DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bar059, doi:10.1093/database/bar059

Original article

Biocurators and Biocuration: surveying the 21st century challenges

Sarah Burge[1,2], Teresa K. Attwood[3], Alex Bateman[2], Tanya Z. Berardini[4], Michael Cherry[5], Claire O'Donovan[1,2], Ioannis Xenarios[6] and Pascale Gaudet[7,*]

[1]European Bioinformatics Institute, [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, [3]Faculty of Life Sciences and School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PT, UK, [4]The Arabidopsis Information Resource, Department of Plant Biology, Carnegie Institute for Science, Stanford, CA 94305, [5]Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA, [6]Swiss-Prot Group and [7]CALIPHO Group, Swiss Institute of Bioinformatics, Centre Médical Universitaire, 1, rue Michel Servet CH-1211 Geneva 4, Switzerland

***Corresponding author:** Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: pascale.gaudet@isb-sib.ch

Submitted 28 September 2011; Revised 25 November 2011; Accepted 28 November 2011

Curated databases are an integral part of the tool set that researchers use on a daily basis for their work. For most users, however, how databases are maintained, and by whom, is rather obscure. The International Society for Biocuration (ISB) represents biocurators, software engineers, developers and researchers with an interest in biocuration. Its goals include fostering communication between biocurators, promoting and describing their work, and highlighting the added value of biocuration to the world. The ISB recently conducted a survey of biocurators to better understand their educational and scientific backgrounds, their motivations for choosing a curatorial job and their career goals. The results are reported here. From the responses received, it is evident that biocuration is performed by highly trained scientists and perceived to be a stimulating career, offering both intellectual challenges and the satisfaction of performing work essential to the modern scientific community. It is also apparent that the ISB has at least a dual role to play to facilitate biocurators' work: (i) to promote biocuration as a career within the greater scientific community; (ii) to aid the development of resources for biomedical research through promotion of nomenclature and data-sharing standards that will allow interconnection of biological databases and better exploit the pivotal contributions that biocurators are making.

Database URL: http://biocurator.org

Introduction

Biocuration involves the analysis, interpretation and integration of biological information into data repositories, primarily to add value by annotating and interconnecting research data and results within a common biological framework. This integration both facilitates the use of data by the wider scientific community and renders them more easily accessible and amenable to computational analysis. A variety of factors, in particular the rapidly increasing pace of data acquisition in the life sciences, have led to the proliferation and wide-spread uptake of biocuration as a full-time career. At the same time, these factors are making the work and careers of biocurators more interesting and challenging every day. Investment in high-throughput technologies, starting with microarray expression analyses in the mid-1990s and continuing with ever improving dissection of the genome, transcriptome, proteome and metabolome, has given rise to a tremendous escalation in the rate of raw biological data production. In turn, this has generated a paradox: on one hand, it has created a pressing need for greater manual annotation and analysis efforts; on the other, it has made it impossible for purely manual efforts to keep up with the scale of data acquisition, creating an urgent need for intelligently designed tools to help automate the conversion of raw data to knowledge and understanding. The challenge for biocurators is clear.

Database, Vol. 2012, Article ID bar059, doi:10.1093/database/bar059

Reflecting the data-driven nature of modern biology, databases have grown considerably both in size and number during the last decade. The exact number of databases is difficult to ascertain. While not exhaustive, the 2011 *Nucleic Acids Research* (*NAR*) online database collection lists 1330 published biodatabases (1), and estimates derived from the ELIXIR database provider survey suggest an approximate annual growth rate of ~12% (2). Globally, the numbers are likely to be significantly higher than those mentioned in the online collection, not least because many are unpublished, or not published in the *NAR* database issue.

Against this background, databases have become a cornerstone of modern biomedical research, and are now being cited in the literature thousands of times per annum. Responsibility for their design, implementation, maintenance, as well as for organizing, annotating, archiving and making their contents publicly available, falls to biocurators and bioinformaticians. As the volumes of data and the number of databases have grown, so too has the biocuration community. In 2009, the International Society for Biocuration (ISB, www.biocurator.org) was formed, to give biocurators a voice and to promote the interests of biocuration. The ISB now counts over 300 members from nearly 150 databases and institutions in 26 different countries. This is a large underestimate: large fractions of the biocuration community are not well-represented in the ISB—in particular, biocurators from commercial databases, as well as researchers, students and post-docs who perform some biocuration work as part of a research project.

While the roles of biocurators in managing and augmenting biomedical data have been increasingly well-documented in the literature (3–6), the nature of their career paths is not well understood, either outside the biocurator community or within it. As part of its mission to advance biocuration as a professional career path, the ISB set out to understand the perceived challenges, concerns and benefits to biocurators of this career choice: specifically, a survey was conducted aiming to gain a qualitative appreciation of biocurators' motivations for entering and remaining in the field, and to comprehend their perceptions of the role of the ISB.

Survey methodology and results

The survey consisted of 37 questions for current biocurators and 13 questions for former biocurators. Questions were a mix of multiple choice, ordinal scale, interval scale and ratio scale. Some questions allowed the respondents to enter a free text reply. There were a total of 257 respondents to the survey. As respondents did not answer every question, the percentages reported correspond to the count of a specific response divided by the total number of responses to a particular question. The survey was publicized through the ISB website, mailing lists and social-networking sites, targeting both current curators and those who had recently left the field. The full survey questions and results are available at (http://biocurator.org/surveys/BiocurationSurveySummary_06292011.pdf). Although the sample size is relatively small, it represents a large fraction of the members of our networks: there are a little over 300 active members in the ISB, over 360 members of the ISB LinkedIn group, and nearly 500 members of the ISB email list (isb@listserv.it.northwestern.edu). Respondents were asked how much time they currently devote to biocuration activities (i.e. up to 10%, up to 50% and up to 100%). The majority of respondents (76%) spent 50–100% of their time on biocuration activities; just over half (53%) were members of the ISB.

Current biocurators

The typical biocurator. Almost 80% of respondents who were currently involved in biocuration were between 31-and 50-years old; 60% were female (Figure 1); and most (71%) were qualified to PhD level. Biocurators come from a range of different scientific backgrounds, most (73%) having previously worked as bench scientists, others (17%) having worked as bioinformaticians, programmers, or in other areas of computational science. Only 11% of respondents described themselves as currently working in industry.

More than half of the respondents (57%) were employed on limited-term contracts, some (25%) of 1–3 years' duration, others (24%) of ≥3 years; 41% were on permanent contracts; and 9% were principal investigators. Notwithstanding the proportion of contract work, 60% of respondents had been in their current role for >4 years, and 82% had been involved in biocuration (in various roles) for ≥7 years.

As shown in Figure 2, the types of data being handled by biocurators were diverse: spanning nucleotide sequences; protein sequences, families, interactions and pathways; small molecules; model organisms; the literature; and

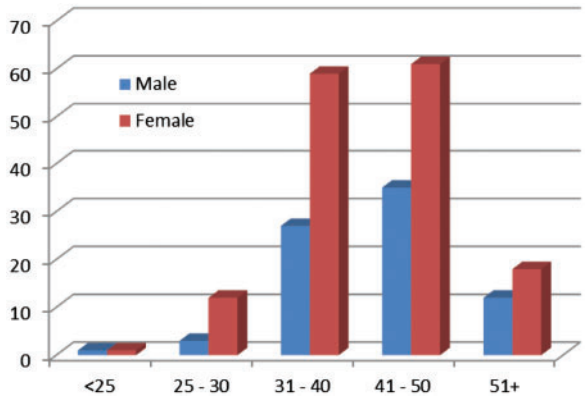

Figure 1. The age and sex distribution of survey respondents.

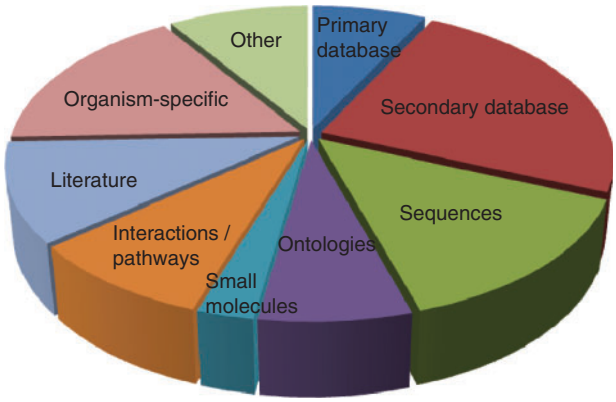

Figure 2. The types of data annotated by biocurators.

Table 1. Main motivations for selecting biocuration as a career

Motivations	%
Wanted to move away from experimental research	48
Intellectual challenge	43
Biocuration is essential for modern science	41
I needed a job	40
Natural transition from previous work	37
The diversity of the work was appealing	33
Previous biocuration experience	11
Knew other biocurators	9
Financial reward	6

biological ontologies. A little over a quarter (26%) of respondents worked with secondary (multi-organism, curated) databases [e.g. UniProtKB (7), IntAct (8), EMAGE (9), 18% with organism-specific resources (e.g. SGD; 10)], dictyBase (11), TAIR (12), etc.]; 16% were involved with sequence analysis; 12% were involved with literature analysis; only 3% worked with chemistry-based resources.

Most biocurators highlighted their appreciation of the team-working aspects of their jobs, with a large fraction collaborating on a regular basis with other scientists, mostly with other curators (78%), and with computer scientists and software engineers (75%); only a minority (17%) of respondents felt isolated in their work. In addition, many respondents were involved in user training/outreach of some sort, whether through posters and talks at meetings (55%), via responses to helpdesk queries (50%), preparation of documentation (45%), or delivery of face-to-face training (34%). For many, these and other activities were associated with some amount of travel, 41% of respondents travelling up to twice a year, and 29% more than 3 times per year. Most (65%) were satisfied with the amount of travelling they did: others (28%) said they would like more travel; 7% wanted less.

Biocuration as a career choice. Respondents were asked what motivated them to become biocurators, by selecting multiple options from a range of pre-set answers. Results are shown in Table 1.

Job satisfaction amongst the surveyed cohort appeared to be high. On a scale of 1–5, where 1 was not at all enjoyable, and 5 was very enjoyable, 77% rated their overall satisfaction at the level of 4 or 5. Moreover, 68% said that they are paid fairly. Nevertheless, only 6% felt that financial reward was a motivating factor in becoming a biocurator. Aspects of the role that particularly contributed to biocurators' enjoyment of their work included the intellectual challenges posed, as well as working extensively with scientific data. Complete results are presented in Table 2.

In the open responses, some respondents highlighted additional benefits they derived from their work, such as enhancement of their analytical thinking, improved ability to critique the literature, and honing of their ability to write concisely.

The survey asked biocurators which aspects of their current work they considered to be important, by ranking the statements on a scale of 1–5 (where 1 was not at all important, and 5 very important); respondents indicated that feeling intellectually challenged, contributing to the direction of the database on which they worked, and keeping abreast of current scientific developments were the most important aspects of their work. They were then asked to consider how strongly they felt that their current roles met these aspirations on a scale of 1–5 (where 1 indicated strong disagreement and 5 as strong agreement). The average answers to those two questions, as well as the differences between the importance of each aspect of the work and how biocuration meets those expectations, are shown in Table 3. Aspects where the differences were largest between expectation and whether biocuration met those expectations were the freedom to choose projects on which biocurators work, as well as recognition from other scientists.

Overall, most respondents derived a sense of accomplishment from their jobs (average, 3.89). Other perceived highlights included learning more and more, and the ability to work remotely, with the consequent lifestyle flexibility that this affords.

These largely positive responses are perhaps reflected in the desire of ~90% of respondents to remain in the field. Nevertheless, 82% expressed concern about future work opportunities, and 60% perceived that the lack of opportunities to move into more senior roles was also a barrier to remain in biocuration. Another concern related to

Database, Vol. 2012, Article ID bar059, doi:10.1093/database/bar059

Table 2. Rewarding aspects of biocuration work

Job aspect	Average rating	Number of 'Enjoyable or very enjoyable' ratings
Intellectual challenges and problem solving	4.39	187
Working with a wide range of scientific data	4.32	183
Working extensively with scientific data	4.21	181
A quantifiable sense of progress	3.87	144
Interaction with end users and data submitters	3.72	130
Scientific work that's not results-driven	3.33	94
Repetitive nature of day-to-day work	2.45	24

Table 3. Important aspects for job satisfaction, and how those aspects are met according to biocurators surveyed

Job aspect	Importance	Job meets expectation	Average difference
Feeling intellectually challenged	4.42	3.88	−0.45
Having an input into the overall direction of your resource	4.27	3.84	−0.36
Keeping abreast of current developments in your scientific area	4.14	3.81	−0.27
Autonomy over work	4.06	3.58	−0.39
Feeling part of a community of scientists	4.06	3.45	−0.48
Recognition from other scientists	3.86	3.30	−0.46
Feeling part of a community of biocurators	3.74	3.24	−0.39
Freedom to choose curation projects	3.73	3.05	−0.54
Freedom to conduct research outside of your core curation responsibilities - both curation-based research and other research	3.36	2.51	−0.70

The numbers presented represent the average score for each aspect, with 1 being the lowest and 5 the highest.

credit: although it was important to feel that their work was recognized by other scientists (average rating, 3.86), there was relatively low confidence that other scientists fully appreciate biocurators' work (average, 3.29).

What makes a good biocurator? Respondents were asked what attributes they thought were important for a biocurator to possess. On a scale of 1–5 (where 1 was not at all important, and 5 very important), respondents indicated that theoretical knowledge (average rating, 4.3), formal scientific training at degree level or above (average, 4.26), good written and verbal communication skills (average, 4.23), and previous experience as an experimental scientist (average, 4.04) were the most important attributes.

On this scale, formal training in data management (average, 2.79) and scripting/programming knowledge (average, 2.58) appear to be less important attributes. However, a significant number of curators did feel that software programming was important, 55% of respondents acknowledging that better training in computer languages would be beneficial, and 43% indicating that they would benefit from better training in bioinformatics. Other aspects that were perceived to be advantageous were improved software (66%), greater automation of routine tasks (61%), and greater adherence to community standards by data submitters (55%).

Career progression

Many biocurators have chosen this career as an alternative to a 'traditional' academic career. However, to attract and retain highly qualified candidates, it is important that opportunities for career progression exist. Although the numbers are far too small to allow us to uncover any trends, 20 of the respondents described themselves as principal investigators, 6 of whom had been in that position for <3 years, which suggests that there are some opportunities for more senior roles within biocuration.

Biocurators who have left the field

The survey also attempted to reach curators who have left the field, to try to gain some understanding of their motivations for doing so. Inevitably, it was challenging to

publicize the survey amongst former biocurators, as they are unlikely to visit the ISB website, or to remain on biocurator-related e-mail lists. Only 10 respondents no longer worked in the field: 9 were aged 41–50 years and held a PhD in a biological science; 6 were women. Two were offered better jobs elsewhere, and although three left because their jobs were no longer sufficiently challenging, five said they had enjoyed their work as a biocurator and five felt that the role had given them general transferable skills or specific skills for their current posts (e.g. critical assessment of publications, and analytical thinking). Three respondents had moved on to work in bioinformatics; four expressed a desire to return to biocuration in the future.

Challenges for the future of biocuration

Respondents were also asked to consider what are the main challenges to biocuration, both by selecting from multiple pre-set answers and with free-text responses. Most (78%) indicated that securing funding to maintain and develop biodatabases was the major threat, and many (71%) also considered that dealing with the increased volumes of data was a significant challenge. Emphasis was also given by many (57%) respondents to the difficulty of impressing on other scientists the importance and hence the need for funding of biocuration. Interestingly, 40% identified with the threat that biocuration might be perceived to be irrelevant if curators cannot keep pace with the current flow of data.

The role of the ISB

The final part of the survey reflected on the role of the ISB in promoting biocuration. Respondents were asked to select those activities they considered most relevant from a range of pre-set answers. Most (88%) felt that the ISB should engage with funding bodies to promote the importance of curation; 80% highlighted the need to engage with journals to encourage the adoption of standard nomenclatures; many (60%) also felt that the ISB should seek out and publicize employment opportunities. Half of the respondents indicated that the ISB should organize and secure funding for regional meetings for curators.

Discussion

Although we recognize that the form of the survey and its results do not lend themselves to rigorous statistical analysis, it has nevertheless yielded some important insights from a fraction of the biocurator community, with respect both to their views on biocuration as a career and to their perceptions of the role of the ISB. Based on the snapshot this survey provided, the career outlook for biocurators seems broadly positive, with high levels of job satisfaction. Respondents generally felt that they benefited from the challenging and problem-solving aspects of their work, yet many highlighted the repetitive nature of the day-to-day job; it is not surprising, therefore, that many respondents highlighted the need for better and more-automated curator-assistant tools, and felt that better training in bioinformatics and software programming would be valuable. Perhaps inevitably, there were concerns about career structure and progression, including the availability of more senior roles and the likelihood of being able to progress into them.

Most of the active biocurators who responded to the survey were >30 years of age. This is consistent with the fact that most respondents held PhDs and had not entered the field directly from their studies, but had held previous posts as bench scientists. This prior experience was clearly considered to be an important attribute for biocurators to possess. Despite the prevalence of contract work, many respondents held permanent posts, and a substantial number had been involved in the field for ≥ 7 years. Biocuration thus appears to lend itself to greater career stability than other scientific fields: the average contract length established for a similar demographic population by the Vitae UK Careers in Research survey was under 3 years, with a majority of those respondents remaining at institutions for <5 years (13).

A secondary aim of this survey was to solicit feedback on curators' perceptions of the role of the ISB in advancing biocuration as a career. Importance was attached to engagement with journals to promote the adoption of standard nomenclature, echoing the view expressed earlier in the survey that adherence to community standards by data submitters would facilitate biocurators' work. The ISB has made substantial commitments to such activities, in collaboration with the BioSharing initiative (14; www.biosharing.org), operating at a global level to build stable linkages between journals and funders, and implementing data-sharing policies and standardization efforts in the biosciences. Members of these two groups have worked in close collaboration with publishers and journals (e.g. Elsevier, Nature Publishing Group, F1000, *Nucleic Acids Research*, *Database*), to develop the BioDBcore standard (15), a proposed uniform system for describing catalogues of databases. Progressively, such efforts will help users to more easily locate and access information dispersed within bio-resources; help shape the data-preservation, data-management and data-sharing policies implemented by journal editors and funders; and encourage software and database developers to embrace and extend community-endorsed standards. In a concrete step towards this goal, BioSharing and the ISB held a workshop at the ISMB meeting in Vienna, in which several journal editors and standards groups stated their commitment to widen participation in, and expedite the implementation of, data-sharing and nomenclature policies

(http://blog.biosharing.org/2011/07/biosharing-at-ismbeccb-2011-vienna.html).

Many respondents also felt that the ISB had a role to play in publicizing employment opportunities and providing opportunities for biocurators to interact with each other. ISB maintains a biocurator job market forum on its website (http://biocurator.org/jobs.shtml), and regularly notifies members of job opportunities via its email lists, through social-networking sites, and the publication of its monthly newsletter (http://biocurator.org/newsletter.shtml). In addition, an international biocuration conference has been held roughly every 18 months since 2005. The ISB has made the support of these conferences part of its mission statement, in order to continue to provide a venue for biocurators and programmers to exchange ideas, discuss their work, improve their methods, and establish collaborations. The Fifth International Biocuration Conference will be held in Georgetown, USA, 2–4 April 2012 (http://pir.georgetown.edu/biocuration2012.html).

Perspective

One concern highlighted by the survey is the possibility that biocuration might, in the future, become irrelevant if biocurators cannot keep up with the onslaught of data. A closely allied fear expressed by a few respondents is the emphasis placed on automatic annotation, and the sense that '[manual] biocuration is meant to be replaced by automated processes'. Given the difficulties of securing funding to support the growing numbers of databases and curators who maintain them, these fears are perhaps understandable. With the pace of data-generation on course to be a million times greater than at present by 2020, there are clearly significant challenges ahead for biocurators. However, without question, the new reality of biological research both demands expert biocurators now in order to make sense of the data deluge, and it assures their role in future, whether at dedicated resources or within research projects; it also argues strongly for continued technological innovation (through deployment of appropriate software, controlled vocabularies, plus data and nomenclature standards) to ensure appropriate use of computers for monotonous high-volume data-processing tasks, releasing biocurators to tackle the current and future intellectual challenges of data management, analysis, interpretation and validation.

It is unfortunate that manual and automatic processes should be considered in opposition, as excluding or superseding each other, or pictured as posing threats to each other. Although many aspects of biocurators' work depend on computation and automation, the development of new tools absolutely requires biologists and bioinformaticians to validate the methods, provide validation tests and ensure their overall usefulness for the community. In addition, several tasks in biocuration can only be performed manually: for instance, the creation of gold standard data sets, and the development of new tools and data models to handle new data types. The research and biocurator communities must work together to ensure that the maximum benefit can be derived from all experimental data being produced. Education of the community on meta-data tagging of data sets, and development of tools to assist with this task, could go a long way to maximizing the utility of data to other researchers. As new areas of biology are explored and new experimental methods are developed, the specific tasks carried out by biocurators may change, but the underlying goal of interpreting, organizing, and making data easily accessible for hypothesis generation and testing will remain essential. The challenges that lie ahead for the biocuration community are not only large, but are also extremely stimulating. We hope that the field will continue to attract innovative and far-sighted scientists to further bridge the gap between data and researchers.

Acknowledgements

We gratefully acknowledge the participation of all the biocurators who responded to the survey.

Funding

Funding for open access charge: International Society for Biocuration.

Conflict of interest. None declared.

References

1. Galperin,M.Y. and Cochrane,G.R. (2011) The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, **39 (**Suppl. 1), D1–D6.
2. Southan,C. and Cameron,G. (2009) *Database Provider Survey. Report for ELXIR Work Package 2, July 2009, Version 30 June 2009*, http://www.elixir-europe.org/prep/bcms/elixir/Documents/reports/WP2_Annex-Provider_Survey_Report.pdf.
3. St Pierre,S. and McQuilton,P. (2009) Inside FlyBase: biocuration as a career. *Fly*, **3**, 112–114.
4. Burkhardt,K., Schneider,B. and Ory,J. (2006) A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLoS Comput. Biol.*, **2**, e99.
5. Salimi,N. and Vita,R. (2006) The biocurator: connecting and enhancing scientific data. *PLoS Comp. Biol.*, **2**, e125.
6. Sanderson,K. (2011) Bioinformatics: curation generation. *Nature*, **470**, 295–296.
7. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**(Database Issue), D214–D219.

8. Aranda,B., Achuthan,P., Alam-Faruque,Y. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**(Database Issue), D525–D531.
9. Richardson,L., Venkataraman,S., Stevenson,P. *et al*. (2010) EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res.*, **38**(Database Issue), D703–D709.
10. Engel,S.R., Balakrishnan,R., Binkley,G. *et al*. (2010) Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res.*, **38**(Database Issue), D433–D436.
11. Gaudet,P., Bairoch,A., Field,D. *et al*.; on behalf of the BioDBCore working group (2011a) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, **39**, D7–D10.
12. Swarbreck,D., Wilks,C., Lamesch,P. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**(Database Issue), D1009–D1014.
13. Mellors-Bourne,R. and Metcalfe,J. (2009) Careers in Research Online Survey (CROS) 2009: analysis of aggregated UK results.
14. Field,D., Sansone,S.A., Collis,A. *et al*. (2009) Megascience. 'Omics data sharing. *Science*, **326**, 234–236.
15. Gaudet,P., Fey,P., Basu,S. *et al*. (2011b) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. *Nucleic Acids Res.*, **39**(Database Issue), D620–D624.