the NCRI organisations to introduce a requirement for scientists to include data sharing strategies in funding applications and for provision to be made by funders to support the cost of data sharing where necessary.

Since the adoption of the data sharing policy the NCRI Informatics Initiative has been working with the NCRI Partner organisations to implement data sharing through developing individual data sharing policies and guidelines that set a clear direction to allow scientists to carefully consider data sharing when applying for funding (see particularly Cancer Research UK’s Terms and Conditions and Administrative Guidelines for Research Grants and Awards launched 2nd February 2009 and the Wellcome Trust’s position on Data Management and Sharing originally published in January 2007, revised in August 2010.)

With regard to genomics: “The genomics research carried out in the UK is funded largely by the Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC), the Natural Environment Research Council (NERC) and the larger biomedical charities such as Cancer Research UK, most of them members of the Association of Medical Research Charities (AMRC has 114 member charities). Of these charities, the Wellcome Trust is the biggest spender on genomics research. The Wellcome Trust led the way for mandating public access to research findings that it had funded, the BBSRC and MRC followed suit, and were joined by the other charities funding medical research. Over 90% of biomedical research in the UK is now carried out under a mandatory open access policy. This applies specifically to articles published in journals, but most of these funders also have a policy on sharing data.”<sup>9</sup>

Some research performed by non-commercial groups is carried out in collaboration with or funded by commercial organisations: “in such cases there may be acknowledged ownership of data by the commercial entity and constraints on how open the data may be made. These situations are relatively rare and in the majority of cases there are no such commercial constraints.”<sup>10</sup>

Some analysts, however, conclude that “the views and practices of life science researchers differ sharply from the strategies being promoted by policymakers and funders, libraries and other information service providers – showing that the attempts to implement such strategies have had only limited impact.”<sup>11</sup>

### 1.1.3. Barriers – incentives and expertise

Discussions of this subject highlight the lack of active incentives (rather than obligations) to share data: “the lack of explicit career rewards, and in particular the perceived failure of the Research Assessment Exercise (RAE) explicitly to recognise and reward the creating and sharing of datasets – as distinct from the publication of papers - are major disincentives.”<sup>12</sup>

<sup>8</sup> *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*

<sup>9</sup> *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*

<sup>10</sup> Ibid.

<sup>11</sup> *Patterns of Information Use and Exchange: case studies of researchers in the life sciences, November 2009 – A report by the Research Information Network and the British Library.*

<sup>12</sup> *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*

As mentioned above when looking at Social and Public Health Sciences the Research Intelligence Network found “scant evidence of researchers wanting to publish datasets. Typically researchers will request data from one or more publicly-available datasets and they will undertake analysis. Often this process leads to the creation of new, derived datasets but these tend not to find their way to the public domain.”<sup>13</sup>

Lack of incentive is blamed for this outcome: “Unlike in some of the other areas we have looked at there are no obvious rewards that accrue to researchers who decide to make their datasets publicly-available – though few deny that sharing datasets produced with public funds is a worthwhile principle..... researchers producing small scale datasets see no reason to invest the time and effort required to make their datasets publicly available. Besides which, some want to control their data, limit the possibility of the data being misrepresented, and limit the scope for competition [our italics].”<sup>14</sup>

“Other disincentives include lack of time and resources; lack of experience and expertise in data management and in matters such as the provision of good metadata; legal and ethical constraints; lack of an appropriate archive service; and fear of exploitation or inappropriate use of the data..... Relatively few researchers have the expertise, resources and inclination to perform themselves all the tasks necessary to make their data not only available, but readily accessible and usable by others.”<sup>15</sup> Many researchers lack the skills to meet the quality standards imposed by data centres without substantial help from specialists.

Additionally, “creating longitudinal datasets is an expensive business and therefore the people responsible for them tend to feel the need to protect them. This is manifested in reported anxiety about commercial organisations using data, deriving slightly or materially different datasets and claiming intellectual property rights over these new datasets.”<sup>16</sup>

Across the biomedical sciences directors and PIs see their restricted data as their intellectual capital: “As with most areas of research, there is competition between researchers to produce the best work in the best journal... Many researchers wish to retain exclusive use of the data they have created until they have extracted all the publication value they can.”<sup>17</sup>

From the perspective of commercial / industry groups data sharing presents many of the same challenges: Intellectual Property and Confidentiality are particularly sensitive issues. In the past big pharmaceuticals organisations have traditionally been conservative over data sharing – concerns include loss of control, cost, other units reaching different conclusions or deriving novel insights which may have commercial value.

Within this field there exists a significant heterogeneity of needs. The complexity and diversity of the biomedical research landscape breeds diversity in tools and methodologies for data capture and analysis, storage, maintenance and curation and this too may fuel confusion and dampen enthusiasm.

---

<sup>13</sup> Ibid.

<sup>14</sup> Ibid.

<sup>15</sup> Ibid.

<sup>16</sup> Ibid.

<sup>17</sup> Ibid.