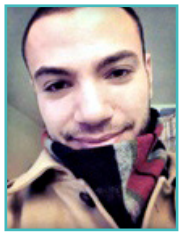
# Longevity of Biological Databases


**Teresa K. Attwood**[1]✉, **Bora Agit**[1], **Lynda B.M. Ellis**[2]

[1]University of Manchester, Manchester, United Kingdom
[2]University of Minnesota, Minnesota, United States

**Received** 23 January 2015; **Accepted** 13 March 2015; **Published** 4 May 2015

Attwood TK *et al.* (2015) *EMBnet.journal* **21**, e803. http://dx.doi.org/10.14806/ej.21.0.803

Competing interests: TKA currently serves on the editorial board of EMBnet.journal; BA none; LBME none

## Abstract

Public Web-based databases are essential for present-day biological research: they i) store the results of past laboratory experiments; ii) guide the focus of future ones; and, iii) allow all to benefit from the wealth of information they contain. Many new databases are born each year; but how long do they live? This study looked at the 18-year survival of 326 databases. Over 60% were dead within that time period, and a further 14% were archived, no longer updated. Those that survived were, for the most part, important to their institution's main focus, and had core institutional support. Database longevity depends on the existence of infrastructures that are underpinned by long-term financial strategies. Researchers and funders need to consider the ramifications for the security of their data, and of the financial investments in them, if they choose to create new databases independently of core infrastructures.

## Introduction

During the last 30 years, since the first public release of resources like the EMBL Data Library (Hamm and Cameron, 1986) and GenBank (Burks *et al.*, 1985), databases have become an indispensable part of the tool-kit of modern biological research: we depend on them to store experimental data of all kinds, to inform our research, and to share the fruits of our collective knowledge with the scientific community. Back in the 1980s, when the field of bioinformatics was just emerging, there was an unwritten rule that biological databases (and their associated analysis software) should be made freely available. In consequence, they became a side-effect of research projects and, each year, many new databases were born and distributed to a voracious community. Indeed, they became such a familiar part of the research landscape that an entire issue of a prestigious journal (*Nucleic Acids Research*) was formed to alert the community to updates and modifications to existing resources and to the appearance of new ones, and a Web-based database was created to catalogue them – DBcat (Discala *et al.*, 1999).

Superficially, this is a success story – life scientists took little persuading that their data benefitted from proper management and analysis. However, no overarching financial strategy underpinned this database revolution – once created, therefore, many struggled to survive. So the question is, how long do they live in reality? In 1998, Ellis and Kalumbi surveyed maintainers of public biological databases listed in DBcat. This survey found that more than two-thirds (68%) of the 153 databases for which information was received (48% response rate) had uncertain near futures (1-5 year funding) (Ellis and Kalumbi, 1998). We, and others, have commented on this shaky future, arguing that a viable, sustainable framework for long-term data stewardship is sorely needed (Ellis and Kalumbi, 1999; Ellis and Attwood, 2001; Abbott, 2009; Bastow and Leonelli, 2010; Baker, 2012; Hayden, 2013).

Fifteen years beyond the original survey, we were curious to know which of those biological databases that were alive at the end of the 20th century had managed to persist into the 21st? In particular, we were keen to understand what distinguishes the survivors from the rest. In an at-

tempt to answer these questions, we planned to return to the DBcat listing that had underpinned the 1998 survey.

## Methods

Ironically, the DBcat database itself died in 2006. However it – and much of the older Web – lives on in the Internet Archive[1]. DBcat was first archived in May 1997, when its home page reported it contained information on 383 databases. This archive[2] was used in the present study. The full data-set used in this study is presented in the spreadsheet, Supplementary File 2[3]; an explanation of the contents of each Sheet in the spreadsheet can be found in Supplementary File 1[4].

DBcat records were examined for each database entry. Eight were duplicates, leaving 375 databases (see Sheet 1 in Supplementary File 2)[3]. Each of these was examined in turn to determine: i) whether it was indeed a public Web-based database; ii) if so, whether it was still 'alive' in the first half of 2015; and, iii) if alive, when it was last updated.

*What is a public web database?* Information in DBcat was, for the most part, entered by the database maintainers themselves. We eliminated five as commercial, two as links to a research group or research centre, and 31 as lists of information lacking even a search function or in other ways not a Web database. Four others, freely available initially but commercial upon re-study, were also eliminated. Some databases might disappear, and their name could be used, knowingly or unknowingly, for a newer database in the same field. In nine situations, we could not determine whether or not this occurred; we classed the state of these databases as **unclear** and removed them from the set (see Sheet 3 in Supplementary File 2[3] for a list of all excluded entries). This left 326 entries (see Sheet 2 in Supplementary File 2)[3].

*What does 'life' mean for a Web-based database?* Determining what constitutes 'life' or 'death' for a Web-based database is non-trivial – answers to the question are not black or white. If the data in a database had been transferred to another, different database, such as the transformation of the collection of 'Modules in Extracellular Proteins', which was published as SMART (Shultz *et al.*, 1998), we classed the original database as **alive–rebranded**.

Some live databases contain notices stating, for example, that they are no longer updated (*e.g.*, the Blocks Database (Henikoff *et al.*, 2000)), or their database history shows that to be the case. Databases that had not been updated since 2012 or earlier but were still functional and searchable, even if only in mirrors, we considered to have been **archived**.

We also found databases whose search function had either disappeared or was non-functional or had lost other key functionality. We counted these as **dead** even if they still existed on the Web. If a mirror site was being updated (*e.g.*, SCOPe at the University of California Berkley (Fox *et al.*, 2014)), the database was classed as **alive**, even if the parent was dead or archived.

In an attempt to gain insight into the relative 'health' of some of these resources, we looked more closely at the 46 databases from the DBcat DNA category included in our analysis. The approach was purely qualitative: databases maintained by large groups or consortia at institutes or organisations whose main mission was service provision at some level, or that were funded privately, we considered to have strong financial support; those that appeared to be maintained by individuals, especially those in academic environments, we considered to have weaker financial support.

The status of these databases changes as we speak: their URLs change; they change their names; their data move. If alive at one moment, they may be archived at the next; if archived, they are eventually likely to die; dead databases might even return to life. Our data and analyses are hence a snapshot of a moving target, and should consequently be read in that light.

## Results

As shown in Table 1 and illustrated in Figure 1, of the 326 entries investigated, we classed 53 as alive, 23 as alive-rebranded, and 47 as archived; according to our criteria, a total of 203 (62%) were dead (see Sheet 2 in Supplementary File 2)[3].

Of the 46 entries in the DBcat DNA category, 21 were alive or alive-rebranded, three were ar-

1 archive.org/
2 web.archive.org/web/19970502044745/http://www.infobiogen.fr/services/dbcat/
3 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1096
4 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1095

Table 1. 18-year survival status of 326 databases from the May 1997 DBcat listing.

| Category | N | Percent |
|---|---|---|
| Alive | 53 | 16.3% |
| Alive - rebranded | 23 | 7.0% |
| Archived | 47 | 14.4% |
| Dead | 203 | 62.3% |
| **TOTAL** | **326** | **100%** |

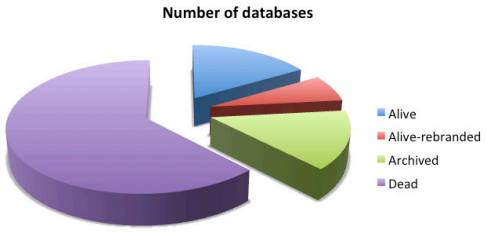



Figure 1. Illustration of the data listed in Table 1, showing the proportions of databases that were alive, dead (or becoming so) after a period of 18 years.

chived, 22 were dead, six were excluded, and one was unclear.

Of the 21 alive or alive-rebranded databases, 17 (81%) were supported by stronger financial infrastructures than the others. Of the 22 dead databases, most (73%) appeared to have had weaker financial support, in the sense of originating from academic environments, or research institutes whose core mission was not service provision (see Sheet 4 in Supplementary File 2)[5].

## Discussion

### Classification

We classified databases as alive according to whether they were updated in 2013 or more recently, and as archived if they were not. We accept that this is an arbitrary cut-off date, but while some archived databases may simply be 'resting' during funding droughts and may resume updates when funds begin to flow again, equally, those that are currently alive may cease to do so if they hit funding deserts – the likelihood is that these numbers will balance.

We excluded databases that were commercial in the original 1997 DBcat data, as they were never public databases. We also excluded those that became commercial after that time. We could have instead classed those as 'dead', as they are no longer public databases.

### Database Longevity

Database longevity depends on finding a continuous funding source. This is possible, say, for a database that supports the main focus of its host institution: for example, the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ) hosts one of the largest microbial culture collections worldwide. Its free Web catalogues[6] will be around as long as the DSMZ exists; they are funded, and updated, as a key part of their institution's mission.

It is sobering how many of the 326 databases were found to be dead (62%) or to exist in an archived state (14%) – the situation may actually be worse than this, as the authors have personal communications of funding problems for some of the databases classed as alive. Regardless, the figures are consistent with the results of the 1998 survey, in which 68% of responding database curators claimed uncertain 1-5 year financial futures for their resources.

### Economic models

Previous work listed several economic models that are, or could be, used for the support of biological databases. We looked at public funding, asymmetric pricing, advertising, deal-making, direct sales and hybrids (Ellis and Kalumbi, 1999); a decade later, several of these models, and their inherent complexities, were also reviewed by Bastow and Leonelli (2010).

Some databases evolve to include more than their database functions, including income-producing endeavours (direct sales), which may help fund database costs: for example, an important focus of the DSMZ catalogues is listing the price of their cultures and how to order them.

Public funding remains the most frequently used financial model, with well-known problems when such funding ceases: for example, in 2009, The Arabidopsis Information Resource (TAIR, (Lamesch *et al.*, 2012)) lost its public funding, generating a relatively large amount of publicity for its plight (Abbott, 2009). Other databases in the alive and archive categories face, or have faced, similar problems (Baker, 2012).

5 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1096

6 www.dsmz.de/catalogues.html

Asymmetric pricing – charging some users more than others – is less frequently used. TAIR, for example, is now funded by subscriptions, charging commercial organisations more than educational institutions or non-profits. It is not yet clear how successful this strategy may be (Hayden, 2013). Other databases may offer some content free and the complete version for a license fee: for example, Transfac (Matys *et al.*, 2006) has a free public version that is more than 10 years older than its commercial version. Commercialisation is only viable for those databases with a sufficiently large subset of users who are willing and able to pay for commercial versions.

Advertising is not used, in part because advertisers are unwilling to pay for display on the relatively low-traffic Web pages of most biological databases. Corporate sponsorship is part advertising and part deal-making: the corporation pays to help support a database that provides value to its potential customers, who may see its logo and a link to its website listed under 'Sponsors', and gains good-will. No biological database has gained appreciable funding through such sponsorship.

## What distinguishes survivors?

It is interesting to reflect on the enormous investment that has been made during the last 20-30 years to establish and sustain so many biological databases, and the energy – the human cost – it has taken to maintain them. More than 60% of Web-based databases available in DBcat in 1997 have died – a significant waste of investment. The persistence of Web-based resources is a known problem: *e.g.*, Hennessey and Ge (2013) found that the median lifespan of Web pages referenced in article abstracts from the Web of Science citation index, published between 1996 and 2010, was around nine years, 62% of them being archived. Similarly, our analysis has shown that while a small number of the 1997 DBcat databases have been able to persist through rebranding exercises, many others are only now accessible in some archived form (in which their value, and future accessibility, is likely to erode further with time). Less than 20% are still actively maintained.

## Case studies

Those databases that do persist today have clearly had winning survival strategies. Many have experienced funding crises, and have had to be rescued from the brink of extinction. Swiss-Prot is a case in point (Bairoch *et al.*, 2004; Bairoch, 2000). In 1996, Swiss-Prot hit a problem: an application for renewal of a grant from the Swiss National Science Foundation (SNSF) was turned down, because the database was being widely used outside Switzerland, and SNSF funds were intended to support primarily national, rather than international, projects; at the same time, an application to the EU was declined, because its infrastructure grants were intended to complement existing local funding, which the SNSF had just declined to provide. To alert users to the problem – at this point, funds existed only for two further months of the biocurators' salaries – an Internet appeal was launched, announcing that Swiss-Prot would disappear on 30 June 1996 if no solution could be found. The *"Internet storm of protest"* that followed did not go unheeded: the Swiss scientific funding agencies recommended that a stable, long-term funding mechanism be sought to sustain the database (Bairoch, 2000). Interim funding was provided on a short-term basis, from 1997-1999; during this time, Bairoch and his colleagues were involved in high-level talks that led to the creation, in 1998, of the SIB Swiss Institute of Bioinformatics as a non-profit foundation, providing the database with a 'permanent' home (Bairoch, 2000).

One consequence of this was that, by Swiss law, the government could only fund up to 50% of the budget of such an institution, the remainder having to be found via other avenues, preferably commercial. Accordingly, a new company – GeneBio – was established as the commercial arm of the SIB. The licensing strategy adopted by GeneBio was, perhaps, unusual. The company's founders wanted to ensure that the methods by which academic and commercial users accessed Swiss-Prot would not change – it was therefore based on trust, relying on commercial users contacting the company to pay an annual licence fee. This system was very successful for several years; however, it was not the end of the story. Additional funding subsequently acquired from the National Institutes of Health (NIH) stipulated that access to the database must be free – Swiss-Prot could therefore no longer be sold commercially.

With this NIH funding, Swiss-Prot was subsumed into UniProtKB (Apweiler *et al.*, 2004; Bairoch *et al.*,

2004)), along with TrEMBL (Bairoch and Apweiler, 1996) and the Protein Information Resource Protein Sequence Database (PIR-PSD) (George *et al.*, 1986). Today, UniProtKB is managed by the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics and the PIR – the UniProt Consortium – and falls under the protective umbrella of Europe's distributed infrastructure for life-science information, ELIXIR (Crosswell and Thornton, 2012). The PIR-PSD's role in this story is interesting, not least because it had competed with Swiss-Prot for many years. In principle, it gained a new lease of life through the creation of UniProtKB. However, for most users, the resource has become largely invisible, archived in UniParc and not overtly visible in UniProtKB except via given entries' database cross-references.

Probably the oldest biological database still in use is the Protein Data Bank (PDB), first launched in 1971 (Anonymous, 1971). Inevitably, during its more than 40-year history, the PDB has faced its share of funding struggles – not least, in the late 1990s, when the funding agencies invited researchers to submit competitive grant proposals in a bid to stabilise the resource and improve its efficiency. This eventually led to a new consortium approach to its management – the so-called Research Collaboratory for Structural Bioinformatics (RCSB) – and with it, a move, in 1999, from its location at the Brookhaven National Laboratories to Rutgers, The State University of New Jersey (Berman *et al.*, 2000), where it remains today.

Aside from UniProtKB and the PDB, amongst the strongest surviving databases are EMBL (now part of ENA (Cochrane *et al.*, 2013)), GenBank (Benson *et al.*, 2014), DDBJ (Kosuge *et al.*, 2014), Ensembl (Flicek *et al.*, 2014) and InterPro (Mitchell *et al.*, 2014). Several of these will benefit from being part of ELIXIR, in which they are 'named services' that may ultimately qualify for core support, whether at the EBI or at designated ELIXIR Nodes across Europe as their host countries ratify ELIXIR's Consortium Agreement. ELIXIR is a pan-European, inter-governmental initiative seeded by the European Strategy Forum on Research Infrastructures (ESFRI), which, in 2002, set out to support the long-term needs of European research communities.

Of course, originating at an institute, organisation or Node with strong financial support is not a guarantee of strong *database* support, and is hence not in itself a guarantee of longevity, especially if the host institution loses its core funding and closes, or undergoes rebranding and mission evolution, or if the key author leaves. For example, of the databases observed to be dead in the DBcat DNA category, ALU (DBC0002) was developed at the NCBI by an individual who moved elsewhere; Genexpress (DBC00007) was developed at Infobiogen, which closed down; the HGMP Primers Database (DBC00280) was developed at the Human Genome Mapping Project Resource Centre, a UK Research Council-funded institute that closed down; and TIGR-AT (DBC00133), EGAD (DBC00197) and HCD (DBC00202) were developed at The Institute for Genome Research (TIGR), which rebranded as the J. Craig Venter Institute (JCVI), and no longer maintains or supports many of TIGR's databases (these databases are marked with an S* comment value in Sheet 4 in Supplementary File 2)[7].

Against this background, recognising the increasing importance of data, or rather, of 'big data', in underpinning advances in biomedicine, a trans-US-NIH initiative – Big Data to Knowledge (BD2K) – was recently launched in the United States (Margolis *et al.*, 2014). BD2K will facilitate biomedical research, in part by supporting a 'data ecosystem' that is able to accelerate knowledge discovery. Discussions of possible interactions between ELIXIR and BD2K are in their infancy, and it will be interesting to see what concerted plans, if any, may emerge for sustaining a data ecosystem globally. Meanwhile, it's clear that European databases that do not belong to ELIXIR Nodes will face much stiffer competition for funds in future, as governments divert resources to sustain their central Nodes. Whether this will be an affordable model remains to be seen. ELIXIR may seem like a light at the end of a long and dark funding tunnel for some databases, but may ultimately cause the lives of many more to be extinguished.

## Access to data in perpetuity?

The last point brings us to the issue of 'biodiversity'. Diverting funds primarily to large, successful databases threatens the existence of smaller but nonetheless valuable resources. Consider, for example, InterPro, which integrates around 12 different databases (including PROSITE (Sigrist

7 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1096

*et al.*, 2013), PRINTS (Attwood *et al.*, 2012), and Pfam (Finn *et al.*, 2014)) and was developed as a key tool for automatic annotation of TrEMBL entries (Apweiler *et al.*, 2001; Mitchell *et al.*, 2014). By virtue of being housed at the EBI, InterPro may achieve some future measure of protection under ELIXIR; however, its source databases that are not maintained at the EBI – most of them – will not. InterPro is thus in danger of losing many of its partners and, with them, much of its diagnostic strength and richness. Ultimately, it is in danger of becoming a mere HMM-based resource, its 'biodatabase biodiversity' completely lost.

Another interesting issue that has emerged in recent years has been the drive to create 'open data repositories'. Just as the Open Access movement drove the creation of Institutional Repositories to archive research papers, similar arguments are pressuring universities into establishing their own research data repositories; there are also moves afoot to create citable 'data papers', to incentivise (rather than mandate) scientists to deposit their data. How this will work in practice is unclear.

One of the drivers behind initiatives like this is the desire to improve research communication by coupling scientific articles more strongly with their research data (Bourne *et al.*, 2011). This will require the research community to *"develop best practices for depositing research data-sets in repositories that enable linking to relevant documents, and that have high compliance levels driven by appropriate incentives, resources and policies."* This vision takes us beyond the problems of how to maintain a few hundred biological databanks, into a world in which we will have to figure out how to archive all published research data such that they will be accessible and searchable for all time. Even if we accept that a static data archive is different from a functional (and evolving) database, if we have not yet solved the sustainability problems for biological databases, it will be interesting to see how archives for *all* research data will be managed in perpetuity.

Regardless, the good news is that, at least at some level, the scientific community and the bodies that fund scientific research have woken up to the importance of organising and archiving research data. Whether this will help to address some of the meatier issues of long-term database maintenance is moot. What remains clear is that this is still very much an unsolved problem, one that the International Society for Biocuration (ISB) is beginning to consider very seriously. The Society has observed that, while research infrastructures are becoming more widespread, securing funding for database maintenance is still problematic, even for well-established databases – although funders are generally keen to support projects that generate yet more data, there is still insufficient recognition of the importance of data curation. This motivated the ISB to launch a survey in order to gain an overview of the financial situation of databases managed by its current members. The results of the survey will be shared at a workshop (*Money for biocuration: strategies, ideas & funding*) to be held at the 8th International Biocuration Conference in Beijing, 23-16 April 2015, in which participants will have the opportunity to discuss what the ISB, and biocurators in general, can do to help. We look forward, with great interest, to the outcomes.

## Conclusion

Much has changed since the 1998 database survey, but there are also several constants. Biological databases are expensive to create and maintain; nevertheless, databases continue to be created afresh each year. Far from stemming the tide of new repositories, some funding bodies are requesting researchers to elaborate 'data management plans' as part of their research proposals. Compelling scientists to explain how their data will be archived and made accessible seems like an important step forward, especially as responsibility for their financial future is being pushed onto institutions. Nevertheless, initiatives like this will not guarantee the long-term sustainability of databases, whose value to the community depends on active update and maintenance schedules rather than passive archiving.

Despite past funding issues, some of the most successful databases have survived by being integrated into larger database federations (ENA, UniProt, InterPro for example). Above all, however, it is clear that institutional support is a key feature in the precarious ups-and-downs of the database-funding landscape. Regardless of their sustainability strategy, databases require the input of skilled biocurators and bioinformaticians, and their ongoing commitment will continue to be costly to support in the long term.

As larger databases battle for their futures, many more smaller, specialist databases are being lost along the way. European infrastructures like ELIXIR and funding initiatives like BD2K will certainly have a significant role to play in securing the long-term future of some key databases, and of the biocurators and bioinformaticians required to manage them. It is too early to tell what the data ecosystem of tomorrow will look like; nevertheless, it is probably safe to say that it will be dominated by many of the most successful databases of today.

### Key Points

- Active maintenance and update of public, Web-based biological databases is time-consuming and costly.
- Without financial sustainability plans, most databases created as outputs of research projects consequently die or are archived within 10-15 years.
- Longevity is a function of core institutional support.
- Researchers should understand these facts before creating a database independently of core infrastructures underpinned by long-term financial strategies.

## Acknowledgements

The authors wish to thank the legions of scientists and educators who have developed and maintained biological Web-based databases, past and present.

## References

Abbott A (2009) Plant genetics database at risk as funds run dry. Nature 462, 258-259. http://dx.doi.org/10.1038/462258b

Anonymous (1971) Protein Data Bank. *Nature New Biology* **233**, 223. http://dx.doi.org/10.1038/newbio233223a0

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**(1), 37-40. http://dx.doi.org/10.1093/nar/29.1.37

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B *et al.*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**(Database issue), D115-119. http://dx.doi.org/10.1093/nar/gkh131

Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB *et al.* (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource – its status in 2012. *Database* (Oxford) 2012:bas019. http://dx.doi.org/10.1093/database/bas019

Bairoch A (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* **16**(1), 48-64. http://dx.doi.org/10.1093/bioinformatics/16.1.48

Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* 24(1), 21-25. http://dx.doi.org/10.1093/nar/24.1.21

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* **5**, 39-55. http://dx.doi.org/10.1093/bib/5.1.39

Baker M (2012) Databases fight funding cuts. *Nature* **489**, 19. http://dx.doi.org/10.1038/489019a

Bastow R, Leonelli S (2010) Sustainable digital infrastructure. *EMBO Rep.* **11**(10), 730-734. http://dx.doi.org/10.1038/embor.2010.145

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J *et al.* (2014) GenBank. *Nucleic Acids Res.* **42**(D1), D32-D37. http://dx.doi.org/10.1093/nar/gku1216

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**(1), 235-242. http://dx.doi.org/10.1093/nar/28.1.235

Bourne P, Clark T, Dale R, de Waard A, Herman I *et al.* (eds.) (2011) The Force11 White Paper: Improving Future Research Communication and e-Scholarship. A publication resulting from the Schloss Dagstuhl Perspectives Workshop: The Future of Research Communication, 15-18 Aug 2011. http://www.force11.org/white_paper

Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S. and Bilofsky, H.S. (1985) The GenBank nucleic acid sequence database. *Comput. Appl. Biosci.* **1**(4), 225-233. http://dx.doi.org/10.1093/bioinformatics/1.4.225

Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.* **41**, D30-D35. http://dx.doi.org/10.1093/nar/gks1175

Crosswell LC1, Thornton JM. (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.* **30**(5), 241-242. http://dx.doi.org/10.1016/j.tibtech.2012.02.002

Discala C, Ninnin M, Achard F, Barillot E, Vaysseix G (1999) DBcat: a catalog of biological databases. *Nucleic Acids Res.* **27**, 10-11. http://dx.doi.org/10.1093/nar/27.1.10

Ellis LBM, Kalumbi D (1998) The demise of public data on the web? *Nature Biotechnology* **16**, 1323-1324. http://dx.doi.org/10.1038/4296

Ellis LBM, Kalumbi D (1999) Financing a Future for Public Biological Data. *Bioinformatics* **15**, 717-722. http://dx.doi.org/10.1093/bioinformatics/15.9.717

Ellis LBM, Attwood TK (2001) Molecular Biology Databases: Today and Tomorrow. Drug Discovery Today **6**, 509-513. http://dx.doi.org/10.1016/S1359-6446(01)01802-5

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.* **42** (D1), D222-D230. http://dx.doi.org/10.1093/nar/gkt1223

Flicek P, Ridwan Amode M, Barrell D, Beal K, Billis K, *et al.* (2014) Ensembl 2014. *Nucl. Acids Res.* **42** (D1), D749-D755. http://dx.doi.org/10.1093/nar/gkt1196

Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins – extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42** (D1), D304-D309. http://dx.doi.org/10.1093/nar/gkt1240

George DG, Barker WC, Hunt LT (1986) The protein identification resource (PIR). *Nucleic Acids Res.* **14**(1), 11-15. http://dx.doi.org/10.1093/nar/14.1.11

Hamm GH, Cameron GN (1986) The EMBL data library. *Nucleic Acids Res.* **14**(1), 5-9. http://dx.doi.org/10.1093/nar/14.1.5

Hayden EC (2013) Popular plant database set to charge users. *Nature News* (31 August 2013) http://dx.doi.org/10.1038/nature.2013.13642

Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res.* **28**, 228-230. http://dx.doi.org/10.1093/nar/28.1.228

Kosuge T, Mashima J, Kodama Y, Fujisawa T, Kaminuma E *et al.* (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Res.* **42** (D1), D44-D49. http://dx.doi.org/10.1093/nar/gkt1066

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40** (D1), D202-D210. http://dx.doi.org/10.1093/nar/gkr1090

Margolis R, Derr L, Dunn M, Huerta M, Larkin J *et al.* (2014) The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J *Am Med Inform Assoc.* **21**, 957-958. http://dx.doi.org/10.1136/amiajnl-2014-002974

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108-D110. http://dx.doi.org/10.1093/nar/gkj143

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43** (D1), D213-D221. http://dx.doi.org/10.1093/nar/gku1243

Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344-D347. http://dx.doi.org/10.1093/nar/gks1067

Schultz J, Milpetz F, Bork P, Ponting CP (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5857-5864.