Nucleic Acids Research, 2005, Vol. 33, Database issue ***D3–D4***
doi:10.1093/nar/gki134

EDITORIAL

Impediments to database interoperation: legal issues and security concerns

The significant growth of this annual over the past decade is testament to the importance of databases to scientific research in general and biomedical sciences in particular. In fact, the centrality of databases within our society at large is demonstrated by the continued debate in Congress regarding the legal protection of databases and the information they contain.

Productive utilization of databases requires interoperability: that is, the precise yet flexible interrelating of information from one database to another. There are, at present, two major impediments to achieving wide-scale interoperability: the state of database protection legislation and computer security issues. While most non-commercial/academic databases may not be overly concerned with the protection of their intellectual property, they still put up barriers to entrance, and consequently interoperability, due to concerns regarding the security of their computing infrastructure. We succinctly outline these two issues below in an effort to raise awareness of the issues among scientists.

The issues surrounding the legal protection of database are intricate and complex; this forum is not the place for complicated legal doctrine, suffice it to say that presently there is no intellectual property model that fully encompasses databases (1). But, what ought to be stressed here is the impact of the present situation of relative anarchy (i.e. no laws expressly guaranteeing the protection of databases here in the United States, in addition to the disparity of legal recourse between here and Europe) on the scientific research front.

Bioinformatics research exemplifies the need to be able to integrate heterogeneous, diverse and distributed large-scale datasets. It attempts to efficiently process, curate, manage, and mine the deluge of biological data available. As it does not produce its own raw data, as is the case with many other fields, it instead must examine and integrate other researchers' data, relying on a culture of sharing to attain this information. This integration, while obviously leading to a better understanding of whatever subject is at hand, also tends to allow for the discovery of new and pertinent information regarding those biological systems: the sum is definitely greater than its parts. The basic requirements for this integration are uniformity and accessibility; data are ineffectual if scattered among incompatible resources.

Unfortunately though, without any legal recourse to protect their databases, many providers have looked toward digital rights management schemes to defend their investments. The protections, digital locks that prevent easy access to data through passwords, complicated web forms, digital watermarks, proprietary data formats, or the delivery of data in piecemeal and limited form also serve to inhibit the interoperability of the databases by limiting the accessibility of each database and creating the need for different interfaces for each database and encumbering the transfer of information to a medium where it can be manipulated and analyzed. Worse still are the database providers that avoid digital protections all together and just refrain from publishing their datasets.

While the issue of protection of intellectual property may not be as pertinent in the academic universe, many academics make great use of commercial databases and also have aspirations to commercialize their own research. Coherent legal protection of databases may provide a way for academics to more easily access commercial information resources. In comparison to legal issues, computer security concerns are a more direct impediment to interoperability for academics.

With the constant and ever growing barrage of malicious internet attacks (worms, viruses, or denial of service attacks, for example) even academic facilities are forced to place limits and restrictions on access to their data. These limitations, like digital rights management, tend to result in situations wherein access to data is limited and convoluted. The end result is the same: different interfaces are required for each database thus hindering the interoperability of these systems. Unfortunately, integration is only one of the casualties of the present situation. There are also significant monetary and opportunity costs involved in protecting of biomedical computing facilities from malevolent and random attacks, i.e., people and resources that could be used towards research as opposed to the mundane maintenance of computers.

There are no easy solutions to either of these problems (2,3). Dueling forces in the Congress may never come to an agreement as to what should be the proper level of protection for databases. But the scientific community must continue to be vocal as to our needs within the continuing database protection debate. And there is no sign of the deluge of internet attacks letting up anytime soon.

While our respective legislatures mire in the political quagmires surrounding these issues that prevent them from resolving these concerns anytime soon, it is up to the scientific community to recognize the need for change and to come together in a joint effort to effect this change. Interoperability standards should be understood to be essential, developed, nurtured and maintained.

Ideally, we would like databases to be freely available and as flexibly interoperable with one another as possible. There should be standards (e.g. based on the various XML technologies) for how databases can be related to one another, and users should be able to create and manipulate integrated views of multiple related disparate datasets. Public databases should have standards-based web services-like interfaces so that users can script complex programs that work and interact across multiple databases distributed across the world, as easily as if they had downloaded all the databases locally to their own site and coalesced them into a common view (which is difficult and time-consuming, but unfortunately the current norm for integrated analyses). Central agencies should invest in either hosting tools and databases or in creating authentication schemes that provide, at the minimum, some added protection against web attacks.

Dov Greenbaum
Andrew Smith
Mark Gerstein

REFERENCES

1. Greenbaum,D. (2003) Commentary: the database debate: in support of an inequitable solution. *Albany Law Journal of Science and Technology*, **13**, 431.
2. Greenbaum,D., Douglas,S.M., Smith,A., Lim,J., Fischer,M., Schultz,M. and Gerstein,M. (2004) Computer security in academia-a potential roadblock to distributed annotation of the human genome. *Nat. Biotechnol.*, **22**, 771–772.
3. Greenbaum,D. and Gerstein,M. (2003) A universal legal framework as a prerequisite for database interoperability. *Nat. Biotechnol.*, **21**, 979–982.