Databases, Biological Information and Collective Action

Tom Dedeurwaerdere
FNRS, Centre for the Philosophy of Law, University of Louvain (Belgium)
Dedeurwaerdere@cpdr.ucl.ac.be

Abstract. Developments within bioinformatics and software for data exchange in the life sciences raise important new questions for social informatics. In this paper, I analyse the role of property rights in information in directing these technological developments in the direction of certain social values. In particular, I focus on initiatives for networking distributed databases, operating both on a global scale (such as the Global Biodiversity Information Facility) and in more single-issue networks (such as the European Human Frozen Tumour Tissue Bank). Three institutional models for developing such distributed networks for sharing information are presented and briefly discussed.

Keywords: information sharing, social informatics, bioinformatics, database governance, knowledge commons.

1 Introduction

As scientists and user groups become better connected with each other (particularly through the Internet), and as research focuses on issues of global importance (such as climate change, human health and biodiversity) there is a growing need to systematically address data access and sharing issues beyond national jurisdictions and thereby create greater value from international cooperation. The goal should be to ensure that both researchers and the broader public receive the optimum return on public investment, and to build on the value chain of investment in research and research data (Stiglitz *et al.*, 2000).

Integrated and combined access to this multifaceted realm of information opens perspectives for the implementation of new applications. In the field of life sciences, new sets of tools for studying biological building blocks and pathways will lay the foundations for ever more complex future projects. These may include the complete mapping of an organism's protein and metabolism networks, as well as the creation of

biological models that can pave the way for theoretical models on bacterial speciation and its complex ecological dynamics (Gevers *et al.*, submitted), or the development of tools for automated species identification. These tools undoubtedly require access to sets of skills that are not typically encountered among systematists or within the departments and institutions in which the bulk of formal taxonomic identifications are conducted. Developing solid approaches requires new collaborations between microbiologists, engineers, mathematicians, computer scientists and people who have significant knowledge of the legal and socio-economic aspects of sharing biological resources and software tools in the public domain.

These new applications of information technologies within the life sciences raise important questions related to the social embedding of information technologies. i.e. for 'social informatics' (Kling, 1996). Indeed technical choices within the field of bioinformatics also depend on social choices, whether in problems such as the building of genomic sequence databases, the design of persistent numerical identifiers for taxonomic information on living organisms, or the integration of clinical data and images coming from brain research. These technological developments reflect social choices on issues such as the protection of privacy, ownership of life, and bioethics. Moreover, the capacity to make these choices depends increasingly on clarifying the property rights to the information, which define who has the right to decide upon the way it is used, managed and exchanged. Open access to the information and shared ownership has become a key condition for connecting the path of development of information technologies to social values and ethical reflection.

Within the life sciences, initiatives for sharing information through networking distributed databases have emerged, operating both on a global scale (such as the consortium for Common Access to Biological Resources and Information (CABRI), connecting world wide microbiological resources) and in more focused networks (such as the European Human Frozen Tumour Tissue Databank (TuBaFrost)). From a governance perspective, these networks face increasing pressure from the development of global markets. In particular, the introduction of new standards of intellectual property protection during the last twenty years has had a profound impact on the sharing of data and resources in the field of the life sciences. Two of the most influential and widely debated changes in this context are the 1980 Bayh–Dole Act in the US (Rai and Eisenberg, 2003) and, more recently, the 1996 EU Database Directive 96/9/EC (Reichman and Uhlir, 1999). The Bayh–Dole act explicitly gives universities the right to seek patent protection on the results of government-sponsored research and to retain patent ownership. As a consequence, in the period from 1980 to 1992, the number of patents granted per year to universities in the US increased from fewer than 250 to almost 2700 (Rai, 1999). The EC Database Directive 96/9/EC was a landmark decision that lowered the standards of eligibility to database protection. Indeed the Database Directive offered copyright protection to databases that were original in the selection or the arrangement of their contents, but also to non-original databases if it could be shown that there had been a substantial investment in either the obtaining, the verification or the presentation of their contents. This extended protection to library catalogues, for instance, but also to biological information facilities that network existing databases.

In this paper, I will analyse the models for the institutional design of information sharing in the context of global intellectual property rights. In particular, I will rely on contemporary insights from new institutional economics that show the necessity of

collective action to deal with both the insufficiencies of market solutions and the limits of the new forms of public regulation (Reichman and Uhlir, 2003; Hess and Ostrom, 2003; 2005a). For instance, within the related field of digital communication, the development of E-print repositories (such as arXiv.org and BioMedCentral) and trusted digital repositories for knowledge of general interest is based on collaboration between groups of scholars and information specialists to build a common knowledge pool. What is new in these initiatives is that researchers are participating in an international epistemic community that is committed to building a global scholarly library – with the aim of obtaining greater joint benefits and reducing their joint harm from the enclosure process. I will build upon these proposals to elaborate a framework for the analysis of institutional choice in the field of the microbiological information commons.

For these reasons I will focus on the following questions:

(1) What are the characteristics of biological information, as a public good that can be exchanged by networking databases, and what are the related incentive problems for the provision and use of this good (Section 2.1)?
(2) What institutional solutions for dealing with these complex incentive problems are currently being proposed (Section 2.2)?
(3) How can we evaluate these propositions from the point of view of their contribution to social informatics (Section 3)?

2 Governance Models for the Microbiological Information Commons

Microbiological information has been characterised as being part of the public domain (Oldham, 2004; Smith *et al.*, 2004), implying appropriate public and regulatory institutions for guaranteeing its provision. However, this characterisation is very broad and, as has been shown in recent research (Kaul *et al.*, 2003), the notion of the public domain covers a heterogeneous set of transaction situations and incentive problems, which demand a more fine-grained approach.

2.1 Microbiological information as a common pool resource

The microbiological information that is managed and exchanged through biological research collections (BRCs) or global information facilities (such as the Global Biodiversity Information Facility) shows characteristics of both public goods and common pool resources. A convenient way to discuss this is to make a distinction between the ideas themselves and the artefacts and facilities through which they are exchanged. In Table 1, I have illustrated these distinctions and the related incentive problems for three components of the knowledge commons: information as an idea; the physical flow units or artefacts through which the information is exchanged; and the resource system or facility storing the ideas and the artefacts (Hess and Ostrom, 2003). Information as an idea clearly has the characteristics of a public good. It is a resource shared by multiple individuals in a non-exclusive way and it is non-depletable. The use of an idea by someone does not subtract from the ability of another individual to use the same idea at the same time. As such, in a similar manner to the self-archiving initiatives

in the field of scholarly communications (Hess and Ostrom, 2003), researchers who participate in building global biological information facilities are building a universal public good for which the more people who have access, the greater the benefit to everyone [*ibid.*]. Positive incentives that play a role in self-archiving initiatives, such as the reduction in costs of publication and access, the scientific recognition and credibility that comes with public disclosure, the increased visibility of information, and instant publication and dissemination (Hess and Ostrom, 2005a), have also been documented in the field of the microbiological information commons (Rai, 1999).

	Examples		
	Contribution of information to a global biological information archive	*Participation in the exchange of tumour tissue data*	*Common web server for storing images*
Bio-physical characteristics	FLOW OF IDEAS (KNOWLEDGE ON BIOLOGICAL DIVERSITY)	Flow of artefacts (images)	Facility (physical storage system)
Type of good	Public good	Common pool resource	Common pool resource
Positive incentives	Visibility, public recognition, instant publication	Access to first-hand, high-quality information related to the data	On-line verification of the diagnosis
Perverse incentives	Under-use: low visibility, lack of use	Misuse: use of the data without contributing to the flow, plagiarism, submitting low quality data	Pollution: storing redundant information that takes a lot of memory space

Table 1. Incentive problems for the public good and common pool resource aspects of the microbiological information commons
(Examples adapted from Hess and Ostrom, 2005b, Table 1; for simplicity of presentation, I have merged production and use incentives)

Information as a physical flow unit or artefact has also been characterised as a depletable resource which presents some of the characteristics of a common pool resource. Indeed, the value of information to users is not only related to the opportunities they have to access a stock or pool of accumulated knowledge somewhere in an encyclopaedia or digital repository, but also to the quality of the flow of the information as it is implemented in the artefacts. By exchanging the information, it is consumed, verified, completed and interlinked with other information. It is this complex process of exchanging artefacts and managing the quality of the information flow that

makes the information valuable to the users of the common knowledge pool. Management of this flow depends on compliance with a set of rules, such as verification of the quality of information submitted to the common pool, appropriate citation of the source of the information, and tools for cross-linking to the information generated by the user-communities in the field of knowledge concerned. Non-compliance with or violation of these rules harms the access to and use of the common knowledge base, and can lead to the information flow drying up (so rendering the resource depletable).

As mentioned above, sharing microbiological information through microbiological information facilities is a complex endeavour that also involves sharing larger physical resources. For example, the TuBaFrost project (which gathers data on high quality frozen tumour tissue samples with an accurate diagnosis which are stored in major European cancer centres and universities) makes information accessible and searchable through an uncomplicated query system on the Internet. A key physical resource that is shared in the TuBaFrost project is the Nanozoomer, which allows representative histology images to be stored in a central database, enlarged 20x or 40x and accessed through the virtual tumour bank. The advantage is that, through the addition of images to the virtual tumour bank, diagnoses can be verified on line. However, this also creates a depletable resource to be shared, the disk space of the central database.

2.2 Institutional solutions to the incentive problems

In the previous section I discussed some of the incentive problems to be solved in the organisation of data sharing in the microbiological commons. In this section, I will analyse some of the collective arrangements that are currently being considered for dealing with these incentive problems, focusing more particularly on the role of property rights and contractual arrangements.

In the field of microbiological commons, three main institutional solutions have been discussed in the literature: a model of free dissemination and two models based on conditional deposits for commercial and non-commercial use. All three are based on a form of decentralised ownership and include a certain level of collective management and exclusion rights. Such an institutional arrangement for the governance of the information flow is in accordance with the results that have been obtained from case studies within the field of natural resource management. Indeed, these studies show that, to deal with collective action problems within a common pool resource, there have to be common rules (at least for exclusion and management). These rules are necessary in order to delimit the boundaries of the common pool and impose graduated sanctions for non-compliance with the rules of use so as to prevent depletion of the resource.

2.2.1 Facilitating free dissemination with decentralised ownership

In a first model of data sharing, ownership – and hence the right to alienation – remains with the individual data providers. However the providers transfer a part of their management and exclusion rights to a common data portal. Some key features of this first model can be analysed using the Global Biodiversity Facility (GBIF) as an example. In the GBIF, data is provided to a collaborative database from a variety of sources; the database in turn makes the data freely available to non-commercial users, as illustrated in Figure 1. The ownership of the data, and any related conditions on its

use, remain with the original providers. This means that GBIF does not assert any intellectual property rights to the data that is made available through its network. Moreover, all the data is made available on the terms and conditions that data providers have identified in the metadata. However, even though GBIF does not assert any ownership rights, each data provider transfers some of the management and exclusion rights to GBIF as specified in the Memorandum of Understanding establishing the organisation. This transfer agreement allows different incentive problems related to the governance of the information flow as a common pool resource to be dealt with:

1. When registering their services with GBIF, the data provider has to sign the GBIF data sharing agreement. This stipulates that the data provider will make reasonable efforts to ensure that the data are accurate and will include a stable and unique identifier with the data (Articles 1.4. and 1.5. of the *Data Sharing Agreement*).
2. The data provider has to be endorsed by a GBIF participant. GBIF participants are the signatories of the Memorandum of Understanding which established GBIF. Data participants maintain stable computer gateways (the data nodes) that make data available through the GBIF network. The GBIF participants maintain services that enable new and existing data providers in their domain to be integrated within the GBIF network (Articles 1.8. and 2.4. of the *Data Sharing Agreement*).

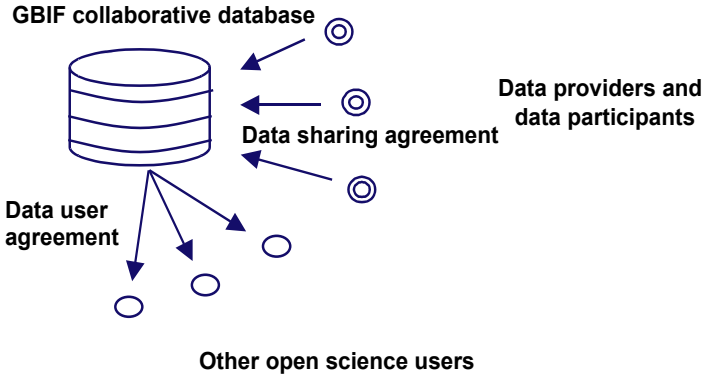

Figure 1. The GBIF model of data sharing

3. The GBIF participants empower the GBIF secretariat to enter into contracts, execute the work programme and maintain central services for the GBIF network. In particular, the GBIF secretariat may provide full or partial data to other users, together with the terms and conditions for use set by the data provider (Article 1.7. of the *Data Sharing Agreement*).
4. Using data through the GBIF network requires agreement to a *Data Use Agreement* when accessing the search engine. This agreement stipulates that users must publicly acknowledge the data providers whose biodiversity data they have used (Article 1.4. of the *Data Use Agreement*).

Through this collective arrangement, GBIF facilitates the free dissemination of biodiversity related data. In practice, GBIF pools data that is, in most cases, already in the public domain or that has been commissioned explicitly for public purposes and can reach a wider audience by being made accessible through the data portal. Elsewhere, more sophisticated two-tiered models have been developed to satisfy both public research interests and commercial opportunities.

2.2.2. Organising the licensing of data through a collective license organisation

The GBIF model is probably not appropriate for all types of microbiological data sharing. Indeed GBIF covers biodiversity-related data (including substantial microbiological databases) but not the wealth of microbiological data that is relevant for research but not directly relevant for biodiversity conservation purposes (such as plasmids, viruses or human cell lines for cancer research). Moreover, certain types of data are relevant both for public research purposes and private R&D and would benefit from a more coordinated approach to the conditions of data licensing to commercial partners.

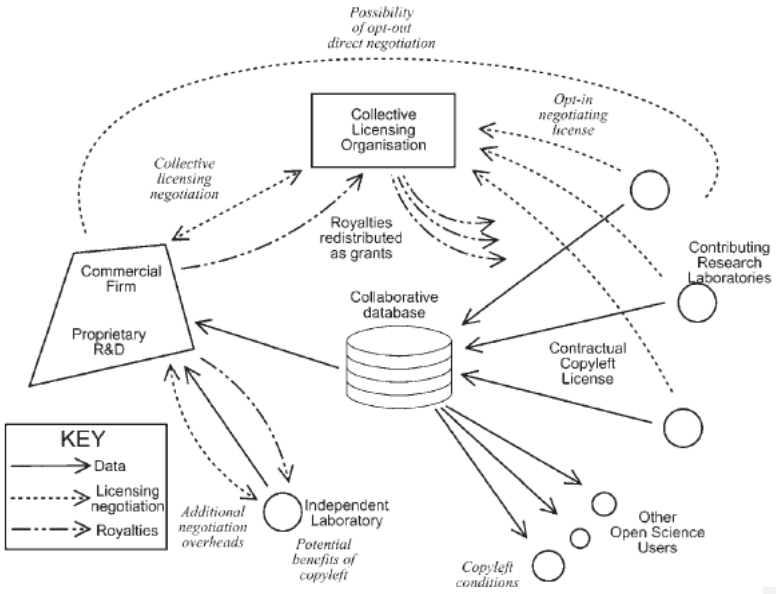

Figure 2. A two-tier system for data sharing based on the transfer of property rights to a collective licensing organisation (Eckersley *et al.*, 2003)

The report of an OECD working group on data sharing in neuro-informatics lists some of the conditions under which a better coordination of the conditions for commercial and non-commercial use of databases is appropriate. For public domain databases and/or in the absence of collective management of the conditions of follow-on use, data

sharing does not always guarantee credit to the researchers who originally produced the data, nor provide them with any reward if extensions to their work are commercialised (Eckersley *et al.*, 2003). Moreover, it only provides weak protection against the broader problem of 'patent thickets' (*ibid*).

Under these conditions, the OECD working group advised that different contractual conditions for access to the database be adopted for commercial and non-commercial use. In this model, which is analogous to the dual licensing model employed by some software developersi, non-commercial distribution is permitted by a copyleft licenseii, with the usual condition that the source of the data must be mentioned (guarantee of credit). Commercial use of the data is only permitted if a specific contract has been negotiated that includes restrictions on this commercial use and specifies a licence fee. Negotiating these ownership licences could be the job of a collective licensing organisation administering the database (Figure 2).

2.2.3 Organising the licensing of data through agreed contractual templates

As Reichman and Uhlir (2003) point out, because of the potential problems of leakage (moral hazard) and enforcement (accountability) in collective licensing organisations, the data providers may very well balk at participating in collectively managed collaborative databases. For this reason they propose a model for data sharing that is in many respects similar to the conditional deposits model illustrated in Figure 2, but where the collective licensing organisation is replaced by a 'soft' agreement, which can be a *memorandum* of understanding, a code of conduct, a common prototype or template specifying the way in which contractual relations with private content providers will be entered into (see Figure 3).

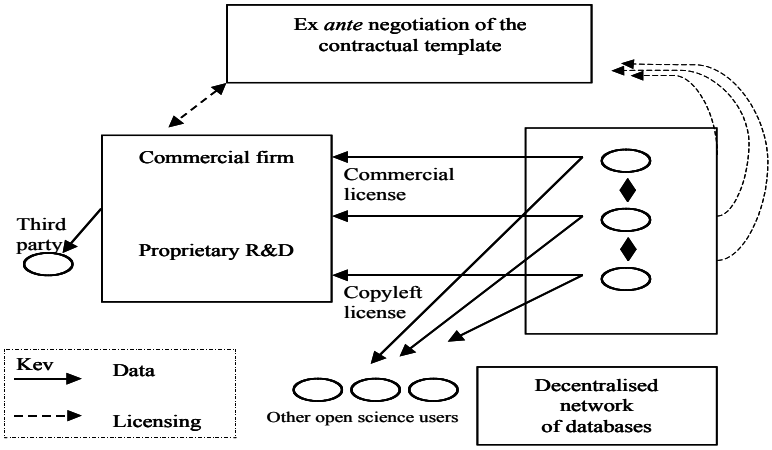

Figure 3. Two-tiered system for data sharing in a decentralised network of providers, based on a multilateral agreement between 'open science' data providers: a copyleft license (for relations with open-science users) and a contractual template (for relations with commercial users) (based on proposals in Reichman and Uhlir, 2003)

Reichman and Uhlir's (2003) proposition is designed to make it possible to restrict the use of basic scientific research data by users who want make a profit from it, or sell it to a third party. In order to prevent a 'race to the bottom' between competing producers of databases in communities that are traditionally dedicated to open science, they propose the adoption of a multilateral negotiated agreement amongst universities and between universities and their funding agencies. These agreements would set the standards for data exchange with other open science users through a kind of copyleft license and the standards for data exchange with commercial users through a common contractual template. To succeed, "these templates must be acceptable to the universities, the funding agencies, the broader scientific community, and the specific sub-committees – all of whom must eventually weigh in to ensure that academics themselves observe the norms that they would thus have collectively implemented" [*ibid*.].

3 Conclusion

The aim of this paper was to discuss a framework for the analysis of the governance of the microbiological information commons, relying on contemporary insights in new institutional economics. I have argued for the importance of considering microbiological databases both as a public good and as a common pool resource, the first referring to them as a common stock of ideas (hence non-subtractable in nature), and the second to the conditions of the organisation of the information flow through the exchange of artefacts and the use of common facilities (resources which are depletable).

Innovative proposals have been made to deal with the complex incentive problems related to the organisation of data sharing, especially in a context where the existing networks have to face increasing pressure from a globalised intellectual property regime. I considered more closely the successful endeavours of the Global Biodiversity Information Facility and the proposals for a two-tiered regime for governing the conditions of follow-on use of the data and related biological resources.

These institutional models offer interesting possibilities for social informatics. As I have argued, retaining some property right in the information, particularly a decision right on the way the information is managed in a certain community, is an important way of embedding the new technologies in the social context. Of course, the question of how these values are put into practice in the different institutional models still has to be evaluated. For instance: do such schemes really prevent enclosure of information in the public domain? do they initiate an effective learning process leading to common beliefs on the social values at stake? The experience of GBIF in this respect is limited. It only connects existing public databases in a distributed network of 'national nodes', without any obligation to decide on a common data policy. The learning that has occurred was mainly technical: through preparatory discussions in OECD, a common data exchange format was agreed upon that benefits the biodiversity conservation community at large. This is an important step forward, but the long-term sustainability of this model, in the absence of a more substantial common policy, can still be questioned. The more centralised propositions by Reichman and the OECD neuro-informatics group go a step

further. Indeed, the creation of more integrated institutions allows a set of common values to be implemented. The centralised organisation proposed by the OECD neuroinformatics group has a certain advantage, in that the governing body has to agree on the common management principles. However, more empirical research is needed to evaluate how these different schemes can strike a balance between the requirements of efficient coordination and the effective implementation of their social values.

References

Benkler Y. (1998), The Commons as a Neglected Factor of Information Policy, Remarks at the Telecommunications Policy Research Conference (Sept. 1998), available at http://www.benkler.org/commons.pdf (last visited July 2005).

Eckersley P. Egan G.F., Amari S., Beltrame F., Bennett R., et al. (2003), Neuroscience Data and Tool Sharing, *Neuroinformatics Journal*, Vol. 1 (2), pp.149–165.

Gevers D., Cohan F.M., Lawrence J.G., Spratt B.G., Coenye T., Feil E.J., Stackebrandt E., Van De Peer Y., Vandamme P., Thompson F.L. and Swings J. (submitted). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*.

Hess C. and Ostrom E. (2003), Ideas, Artefacts, and Facilities: Information as a Common-Pool Resource, *Law and Contemporary Problems*, Vol. 66(1/2), pp. 111–146.

Hess C. and Ostrom E. (2005a), A Framework for Analyzing the Knowledge Commons, in Ostrom E. and Hess C. (eds), *Understanding Knowledge as a Commons*, forthcoming.

Hess C. and Ostrom E. (2005b), A Framework for Analyzing Governance and Collective Action in the Microbiological Commons, paper presented at the Workshop on 'Exploring and Exploiting the Microbiological Commons', Brussels, July 7–8, 2005.

Kaul I., Conceiçao P., Le Goulven K. and Mendoza R.U. (eds.) (2003), *Providing Global Public Goods. Managing Globalization*, Oxford University Press, New York/Oxford.

Kling R. (ed.) (1996), *Computerization and Controversy: Value Conflicts and Social Choices,* second edition, Academic Press, New York.

Oldham P. (2004), Global Status and Trends in Intellectual Property Claims: Genomics, Proteomics and Biotechnology, Centre for Economic and Social Aspects of Genomics, 60pp.

Rai A.K. (1999), Regulating Scientific Research: Intellectual Property Rights and The Norms of Science, *Northwest University Law Review*, Vol. 77, pp. 77–152.

Rai A.K. and Eisenberg R.S. (2003), Bayh–Dole Reform and the Progress of Biomedicine, Vol. 66, (1/2), *Law and Contemporary Problems*, pp. 289–314.

Reichman J. (2002), Database Protection in a Global Economy, *Revue Internationale de Droit Economique*, pp. 455–504.

Reichman J. and Uhlir P.F. (1999), Database Protection at the Crossroads: Recent Developments and Their Impact on Science and Technology¸ *Berkeley Technology Law Journal*, 14, Vol. 793, pp. 799–821.

Reichman J. and Uhlir P.F. (2003), A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment, *Law and Contemporary Problems*, Vol. 66, pp. 315–440.

Matrice UCL 31/3/06 12:36
Formatted: Bullets and Numbering

Smith R., Thorsteindottir H., Daar S.A. and Singer P.A. (2004), Genomics Knowledge and Equity: A Global Public Goods Perspective on the Patent System, *Bulletin of the World Health Organization*, pp. 384–389.

Stiglitz J., Orszag P.R. and Orszag J.M. (2000), The Role of Government in a Digital Age, Report for the Computer and Communications Industry Association, United States.

---

i See, for example, the successful MySQL database software.

ii Under a copyleft regime for software, all users have the right to modify and adapt the program freely, on condition that their resulting development is also made freely available for use and further adaptation. The OECD working group's proposal is to apply the same license provision to the non-commercial use of databases.