doi 10.4436/JASS.92006

JASs *Proceeding Paper*

Journal of Anthropological Sciences
Vol. 92 (2014), pp. 179-200

Perspectives on *Open Science* and scientific data sharing: an interdisciplinary workshop

Giovanni Destro Bisol[1,2], Paolo Anagnostou[1,2], Marco Capocasa[1], Silvia Bencivelli[3], Andrea Cerroni[4], Jorge Contreras[5], Neela Enke[6], Bernardino Fantini[1,7], Pietro Greco[3], Catherine Heeney[8], Daniela Luzi[9], Paolo Manghi[10], Deborah Mascalzoni[11], Jennifer C. Molloy[12], Fabio Parenti[13], Jelte M. Wicherts[14] & Geoffrey Boulton[15]

1) Istituto Italiano di Antropologia, Rome, Italy
e-mail: giovanni.destrobisol@uniroma1.it

2) Università di Roma "La Sapienza", Dipartimento di Biologia Ambientale, Rome, Italy

3) Free lance scientific journalist

4) University of Milan, Bicocca, Italy

5) American University, Washington DC, USA

6) Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Germany

7) University of Geneva, Switzerland

8) Consejo Superior de Investigaciones Científicas, Madrid, Spain

9) Consiglio Nazionale delle Ricerche, Rome, Italy

10) Consiglio Nazionale delle Ricerche, Pisa, Italy

11) University of Uppsala, Sweden and EURAC research, Bozen, Italy

12 University of Oxford, United Kingdom

13) Istituto Italiano di Paleontologia Umana, Rome, Italy

14) Tilburg University, The Netherlands

15) University of Edinburgh, United Kingdom

Summary - *Looking at Open Science and Open Data from a broad perspective. This is the idea behind "Scientific data sharing: an interdisciplinary workshop", an initiative designed to foster dialogue between scholars from different scientific domains which was organized by the* Istituto Italiano di Antropologia *in Anagni, Italy, 2-4 September 2013.We here report summaries of the presentations and discussions at the meeting. They deal with four sets of issues: (i) setting a common framework, a general discussion of open data principles, values and opportunities; (ii) insights into scientific practices, a view of the way in which the open data movement is developing in a variety of scientific domains (biology, psychology, epidemiology and archaeology); (iii) a case study of human genomics, which was a trail-blazer in data sharing, and which encapsulates the tension that can occur between large-scale data sharing and one of the boundaries of openness, the protection of individual data; (iv) open science and the public, based on a round table discussion about the public communication of science and the societal implications of open science. There*

were three proposals for the planning of further interdisciplinary initiatives on open science. Firstly, there is a need to integrate top-down initiatives by governments, institutions and journals with bottom-up approaches from the scientific community. Secondly, more should be done to popularize the societal benefits of open science, not only in providing the evidence needed by citizens to draw their own conclusions on scientific issues that are of concern to them, but also explaining the direct benefits of data sharing in areas such as the control of infectious disease. Finally, introducing arguments from social sciences and humanities in the educational dissemination of open data may help students become more profoundly engaged with Open Science and look at science from a broader perspective.

Keywords – *Data sharing, Biobanks, Metadata, Science and Society.*

The concept of *Open Science*

The advent of the World Wide Web and associated technologies has brought an explosion of free online information, which has had a deep impact on most aspects of our daily lives (Hendriks, 1999; Morrison *et al.*, 2001). This new data-rich era of instantaneous communication creates novel challenges and opportunities for research, one that researchers cannot avoid confronting. The process of doing so is vigorously underway in many research fields that are making scientific knowledge and underlying data available to the whole scientific community and the public, owing to the combined efforts of researchers, science communicators and other stakeholders (Neylon & Wu, 2009; Boulton *et al.*, 2012). There is now a growing international movement for "open science", by which is meant making publication of scientific concepts and the data on which they are based readily accessible to all, together with procedures for sharing important data sets. This trend is not only limited to technical and IT aspects, but extends to epistemological, sociological and political issues (Fecher & Friesike, 2013; Mauthner & Parry, 2013; Velden, 2013) and to governmental initiatives to open official data both to citizens and to entrepreneurs able to offer new data-based services. This widening horizon creates many opportunities for collaboration to scholars from a wide diversity of disciplines. Such cooperative efforts are essential if open science principles are to be adapted effectively to the needs of different knowledge domains, and if they are to be successful in achieving deeper involvement of the public in science. Effective and creative cross-fertilization will not only depend upon theoretical engagement but also on addressing infrastructural, economic and motivational barriers between disciplines. The impact of digital technologies is not restricted to science, but creates challenges for the whole range of research and scholarship. In the "digital humanities" for example, research often entails new methodologies and intellectual strategies that are nonetheless grounded in traditional humanistic areas of focus (the nature of authorship, continuity of concepts over time, the social context of artistic expression). The challenges not only apply to data that are born digitally, but also across large corpora of text, as well as visual, aural, audiovisual, sensory, neurological and even kinesthetic forms of information.

It was in this context that a meeting entitled "Scientific data sharing: an interdisciplinary workshop" (Anagni, Italy, 2-4 September 2013) was conceived, to explore open science from a broad perspective, rather than focusing on field-specific issues (e.g. Schofield *et al.*, 2009; Dalgleish *et al.*, 2012). Its intention was to stimulate dialogue between different perspectives of "open data" and to make a first assessment of common needs and opportunities as a starting point for further interdisciplinary initiatives. The meeting was organized by Giovanni Destro Bisol, Paolo Anagnostou and Marco Capocasa on behalf of the *Istituto Italiano di Antropologia.*

The main sections of this report discuss the conceptual and practical framework for open

science, explore open scientific practices, present studies of the particular case of human genomics which ushered in the modern rationale for open data and finally present the outcome of a round table discussion about the public dimension of open science.

Summaries by their authors of the presentations at the meeting form individual sub-sections. In two cases, authors who were not able to attend the meeting (N. Enke and D. Mascalzoni) also submitted summaries.

Setting a common framework

In order to establish a common framework and facilitate interactions among the participants, the first part of the meeting was dedicated to the discussion of basic aspects of Open Science. This was organized with a top-down logic, starting with an outline of the relationships between Science and Society, followed by an overview of Open Science, and finally, an introduction to key concepts and definitions for open data.

Science within social change: knowledge, communication and the knowledge society[1]

Societies worldwide are currently undergoing rapid and fundamental changes (e.g. Stehr, 1994), firstly through greater involvement in governance by citizens who expect their rights to be increasingly guaranteed (e.g. Elias, 1991) and secondly through the rapid evolution of a knowledge economy, where knowledge is a key-factor in a growing number of socio-economic exchanges (e.g. Foray, 2004). A corollary is that the maintenance and extension of democracy depend up scientific knowledge being more deeply embedded within society. In this setting, the greatest inequalities are determined by the quantity and quality of information that is available to citizens, by cognitive capital, and by the capacity of individuals to relate their own circumstances and history to the dynamics of local and global society.

What then do we mean by "the knowledge society"? We should distinguish three logical levels: the individual, the collective and knowledge (Cerroni, 2006, 2007). The *individual* is linked to the complex processes of acting and reasoning. The *collective* concerns both the reference knowledge-field (with its social networks and institutions) and the knowledge-society (with its national and worldwide institutions and public opinion). Due to its inherent complexity (e.g. Collins, 2010), the *knowledge level* is better understood if we distinguish three families of knowledge: the *intellectual family* (both *explicit* and *implicit*); the *practical family* (to know how-to-do and how to be or behave in specific social contexts); the *objectified family* (artificial products and environments encapsulating encrypted knowledge). Integrating the three logical levels described above in a theoretical model gives rise to four logical phases (one for each transition from one level to the closest): knowledge *generation* (individual-collective), *institutionalization* (collective-knowledge), *diffusion* (knowledge-collective) and *socialization* (collective-individual). The first phase is concerned with the creative participation of the individual, with her/his resources, strategies and aims; the second with the recognition of a collective value of the generated knowledge and its incorporation among the stable values of society; the third with the widespread diffusion of knowledge across society; the fourth with science regulation and individual socialization (*knowledge-able citizens*). The first two phases can be considered as internal communication, and the last two as external communication (Fig. 1). It should be noted that the model implies neither causal nor linear processes. What emerges from the combination of processes of the four phases is what we define as "innovation", i.e. the new knowledge which spreads (albeit not homogeneously) across society.

There is a current priority in many countries to broaden the production of new knowledge and make its institutionalization an open and participatory process, by supporting "open access publication". Open communication of scientific knowledge is regarded as an essential attribute of

[1] Lecture by **Andrea Cerroni,** andrea.cerroni@unimib.it.

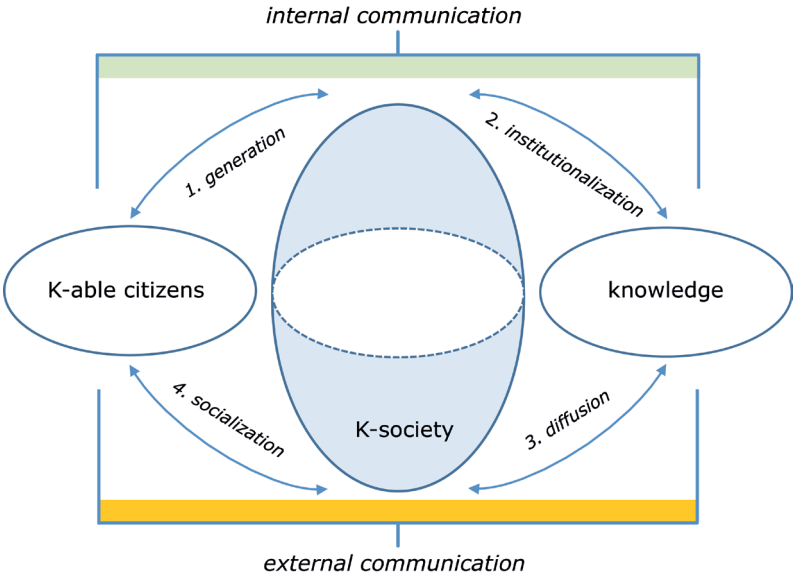

Fig. 1 - Theoretical model describing the four phases* (generation, institutionalization, diffusion and socialization) *by which knowledge and innovations are produced and spread across society. The colour version of this figure is available at the JASs web site.

a modern democratic society, despite open science practices being neither widespread nor theoretically well grounded. Communication is neither a one-way process (top-down) nor an easy two-way process. It requires science to be open to society and, *vice versa*, society to science. In the former case, science has to be more responsible of its role in front of an active societal participation, and in the latter, citizens have to be more aware of the significance of knowledge and engaged in a science-based public choice. Open science, in other words, influences the quality of both scientific knowledge and democracy within the knowledge-society.

Science as an open enterprise[2]

Open publication of scientific concepts together with the evidence (the data) on which they were based was the bedrock on which the scientific revolutions of the 18th and 19th centuries were built and is fundamental to the future progress of science. It allows scrutiny of the evidence on which the concept is based and the logic of the argument connecting them. It permits replication of observations or experiments or their refutation, and has been the basis of the principle of scientific self-correction that ensures that scientific understanding is cumulative. The "data storm" of recent decades has seriously undermined these fundamental processes. In many fields, it is no longer the norm that the data on which a concept is based are published concurrently with the concept. This is reflected in the increasing incidence of results that are not reproducible, not necessarily because of error, but because the data and/or metadata are absent or inadequate (Alsheikh-Ali *et al.,* 2011; Freese, 2007; Savage & Vickers, 2009; Wicherts *et al.*, 2006), and in some cases because data has been lost (Vines *et al.*, 2014). This is a crucial issue for the future of science, potentially putting the credibility of the scientific process at stake. Failure to release "intelligently open data" (see below) concurrently with the publication of the concept behind it should come to be regarded as scientific malpractice.

[2] Lecture by **Geoffrey Boulton**, G.Boulton@ed.ac.uk.

However, the "data storm" also offers enormous opportunities. Linking databases in ways that can integrate their contents creates powerful new means of identifying patterns in phenomena that were previously beyond our horizon. Open data and data sharing in ways that reinvent open science for the modern age have the potential to improve the role of science in facing global challenges, combat fraud and malpractice, and engage with citizens in ways which can potentially change the social dynamics of science, by making it a public enterprise rather than a private one conducted behind closed laboratory doors.

However, merely dumping data into databases is not sufficient. Effective open science requires "intelligent openness" (Boulton *et al.*, 2012), which means that data and metadata must be:

a) Discoverable - how can you find out that they exist?
b) Accessible - can you obtain them?
c) Intelligible - can they be understood?
d) Assessable - e.g are the originators trustworthy?
e) Re-usable - can the data be used for replication or re-purposing?.

These are the fundamental criteria for truly open data.

Although the default position for scientific data derived from research funded by the public purse is that it should be "intelligenly open", there are three legitimate boundaries to openness:

1) Commercial actitivies where the business model does not favour openness and where there is an overriding public interest in deriving economic benefit. Complications arise in this context from public/private partnerships in funding research.
2) Where personal data is involved, it is important that personal privacy and confidentiality should not be infringed in the public domain. However there is a proportional balance to be struck between protecting the privacy of the individual and the wider public good, for example in using statistics derived from national medical records to set public health priorities. There is every sign that the European Parliament will shortly enact a regulation that could seriously inhibit much medical research, implying that the Parliament has chosen to prioritise individual privacy over the broader public good.
3) To protect safety and security, for example when a scientific discovery has dual use, where knowledge could be beneficial in some hands but could threaten individual or population safety or security in malign hands.

It is important to note however that these boundaries are not so sharply defined in a way that could be readily prescribed by a few generic rules. The boundaries tend to be fuzzy and complex, and require the exercise of much judgement.

In conclusion, where do responsibilities lie in implementing the changes referred to above? They lie with scientists in accepting that concurrent data publication is intrinsic to science; with universities in taking responsibilities for the knowledge they create; with funders of research in recognizing that open data is part of the fundamental process of science and not a voluntary add-on; with academic publishers in insisting on the concurrent publication of intelligently open data; with the learned societies that set the principles and priorities of individual disciplines in advocating open data as a norm for their discipline; and, finally, with national governments in ensuring development of effective infrastructures and support of an open science *ethos* (see also "Concluding Comments").

The Open Knowledge definition and principles for open data in science[3]

The Open Knowledge Definition (OKD; http://opendefinition.org/od/) and Panton Principles for Open Data in Science (Murray-Rust *et al.*, 2010; Molloy, 2011) specify how accessible and reusable knowledge must be in order to be described as "open". In the view of the authors: "A piece of data or content is open if anyone is free to use, reuse and redistribute it – subject only, at most, to the requirement to attribute

[3] Lecture by **Jenny Molloy**, jenny.molloy@okfn.org.

and/or share-alike". This involves inclusiveness in terms of both users and uses, explicitly permitting commercial reuse and also elaborating technical requirements such as the ability to bulk download data in a reusable digital format. An accompanying curated list of OKD-compliant licenses enables copyright holders to clearly express their wishes and maximize the interoperability of data and content within the global knowledge commons.

The OKD will not always correspond to personal definitions of openness, which is a normative and multi-faceted concept. Even concentrating solely on legal and technical aspects, many individuals and organisations choose a different cut-off point along the spectrum of openness to demarcate as 'open'. Therefore, some may consider subsets of the OKD's fourteen clauses to be 'too open' or too specific for their requirements and aims. However, even in these cases the OKD serves a useful role in outlining parameters around which domain specific discussions can take place. In relation to science, the OKD is consistent with the Budapest Open Access Initiative's (BOAI's) stance on inclusiveness of access to research articles, but recognizes that scientific knowledge also exists in other forms, as affirmed in the 2012 BOAI10 recommendations a decade on from the original declaration (http://www.budapestopenaccessinitiative.org/boai-10-recommendations). Access to raw data underlying published research is one of these forms and constitutes a separate and important issue, with few datasets currently being made available online and significant confusion regarding the reuse of those that are.

The Panton Principles for Open Data in Science (Murray-Rust *et al.*, 2010) directly address the issue of access to research data, emphasising that data are the currency of science on which all future scientific enquiry builds, so the ability to validate, reuse and remix data is crucial to the scientific endeavour. The Principles take a stronger stance on legal openness than the OKD, for several reasons: the social norm of attribution that already exists in the scientific community, the inappropriateness of applying many open content licenses to data and the major issue of interoperability of scientific datasets. For these reasons, they advocate that all scientific data is clearly placed in the public domain via a copyright waiver or appropriate license, maximising clarity for both producers and consumers. Boulton's call for 'intelligent openness', which encompasses a broader view beyond the technical and legal nuances of how data is made available, is crucial to changing norms of data availability in science. However, the strong and unambiguous statements of the OKD and Panton Principles still serve as a benchmark and focus for discussion on the meaning of 'open data' to different scientific communities and their stakeholders.

Insights into scientific practices

There have been three categories of study in the last decade that evaluate data sharing practices:

1) studies aimed at assessing data sharing rates based on the analysis of scientific literature;
2) investigations of attitudes towards data sharing across different researcher groupings (communities, disciplines, institutions) by using *ad hoc* questionnaires;
3) analyses of specific case-studies which illustrate the potential of data sharing for matters of general interest (e.g. human health) or the need for open data approaches in areas which are traditionally less inclined towards the use of IT tools.

The following sections explore these categories in some depth.

The willingness to share scientific data in psychological research[4]

A recent fraud case has sparked debate on data sharing and reproducibility in the field of psychology (Wicherts, 2011). Despite the fact that professional guidelines clearly state that psychological data should be shared for verification purposes (on condition that privacy of

[4] Lecture by **Jelte Wicherts**, J.M.Wicherts@uvt.nl.

human participants is protected), and despite the use of forms that stipulate those guidelines for published articles, only 27% of corresponding authors of recent papers in top psychology journals shared data on request (Wicherts *et al.*, 2006). It is also common for statistical analyses of data to be prone to error (Bakker & Wicherts, 2011) and for unwillingness to share data to be associated with the prevalence of errors in papers from which data were requested (Wicherts *et al.*, 2011). In addition, psychological researchers who present less convincing statistical evidence (against the null hypothesis of no effect) are less inclined to share data for re-analyses by peers than researchers who present stronger evidence (against the null hypothesis). Lack of access to the data underlying scientific claims makes appropriate scrutiny of those claims difficult or impossible, and increases the incidence of false claims.

These results highlight not only relatively poor and error-prone practices of data documentation but also tendencies to present the "best" of a number of possible statistical outcomes. Taken together, poor availability and substandard documentation of data may lead to more error and bias in the presented results, suggesting that sharing of data can enhance the reproducibility of published results. Moreover, willingness to share data could be seen to reflect a principled and collaborative commitment to the creation of new scientific knowledge rather than an exclusive concern with personal reputation. Finally, data documentation is currently given little attention in the curriculum of (under)graduates in psychology and related fields, something that is likely to change as the importance of open data as a basic principle of good science is re-established.

Data sharing, cooperation and empirical approaches[5]

The impact of open science is greatest when it operates as a collaborative process. This was idealized by René Descartes in *Discourse on the method of rightly directing one's Reason and of seeking Truth in the Sciences (1637)* where he writes:

> *"The best minds would be led to contribute to further progress, each one according to his bent and ability, in the necessary experiments, and would communicate to the public whatever they learned, so that one man might begin where another left off; and thus, in the combined lifetimes and labours of many, much more progress would be made by all together than anyone could make by himself."*

This may be the first published statement of the importance of cooperation in science and the starting point for analyses of the processes of scientific production. Those processes no longer involve only "the best minds" but now include a heterogeneous category of stakeholders, including the public. Rather than being restricted to initiatives carried out by individuals or groups of researchers, science production is increasingly seen as a cooperative system (see Velden, 2013). Advancing knowledge and exploiting that knowledge to create novel applications are the two main purposes of scientific research, which combines empirical observation and experiment together with explanatory theories and models. Actors, targets and tools in these processes should not be conceived as closed entities, but entities with internal synergies (e.g. between basic and applied research). Effective communication processes, for example between researchers and stakeholders, and continuous redefinition of theories due to the production of new knowledge are essential for the proper functioning of scientific production. In turn, this requires transparency of practice and an unconstrained information flow between all parts of the system.

A third important aspect lies in optimizing system function to maximize use of human and other resources through cooperation: i) within and across communities, research fields and domains; (ii) in the form of a vertical transfer, i.e. among researchers and trainees; (iii) between researchers and the public, e.g. under the umbrella of the so-called citizen science. The question then arises whether current strategies for data sharing maximize opportunities for cooperation?. Most current strategies are based

[5] Lecture by **Giovanni Destro Bisol, Paolo Anagnostou** and **Marco Capocasa**.

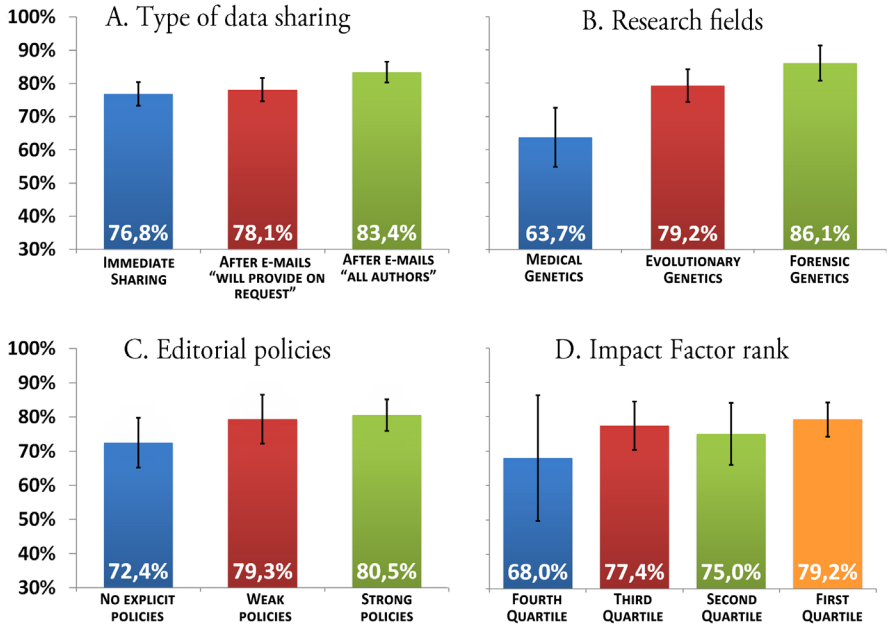

Fig. 2 - Sharing rates of published datasets regarding human genetic variation (from Milia et al., 2012). The analysis was carried out on a total of 253 mitochondrial and 290 Y-chromosomal datasets which were extracted from 508 papers indexed in the Pubmed database between 1st January 2008 and 31st December 2011. Vertical bars indicate 95% confidence intervals. (A) In the "Immediate sharing" category, we reported the rate of datasets shared in the main text, its supplementary material or online databases which were explicitly indicated in the paper; E-mails "will provide on request" were sent to the corresponding authors to request information from papers where data availability upon request is explicitly declared; E-mails "all authors" were sent to all corresponding authors who withheld datasets. The results reported in frames B and C were obtained using the sharing rates obtained including the positive answers to E-mails "will provide on request". See Milia et al. (2012) for further details. The colour version of this figure is available at the JASs web site.

on top-down approaches that aim to provide researchers with tools (e.g. infrastructure and standards), norms (policies and guidelines) and motivations (moral suasion, incentives) without any broad involvement of the scientific community. Such a limited approach implies that current strategies are sub-optimal. However, empirical studies of data sharing practices, both via *questionnaires* or analysis of scientific literature, may provide quantitative answers to questions about the efficacy of norms (are the policies really successful?), identification of motivation for sharing behaviour (why are data shared or withheld?) and adequacy of tools (how often and what databases and standards are used to share data?) (Destro Bisol *et al.*, 2013; see also Congiu et al., 2012). Empirical studies may also reveal effective, informal, sharing practices and barriers to sharing that are widespread in a particular research field, rather than necessarily being prescribed by funding bodies and institutions, and may assist in establishing the most effective data sharing strategies. A relevant example is contained in a recent study of publications in three sub-fields of human population genetics (evolutionary, medical and forensic genetics; Fig. 2) (Milia *et al.*, 2012) which showed that an effective and robust form of data sharing is not yet common practice even in research fields where the nature of the data (codified DNA data, relative simplicity

of metadata and availability of infrastructures) would make this easier to achieve. Our investigation produced three evidence-based proposals on how to increase data sharing: (i) exploring different approaches in closely related research fields; (ii) mandating data sharing before a paper is finally accepted for publication, rather sharing merely being a recommendation, given the difficulties encountered in recovering withheld data after publication of the article based on the data; (iii) extending the practice of collaboration between laboratories (already undertaken by forensic geneticists), a training which may help make researchers more conscious of the fundamental value of data quality and reproducibility, while promoting a climate of trust, transparency and teamwork (Anagnostou *et al.*, 2013).

Data sharing and the impact of technology on the spread of knowledge[6]

A questionnaire-based survey was recently carried out among the researchers of the Department of Health and Environment of the *Consiglio Nazionale delle Ricerche* (CNR) (Luzi *et al.*, 2013) to explore researchers' attitudes to data sharing along with practices of data acquisition and management (see also Parse.Insight, 2009; Tenopir *et. al.*, 2011; Enke *et al.*, 2012). The majority of the 523 respondents to the questionnaire (48% response rate) tend to be very cautious in sharing data. They select which data they are willing to share. Thirty-six percent of them affirm they share some data without restriction in their Institute's website, while 44% assert they do so in national and international networks. Interestingly this percentage increases when researchers use data produced by others (40%) and especially when there are local archives where research data can be submitted (52%). This tendency is more evident when there are national (60.8%) and international networks (50,9%). Moreover, the presence of international networks also slightly increases the percentage of researchers who make the majority of their data available (21.4%) (see Fig. 3).

The researchers' cautious attitude is confirmed by their request to maintain control over research data even after their submission, to be able to update them and know who is using them, when and for which purpose. These aspects are closely related to the lack of formal recognition of the efforts connected with data sharing. The motivation to share data would be significantly strengthened if these activities were evaluated in the same way as they are in producting scientific publications, and if citation for data re-use became routine.

CNR researchers are not widely aware of the standards used by their community of reference. However, they use metadata (e.g. georeferencing information, codes, instrument setting, etc.) which make collected data more easily reusable. Metadata creation tends to be done on a personal/research team basis rather than by applying institutional norms, as few institutes have established common management plans to preserve data, although an encouraging percentage of them intend to implement management plans in the near future. Moreover, almost 85% of researchers stated that there is no-one in their institutes who is trained in or responsible for data management and preservation. Lack of technical support was mentioned as one of the main obstacles to data sharing. On a very positive note, the high rate of responses to the questionnaire as well as researchers' opinions on the importance of research data indicate a high level of awareness and an encouraging willingness to share data if that were facilitated by appropriate policies and support, and by the development of e-infrastructures tailored to researchers' needs.

Reasons for the reluctance to share data in biodiversity science[7]

A survey has been carried out to explore the reasons for reluctance in sharing data in biodiversity science. This included interviews with over 60 researchers and an online survey with over 700 participants. The participants were mainly from the EU, USA, and Canada, but with some from other countries (e.g. Brazil, China, Fiji). The

[6] Lecture by **Daniela Luzi**, d.luzi@irpps.cnr.it.

[7] Summary by **Neela Enke**, n.enke@bgbm.org.

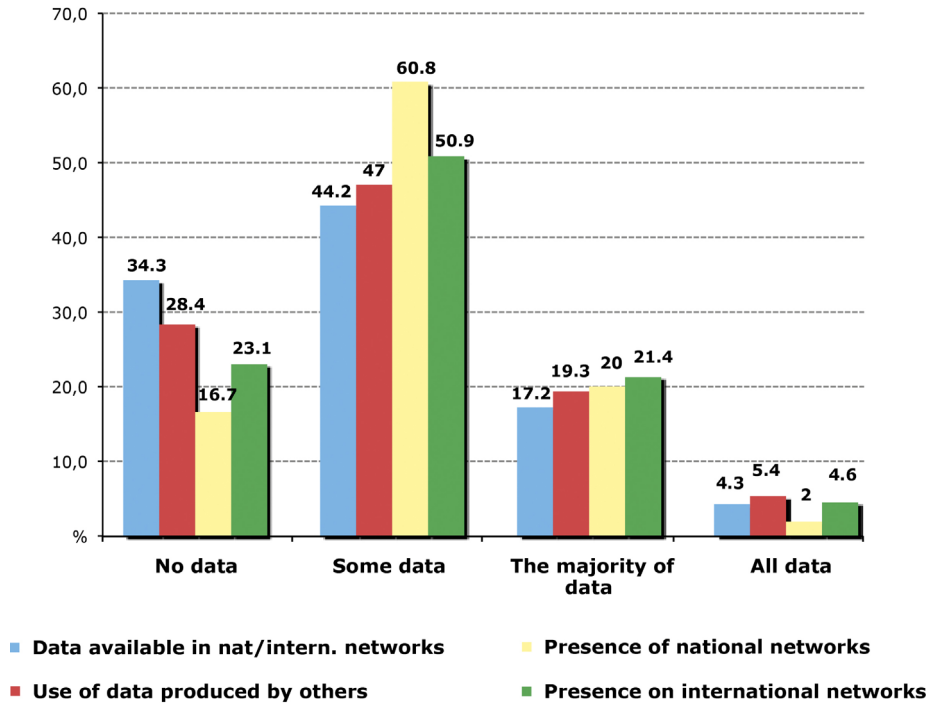

Fig. 3 - Data sharing by researchers using data produced by others and who can deposit data in national and international networks. The colour version of this figure is available at the JASs web site.

participating scientists came from a wide range of fields within biodiversity research: e.g. Systematics, Taxonomy, Ecology, and Climatology. A complete overview of all participants and their answers (anonymized) is given in Enke *et al.* (2012). The study showed that even though nearly 80% of participants were willing to share data, only very few did so. The main reason for not sharing data was the fear of losing control over the data. The effort required to edit data in the way necessary for for sharing was a huge impediment, especially because participants felt that they would not receive any professional acknowledgement for their effort. Instituting clear guidelines about the re-use of published data together with a requirement for data to be formally cited when re-used are important priorities. Depending on the field of research, there was also a lack of data standards and repositories which are vital if sharing is to become the norm.

The interviews showed that the time and effort required to edit and reformat data for re-use was a major impediment to re-use. The creation of institutional data management plans are one way to tackle this issue. The results of the study implied that if the data was inserted into a structured data repository at an early stage in research (ideally at the moment it is collected) and if researchers were supported in managing their data, the amount of time needed in editing data for release could be significantly decreased. However, especially in the EU, most researchers (~70%) never came into contact with any kind of data management plan. The situation is better in the USA where over 50% of researchers were at least aware that their institution had a data management plan (Enke *et al.*, 2012).

In conclusion, the main priorities for reducing the reluctance to share data would be:

a) to promote existing infrastructures and expand them where needed;
b) to integrate data management into every day work routines (e.g. through virtual research environments);
c) to educate students/researchers at a very early stage of their careers on the importance of data sharing;
d) to increase the pressure through journals and funding agencies to publish the data sets themselves along with the results;
e) the development of systems of professional rewards for sharing data.

Data sharing and control of emerging viruses[8]

Data sharing is a vital part of controlling the emergence of infectious diseases, as exemplified by the response to the outbreak of SARS (Severe Acute Respiratory Syndrome). This disease emerged in February 2003 and spread rapidly across three continents, producing a total of 7761 cases and 623 deaths. The SARS epidemic was controlled only four months later by an international effort coordinated by the WHO (World Health Organization) based on extensive data sharing.

The history of SARS is brief, but dense in significance. On February 14, a small notice in the *Weekly Epidemiological Record* reported 305 cases and 5 deaths from an unknown acute respiratory syndrome that had occurred in the Guangdong Province of China. One month later, WHO issued a first alert and the new syndrome was designated as "severe acute respiratory syndrome". On 28 February, Dr Carlo Urbani, a WHO official based in Viet Nam was asked to assist a case of atypical pneumonia in the French Hospital in Hanoi. Realising that the disease was new and potentially very dangerous, he notified the WHO Regional Office, asking for a heightened state of alert. After continuing to treat cases of SARS in Hanoi, on 11 March Dr Urbani left for Bangkok, for a conference on tropical diseases. He was ill upon arrival and asked to be immediately hospitalized. He died of SARS, the disease he had discovered, on March 29th. On March 26, the WHO organized the first "virtual round table" on the clinical and therapeutic aspects of SARS. The « electronic conference » brought together 80 clinicians from 13 countries and a summary of the discussion was published on the page dedicated to SARS on the WHO web site. At the same time, the WHO asked 11 laboratories of excellence in nine countries to set up a network for multicentre research on the aetiology of SARS, and at the same time to collaborate on the development of a diagnostic test. The laboratory network created by WHO took advantage of the new communication technologies (e-mail, secure websites) so that search results on clinical samples from SARS cases could be shared in real time. On the secure website, the various network members shared images of viruses obtained from an electron microscope, sequences of genetic material for the identification of the virus and its typing, viral isolates and samples of various types taken from patients or during post-mortem examinations. The samples from a patient could be analysed in parallel by various laboratories and the results were disclosed in real time. The objectives of the identification of the causal agent of SARS and the development of a diagnostic test were obtained within only a few weeks. On March 21, the Centers for Disease Control and Prevention (CDC) published the first clinical description of SARS, and on April 16 the WHO announced that the cause of SARS was a new pathogen, a member of the coronavirus family that had never been seen before in humans. On May 1, two research groups published the complete genome sequence of the SARS virus in *Science* (Rota *et al.*, 2003).

Based on epidemiological and virological data and clinical evidence, a series of very effective public health measures were rapidly introduced in all interested countries. As a result, the spread of the new disease was stopped. On 5 July, the WHO declared the end of the pandemic because the last human chain of transmission of SARS had been broken.

[8] Lecture by **Bernardino Fantini**, Bernardino.Fantini@unige.ch.

The history of SARS presents a paradox. The SARS disease was the direct result *of* globalization: in any previous historical period, the disease would have caused a few sporadic and isolated cases, without any serious consequences outside its small geographical context. At the same time, globalization has furnished the main scientific tools for fighting epidemics. Furthermore, because there was no vaccine or treatment, health authorities had to resort to control tools dating back to ancient times: early tracking, identification and isolation of patients (chain of transmission), management of close contacts (quarantine), travel restrictions, public information, education to encourage prompt reporting. These traditional methods were, however, integrated with a wide use of modern information technology and data sharing. As a consequence, SARS dramatically demonstrates the global havoc that can be wreaked by a newly emerging infectious disease. At the same time, it dramatically highlighted the extreme importance of open data to fight against the risk of such sudden global epidemics.

An on-line database project for archaeological sites in Italy[9]

Awareness of the importance of open data is growing among archaeologists and several initiatives are underway (Kintigh, 2006; Beck & Neylon, 2012; Kansa & Kansa, 2013). Sharing data and information about prehistoric landscapes is very important for site management, research and conservation, particularly as prehistoric sites tend not to be highly visible, making their destruction more probable because of farming, infrastructures and housing. *PreBiblio,* a bibliographic and topographic database of prehistoric and paleontological sites of Italian prehistory, covering the period from 2 million years BP to the ninth century BC, began to be developed in 2005. *PreBiblio* is a relational database with 15,000 sites and 6,000 references, now on-line on the server of *DigiLab* of the Sapienza University of Rome at http://prebiblio.uniroma1.it/. In order to combine bibliographic information with a geographical information system (GIS), references are now being located on a topographic map.

As in any other long term enterprise in Italy, *PreBiblio* now faces two major obstacles: a) the lack of funding in the context of a general surrender of public control over the landscape and heritage; b) the lack of a common scientific language for sharing information in archaeological disciplines that is as fast, effective and useful as in genomics or biological sciences. The database does however provide a general cover of the distant past of human settlements in Italy, although there are three main obstacles for the dataset:

a) The dilemma of conservation. Officers of the Ministry of Culture think that the widespread availability of the precise locations of sites could be a threat to conservation. However, we need to assess the balance between the danger that maps would be used by site looters against the benefits they offer to conservation of prehistoric and paleontological heritage.
b) If popular access to science is a duty, do sites need to be open to the public? What of the large majority of "minor" sites?
c) The misuse of sites because of tourism. Many public administrators aim only to explore them in order to have a short term income, and refuse to cover the expenses of maintenance, which tends to be greater than the profits from tourism.

Data sharing is also crucial because of three factors that influence practices of landscape management: a) the speed and effectiveness of rescue when a site is being developed for public or private ventures; b) the irremovability of many such "archives"; c) that an archaeological site is a unique phenomenon: excavation being the controlled destruction of an archaeological deposit. It is also important to recognize the vital role of GIS, not only for the immediate conservation of "heritage", but also for the diffusion of knowledge regarding the history of our species.

[9] Lecture by **Fabio Parenti**, scriptoriumparentii@gmail.com.

The European Commission: towards publications and data sharing[10]

The EC recently promoted two initiatives aimed at ensuring that research results funded by EU citizen are made freely available to the population at large in order to increase the EU's return on research and development (R&D) investment. The European Research Council (ERC) published its Guidelines for Open Access in December 2007 (http://www.openaire.eu/en/component/attachments/download/3), as a follow up to the 2006 Statement on Open Access (OA). This mandate requires researchers to provide open access – within a specified time period, typically six months – to articles resulting from EC-funded research. In August 2008, the European Commission launched the Open Access Pilot in the Seventh Framework Programme (FP7, 2007-2013; http://www.openaire.eu/en/component/attachments/download/4.html), which will run until the end of the Framework Programme. The Pilot's goal is to monitor the impact of the ERC OA mandates by means of statistics of issues such as OA vs. non-OA peer-reviewed publications per project and FP7 program.

Emerging approaches are illustrated by the motivation and goals of the OpenAIRE infrastructure (Open Access Infrastructure for Research in Europe, http://www.openaire.eu), which is currently being implemented according to EC requirements. The infrastructure, which was funded by the European Commission as part of the OpenAIRE project (Dec 2009 - Nov 2012) and the OpenAIREPlus project (Dec 2011 - May 2014), delivers a *networking* infrastructure and a *technical* infrastructure (Manghi *et al.*, 2010). At networking level, the project operates a European Helpdesk System, comprising a European Centre and National Open Access Desk liaison offices (NOADs), which serves the EU in its entirety by engaging people and scientific repositories in almost all 27 member states (Rettberg & Schmidt, 2012). The NOADs liaise with other Open Access and repository-related activities in Europe (e.g., COAR, SPARC Europe, LIBER) and exploit their hierarchical organization in order to disseminate best practices efficiently using OpenAIRE guidelines on how to export and share data, initiatives, and events related to OA among local decision makers and research organizations.

At a technical level, the project operates a data infrastructure capable of providing a unique access point for European (and beyond) research outcomes by monitoring the Open Access trends in projects supported by the EC and national funding agencies (Manghi *et al.*, 2012). Figure 4 shows a graph obtained by combining data from scientific papers and data inferred using algorithms. In particular, it offers services for:

a) Collecting contents from publication repositories, dataset repositories, CRIS systems (metadata about projects, organizations, and people involved, e.g. CORDA for EC FP7 projects), and "entity registries" (directories of entities, e.g. OpenDOAR for Publication Repositories, re3data for Data Repositories).
b) Automatically identifying semantic relationships between publications, datasets, and projects.
c) The support of end-users (e.g. researchers, EC officers, National funding agencies officers, project coordinators) through an on-line portal which can be utilized to search, browse and check the statistics of research results. In addition, the portal allows users to make public the relationships between their publications or datasets and relative projects.

Finally, the project also offers the Zenodo repository (http://zenodo.org), which supports communities and/or researchers who do not have a repository of reference for depositing publications or datasets.

The example of human genomics

One of the first scientific areas in which data sharing became a significant issue, and in which

[10] Lecture by **Paolo Manghi,** paolo.manghi@isti.cnr.it.

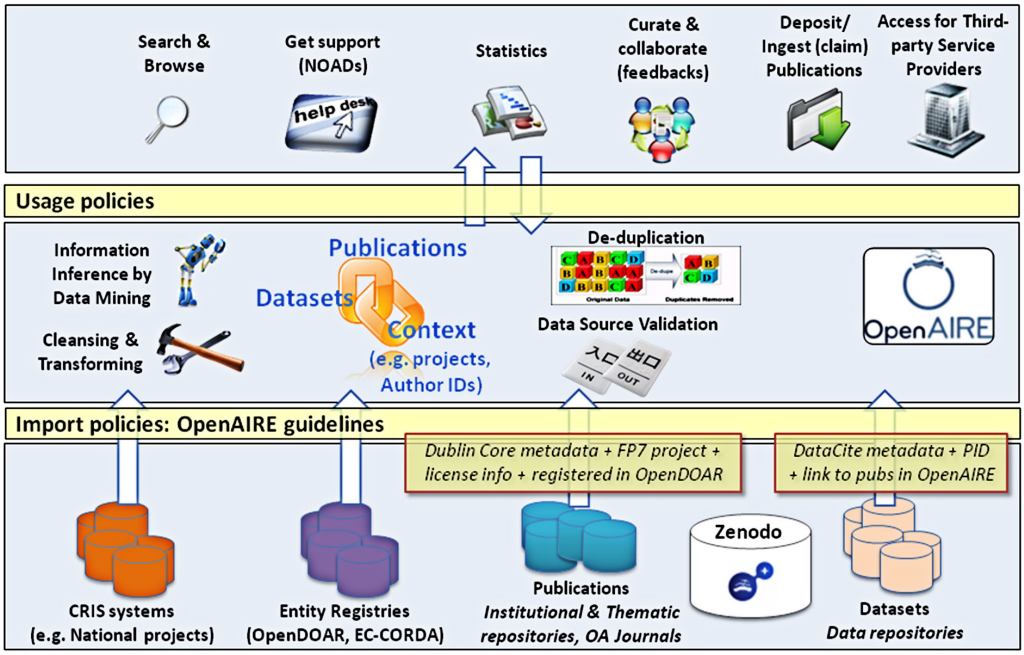

Fig. 4 - Functional areas of the OpenAIRE infrastructure. The colour version of this figure is available at the JASs web site.

large-scale data sharing among research groups was successfully implemented, is that of human genomics. However, the public release of human genomic data also raises issues relating to the protection and privacy of individual data, as well as to scientific priority and credit, both of which have given rise to counter-currents pushing against the broad release of this data. For these reasons, we dedicated a specific section to this topic.

The legacy of the Human Genome Project for a public human genome database[11]

The data sharing framework of the Human Genome Programme (HGP) grew out of earlier U.S. government policies requiring the release of scientific data generated from federally-funded research upon the publication of the results of this research (Contreras, 2011). However, because of the need for close coordination between multiple sites, a desire to enable the use of HGP data in scientific discovery as rapidly as possible, and a concern that increasing portions of the human genome were being patented by private parties, the organizers of the HGP met in Bermuda in 1996 to adopt a new data release policy. The resulting "Bermuda Principles" were radical in their scope and effect, requiring participating sequencing centers to release all genomic sequence data into public databases a mere 24 hours after generation. The rapid, pre-publication data release model established by the Bermuda Principles has prevailed in the field of genomic research and, through a series of subsequent refinements, has become the norm for genomic (and related "omics") fields (Contreras, 2010, 2011; Kaye, 2012a).

But while the rapid release of genomic data has been shown to have had positive effects on follow-on discovery and innovation (Williams, 2013) and to have reduced the patenting of

[11] Lecture by **Jorge Contreras**, contreras@wcl.american.edu.

DNA sequence data (Contreras, 2011), countervailing policy considerations have also emerged. The most prominent of these among the scientific community is a reluctance of researchers to reveal their data prior to having a chance to analyze it fully and then prepare publications based on it. This reluctance led to the adoption of the 2003 Fort Lauderdale Principles, which re-affirmed the commitment of the Bermuda Principles to rapid release of data whilst also recognizing the need for researchers to receive recognition for the discoveries made using data that they have generated (http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf). As a result, large-scale genomics policies have recently implemented rules that place embargoes on publication of results based on data publicly-released by a researcher for periods generally ranging from 6-12 months in order to give the data-generating researcher a "head start" in analysis of data and publishing the results of that analysis (Contreras, 2010). Another significant concern with the release of human genomic data centers on the potential identifiability of human DNA donors, and a loss of privacy associated with such genetic material (eg. see McGuire & Gibbs, 2006; Homer *et al.*, 2008; Gymrek *et al.*, 2013). Accordingly, current data usage policies for genomic data typically prohibit attempted re-identification of human subjects and use of data for non-research purposes (Kaye, 2012a). However, the legal and practical enforceability of these restrictions has not yet been tested.

The protection of privacy in genomic research[12]

One of the most difficult ethical challenges for open science relates to the privacy of human subjects and the confidentiality of data about them. The enormous scientific achievement of mapping the human genome is often attributed to the way the scientists on the Human Genome Project collaborated by sharing their data daily over the Internet. One of the wonderful things about open data is that it allows the reuse of data in order to answer an array of possible questions and in combining and integrating it with other data to reveal unexpected relationships. However, where human subjects are involved, the context in which data is collected has important implications regarding expectations about its future uses (Nissenbaum, 2010). Open science involves processes of sharing that cross traditional boundaries of use to allow faster, better analysis of larger datasets and may conflict with or be incompatible with traditional processes that protect privacy (Rule, 2009). There is a legal barrier between anonymised or non-identifiable data and personal data or data which identifies a named individual. This dichotomy is becoming increasingly unreliable as a means of protecting individual privacy in human genomics (Heeney *et al.*, 2010; Heeney, 2012) as a consequence of processes of data profiling (Brown *et al.*, 2011; Heeney, 2012). Whilst there have been increasing concerns and acknowledgement of the impact of profiling of publicly funded databases, there is a strong private market in profiling (House of Lords, 2009) which is part of the core business of some companies (Lyon, 2003). The proliferation of data sources as well as the tools for using data to categorize and monitor groups and individuals has taken us into an era of 'liquid surveillance' (Bauman & Lyon, 2012) whereby the power of surveillance is no longer concentrated in the hands of a few carefully controlled state actors.

A "data environment" (Elliot *et al.*, 2011) remains an excellent way in which to think about data release, as in any environment, new data will interact with what is already available. This poses great challenges for the protection of privacy within the current legal system and until this is addressed, moves towards open science and open circulation of data may serve to acerbate the problem. Although there are strong arguments for open science, a question hangs over those areas where personal data is an important part of the research process, and whether and under what circumstances, public benefit from new knowledge generated by scientific research outweighs consequent threats to personal privacy. It could be that personal human data should be treated separately from other types of data.

[12] Lecture by **Catherine Heeney**, catherine.heeney@cchs.csic.es.

Managing data in biobanks[13]

The importance of data sharing in Genomics is directly related to the increase in statistical power inherent in many larger and diverse sets of data. The call for widespread data-sharing led to the proposal of the Global Alliance (http://news.sciencemag.org/people-events/2013/06/qa-david-altshuler-how-share-millions-human-genomes) for responsible sharing of genomic and clinical data and in specific projects funded by the EC. These include the "biobank cloud", or "Rare Diseases-Connect" that aim to build platforms for storage, analysis and sharing of digital genomic data (http://rd-connect.eu/). This top-down approach seems not to take into account the difficult position of biobank, registry and bio-repository directors or managers who are asked to share data collected at a high cost to the institution, and the loss of control over this data by the "donor/patient". In practice, sharing tends to occur in a more controlled fashion, using detailed collaboration contracts (data transfer agreements - DTAs) that regulate how data can be used and the limits of their exploitation by partners. Although the ideal of sharing for the benefit of science remains an important principle for scientists, institutions are more reticent, fearing that their work will be exploited by others without proper recognition (Shamoo & Resnik, 2009).

A donor or patient that is asked to give up all control over their data is implicitly asked to trust the institution that receives their data and take all decisions on data use without recourse to the data subject. Although new technologies offer opportunities for greater involvement, participants are often not allowed to play a decisive role in managing their own data (Kaye, 2012b). The normal application of the principle of "informed consent" gives little choice or control to the participant, and little or no choice about whether she/he approves the sharing of their data with national and/or international partners. It gives freedom to share data by the hosting institution, but risks losing participants who want to maintain control of their data. In contrast, electronic "dynamic consent"makes it possible to host tiered consent on a personal WEB account. For example, the CHRIS (Cooperative Health Research In South Tyrol) Project, a prospective epidemiological research study carried out by the European Academy (EURAC, Bolzano, Italy) offered an opportunity to test the electronic Consent Tool in approximately 4,000 individuals. Before compiling their details in the database, participants are extensively informed through old (brochure) and new technologies (informative movie) and may interact with a nurse. Ongoing information through annual newsletters and occasional e-mails ensure participants are updated about possible changes. By providing their consent electronically, they can decide on major issues regarding whether and how widely the research data can be shared with other institutions or inserted into public databases such as the American DBGap. Furthermore, they can decide whether or not to leave their data for research in case of death or mental incapacity and ask to be re-contacted in case of incidental findings. Participants keep control over time of their choices and are entitled access to their consent options via WEB and modify some of them should the original conditions change. This approach appears to reassure participants about the threat to privacy, so that they proved to be very open to data sharing. Comprehensive information about research conduct and the option of taking decisions about the use of their own data persuaded 97% (4000) of donors to allow their data to be made available in public databases (Mascalzoni, 2013). During qualitative interviews, they declared that the data management scheme adopted by CHRIS fostered their trust in the research institution and helped them better understand how science worked. The CHRIS project participants showed how patients are prepared to trust the scientific community on condition that they receive appropriate information and believe that their rights are respected. Biobanks should consider adopting such procedures.

[13] Summary by **Deborah Mascalzoni**, deborah.mascalzoni@crb.uu.se.

Opening science to society

A round table[14] discussed the potential role of open data practices in the relationship between science and society, with particular attention paid to the role played by science communicators. Traditionally, professional science communicators have acted as gatekeepers between the scientific production process and the public. Today, intermediators able to look at research data in depth are becoming more important due to the growing complexity of research practices. At the same time, the general role of science communicators is being challenged. In fact, the increasing diffusion of scientific blogs and social networks makes it possible for many other voices to express themselves in the public arena. This trend, which accelerates with each novel way of exploiting the borderless communication potential of the Internet, is one that poses challenges and opportunities for scientific institutions, newspapers and professional communicators and for the development of data-sharing practices. It is also beginning to broaden the institutional and professional community involved in science communication and dialogue to public administrators, educators, non-specialized journalists, commercial companies (e.g. food and pharmaceutical) which may use or claim to use scientific information to attract the interest of the public.

The Internet is making an unprecedented wealth of information accessible to public scrutiny, including cutting-edge research results. But it also fails to discriminate between the rigorous, the tendentious and the fallacious. The difficulty of discrimination is increasing through the plethora of on-line journals which offer opportunities for publication whilst not always offering high standards of editorial and reviewing rigour (Bohannon, 2013). Whilst it is possible that this increasing volume of inadequately scrutinized work could undermine the public's faith in the credibility of science, discourse in the public domain has always been uncontrolled and contentious compared with discourse within the scientific community that has hitherto been controlled and relatively restrained. But we cannot role back the clock. The open, instantaneous communication enabled by the Internet and the web, coupled with a more democratic spirit where citizens wish to explore the implications of science that are of interest or concern to them rather than simply accepting the pronouncements of scientists, are producing a new world of discourse and dialogue, of information and misinformation, to which the scientific community must adapt rather than hide from.

It is important therefore that there is an interdisciplinary debate on Open Science and how its development can integrate important issues of data production, science communication, education and more effective public engagement with science.

Concluding comments

Participants were asked to identify three points which should be taken into account when planning further interdisciplinary initiatives on open science, and from which the following issues arose.

The achievement of greater, more widespread and effective openness in science requires the synergic action of various players, including researchers, those who employ them, those who fund them, those who publish their work and governments. Governments as the ultimate funder of science have a major role to play. For example, in the USA, a recent government initiative insists on the public release of all data generated by U.S. federally-funded research (Office of Management and Budget, 2013), whilst the U.S. National Institutes of Health (NIH) has initiated the development of a uniform and searchable data catalogue for NIH-funded biomedical data sets (Kuehn, 2013) It is important however that national strategies are not so top-down that they suppress bottom-up creativity, which is usually a source of novel solutions for hitherto intractable problems. Such bottom-up efforts are important for example in: adapting open data approaches to the features of specific research fields; refining policies on the basis of their effectiveness;

[14] Coordinated by **Silvia Bencivelli**, sbencivu@gmail.com.

identifying optimal incentives for data sharing in specific research fields; involving "small science" more in the open data movement.

Moving on to the societal implications of Open Science, an important message is that we need to devise strategies which may more effectively involve the public in the discourse about open data. There are well-documented examples of the advantages of open data (Boulton, 2012). However, the experience of some participants suggested that these may lose efficacy when used for dissemination to the public. Nonetheless, the case studies about the control of emerging viruses and, to some extent, that regarding management of information about prehistoric sites in Italy, seem to exemplify the idea that that opening up research data may have a positive, rapid and easily recognizable impact on our daily lives. This could be an effective argument also for a non-specialized audience.

Several speakers pointed to the importance of making students and researchers at the beginning of their careers more aware of the value and benefits of Open Science. This idea is not new. The value of unrestricted access to experimental data for the more effective training of young people, as well as the negative effect of exposure to data withholding on their future sharing behavior have been already discussed (Vogeli *et al.*, 2006; Feldman, 2012; Barr & Onnela, 2012; Tenopir *et al.*, 2011). However, there are ways to look at the relations between Open Science and scientific education which seem worth exploring in more detail. Introducing arguments from social sciences and humanities in the educational dissemination of open data may have a double benefit: making students more profoundly engaged with Open Science and helping them look at science from a broader perspective. Two examples are suggested. Firstly, reviewing ethnographic studies of research groups may be useful to understand how and why openness varies across research domains. In a recent work, Velden (2013) has shown that researchers in synthetic chemistry are less willing to share their data than those in experimental physics, a difference which seems to be related to the prevalence of a more individualistic research culture in the former and a team-oriented research culture in the latter. This finding may be a starting point for a discussion of the relationship between the intellectual and social environments of research and data sharing behaviour. Secondly, looking at the dichotomy between sharing and withholding data from a philosophical point of view may be a means to bring epistemological aspects into the discussion about Open Science. In a recently published commentary (Boniolo & Vaccari, 2012) it was advocated that "Science should be available for evaluation by other scientists and for public scrutiny, just as it has been since Galileo's time" and concluded that withholding scientific data may be an "epistemological suicide" dictated by "vested interests or a creeping loss of awareness of the theory of knowledge". Considering such a radical viewpoint provides an opportunity to debate with students whether classical ethical principles of Science (Merton, 1942) are still suitable for the current processes of scientific data production or whether they should be revisited in the light of recent issues, such as the protection of personal data in bio-medical studies (Kaye, 2012a) and respect of confidentiality in social studies (Bishop, 2009).

In conclusion, it is to be hoped that the overview the meeting on "Scientific data sharing: an interdisciplinary workshop" convinces readers that merging experiences from biology, genomics, psychology and archaeology is a worthwhile effort. In general, it was felt that fostering interdisciplinary dialogue and synthesising apparently different perspectives could lead to new and promising avenues of a more Open science.

Acknowledgements

The meeting "Scientific data sharing: an interdisciplinary workshop" was supported by the Ministero dei Beni e delle Attività Culturali e del Turismo (Direzione generale per le Biblioteche, gli Istituti Culturali ed il Diritto d'autore, DGBID). We would like to thank the coordinators and the staff of "Science on the net" for their help with the dissemination of the initiative.

Info on the web

http://opendefinition.org/
Website of the Open Knowledge Definition containing the full text of the OKD.

http://www.scienceonthenet.eu/en
The portal dedicated to Italian research in Europe and in the world, with a section dedicated to Open Access.

http://www.sciencewise-erc.org.uk/cms/public-dialogue-on-data-openness-data-re-use-and-data-management/
Website for a UK based public dialogue on data openness, reuse and management in science.

https://sites.google.com/site/openingsciencetosociety/
ISItA web site dedicated to open data (Destro Bisol et al., 2013).

https://sites.google.com/site/scientificdatasharing/Presentations.
Slides and videos of lectures presented during the meeting are available at this address.

http://www.pijip.org/scientific-data-release-and-genome-commons/
Site hosted at American University, Washington, DC, USA, dedicated to Scientific Data Release and Genome Commons.

Authors contribution

GDB organized the structure of the paper and wrote the general parts; GB revised and edited extensively the whole manuscript; GDB, GB, PA, MC, SB, AC, JC, NE, BF, CH, DL, PM, DM, JCM, FP, JMW and GB wrote the summary of their presentations; PG helped with the dissemination of the initiative. All authors read, provided feedback and approved the manuscript.

References

Alsheikh-Ali A., Qureshi W., Al-Mallah M.H. & Ioannidis J.P.A. 2011. Public availability of published research data in high-impact journals. *Plos One*, 6: e24357.

Anagnostou P., Capocasa M., Milia N. & Destro-Bisol G. 2013. Research data sharing: Lessons from forensic genetics. *For. Sci. Int. Genet.*, 7: e117-e119.

Bakker M. & Wicherts J.M. 2011. The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods*, 43: 666-678.

Barr C.D. & Onnela J.P. 2012. Establishing a culture of reproducibility and openness in medical research with an Emphasis on the training years. *Chance*, 25: 8-10.

Bauman Z. & Lyon D. 2013. *Liquid Surveillance: A conversation*. Wiley, Oxford.

Beck A. & Neylon C. 2012. A vision for Open Archaeology. *World Archaeol.*, 44: 479-497.

Bishop L. 2009. Ethical sharing and reuse of qualitative data. *Aust. J. Soc. Issues*, 44: 255–272.

Bohannon J. 2013. Who's Afraid of Peer Review? *Science*, 342: 60-65.

Boniolo G. & Vaccari T. 2012. Publishing: alarming shift away from sharing results. *Nature*, 488: 157.

Boulton G. 2012. Open your minds and share your results. *Nature*, 486: 441.

Boulton G., Campbell P., Collins B., Elias P., Hall W., Laurie G., O'Neill O., Rawlins M., Thornton J., Vallance P. & Walport W. 2012. *Science as an open enterprise*. The Royal Society, London.

Brown I., Brown L. & Korff D. 2011. The limits of anonymisation in NHS data systems. *Brit. Med. J.*, 342: d973.

Cerroni A. 2006. *Scienza e società della conoscenza*. Utet, Torino.

Cerroni A. 2007. Individuals, knowledge and governance in the 21st century society. *Journal of Science Communication*, 6: 1-9.

Collins H. 2010. *Tacit and explicit knowledge*. Chicago University Press, Chicago.

Congiu A., Anagnostou P., Milia N., Capocasa M., Montinaro F. & Destro Bisol G. 2012. Online databases for mtDNA and Y chromosome polymorphisms in human populations. *J. Anthropol. Sci.*, 90: 201-215.

Contreras J.L. 2010. Prepublication Data Release, Latency, and Genome Commons. *Science*, 329: 393.

Contreras J.L. 2011. Bermuda's Legacy: Patents, Policy and the Design of the Genome Commons. *Minn. J.L. Sci. & Tech.*, 12: 61.

Dalgleish R., Molero E., Kidd R., Jansen M., Past D., Robl A., Mons B., Diaz C., Mons A. & Brookes A.J. 2012. Solving bottlenecks in data sharing in the life sciences. *Hum. Mutat.*, 33: 1494-1496.

Descartes R. 1637. *Discours de la méthode pour bien conduire sa raison, et chercher la verité dans les sciences. Plus la Dioptrique. Les Meteores. Et la Geometrie. Qui sont des essais de cette Methode.* De l'Imprimerie de Ian Maire, Leyde.

Destro Bisol G., Capocasa M., Anagnostou P. & Greco P. 2013. Opening Science to Society, a new initiative of the Istituto Italiano di Antropologia. *J. Anthropol. Sci.*, 91: 233-235.

Elias N. 1991. *The society of individuals*. Blackwell, Oxford.

Elliot M., Lomax S., Mackey E. & Purdam K. 2011. Data environment analysis and the key variable mapping system. *Lect. Not. Comput. Sc.*, 6344: 138–147.

Enke N., Thessen A., Bach K., Bendix J., Seeger B., Gemeinholzer B. 2012. The user's view on biodiversity data sharing - Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, 11: 25-33.

Fecher B. & Friesike S. 2013. *Open Science: One Term, Five Schools of Thought*. Proceedings of The 1st International Conference on Internet Science, Brussels.

Feldman L., Patel D., Ortmann L., Robinson K. & Popovic T. Educating for the future: another important benefit of data sharing. *Lancet*, 379: 1877-1878.

Foray D. 2004. *The economics of knowledge*. MIT Press, Boston.

Freese J. 2007. Replication standards for quantitative social science. *Sociol. Method. Res.*, 36: 153-172.

Gymrek M., McGuire A.L., Golan D., Halperin E. & Erlich Y. 2013. Identifying personal genomes by surname inference. *Science*, 339: 321-324.

Heeney C. 2012. Breaching the contract? Privacy and the UK Census. *J. Inform. Soc.*, 28: 316-328.

Heeney C., Hawkins N., de Vries J., Boddington P. & Kaye J. 2010. Assessing the privacy risks of data sharing in genomics. *Public Health Genomics*, 14: 17-25.

Hendriks P. 1999. Why share knowledge? The influence of ICT on the motivation for knowledge sharing. *Knowledge and Process Management*, 6: 91–100.

Homer N., Szelinger S., Redman M., Duggan D., Tembe W., Muehling J., Pearson J.V., Stephan D.A., Nelson S.F. & Craig D.W. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *Plos Genet.*, 4: e1000167.

House of Lords Select Committee on the Constitution. 2009. *Surveillance: Citizens and the state, Volume I: Report*. The Stationery Office Limited, London.

Kaye J. 2012a. The tension between data sharing and the protection of privacy in genomics research. *Annu. Rev. Genomics Hum. Genet.*, 13: 415-431.

Kaye J. 2012b. From patients to partners: participant-centric initiatives in biomedical research. *Nat. Rev. Genet.,* 13: 371-376.

Kintigh K. 2006. The promise and challenge of archaeological data integration. *Am. Antiquity*, 71: 567-578.

Kansa E.C. & Kansa S.W. 2013. Additional Thoughts on Sustaining and Promoting Open Data in Archaeology. *Journal of Eastern Mediterranean Archaeology and Heritage Studies*, *1*: 102-103.

Kuehn B.M. 2013. NIH Recruits Centers to Lead Effort to Leverage "Big Data". *JAMA*, 310: 787-787.

Luzi D., Ruggieri R., Biagioni S. & Schiano E. 2013. Data sharing in environmental sciences: A survey of CNR researchers. *International Journal of Grey Literature*, 9: 69-81.

Lyon D. 2003. Surveillance as social sorting: Computer codes and mobile bodies. In D. Lyon (ed): *Surveillance as social sorting. Privacy, risk, and digital discrimination*, pp. 13–30. Routledge, London.

Manghi P., Bolikowski L., Manold N., Schirrwagen J. & Smith T. 2012. Openaireplus: the european scholarly communication data infrastructure. *D-Lib Magazine*, 18: 1.

Manghi P., Manola N., Horstmann W. & Peters D. 2010. *An infrastructure for managing EC funded research output.* The OpenAIRE Project.

Mascalzoni D. 2013. *Poster Informed consent: A Challenge Won by Trust* 8. 12-14 June 2013, Geneva, Brocher Foundation, Exploring innovative mechanisms to build trust in human health research biobanking.

Mauthner N.S. & Parry O. 2013. Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology*, 27: 47-63.

McGuire A.L. & Gibbs R.A. 2006. No longer de-identified. *Science*, 312: 370-371.

Merton R.K. 1942. Science and technology in a democratic order. *Journal of Legal and Political Sociology*, 1: 115-126.

Milia M., Congiu A., Anagnostou P., Montinaro F., Capocasa M., Sanna E. & Destro-Bisol G. 2012. Mine, yours, ours? Sharing data on human genetic variation. *Plos One*, 7: e37552.

Molloy J.C. 2011. The Open Knowledge Foundation: Open Data Means Better Science. *Plos Biol.*, 9: e1001195.

Morrison J.B., Pirolli P. & Card S.K. 2001. A taxonomic analysis of what world wide web activities significantly impact people's decisions and actions. *Proceedings CHI 2001*, 163-164.

Murray-Rust P., Neylon C., Pollock R. & Wilbanks J. 2010. *Panton Principles, Principles for open data in science.* Retrieved from http://pantonprinciples.org/.

Neylon C. & Wu S. 2009. Open science: tools, approaches, and implications. *Pacific Symposium on Biocomputing*, 14: 540-544.

Nissenbaum H. 2010. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, Stanford.

Office of Management and Budget. 2013. *Open Data Policy - Managing Information as an Asset*. URL: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

PARSE. Insight 2009. *Insight into issues of Permanent Access to the Records of Science in Europe*. Deliverable 3.4 Survey Report. URL: http://www.parse-insight.eu/publications.php.

Rettberg N. & Schmidt B. 2012. Repository communities in OpenAIRE: Experiences in building up an Open Access Infrastructure for European research. *Open Repositories 2012.*

Rota P.A., Oberste M.S., Monroe S.S., Nix W.A., Campagnoli R., Icenogle J.P., Peñaranda S., Bankamp B., Maher K., Chen M.H., Tong S., Tamin A., Lowe L., Frace M., DeRisi J.L., Chen Q., Wang D., Erdman D.D., Peret T.C., Burns C., Ksiazek T.G., Rollin P.E., Sanchez A., Liffick S., Holloway B., Limor J., McCaustland K., Olsen-Rasmussen M., Fouchier R., Günther S., Osterhaus A.D., Drosten C., Pallansch M.A., Anderson L.J. & Bellini W.J. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300: 1394-1399.

Rule J.B. 2009. The limits of privacy protection. In J. Goold & D. Neyland (eds): *New Directions in* Surveillance and Privacy, pp. 3-17. Willan Publishing, Collumpton.

Savage C.J. & Vickers A.J. 2009. Empirical study of data sharing by authors. *Plos One*, 4: e7078.

Schofield P.N., Bubela T., Weaver T., Portilla L., Brown S.D., Hancock J.M., Einhorn D., Tocchini-Valentini G., Hrabe de Angelis M., Rosenthal N. & CASIMIR Rome Meeting

participants. 2009. Post-publication sharing of data and tools. *Nature*, 461: 171-173.

Shamoo A.E. & Resnik D.B. 2009. *Responsible conduct of research 2nd ed*. Oxford University Press, New York.

Stehr N. 1994. *Knowledge Societies*. Sage, London.

Tenopir C., Allard S., Douglass K., Aydinoglu A.U., Wu L., Read E., Manoff M. & Frame M. 2011. Data sharing by scientists: practices and perceptions. *Plos One*, 6: e21101.

Velden T. 2013. Explaining field differences in openness and sharing in scientific communities. *Proceedings of the 2013 conference on Computer supported cooperative work*, 445-458.

Vines T.H., Albert A.Y., Andrew R.L., Débarre F., Bock D.G., Franklin M.T., Gilbert K.J., Moore J.S., Renaut S. & Rennison D.J. 2014. The availability of research data declines rapidly with article age. *Curr. Biol.*, 24: 94-97.

Vogeli C., Yucel R., Bendavid E., Jones L.M., Anderson M.S., Louis K.S. & Campbell E.G. 2006. Data withholding and the next generation of scientists: results of a national survey. *Acad. Med.*, 81: 128–136.

Wicherts J.M., Borsboom D., Kats J. & Molenaar D., 2006. The poor availability of psychological research data for reanalysis. *Am. Psychol.*, 61: 726-728.

Wicherts J.M. 2011. Psychology must learn a lesson from fraud case. *Nature*, 480: 7.

Wicherts J.M., Bakker M. & Molenaar D. 2011. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *Plos One*, 6: e26828.

Williams, H.L. 2013. Intellectual Property Rights and Innovation: Evidence from the Human Genome. *Journal of Political Economy* 121:1.

Associate Editor, Rita Vargiu