Beyond Sharing and Re-using:

Toward Global Data Networking

With ideas provided by (in alphabetical order)

Fred Friend, Jean-Claude Guédon, and Herbert Van de Sompel

A. General purpose of this text

1. How data can be networked;
2. How to envision and set up data governance on a global scale;
3. How the European Union can play a **leading** role in helping start and steer this global trend. Clearly, such a process will work only if the majority of the major players are positively engaged in it; at the same time, Europe may well shape the process in favourable ways by taking some well-defined and judicious initiatives.

What follows has been largely influenced and inspired by the history of the Internet, and the development of the Internet Engineering Task Force (along with the Internet Architecture Board). In particular, the ability to draw ideas from any researcher through an iterative process based on "RFC's" (Requests for Comments) appears eminently wise, all the more so that it transposes the wisdom of science publishing itself. *This document, therefore, offers a vision for the future of data curation, preservation and sharing, and it is based on the belief, also inspired by the Internet, that the* ***best top-down approach is one that catalyses and sustains a bottom-up approach****.*

Some lessons from the IETF

1. It was based on a clear network project : it actually succeeded to a more generic and general structure called "Gateway algorithms and data structures" (GADS), but split from the purely research concerns of some of the GADS members to focus on the messy, practical issues of deployment and implementation;
2. It started by invitation, but once it found its pace and a culture began to grow out of these meetings, it opened up to anyone;
3. It is known to obey brutal frankness and a series of common-sense rules that have made it ruthlessly efficient. One of these rules is the famous (and highly debated) "implementation precedes standardization". It is this last rule that allowed the Internet bundle of protocols to pull ahead of the OSI model that the International Standards Organization (ISO) was proposing at about the same time;
4. Its basic working mode is to describe an implementation proposal in a paper called an Internet Draft that is then submitted to the scrutiny of the interested experts[1]. These Internet Drafts go through various iterations to eventually become Request for Comments

[1] See http://en.wikipedia.org/wiki/Request_for_Comments for much fuller details about RFC's.

(RFC). Each RFC is assigned a status of Informational, Best Current Practice, Experimental, Standards Track (with various sub-types) or Historic. One should note that RFC's actually play the role of published papers: the engineering community as a whole acts as a steward for the Internet Drafts and RFC's.

B. The case of data: some preliminary remarks

Means are needed to monitor and survey the growth of data on a global scale. Tools to classify, catalogue and preserve data are also needed if data is to be useful beyond the sites where they are produced.

1. The quantity of data being produced grows very quickly;
2. Data diversity and complexity is also growing very fast;

In parallel, the nature of scientific research is evolving. This has been diversely described, but we will use here the concept of "mode 2" of knowledge production found in Michael Gibbons et al.'s vocabulary[2]

1. The need to address ever more complex problems increases the need for interdisciplinary work;
2. The most significant innovations tend to appear either at the boundaries of two or more disciplines, or by importing concepts from one discipline into another.

As a consequence of what precedes, data must be recognizable across disciplines and specialities. This means that a common (metadata) framework for data identification, description, and discovery needs to emerge, and it also means that some basic degree of data level interoperability must be obtained.

Interoperability also provides the opportunity to apply Robert K. Merton's "norms of science"[3] – particularly "communalism", "universalism" and "organized skepticism" – to research data as well as to text. Technical interoperability fits well with openness to the sharing and re-use of research data that is emerging within the research community.

The new data situation and the new ways in which knowledge is being produced lead to the following requirements:

1. Monitoring the growth of data and their provenance;
2. Monitoring the formats in which data are produced;
3. Surveying the tools to ensure interoperability of data, and their use across disciplines: curation;

[2] Gibbons, Michael; Camille Limoges, Helga Nowotny, Simon Schwartzman, Peter Scott, & Martin Trow (1994). *The New Production of Knowledge: the Dynamics of Science and Research in Contemporary Societies*. London: Sage.

[3] Merton, R.K. (1942) *The Normative Structure of Science* In: Merton, Robert King (1973) *The Sociology of Science: Theoretical and Empirical Investigations.* Chicago: University of Chicago Press. ISBN 9780226520919.

4. Classifying kinds of data: metadata formats
5. Creating harvestable, distributed databases of metadata to support the creation of cross-database services, e.g. discovery;
6. Tackling the data preservation challenge

Irrespective of the rights or sensitivity constraints that may apply to data, some functionalities of the interoperable data infrastructure should still apply. For example, identification, discovery should work irrespective of such constraints.

C. The present data landscape

1. It is presently highly fragmented, by disciplines or by domains (oceanography, health, agriculture, space, climate, ecology, etc.);
2. A variety of institutions, some national, some international, strive to deal with some aspects of data, but no site exists where some degree of coherence is achieved or even sought.

The present data situation corresponds roughly to one where various national and international bodies would be trying to develop various network standards without much attention to other networks, or even as a way to fend off competition from other networks. This looks very much like the situation that prevailed in the videotex phase of computer networks (Minitel - France, Prestel - UK, BTX – Germany , Captain – Japan);

3. The stewardship of data remains uncertain: beyond the scientist(s) or laboratories that produce data, no specific group or profession (such as librarians for publications) is in charge of the preservation and organization of data. This makes data very vulnerable whenever institutional shifts occur (without mentioning broader concerns such as political or economic instability);
4. Some domains are experiencing exponential forms of growth with doubling rates that can be as short as a few months (seven months in the case of second generation sequencing of genes), while others plan new instruments that will suddenly produce enormous amounts of data (this is sometimes referred to as a “data tsunami”).

The present situation of data also generates a series of consequences that may deeply and adversely affect the very evolution of science:

1. The quality of much scientific work would increase if the associated data were exposed to the scrutiny of more than one team (free software wisdom, through the voice of Eric Raymond, contends that **“with many eyeballs, all bugs are shallow”**). Presently, the replication of complex and costly experiments is so difficult as to be very rare. What this means for the evolution of scientific knowledge is rarely discussed although it subverts the very methodology underpinning scientific epistemology;
2. The quality of scientific education would increase if students were to establish a first-hand relationship with data and thus come to understand that science is not a question of truth and belief, but rather one of trust and method.

D. **The suggested approach**

1. Given what precedes, a top-down approach to achieve global sharing of data is not likely to succeed; if perchance it should, it will initially meet with a great deal of resistance and this will delay the whole process, especially in the absence of a global governance structure with real power of implementation;

2. Because the present situation reveals a complex landscape of fragmented and disparate data sets, working toward an ever denser set of linkages between them appears best. Such a set of linkages will grow best if they grow from actual needs, in particular those of various research communities. At the same time, research communities should easily find their way to sets of evolving guidelines that would ensure a growing degree of coherence across all boundaries.

3. The experience of the Internet and its evolution suggests that an approach based on layered thinking is best. This source of inspiration is all the more compelling that, like the Internet, data is part of a vast communication system, that of science, and that it is presently expressed and processed in the form of digital bits.

4. The experience of the Internet also suggests that no grand Cartesian architecture will prevail in such an environment. The data landscape, very much like the network landscapes at the end of the sixties and the early seventies, is much too fluid and much beyond the full control of any one entity to fit in any Procrustean bed anyone would want to design for it. It is better to adopt an iterative approach based on modest gains that gradually add up. Some working principles of the Internet deserve being recalled in this context: **rough consensus and working code are to be preferred to formal negotiations**; **releasing code soon and often is preferable to trying to come up with a fully worked out solution from the outset**; **favouring implementation over standardization in order to achieve the latter has already been mentioned;**

5. The impetus must come from the research communities themselves. Consequently, the best starting point for the European Community is:

   a. To begin work with some of the research communities they directly support and use this lever to nudge them into the desired direction. In effect, they would be the seed groups that emerged with GATS before the IETF in the case of the Internet.

   b. Relevant interlocutors from other regions or countries would be invited to do the same.

   c. Small, by invitation only, workshops in various countries and regions (e.g. The E.C.) would ignite the process aiming at testing the various possibilities available and begin working towards **networking of data** across communities.

   d. Each workshop would identify and prioritize work items and discuss essential characteristics of possible technical solution to address them, always keeping in mind the relevant prior work in the problem domain. This workshop-level consensus should form the basis for compiling Draft Specifications that address the challenge posed by each work item. These Draft Specifications can then be shared, commented upon, form the basis for consecutive Reference Implementations, and eventually be published as Data Interoperability RFCs on a suitable web site. The humanities and the social sciences should not be forgotten at this stage. Neither should the library profession and computer specialists. But the objective of **networking data** should be the dominant metaphor guiding the creativity of these groups.

e. A broader meeting, ideally convened by the E. C., would then work toward making these RFC's cohere, which would allow for test implementations that tackle multiple work item challenges at once.

f. The above process could be repeated, with new challenges being tackled in each cycle. Eventually, the suite of Data Interoperability RFCs would provide guidance on an increasing amount of data networking challenges.

Both the Draft Specifications and the Data Interoperability RFCs should be public in order to maximize the impact of the effort:

a. It would obviously publicize the data effort in the wider world and begin the process of globalizing **data networking**;

b. Interested parties that were not initially involved would have a chance to manifest themselves. As a result, the process of growing a community concerned with this particular layer of activity would begin;

c. It would obviously accelerate the evolution of RFC's

6. The stewardship of the RFC's could be attributed to some individual(s) that would work with the resources of a well-established European institution. OpenAIRE could be such an institution;
7. An “evangelical” phase could follow to broaden the interest and make the effort ever more visible and relevant to ever more communities, thus ensuring the coordinating role of the RFC stewards.
8. Once the “evangelical” phase is strongly underway, an open, very visible, and even prestigious conference should be organized, in effect to launch the movement publicly.
9. By then, the movement in favour of good data networking would be well on its way.

October 2011