SP Capacities - Research Infrastructures

Project no. 211601

ELIXIR

European Life-science Infrastructure for Biological Information

Preparatory Phase Project - Combination of CP-CSA

Start date of project: 1st November 2007
Duration: 50 Months

D2.1: Database Provider Survey report for ELIXIR Work Package 2

Due date of deliverable: 30th Jun 2009
Actual submission date: 30th Jun 2009

Workpackage WP2, Task n.:T2.21
Organisation name of lead contractor for this deliverable: EMBL-EBI

Version 1.0

Project co-funded by the European Commission within the Seven Framework Programme		
Dissemination Level		
PU	Public	√
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Database Provider Survey

Report for ELIXIR Work Package 2

Christopher Southan ELIXIR Database Survey Coordinator EMBL-EBI

Graham Cameron, Associate Director EMBL-EBI and Chair of Work Package 2

Contents

INTRODUCTION

Data Resources are the focus of ELIXIR Work Package 2 (WP2). Standards, annotations and tools are explored in other work packages. These data resources are the foundation on which research and applications in the life sciences increasingly depend. It has become clear that criteria are needed for judging the necessary support model for these resources. Such criteria are also essential for the purposes of comparison for example between core resources, which aim for completeness, are standardised and universally used against specialist resources, which serve a more limited community, but are nevertheless of vey high value. Currently core resources usually receive their major funding through European or international support, although they may also receive some national funding. In contrast specialist resources are usually funded nationally, normally from research funds, or through small Commission grants to establish them, but not to maintain them. Some of these specialist resources merit longer term support and closer network integration with the core resources. There is therefore a pressing need for criteria both for prioritising established investments as well as starting new ones.

Core biomolecular resources in Europe include those for nucleotide sequences and genomes, protein sequences, protein structures, protein-protein interactions and expression data. These data resources are mainly based at EBI, though several involve major collaborations with partners elsewhere in Europe (e.g. the Swiss Institute of Bioinformatics). In contrast, specialist data resources are widely distributed and are complementary to the core databases. As a key component of WP2 the Database Providers Survey was conceived to assess, in the most comprehensive way we could manage, the *status quo* across biomolecular databases in Europe. The survey was designed to gather information about content, operational details, standards, usage, funding, team size, and sustainability. The provision of this information both qualitatively and quantitatively where possible is crucial to inform future ELIXIR planning and to help define processes associated with long term funding, e.g. the possible transitions of selected databases from non-core to core. The Data Provider Survey results also compliment the User Survey results from WP3. The general topic of database sustainability has been the subject of a number of publications and reports, both prior to ELIXIR as well as some appearing in 2008/9. These publications have been reviewed in **Appendix I** (Database_literature_review_May09.doc)

ASSESSMENT OF DATABASE NUMBERS

Identification and Scope of ELIXIR-relevant databases

A number of basic questions arose for planning the survey of data resources. The principal ones were;

- What types of resources are within scope?
- How can we identify these?
- How many are there?
- Which ones are relevant for ELIXIR consideration?
- How can we compile e-mail lists?

The predominant type of resource to be considered by WP2 are bioinformatics databases. We can obviate the necessity for any theoretical specification on what these are by using a pragmatic definition i.e. that they conform to the collective characteristics of those resources published in the *Nucleic Acids Research* (NAR) annual database issue (see the 2008 volume). While the focus of ELIXIR is clearly on the utility of resources rather than any strict classification, in September 2008 the WP2 committee endorsed the biomolecular focus and cautioned against over-expansion, e.g. into the clinical arena, but recommended the inclusion of public domain bioactive chemical resources.

It is thus useful to define exclusions even if these are pragmatic considerations e.g. certain resources cannot be surveyed by questionnaire. On-line supplementary data from journal articles would fall into this category. They can have some technical characteristics of databases but are general one-off specific compilations that are neither updated nor have the developed interface that of a queryable one-line resource. However, mechanisms by which this supplementary data can be aggregated into databases are relevant.

Metadatabases, i.e. aggregates of individual databases under a common front-end, are also excluded in cases where these databases exist in their own right. The NAR db issue also includes a few "databases-of-databases" the utility of which are mentioned later, but cannot be surveyed. The position of Wiki-based resources of biological information is less clear. While they do not have the underlying schema or advanced query options characteristic of databases they are becoming increasing important as repositories of annotated data types.

Efforts to count data resources in the literature (reviewed below) clearly show the predominance of what can be generally classified as clinical

databases. These can vary is size from small specialist case collections up to national health information repositories. The WP2 committee agreed for these to be out of scope, particularly because of their potential to swamp out bioinformatic resources. However they also suggested a watching brief for emerging biomolecular and/or ontological connectivity between clinical and bioinformatic databases e.g., in the areas of biomarkers and translational research.

Compilations and Growth Rate

There have been many attempts in the past to maintain bioinformatics database and tool compilations but almost all have given up any pretence at being current or complete. One of the earliest of these, Pedro's BioMolecular Research Tools gave up in 1995 (but the URL still exists). The latest to (probably) have given up is the Database of Databases (PMID 18188423) that has not made it beyond 2007. The largest single source that appears to be updated is the Online Bioinformatics Resources Collection (OBRC) hosted by the Health Sciences Library System at the University of Pittsburgh, This contains annotations and links for 2394 bioinformatics databases and software tools (see PMID 17108360). In fact this draws on two better known compilations, the Bioinformatics Links Directory and the Nucleic Acids Research online Molecular Biology Database Collection (MBDC). While the former is focused on tools the latter lists publicly available databases described in the Nucleic Acids Research annual database issues, as well as a selection from other journals. The 2009 update includes 1078 databases, 192 more than 2008.

For the ELIXIR database provider survey we have used the MBDC compilation as our core listing. The reasons are described in the editorial for the 2009 collection (PMID 19033364). These include its expert curation, update frequency, non-redundancy, peer-reviewed resources, classification system, affiliation information, contact e-mails, the courtesy of Dr Galperin in providing selected list extractions, and, apart from some aggregated reports, most of the individual databases are within the

ELIXIR scope. The growth rate of these compilations is shown below.

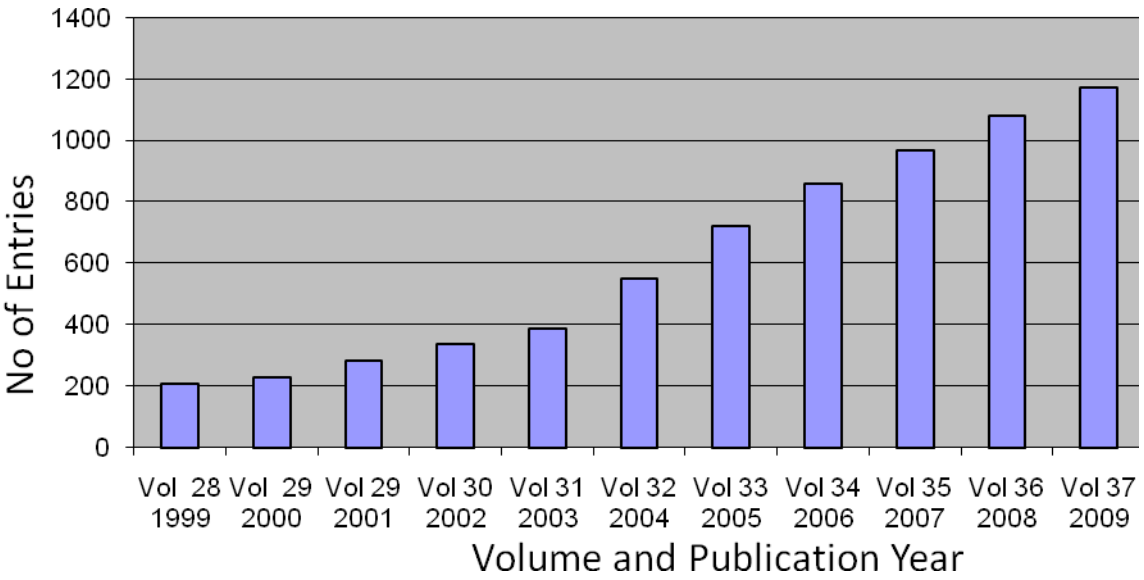

Figure 1. Database totals for the MBDC from 1999 to 2009. The latest entries suggest an annual growth rate of approximately 11%.

Affiliation Selection

ELIXIR-affiliated countries include members of the European Union plus Norway, Israel and Switzerland. Because of the global nature of bioinformatics collaborations ascribing affiliations is not always straightforward. The inclusive option was taken of either using the URL domain or the database contact e-mail domain to assign affiliation. From the 2008 collection this gave 410 for ELIXIR countries and 722 other i.e. 36%. This listing was expanded during the survey, primarily due to new databases being published over 2008/2009 (see post-survey compilation below. The final distribution for the ELXIR countries is shown below

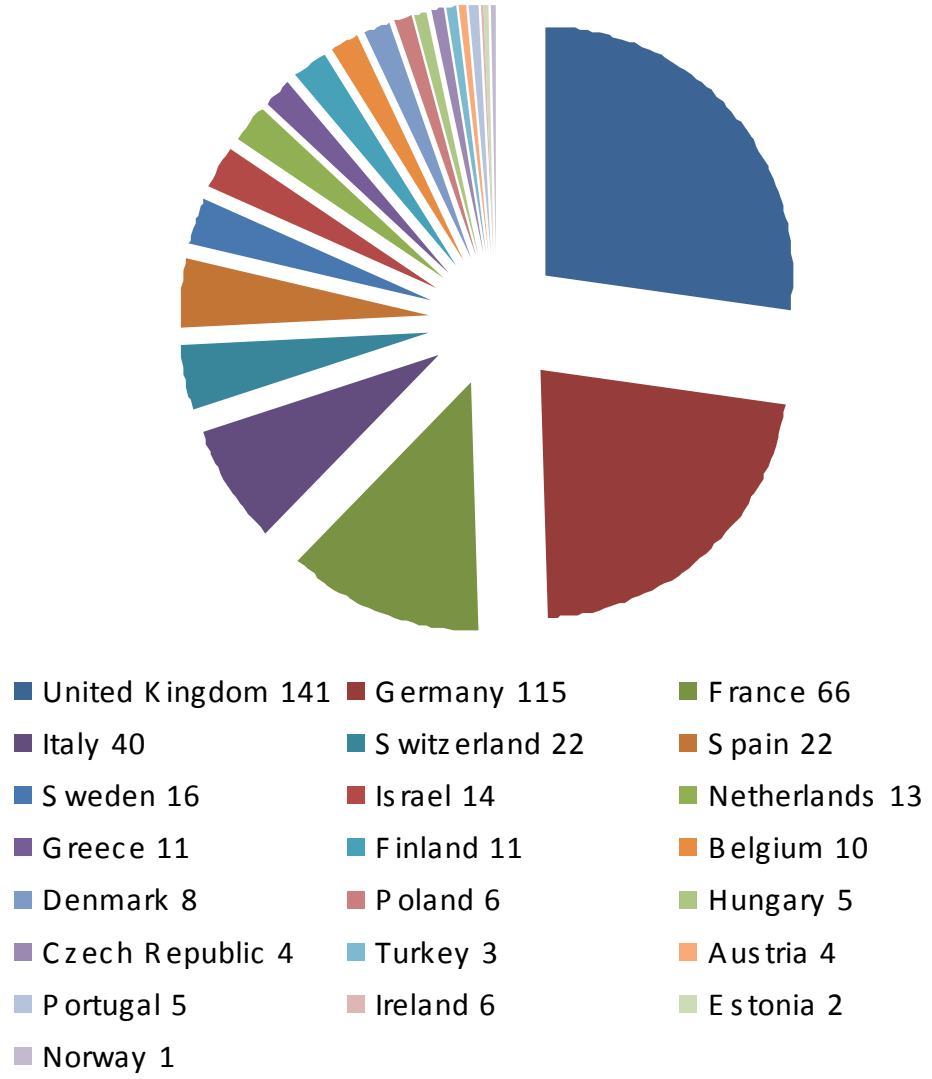

Figure 2 Distribution of surveyed databases by ELIXIR country

Examination of the broad data type classifications used in the MBDC collection (courtesy of Rafael Najmanovich) indicated no significant differences in distribution between ELIXIR and non-ELIXIR countries.

Database Counts in PubMed

Because the MBDC does not claim to capture all published databases searches in PubMed were explored to see if more databases could be retrieved and what proportion might be ELIXIR-relevant. An obvious first step was to search with the word "database" in the title and this was combined with "published in the last 10 years". At just over 6000 matches this exceeds MBDC by 5-fold. Even a cursory inspection shows the usual problems of specificity and recall that concerns false-positives and false-negatives, respectively. Dealing with specificity first it was clear the false-positives, in the ELIXIR context, were predominantly

clinical databases, a good example being "Hoof kick injuries in unmounted equestrians. Improving accident analysis and prevention by introducing an accident and emergency based relational database" (PMID 12421795). Arguably "Oral contraceptives and venous thromboembolic disease. Analyses of the UK General Practice Research Database and the UK Mediplus database" (PMID 10652979) is of more biomedical relevance but still out of scope.

On the recall side the database/title query was effective at identifying Epubs ahead of print from the 2009 NAR annual database issue that would eventually be captured in the MBDC. It was thus clear that a journal filter would improve specificity without too much loss of recall. From a manual inspection of the "sort by journal" it was straight forward to pick the highly represented journals and cross-check with an individual journal search. The top-ten with database title hits (as of Nov 2008) are shown below in fig.3

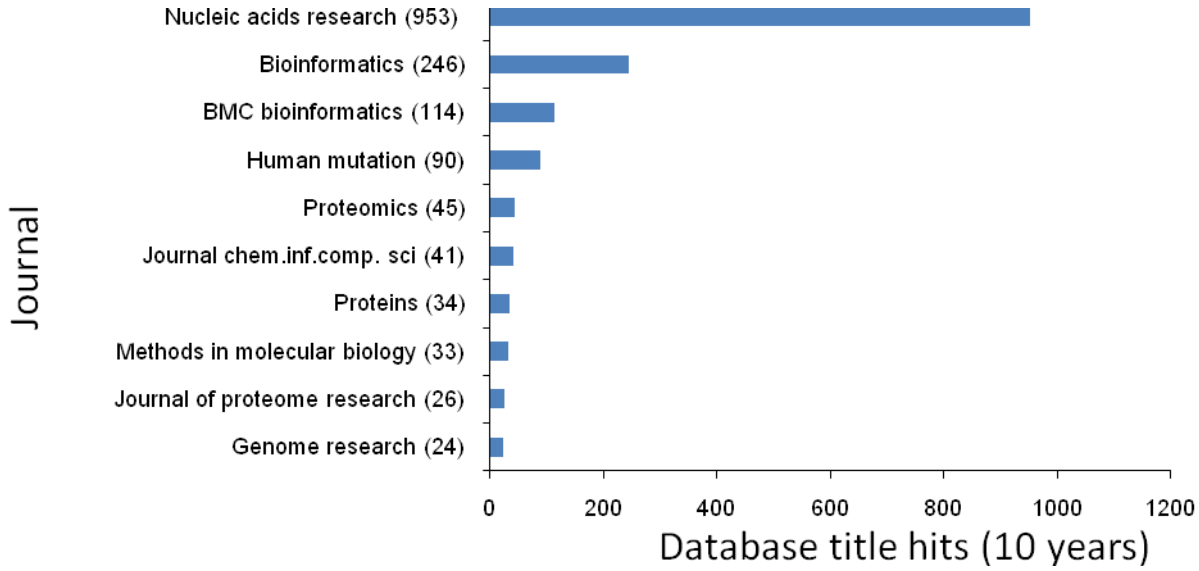

Figure 3. Top-ten biomedical journals with database in titles over the last 10 years.

In fact three of these journals, *Proteomics*, JPR, and JCIM showed a low ELIXIR relevance in their database articles so, in the final journal filter these were substituted with *Plant physiology, in silico biology* and *Gene* which each had over 20 title hits. On its own the 10-journal filter gave over 100,000 articles but combining this with the database title and 10 year cut gave 1705 i.e. about 50% larger that the MBDC collection. Inspection of this listing indicated an approximate 5% false-positive rate, most of which came from use-of-database articles and repeat publications e.g. both PROSITE and Pfam (and database in title) each have 6 hits of both types.

By definition finding the false-negatives to estimate recall is more difficult. From searches on the NAR journal site it was established that the 2008 database issue had a 45% false-negative rate for database in the title i.e. 84 of the 185 articles would not be recalled by this query. While a further 69 would have been recalled with "database" in abstract but not in title this had an unusable low specificity in PubMed. Examples of database title false negatives include "The Universal Protein Resource (UniProt)" "Gene3D: comprehensive structural and functional annotation of genomes" and "Ensembl 2008". While these would be included in the MBDC and therefore not lost to ELIXIR, it does expose a significant false-negative rate for database title search. However, there is some indication that outside the NAR database issue this false negative rate is lower than within it.

The final filter used in this assessment was by using an ELIXIR-affiliated countries list to search the affiliation field. The combinations and figures for June 2009 were as follows

1. Database in title field (all years) = **8816**
2. Database in title – last 20 years = **8545**
3. Database in title – last 10 years = **6330**
4. In the top-ten journals for database relevance = **105447**
5. Having at least one ELIXIR affiliated country affiliation = **830298**
6. 3 AND 4 – all bioinformatics dbs = **1715**
7. 3 AND 4 AND 5 – all ELIXIR-relevant bioinformatics dbs = **609**

Thus, by affiliation, the ELIXIR proportion was 35% i.e. in close agreement with the 36% calculated from the MBDC. The PubMed approach clearly has a lower specificity but could be used to find ELIXIR-relevant databases that are not yet in MBDC. The combined query in 7 was run as a monthly alert during the course of the survey.

These queries, including the journal restriction, were run retrospectively by year thus making them comparable to the annual figures for the MBDC compilation. The results are shown in fig.3 below.

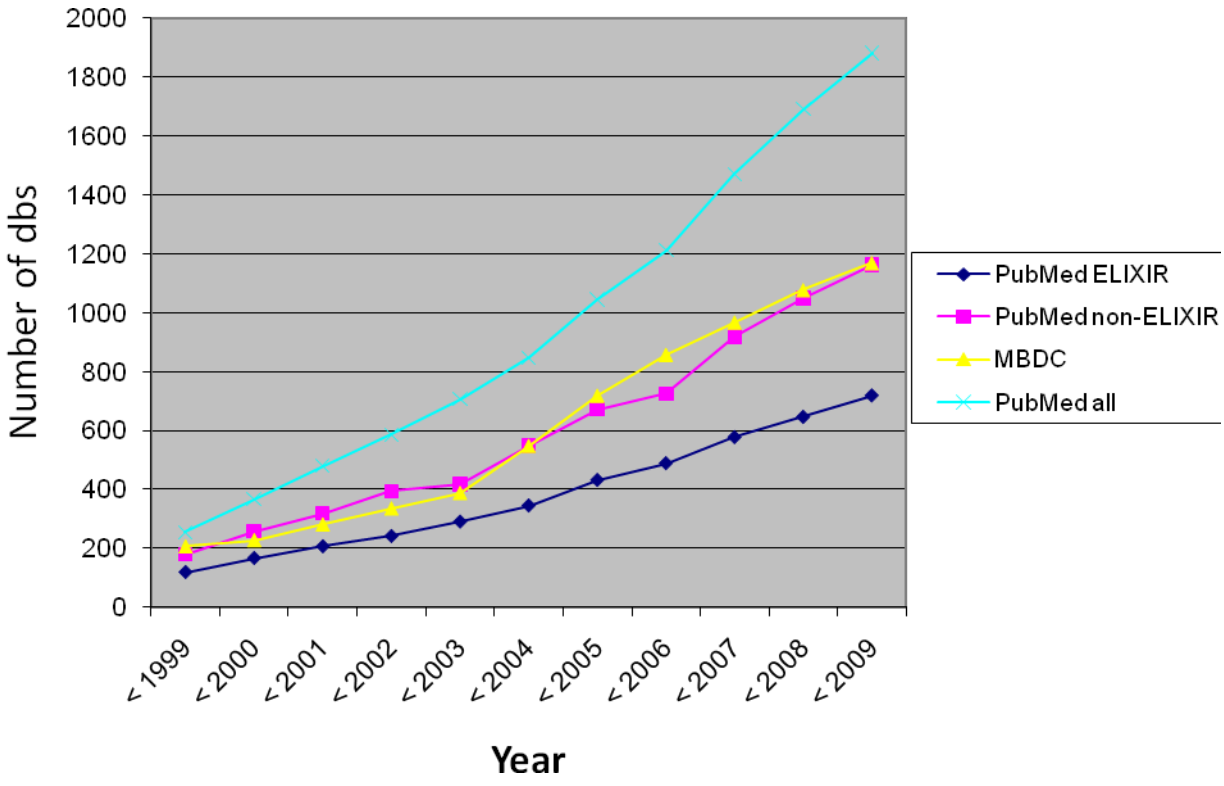

Figure 4. Comparison of database numbers from PubMed queries and the MDBC.

This shows that the growth of ELIXIR-affiliated databases is slightly lower that the overall rate.

While there has been no explicit exclusion of unpublished data resources from consideration by ELIXIR there has been a *de facto* focus on published databases. This is not only because of the primacy of peer-review but also because there is no comprehensive index of unpublished ones. Estimates have been made that the numbers for these could be as high as those for published ones (Galperin, personal communication). As is reported below we encouraged the dissemination of the survey but the returns from unpublished databases were low.

Database Mention Counts in Medline

A complimentary approach to assessing database numbers was carried out by Dietrich Rebholz-Schuhmann and Antonio Jose Jimeno Yepes (EBI) using queries of Medline. The first stage of this was to run the search term "database" and then use proximity-based syntactic rules to parse out the database titles. This was done for the whole of Medline to give maximum recall (termed "all" in the charts below) and then filtered for a list of biomedical informatics journals (termed "bm" in the charts below).

This approach captures the first mention i.e. when, where and who publishes the database for the first time. Subsequent mentions of the database name can then be tracked as a citation (used in this context as a mentioned occurrence in a Medline title or abstract). European was distinguished from non-European by the country affiliation of the first author. The results of this "first mention" search are shown below.

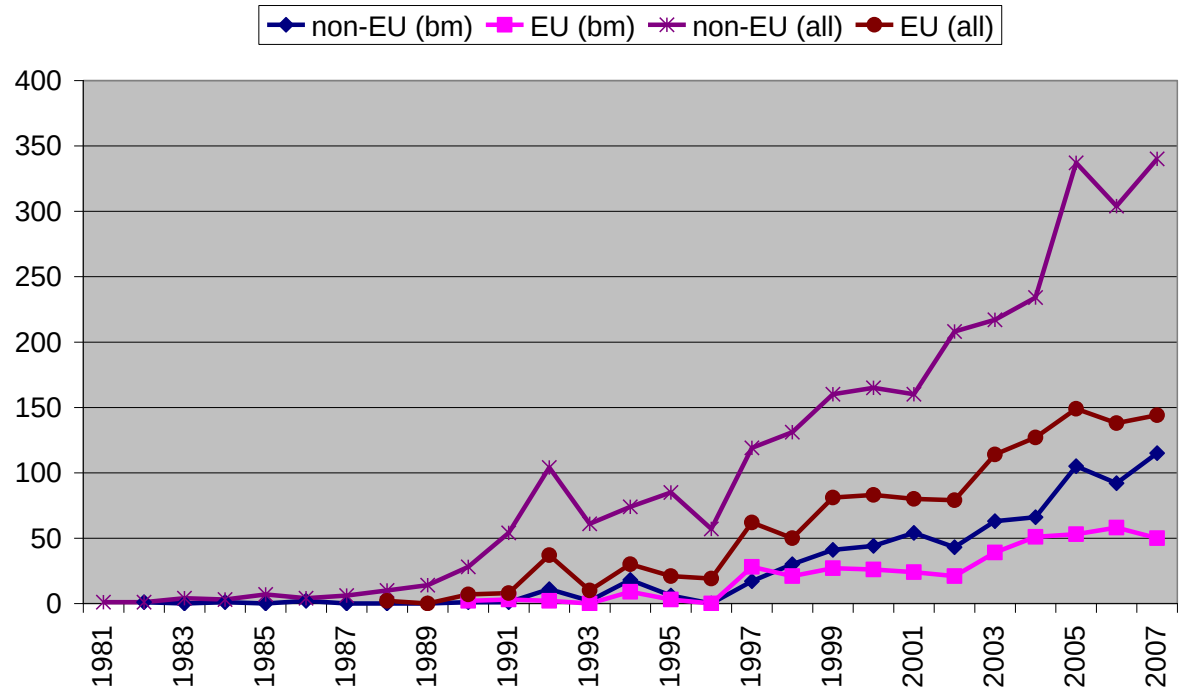

Figure 5. Plot of the number of databases by first mention

This search extracted 6000 database names but 5000 of these were post-1980. Of these 2390 (39%) are European, in good agreement with the ratios given above calculated from PubMed and MBDC. Evidence for the retrieval specificity was that 731 results came from NAR. From identification by first mention database names can be tracked by subsequent citations as shown below.

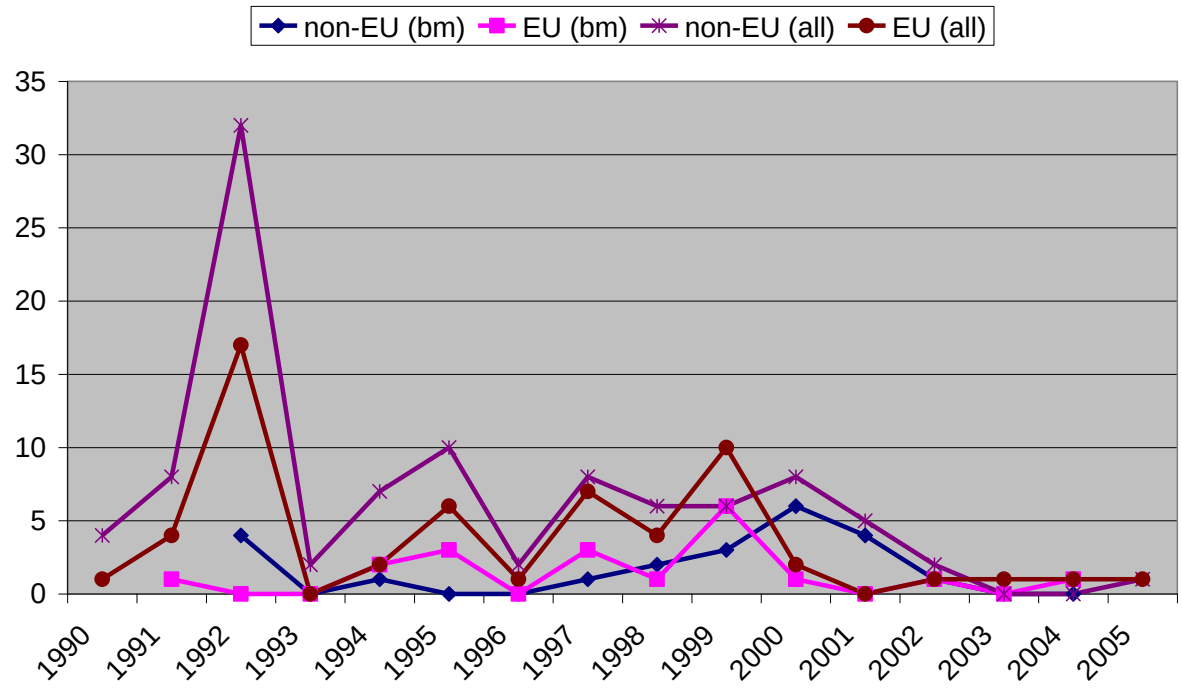

Figure 6. Databases with first mention and at least 20 citations

The trends show a "burst" of highly cited databases occurring in the early '90s and emphasises the long lag time necessary to accumulate citations.

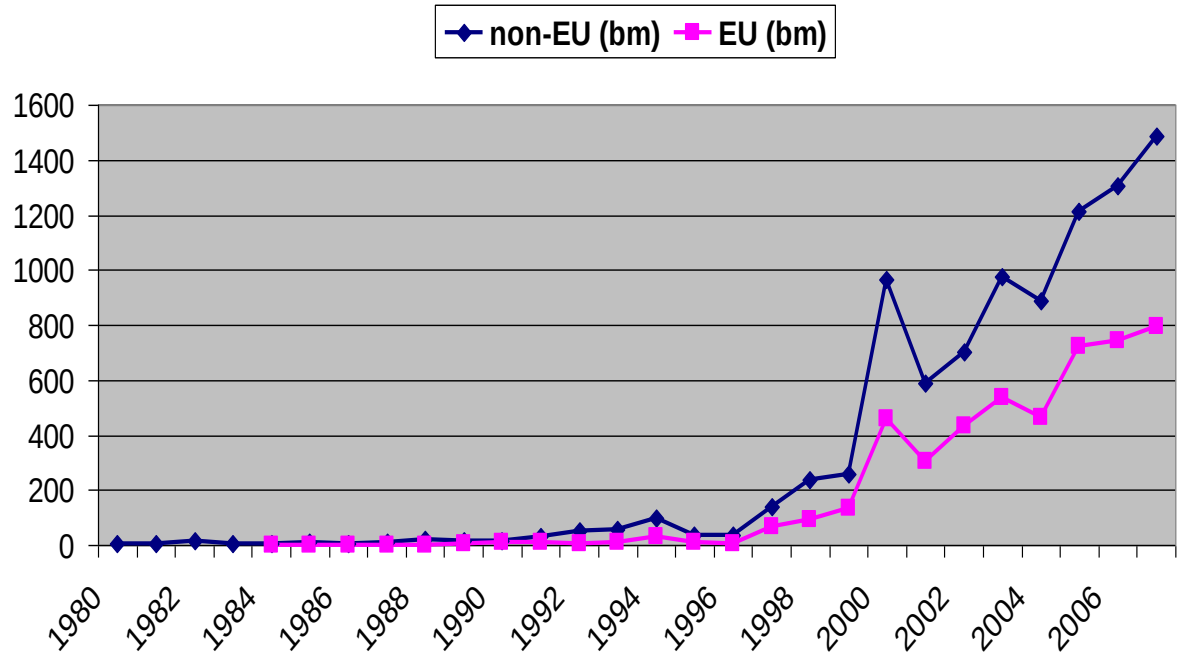

Figure 7. Number of citations for all databases.

The conclusions of this text mining study were that the number of biomedical databases cited in the scientific literature has grown strongly since 1999 to today, almost all European countries have provided some of these in the last 15 years and that they now form 40% of the total. In addition those with at least 20 mentions make up 80% of all mentions in Medline. Thus, these three different approaches to assessing the number of ELXIR-relevant databases I) analysing the MBDC, II) PubMed queries and III) text mining in Medline, gave broadly congruent and complimentary results. These were:

1. The number of ELIXIR-relevant databases published over the last 10 years was approximately 600, although the propensity for both false positives and false negatives preclude an exact count.

2. Those not including "database" in the title, together with a long tail of journals, suggest this is an undercount.

3. The proportion with ELIXIR national affiliations is approximately 36%

4. The current growth rate of approximately 12% per year is slightly lower than the overall increase in non-ELIXIR countries.

5. The majority of European countries have published databases

6. Databases with over 20 mentions in Medline tend to have been published in the early 90's and account for 80% of all mentions.

SURVEY DEVELOPMENT

Evolution of the Questionnaire

The exercises outlined above were direct towards one of the main objective to collect detailed information about existing databases and resources within ELIXIR-affiliated countries by conducting an on-line questionnaire. Due to its already proven utility for EBI-outreach activities and access to a professional account, this was conducted using the SurveyMonkey tool (http://www.surveymonkey.com). The outline plan was as follows:

- Frame a set of questions to capture ELIXIR-relevant data and metrics for reviewing data sources
- Iterate several versions of the survey through a local expert group
- Pilot this survey with a small set of database providers
- Analyse the results of the pilot
- Optimise the final questionnaire based on pilot feedback, WP2 committee input and another round of iteration with local experts
- Distribute the final questionnaire to ELIXIR-relevant e-mail contacts
- Circulate reminders and field any technical queries
- Return pilotee's results if requested to save them time on the 2nd response
- Analyse the results and review in the context of other relevant information
- Present to, and seek feedback from, WP2 committee
- Compile a spreadsheet of live URLs and correct contact e-mails for respondent
- Distribute summaries back to respondents
- Incorporate results and inferences into interim ELIXIR documentation
- Decide on necessary follow-up analysis or further data-collection
- Prepare a summary for publication

Challenges

This undertaking presented a number of constitutive challenges that can be summarized as follows:

- It was necessity to define the relevant questions and metrics before the implementation process of ELIXIR decisions become clear
- Mastering "Survey Monkey " question design, collection parameters and result analysis
- Balancing the depth and breadth of the questionnaire against respondents' knowledge, compliance and sensitivity (especially for funding questions)
- Assessing funding, resources, value, impact and usage of databases via self-reporting
- Defining the limits of which resources to include
- Extracted e-mail contacts from MBDC not being current
- Mining the comment fields
- Content and format of reports and presentations
- Deciding follow-up, gap-filling and complimentary data collection (e.g. a standardised citation analysis)

A number of strategic decisions were made during the pilot phase and for the final iterations. A covering letter explaining the context of the survey and some guidelines for completing it was careful drafted and endorsed by Janet Thornton, Graham Cameron, and Andrew Lyall. While there are many options in SurveyMonkey to constrain responses to produce cleaner data no questions were set as compulsory either by necessity to answer or forcing format compliance. This does allow noise to creep in and increases the need for data clean up but this is balanced against minimising irritation for respondents.

In balancing increased length against the possible consequences of a lower response rate we opted for the longer set of questions for a number of reasons. a) given the investment in the undertaking the return of a smaller number of more in-depth responses was deemed preferable to a larger number of "lite" returns., b) whilst the database contacts are doubtless very busy they would not be expected to be a particularly "survey fatigued" community, c) the depth and scope of questions conveys an implicit credibility and seriousness of intent, d) there were no complaints about being "too long" from the pilot e) the inclusion of "other" options and open comment fields gave respondents an opportunity to "air their views" that they may actually appreciate, f) it is

conceivable that the survey may be re-cycled in some form e.g. to be used outside Europe or as basis for a tool-provider survey.

Messages

In accordance with the principle that most attempts to observe a system also perturb it, it became apparent that the questionnaire encodes a number of messages. The first was of course increasing ELIXR awareness in the database community and the answer to question 4 (see Results section) showed this had been effective. Other messages arise because any type of "do-you-do-this?" question is an implicit recommendation that this might be a good idea e.g. "do you have web services?" and "will you have them in 12 months?" not to mention "how often do you update". Similarly the whole set of questions on standards and connectivity give a clear message on their future importance. Last but not least several respondents gave us informal feedback that they appreciated the stimulus given by the survey for them to "step back" and prepare a detailed overview of what they were about.

The Pilot Survey

The version 1.9 of the questionnaire was used as pilot survey; this was sent out to predominantly EBI and Sanger Centre databases but also a spread of other institutions. At the end of June 50 were sent out. By the end of July 28 had been initiated and reminders were sent to non-respondents. There were 31 completions before closing the pilot survey at the end of August i.e. a 65% completion rate. The geographic responses were UK 20, Germany 6, France 2, Belgium 2, Denmark 1, Netherlands 1, Norway 1, and Greece 1. The main survey now supersedes the pilot and is expected to generate different statistics but the following points were extracted from the pilot

- 85 % incorporate hand-curation or manual annotation
- 25% of Web Services/programmatic access not documented
- 40% plan to introduce web services/programmatic access
- 30% skipped the funding metrics
- 60% reported closely-related databases
- Impact assessment split between citations, peer review and user feedback
- There was no clear consensus on usage statistics but hits will do
- High "honesty response" of 80% reporting user problems
- 70% would like more reciprocal links
- 80% planed to incorporate new data sources
- 40% had major funding concerns

Optimisation and Distribution of the Final version

Feedback from the pilot was invaluable for optimising the final version. Highlighted ambiguities and redundancies in questions were removed and certain sections were merged or cut. The most significant change was from open numerical options to ranged bins or dropdown menus, e.g. rather than "we get [5367] – hits per month" the reply option is "hits per month are: 2K to 5K or 5K to 10K" etc. The former is more precise, at least where respondents have exact numbers but always requires data clean-up for inconsistencies in units or spacing. The latter is much easier to fill in and see the ranges in advance but there is a loss of precision for any arithmetical processing of the results e.g. for summing or averaging. The final full list of 72 questions, including the introduction, in the format seen by the respondents is available in PDF format from the ELIXIR website.

http://www.elixir-europe.org/files/documents/ELIXIR_provider_survey_2.2.pdf

There is also a summary list of questions (as used in the results section below) as a word file

http://www.elixir-europe.org/files/documents/ELIXIR_Database_Providers_Survey_Question_List.doc

Parsing the database titles, URLs, country of origin and e-mails from the 410 ELIXIR-relevant databases extracted from MBDC (courtesy of Rafael Najmanovich and Andrew Lyall) produced a list of 383 with e-mails From these a list of 327 unique e-mails was extracted i.e. there were some where the same contact was given for more than one database. This list was merged with the pilot e-mail list of 48. An update search in PubMed, carried out as described above added 54 new databases either new since Jan 2008 or from journals other than NAR. This produced a final de-duplicated list of 377 uploaded to Mailman. A covering mail was prepared, copied from the introduction page of the questionnaire and the SurveyMonkey URL pasted in. This link was also put onto the ELIXIR website along with a link to a PDF and Word copies of the questions. This gave respondents the opportunity to peruse and prepare answers to the questions before filing it in on-line. The option was set for re-entry from the same IP address i.e. even after completion to allow updates as long as the survey was left open. An additional communication was sent to the pilot addresses offering to copy back their replies from the July pilot to save time filling it in a second time, although they were cautioned that some questions had been changed for the final version. The first tranche

of 360 invitations to complete the survey were sent out on October 20th 2008. This list was supplemented with both new dbs published between 3Q 2008 and 2Q2009 and, with the inclusion of respondents that had not been circulated (i.e. had received it by pass-along or picked up from the website) reached a final number of 516 contact e-mails. Reminders were sent out approximately quarterly. The survey was closed on April 6th 2009.

Post-Survey ELIXIR Database Status listing

Both as a response to requests by survey participants and because of its utility for assessing the survey results it was decided to make a complete list of all the dbs circulated and that returned surveys. These were extracted from the following sources

- The 2008 MBDC collection used for the original e-mail list
- New publications appearing in 2008/9 picked up via the PubMed queries described above
- Dbs not included in the above sources but were captured from the survey returns; either via pass-along or having been picked up directly from the ELIXIR website

The merged list was checked for the following:

1. Removal of duplicates arising mainly from merging the survey response data with the original lists. These rarely returned identical wording on descriptions, names or URLS (e.g., with or without http ://). Thus the initial merge of 640 had to be manually de-duplicated to 509
2. Checking if the URL was live, dead, moved, down for maintenance, or had a re-direct
3. Printing off a front page for a hard-copy compilation
4. Searching for a town name that could be converted to lat & long for automated map display. In many cases this was only possible by following links to the publication.
5. Ascertainment of update status. This necessarily had to be capped at a few minutes search on the front page and following obvious links such as "news", "statistics" or "version history". Frequently the update status could simply not be divined from the web pages.
6. Ascertainment of ELIXIR relevance, e.g. that it was a db rather than a portal, tool or one-off data set
7. Recording of response status from merging in the details from completed survey responses

This compilation is a valuable resource from a number of aspects.

- The information captured constitutes an unbiased survey in its own right with important orthogonality and complementarity to the questionnaire returns
- Such a listing was requested from a number of sources including survey respondents.
- When posted on the ELIXIR website (after removal of e-mails and dead URLs) this has already proved a welcome resource.
- The location data has been fed into automated map displays
- It could be extended by adding other data derived directly from the websites (e.g. institutions and a link for the most recent publication) or indirectly e.g. inclusion in web services directories or Google rankings.
- At just over 500 entries is it manually browsable and provides an overview, including short descriptions, that would be difficult to obtain otherwise
- Because it includes new dbs not published NAR and some unpublished that submitted survey returns it has content that will not appear in the MBDC.

The results from this compilation cover a number of themes that are either directly relevant to the survey or to the ELIXIR undertaking.

- *Update Status*. For no less than 11% of the dbs there was evidence that they had not been updated from anywhere between two and ten years and for a further 12% the status was unclear

- *The status of the NAR-MBDC collection*. While the value and overall quality of this resource is not in doubt this survey suggests the level of curation for the older entries is low. Realistically, beyond removing dead URLS, the MBDC collection editors would neither be expected to neither prune the list by identifying update recalcitrants nor remove those that, in the ELXIR survey context, we might exclude because they were unusable, tools, aggregation portals or locally installable dbs. Arguably we could have pruned our list before the survey as this would have certainly increased our relative return rate. However it was a strategic decision to cast a maximally inclusive net not only to get the highest absolute number of returns but also to give some benefit of doubt e.g. A db might decide to update as a consequence of being alerted to ELIXIR by the survey.

- *Deadlinks*. The deadlink count was 3% with three sites nominally down for maintenance. This is actually lower than figures published in the literature (see PMID: 17238638 and PMID: 16779199) and is clearly a combined result of the selectivity of the MBDC and our inclusion of new databases.

- *Redundancy*. This problem can clearly be discerned even at the level of database titles and seems particularly prevalent in the protein sequence clustering and 3D structure areas.

- *Substitution by de-novo generation or federated queries*. The list reveals a number of cases where the db content could be generated on-the-fly either directly from the core data resources e.g., Ensembl or from federated sources via Biomart and/or Taverna. However, it is important to discern which dbs are consistently incorporating manual curation as a complement to automated extraction. The paradox is the intrinsic high value of this activity is perhaps the biggest reason for the inability to keep pace with updating.

- *Design and usability*. Manual inspection revealed a spectrum from the sublime to ridiculous on database entry pages. Consequently many dbs are presenting high navigation barriers and/or a paucity of contextual information that renders them close to unusable even if data structures of high utility were hidden behind them. Other than the question we included about documentation it is difficult to assess this important aspect by self-reported survey.

The complete compilation is available as a **Appendix II** (ELIXIR_merged_ database_list_Jun09.xls). A trimmed version without e-mail addresses and dead linked dbs removed is included on the ELIXIR web site.

http://www.elixir-europe.org/files/documents/ELIXIR_survey_list.xls

The town locations in this list were converted to latitude and longitude to enable their location to be display in Google maps (courtesy of Bren Vaughn). The response status and database name is also included in this display

http://www.elixir-europe.org/page.php?page=survey_dots

RESULTS AND ANALYSIS

The result summary PDF automatically generated by Survey Monkey has been posted on the ELIXIR website

http://www.elixir-europe.org/files/documents/ELIXIR_database_providers_survey.pdf

A brief overview poster has also been presented at ISMB 2009

http://www.slideshare.net/cdsouthan/elixirposter

Data Clean-up

This is an essential prerequisite for optimising the quality of data analysis and the conclusions and is discussed in more detail in the WP2 Report. The process was continous for the survey duration and began by pruning a small number of "ghost" replies and those who had not got past the first page. We received about 20 partial replies over the course of the survey from respondents who had made progress but not completed. These were all sent individual encouraging e-mails that resulted in successful completion (some were due to technical issues such as changing IP addresses) in approximately 50% of cases. Towards the end of the survey most partials were removed with the exception of a few who had clearly gone more than 60-70% in to the responses but had not exited properly. A number of "aggregators" were identified where respondents had filled in one questionnaire for several individual databases. In all cases these were understandingly cooperative and resubmitted responses for the individual databases and the original aggregate entries were then deleted.

General Information

Q1. Basic information

Answer Options	Response Frequency
The full name of the database	100.0%
Acronym/abbreviation	87.6%
URL	99.5%
Your name	99.0%
Your e-mail	99.0%
Name of your institution (if at EBI, Sanger or MIPS please use only these exact spellings)	98.1%

The only note here is that 12% of dbs offered no abbreviation

Q2. Are you the same contact person we e-mailed?

Answer Options	Response Frequency
Yes	63.8%
No, I am the new contact person the this has been passed on to	30.9%
No, I have picked up the link from the ELIXIR website	5.3%

This showed that asking the initial set of contacts to forward the questionnaire link to the appropriate contact had paid off. However, the answer cannot discriminate between genuine change of contact or a delegation pass-along. Relatively few dbs had picked up the survey from the website.

Q3. Which European or ELIXIR-affiliated country is the primary location of the database?

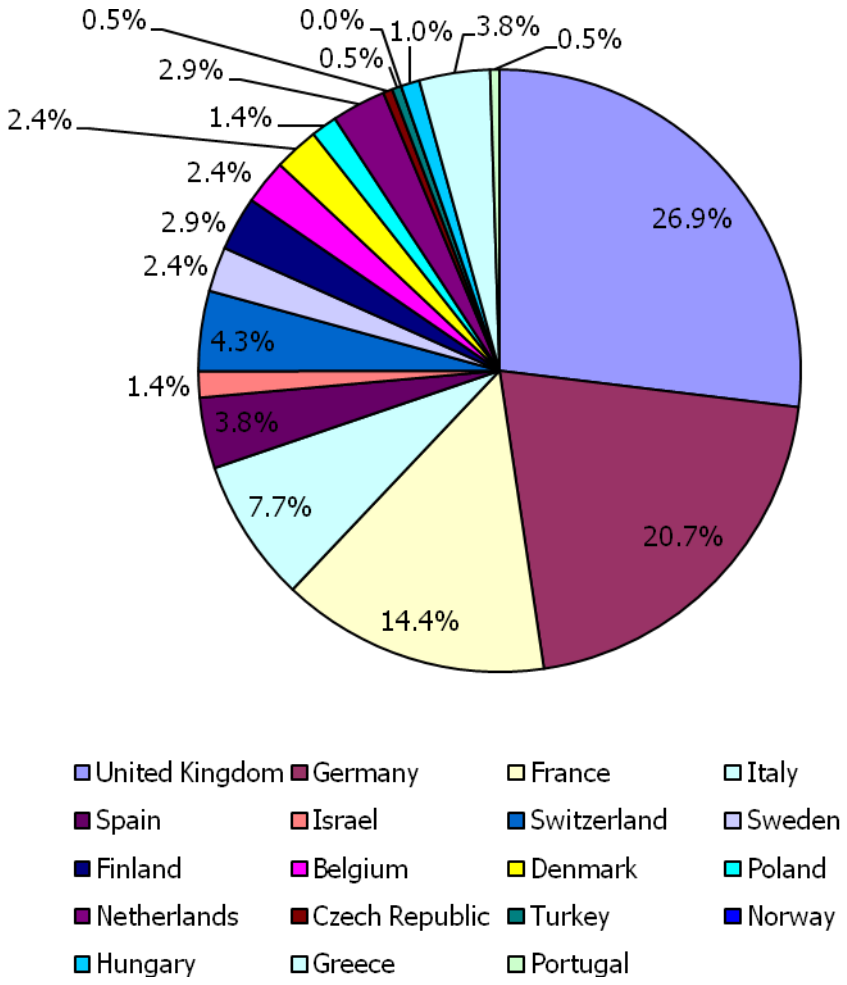

Figure 8. Db number vs. country

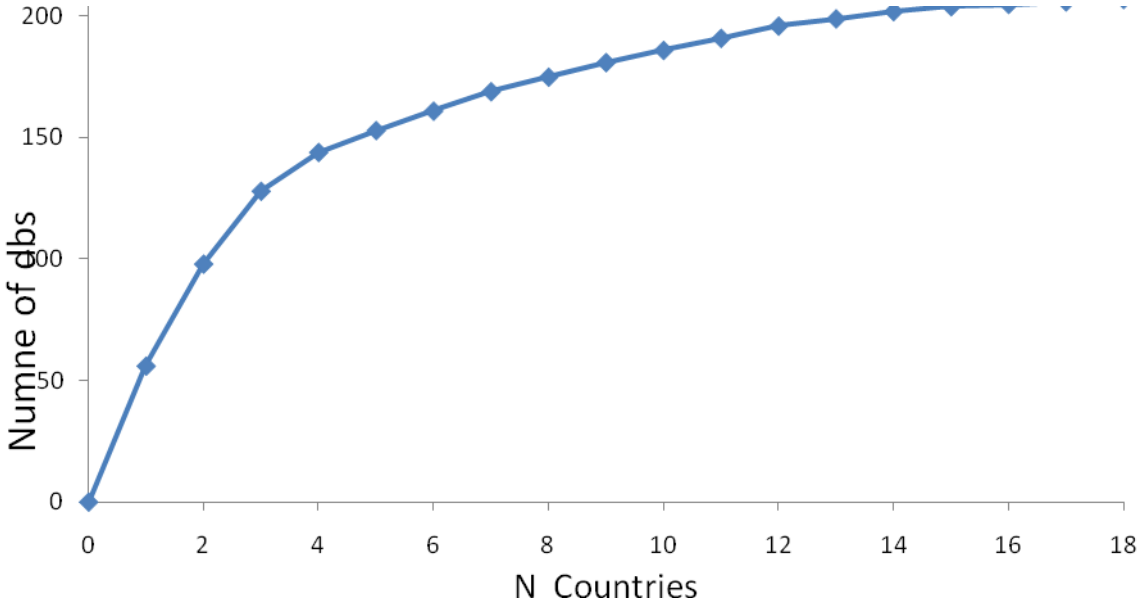

Figure 9. Cumulative plot of db number vs. country

This shows approximately 75% of the dbs are from the top-five countries

Q4. Had you heard about the ELIXIR project before this survey was sent to you?

Answer Options	Response Frequency
No	29.7%
Yes, but in outline only	25.4%
Yes, in some detail	45.0%

This 100% response indicated that circulating the survey was of itself a useful instrument for increasing ELIXIR awareness not only in the db community but possibly also their host institutions.

Information about the Database

Q5. Please provide a count of the number of known mirror URLs

Answer Options	Response Frequency
None	77.2%
1	9.9%
2	3.0%
3	2.0%
4	0.5%
5	2.0%
6	2.5%
7	0.0%
8	0.0%
9	0.0%
10	1.0%
<10	2.0%

These answers, from 202 respondents, show mirroring is the exception rather than the rule but a few are highly mirrored

Q6. Please estimate approximate size in Gigabytes

Answer Options	Response Frequency
0-0.5Gb	23.5%
0.5-1	16.0%
1-2	11.0%
2-4	6.5%
4-6	3.0%
6-8	2.5%
8-10	3.5%
10-20	4.0%
20-50	8.5%
50-100	7.0%
100-200	3.0%
200-500	4.0%
500-1000 Gb	3.0%
1000-2000	1.0%
2000-3000	1.0%
3000-4000	0.0%
4000-5000	2.5%
answered question	
skipped question	

The raw data from the table above, from 200 respondents, was converted into the plot below

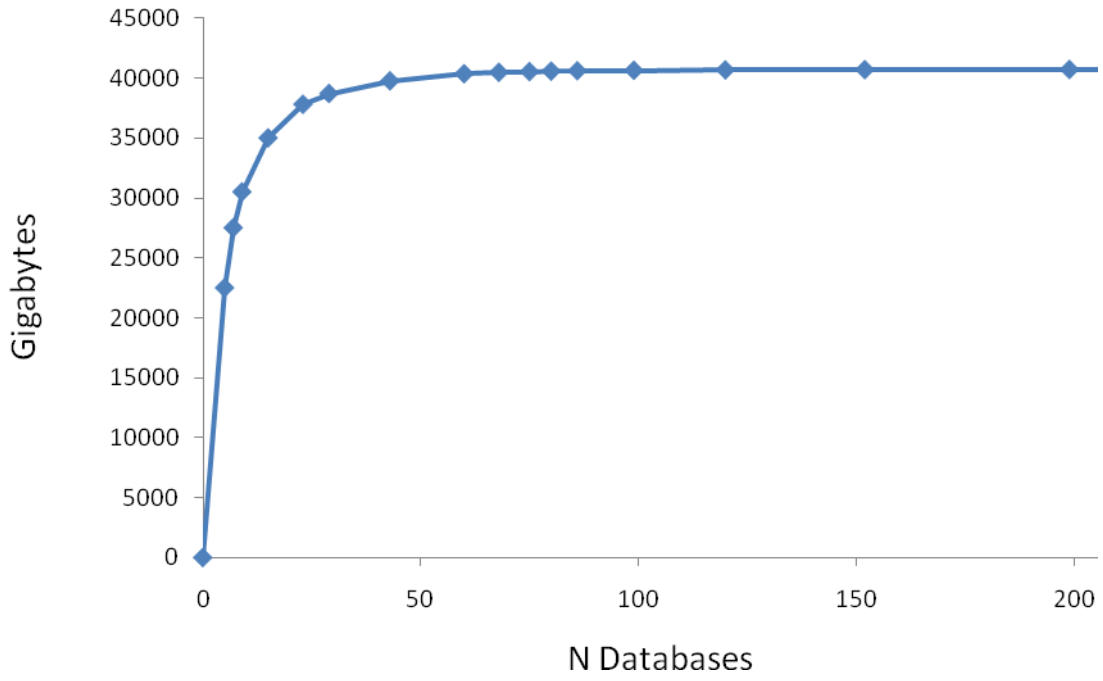

Figure 10. Cumulative plot of Gigabytes per db.

This shows that over 90% of the total capacity is covered by the top 50 followed by a longer tail of smaller dbs

Q7. Please estimate approximate total number of entries:

While the number of entries may not be as good a representation of db size as the total storage in Q6 the answers to this question are represented in the chart below.

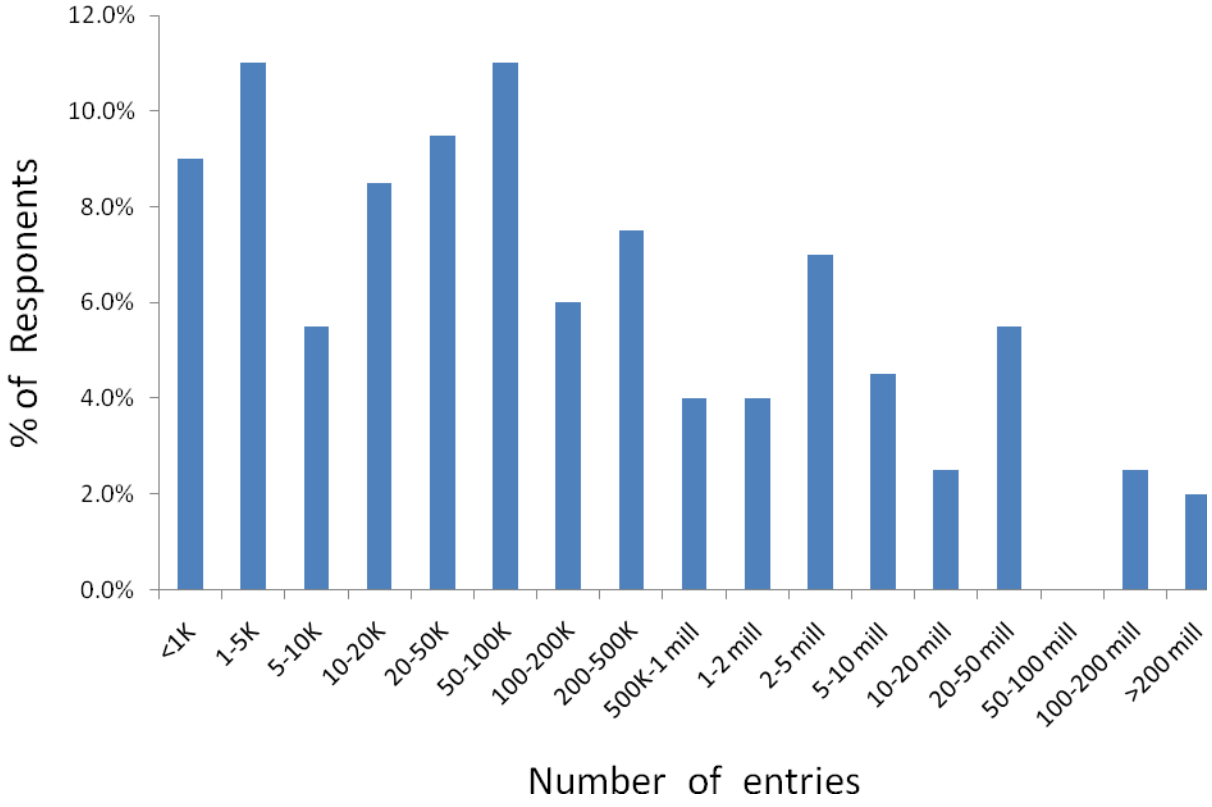

Figure 11. Total entries per db vs. the % of respondents.

Q8. Please indicate the major data types and keywords relevant to your database

The ranked distribution of data types and keywords is shown in the table below, expressed as the % response from 205 replies.

Eukaryotic	41.5%	Mus musculus	16.1%
Protein sequence	38.5%	Predicted DNA features	15.6%
DNA sequence	37.6%	Disease association data	15.1%
Gene names	35.1%	Transcript expression data	12.7%
Publications	34.1%	Predicted RNA features	12.7%
Genomic sequence	32.2%	Phylogenetic group specific	12.2%
Species specific	30.7%	Experimental DNA features	12.2%
Protein domains	29.3%	Drosophila melanogaster	11.2%
Predicted protein features	28.3%	Saccharomyces cerevisiae	10.7%
Homo sapiens	27.8%	Images	10.2%
Ontologies	24.9%	Experimental RNA features	9.3%
Protein 3D molecular structures	22.9%	Genotyping	9.3%
Predicted protein function	22.4%	Caenorhabditis elegans	8.8%
Transcribed sequence	21.5%	Protein expression data	7.3%
Mammalian	21.5%	Small molecule chemical structures	7.3%
Prokaryotic	21.0%		
Protein-protein interactions	20.5%	Locus specific	6.8%
RNA sequence	19.0%	Clinical data	5.9%
Sequence polymorphisms	19.0%	Mass-spectrometry data	4.4%
Experimental protein features	18.0%	Virus specific	3.9%
Protein family specific	18.0%	NMR	2.9%
Enzymes	17.1%	Patents	2.0%
Experimental protein function	16.6%	Tomography	1.0%
Vertebrate	16.1%	EM	1.0%

Q9. Please add any additional major data types and keywords

Well over 200 additional terms were offered here. Although most of these were singletons "pathways" was mentioned frequently.

Q10. Where does your data come from? (Please tick all that apply)

The answers are shown below. Of the 30 categories mentioned in "other" submissions from users were the most common.

Answer Options	Response Frequency
Derived from one primary source (e.g. a genome database)	19.3%
Derived by combing data from several other databases	65.2%
Extracted from the literature	51.2%
From experimental data	39.1%
From a collaboration	23.2%
Tool-derived data	29.5%
Unique data generated your laboratory	22.7%
Other (please list)	14.0%

Q11. Does you database incorporate manual curation or annotation (biocuration)

While some Biocuration was reported by 28% the number for extensive incorporation was 44% with the remaining 28% answering no

Q12. Would you consider your database as?

This shows a fairly even three-way split between the categories.

Answer Options	Response Frequency
A specialist resource	35.0%
Moderately specialised	32.5%
Of broad utility	32.5%

Q13. Please provide a short description of unique content

Q14. Please provide a short description of biological utility

Q15. Please provide a short description of scientific impact

This series of questions produced a valuable free-text set of answers and these have been captured in **Appendix III** (Q13-15_free_text_answers)

Interestingly there was a drop off across the set with the responses at 95% for Q13, 93% for Q14 and 87% for Q15, suggesting that impact was more challenging to describe compared to content.

Data Access and Re-usage Policies

Q16. In what ways can users access the data?

Answer Options	Response Frequency
Web browser queries	99.0%
E-mail queries	14.4%
Downloads	71.8%
Programmatic access	33.0%

Q17. Can the data be downloaded in their entirety?

Answer Options	Response Frequency	Response Count
Yes	54.3%	113
No - but only because of technical limitations	32.2%	67
No - because there are restrictions associated with the data	13.5%	28
Please specify the data restrictions		40

Q18. In the case of allowing downloads do you impose any restrictions on re-use of the data?

Answer Options	Response Frequency	Response Count
No	74.3%	139
Yes	25.7%	48
If yes please specify the re-use restrictions		53

Q19. Are there any confidentiality issues in relation to the data?

Answer Options	Response Frequency	Response Count
No	89.7%	183
Yes	10.3%	21
If yes please specify the issues		23

Funding

Q20. The database is

Answer Options	Response Frequency
Free to all	84.7%
Academic but charges commercial users	15.3%
Commercial	0.0%
Commercial with an open "lite" version	0.0%

A small number of commercial databases had been sent a survey but none had returned.

Q21. What type of funding does your database have? (if mixed please tick multiple boxes)

Answer Options	Response Frequency
Funding outside Europe	11.7%
Institutional	48.8%
National grant	37.6%
Rolling funding	2.0%
European funding	35.6%
Intermittent	10.2%
Commercial	7.8%
No formally specified funding (e.g. the database was a by-product of a research project)	30.7%
Other funding type	3.4%
Please outline other funding type	(31 replies)

Q22. If you ticked European funding above please indicate the proportion of support it supplies and the duration

Answer Options	1 year	2 yr	3 yr	4 yr	5 yr	> 5 yr	Rating Average	Response Count
<20%	5	5	9	3	3	4	3.21	29
20%-40%	0	2	6	2	1	1	3.42	12
40%-60%	2	0	3	2	3	1	3.64	11
60%-80%	0	2	5	2	2	1	3.58	12
100%	1	0	4	0	1	1	3.43	7
							answered question	69
							skipped question	140

These answers show the pattern of European funding is variable.

Q23. Please list your funding sources

A listing of over 250 sources was returned from this question.

Q24. Please give the level of funding used for your database, in thousands or millions of Euros, including institutional overheads (these may be rough estimates but please try to provide something)

The raw data is shown in the table below.

Answer Options	<5K	5-10K	10-25K	25-50K	50-100	100-200	200-500	500K - ...	1-2 mill	2-5	5-10	10-20	20-50	50-100	100-200	> 200 ...	Rating	Respon
Initial	28	13	16	27	27	19	23	8	2	3	1	0	0	0	0	0	4.45	167
Cumulativ	21	10	14	8	14	16	15	16	13	6	4	2	2	2	0	0	5.69	143
Maintenan	51	15	17	11	11	21	18	2	2	3	0	1	0	0	0	0	3.68	152
																answered question		174
																skipped question		35

The same data has been transformed into the cumulative representation below.

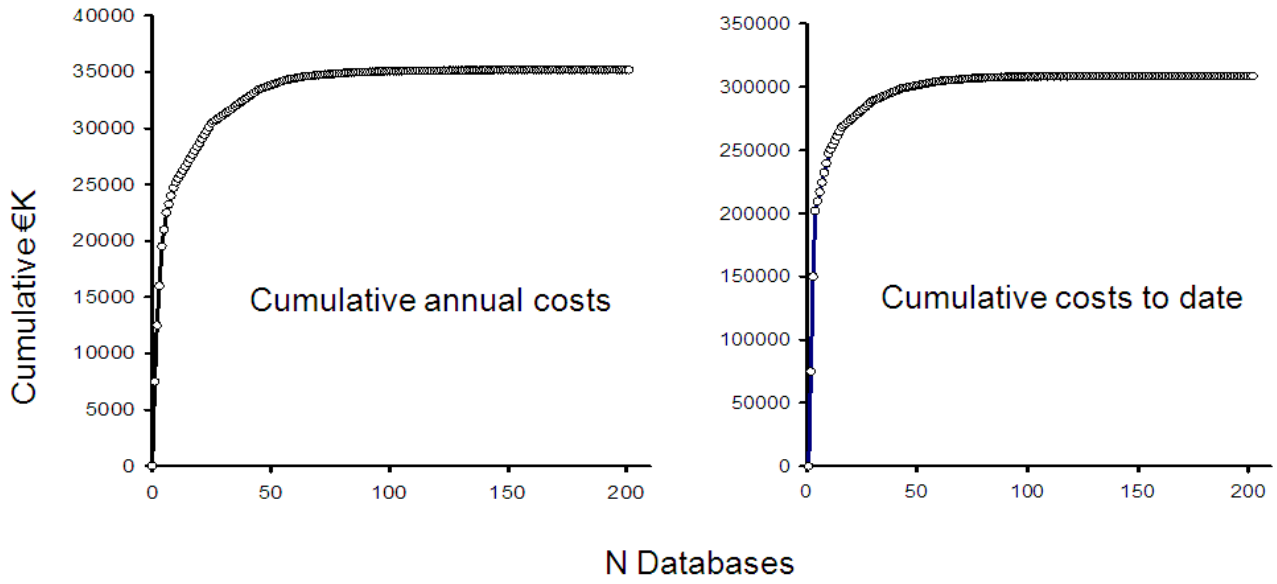

Figure 12. Plot of Costs per database in 1000 Euros.

Q25. The current funding of the database is

Answer Options	Response Frequency
Not assured	36.3%
Assured for at least 1 year	29.9%
Assured for at least 3 years	26.9%
Assured for at least 5 years	3.0%
Assured for more than 5 years	3.5%
We are considering commercialisation	0.5%
answered question	*201*
skipped question	*8*

Q26. Please rate your level of concern for the long-term sustainability of your database as a European resource (on a scale up to 5 = very concerned)

Answer Options	Response Frequency	Response Count
Not concerned	8.0%	16
1	5.0%	10
2	10.0%	20
3	22.9%	46
4	22.4%	45
5	31.8%	64
answered question		201

skipped question	**8**

Infrastructure

Q27. Development of your database incorporated input from:

Answer Options	Response Frequency	Response Count
Software/database developers	74.9%	155
Web interface designers	50.7%	105
Bioinformaticians/Computational biologists	91.3%	189
Computer scientists	36.2%	75
Bench scientists	46.9%	97
Other (please specify)		24
answered question		**207**
skipped question		**2**

Q28. Do you collaborate with other groups in the development of your database

Answer Options	Response Frequency	Response Count
No	40.0%	82
With one other group	22.9%	47
With two groups	15.1%	31
Three	7.3%	15
Four	2.0%	4
Five	2.0%	4
More than five	10.7%	22
answered question		**205**
skipped question		**4**

Q29. Do you know of other databases that are closely related to yours in concept, content and utility?

Answer Options	Response Frequency	Response Count
No - we judge ours to be unique	41.8%	87
Yes	58.2%	121
answered question		**208**
skipped question		**1**

Q30. If yes, how many of those closely related databases are:

Answer Options	1	2	3	4	5	> 5	Response Count
In Europe	34	26	9	1	2	6	78
Outside Europe	43	27	11	4	0	10	95
answered question							120
skipped question							89

Q31. Please list the names of the closely related databases you know of

The replies from 123 respondents listed 312 related dbs. Many of these are multiple related databases as in the table below

Answer Options	Response Frequency	Response Count
1	100.0%	123
2	68.3%	84
3	43.9%	54
4	25.2%	31
5	16.3%	20
answered question		123
skipped question		86

Q32. Do you collaborate with these closely related databases for example by data exchange?

Answer Options	Response Frequency	Response Count
No	48.3%	72
Yes with one	32.2%	48
Yes with more than one	19.5%	29
answered question		149
skipped question		60

Interoperability and Standards

Q33. Does your database have Webservices (SOAP, REST, WSDL etc)?

Answer Options	Response Frequency	Response Count
Yes	32.5%	67
No	47.1%	97
No; but we plan to introduce it within approximately 12 months	20.4%	42
answered question		206
skipped question		3

Q34. Do you exchange data with other databases? (unidirectional or reciprocal)

Answer Options	Response Frequency	Response Count
No	33.5%	69
Yes with 1	20.4%	42
Yes with 2-4	29.1%	60
Yes with 5 or more	17.0%	35
If yes please list the format(s) you use		122
answered question		206
skipped question		3

Q35. Does your database conform to specified vocabularies/ontologies?

Answer Options	Response Frequency	Response Count
No - we are unaware of standards applicable to our data types	37.3%	76
No - but we plan to within approximately 12 months	11.8%	24
Yes - but not OBO specified	25.5%	52
Yes - we use those specified under the OBO umbrella (www.obofoundry.org/)	25.5%	52
If yes to OBO please list the types		55
answered question		204
skipped question		5

Q36. Does your database content conform to specified minimum information standards?

Answer Options	Response Frequency	Response Count
No - we are unaware of any applicable to our data types	61.6%	125
No - but we plan to within approximately 12 months	8.4%	17

Yes - but not MIBBI specified	21.2%	43
Yes - we use those specified at MIBBI (www.mibbi.org/index.php/MIBBI_portal)	8.9%	18
If yes to MIBBI please list the types		21
answered question		**203**
skipped question		**6**

Q37. Are you involved in the development of standards?

Answer Options	Response Frequency	Response Count
No	58.3%	120
Yes	41.7%	86
If yes please specify this involvement		79
answered question		**206**
skipped question		**3**

Q38. The broad area of interoperability and federation is important for the future. Beyond webservices and standards what other steps in this direction have you implemented?

Answer Options	Response Frequency	Response Count
DAS	62.3%	43
Biomart	30.4%	21
Biomoby	17.4%	12
Grid compatibility	15.9%	11
Workflows	33.3%	23
Other (please specify)		21
answered question		**69**
skipped question		**140**

Outreach

Q39. For how many years has your database been publicly accessible?

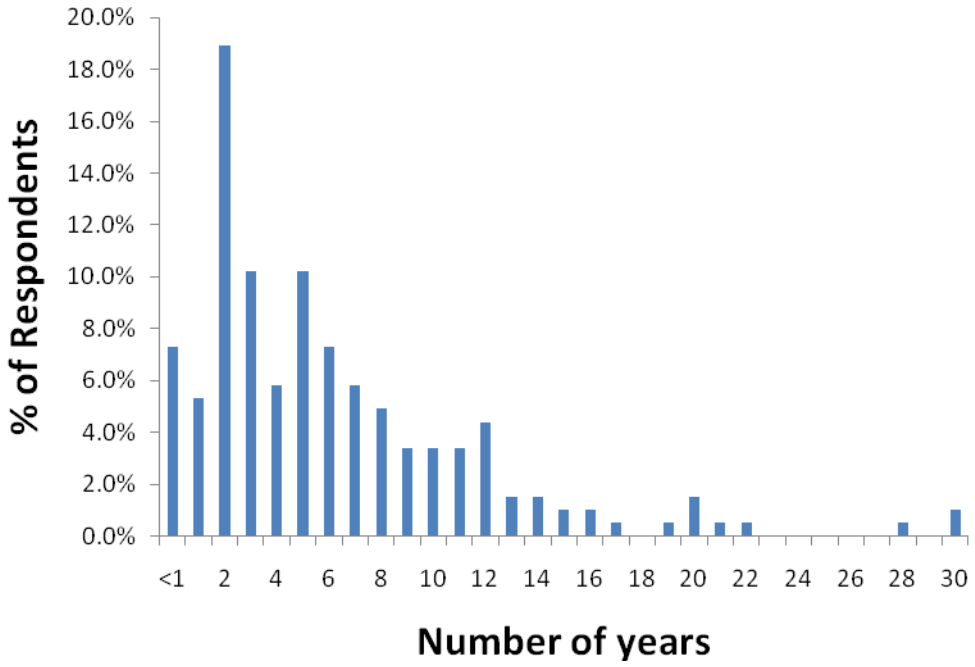

Figure 13. Plot of database age.

Q40. Do you have adequate user documentation/database help facilities?

Answer Options	Response Frequency	Response Count
Yes	80.0%	164
No	20.0%	41
answered question		205
skipped question		4

Q41. Is a description of your database published in a journal article?

Answer Options	Response Frequency	Response Count
No	9.2%	19
Yes	90.8%	188
answered question		207
skipped question		2

Q42. If your database has been published please indicate if the journal was:

Answer Options	Response Frequency	Response Count
The Nucleic Acids Research annual database issue	42.8%	80
Another journal	26.2%	49
Both i.e. in the NAR database issue and another journal(s)	31.0%	58
answered question		187
skipped question		22

Q43. Please provide the PubMed IDs (or references if not in PubMed) for the publications describing your database

Answer Options	Response Frequency	Response Count
1	100.0%	187
2	57.2%	107
3	32.1%	60
4	21.9%	41
5	16.0%	30
answered question		187
skipped question		22

Q44. If the description is published, how many citations have those paper(s) had? (any source will do e.g. Citation Index, Google Scholar, Scopus etc) If there are multiple papers describing your database you can provide a cumulative total but try to exclude self-citations

Answer Options	1	2	3-5	5-10	10-20	20-50	50-100	100-200	200-500	500-1000	>1000	Response Count
Total citations	9	4	14	15	17	24	13	14	20	11	17	158
Citations per-year	5	8	19	13	12	21	13	6	3	4	0	104
answered question												159
skipped question												50

Q45. For citations of the use of your database, please give the PubMed IDs (or reference if not in PubMed) for those papers where you feel its utility has been best highlighted

Answer Options	Response Frequency	Response Count
1	100.0%	96
2	67.7%	65
3	44.8%	43
answered question		96
skipped question		113

Q46. What strategies have you used to promote usage of your database?

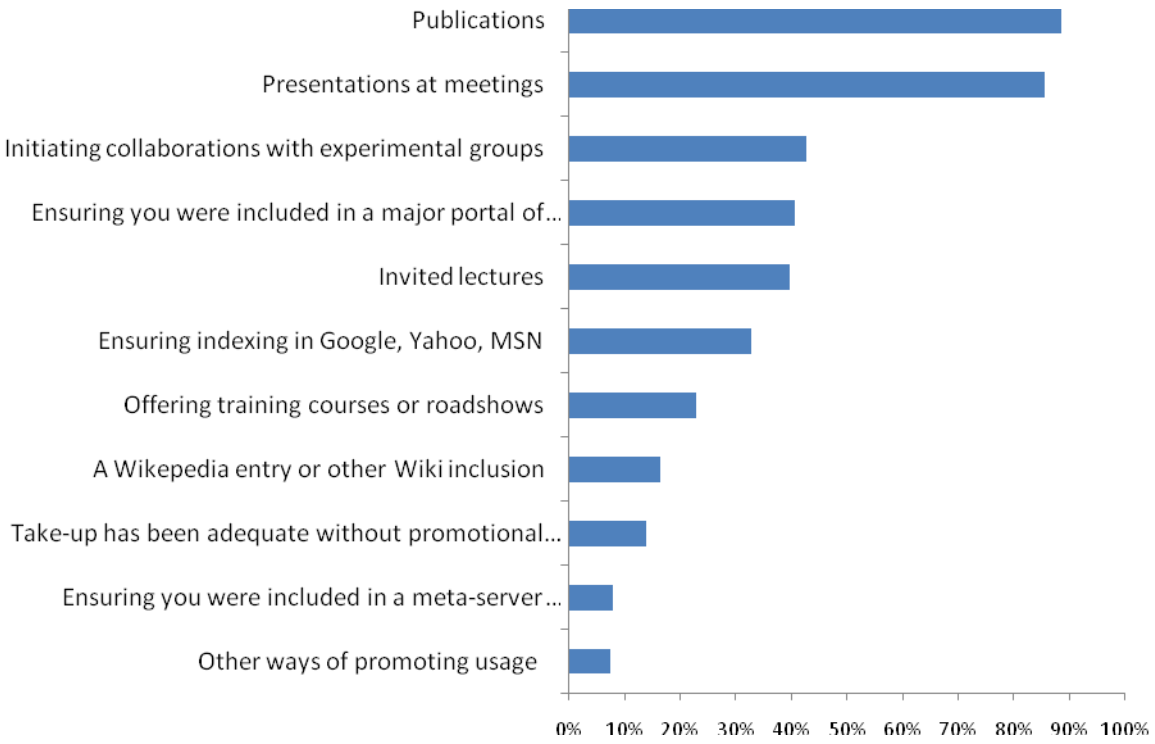

Figure 14. Plot of ranked strategies for database promotion.

Q47. Please indicate your rating of these usage promotion methods, on a scale of; 1 slightly effective, up to 5 very effective. (you can still rate them even if you don't actually use them)

Answer Options	not effective	1	2	3	4	5	Rating Average	Response Count
Publications	0	9	19	45	51	53	4.68	177
Presentations	0	2	22	40	82	23	4.60	169
Collaborations	3	7	21	41	51	22	4.35	145
Indexing	4	5	16	27	41	18	4.35	111
Lectures	3	11	20	49	35	8	4.00	126
Portals	5	11	24	27	29	12	3.93	108
Training	6	7	10	48	28	11	4.07	110
Meta-servers	9	7	22	22	22	10	3.77	92
Wikis	11	15	18	33	17	10	3.58	104
							answered question	183
							skipped question	26

Q48. You assess the scientific impact of your database by:

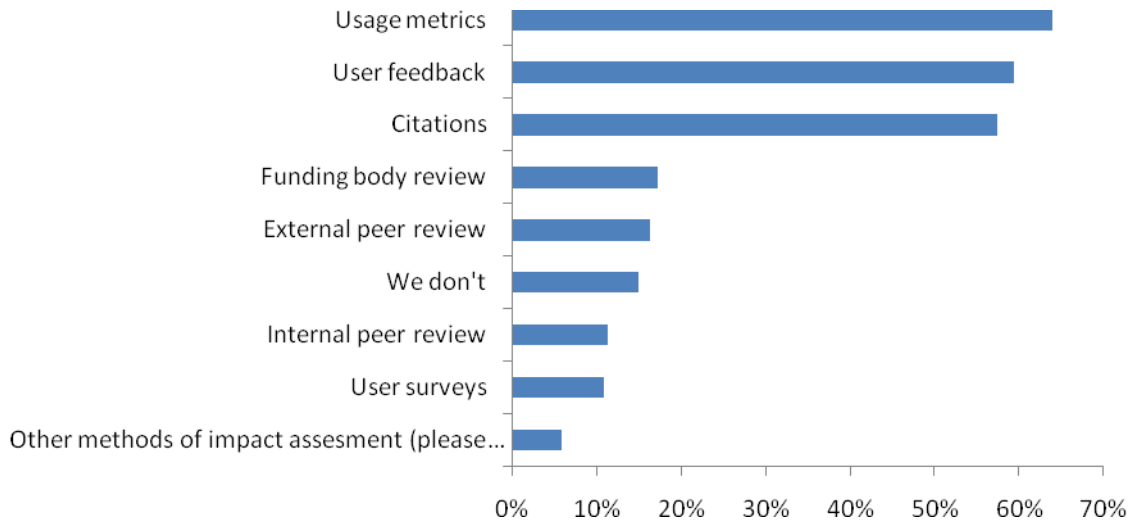

Figure 15. Plot of used impact assessment methods.

Q49. Please indicate your rating of scientific impact assessment methods on a scale of; 1 slightly effective, up to 5 very effective. (you can still rate them even if you don't actually implement them)

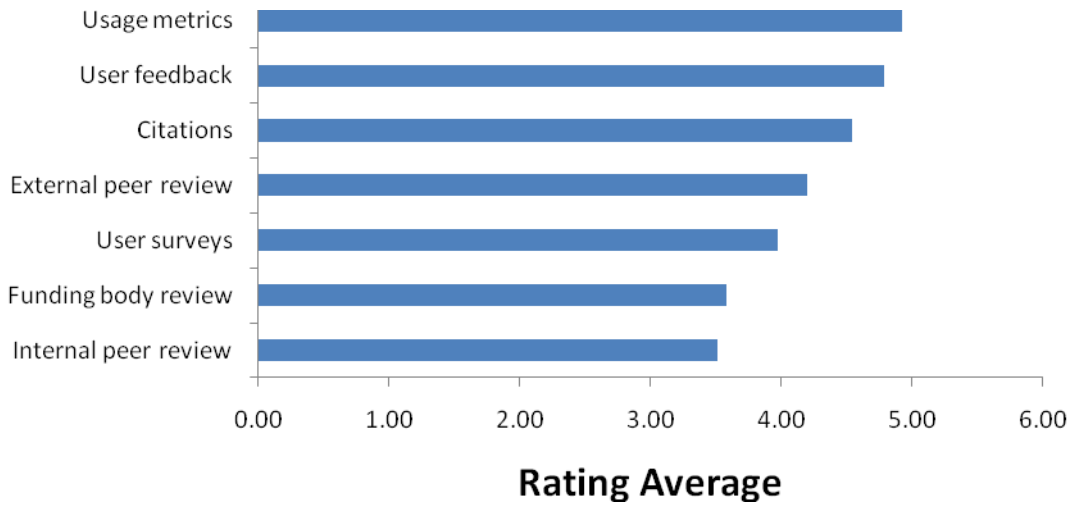

Figure 16. Plot of impact assessment method ranking.

Q50. Who do you think uses your database?

Answer Options	Response Frequency	Response Count
Bioinformaticians/computational biologists	88.3%	181
Bench scientists Bench scientists	72.2%	148
Biologists	68.3%	140
Geneticists	49.3%	101
Biochemists	42.4%	87
Pharmaceutical/biotech industry	42.4%	87
Pharmacologists	20.0%	41
Clinical specialists	16.6%	34
Chemical biologists	16.1%	33
Environmental scientists	15.6%	32
Medicinal chemists	12.7%	26
General public	9.8%	20
Other (please specify)		19
answered question		205
skipped question		4

Q51. What are they using the database for?

Answer Options	Response Frequency	Response Count
Searching for specific biological information	91.0%	181
Downloading data sets for their own analysis	67.8%	135
Analysing their experimental results	50.3%	100
Downloading benchmarking data sets	24.6%	49
Other (please specify)		19
answered question		199
skipped question		10

Q52. What are the most common problems reported by users of your database (honesty is useful here please!)

Answer Options	Response Frequency	Response Count
Lacking certain features	45.1%	79
The data need updating	41.1%	72
The data are incomplete	28.6%	50
The inherent complexity of the data is challenging	28.0%	49
The data contain errors	22.3%	39
Speed/network problems	17.1%	30
User help/documentation is inadequate	17.1%	30
Download formats not optimal	12.0%	21
The site is down	12.0%	21
Webservices not optimal	12.0%	21
Site navigation problems	11.4%	20
No download options	8.6%	15
Dead links	4.6%	8
Other problems (please specify)		25
answered question		175
skipped question		34

Usage Metrics

Q53. Do you collect usage metrics?

Answer Options	Response Frequency	Response Count
No	20.3%	42
Yes; basic ones	44.0%	91
Yes; in some detail	29.5%	61
Yes; in extensive detail	6.3%	13
answered question		207
skipped question		2

Q54. Do users need to register?

Answer Options	Response Frequency	Response Count
Yes	5.9%	12
No	94.1%	193
answered question		205
skipped question		4

Q55. Where you can, please supply web hits per-month (excluding web-crawling)

There were 147 responses to this question.

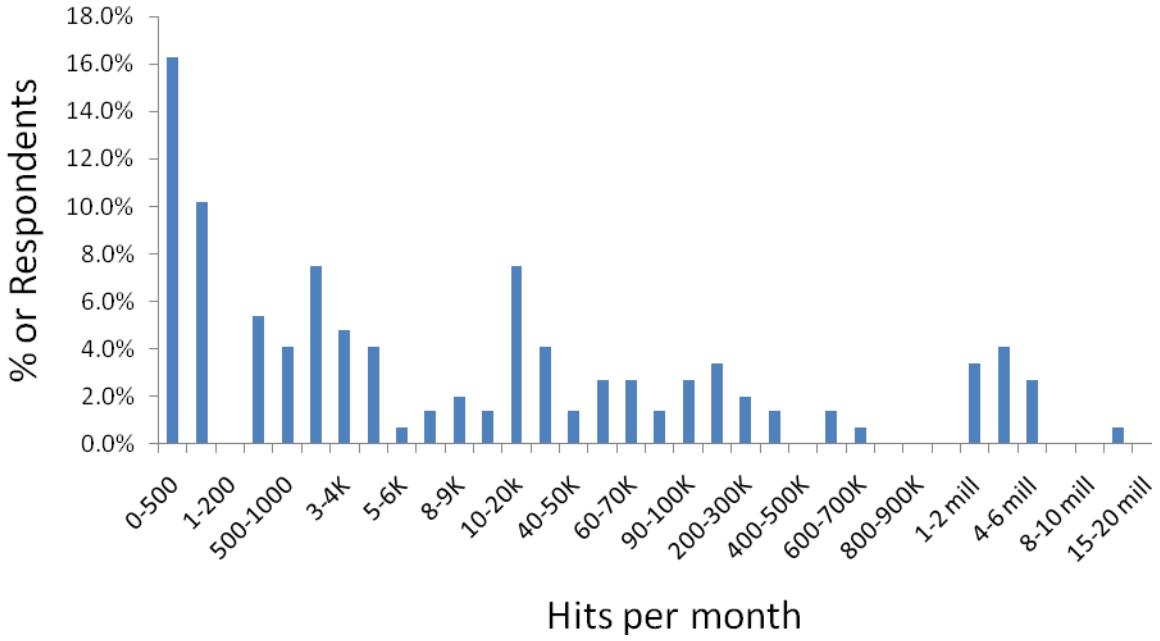

Figure 17. Plot of web-hits-per-month.

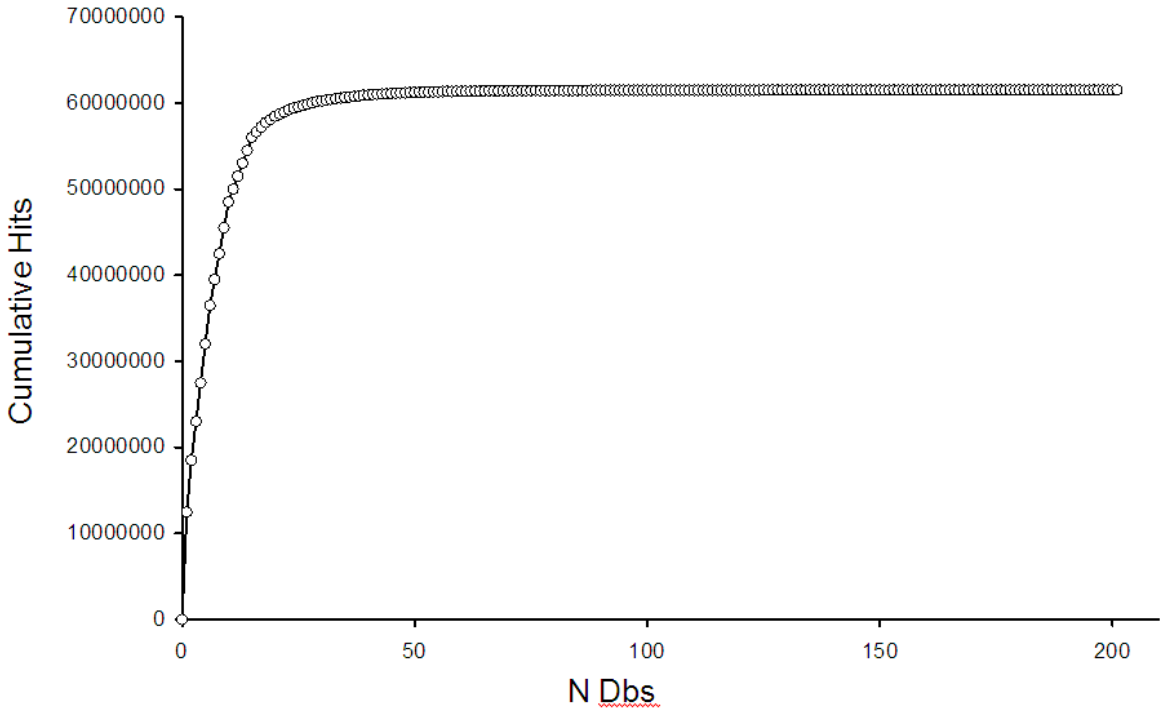

Figure 18. Cumulative plot of web-hits-per-month.

Q56. Where you can, please supply the number of unique users per-month

There were 127 responses to this question.

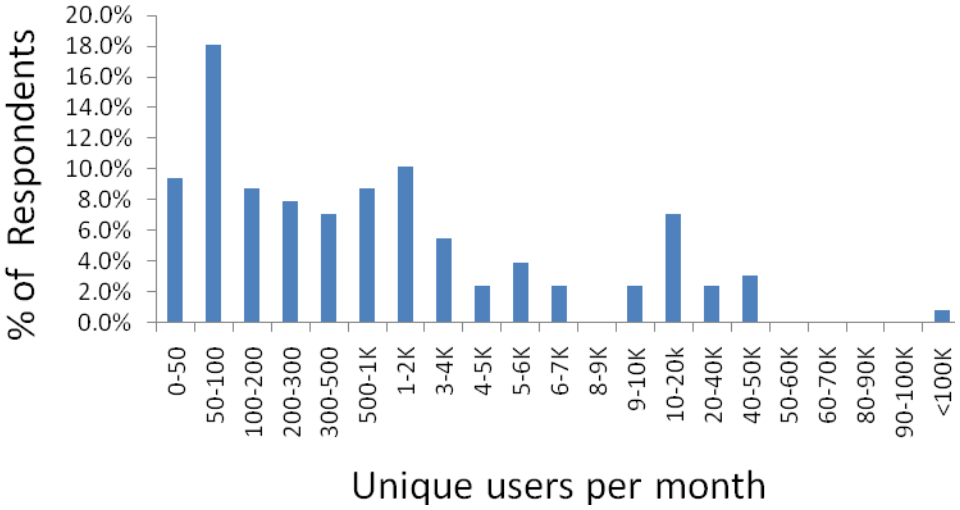

Figure 19. Plot of unique-users-per-month.

Q57. Were you can, please supply additional metrics for the following:

There were a range of figures given for these metrics but there were less than half of the number of respondents for hits and users above.

Answer Options	Response Frequency	Response Count
User logins or sessions per-month	46.0%	29
New users per-month (% increase)	34.9%	22
Total users	38.1%	24
Programmatic access usage (please specify the type of metric)	25.4%	16
Downloads per-month	30.2%	19
Downloaded Gigabytes per-month	46.0%	29
answered question		63
skipped question		146

Q58. How do you rate these as reflecting real-world usage?

Answer Options	1	2	3	4	5	Rating Average	Response Count
Unique users per-month	4	5	22	61	24	3.83	116
Logins or sessions per-month	9	8	27	21	17	3.35	82
Total users	9	12	16	37	11	3.34	85
Site hits per-month	10	18	35	39	18	3.31	120
New users per-month (% increase)	10	17	30	23	7	3.00	87
Downloads per-month	19	12	19	26	8	2.90	84
Programmatic access usage	13	15	19	13	9	2.86	69
Downloaded Gigabytes per-month	23	16	22	11	4	2.43	76
You may provide evidence or opinions for the choice above							23
answered question							126
skipped question							83

Q59. Compared to what you expected your assessment is that usage of your database is:

Answer Options	Response Frequency	Response Count
We did not have any expectations	22.8%	42
Less than we expected	11.4%	21
About what we expected	48.4%	89
Exceeded our expectations	17.4%	32
If usage was very different from your expectations can you speculate why?		26
answered question		184
skipped question		25

Sources, Dependencies and Links

Q60. If your essential data comes from other databases, please give their names

Answer Options	Response Frequency	Response Count
1	100.0%	127
2	72.4%	92
3	52.8%	67
4	37.8%	48
5	26.8%	34
answered question		127
skipped question		82

Q61. Approximately how many databases do you know you are linked with?

Answer Options	1	2	3	4	5	5-10	10-20	20-30	30-40	40-50	> 50	Rating Ave	Resp Count
That you only link out to	9	15	14	11	21	40	25	7	3	0	10	5.41	155
That only link in to you	20	17	11	5	5	14	4	1	1	0	9	4.14	87
Reciprocal links	27	22	13	3	4	18	1	4	1	0	6	3.68	99
answered question													175
skipped question													34

Q62. Where you can please list your major linking URLs

There were 106 responses to this question.

Q63. Would you like your database to have more reciprocal links?

Answer Options	Response Frequency	Response Count
Yes	77.2%	139
No	22.8%	41
If yes could you name those you would most like to have links with		76
answered question		180
skipped question		29

Resources

Q64. Approximately how many Full-time equivalent (FTE) "person-years" has your database needed?

There were 195 responses to this question. In the table below the top row is for development of the db with 193 responses and an average of 6 FTEs. The second row is for maintenance in 2007 with an average of four FTEs

	<0.2	0.2-0.5	0.5-1	1	2	3	4	5	6-10	10-20	20-30	30-40	40-50	50-60	60-70	<70	Resp
Dev	5	14	21	42	29	11	12	13	23	8	6	4	2	0	0	3	193
Maint	48	32	22	27	25	10	3	5	7	4	2	0	0	0	0	0	185

Q65. What is the active team size in Full-time equivalents (FTEs)

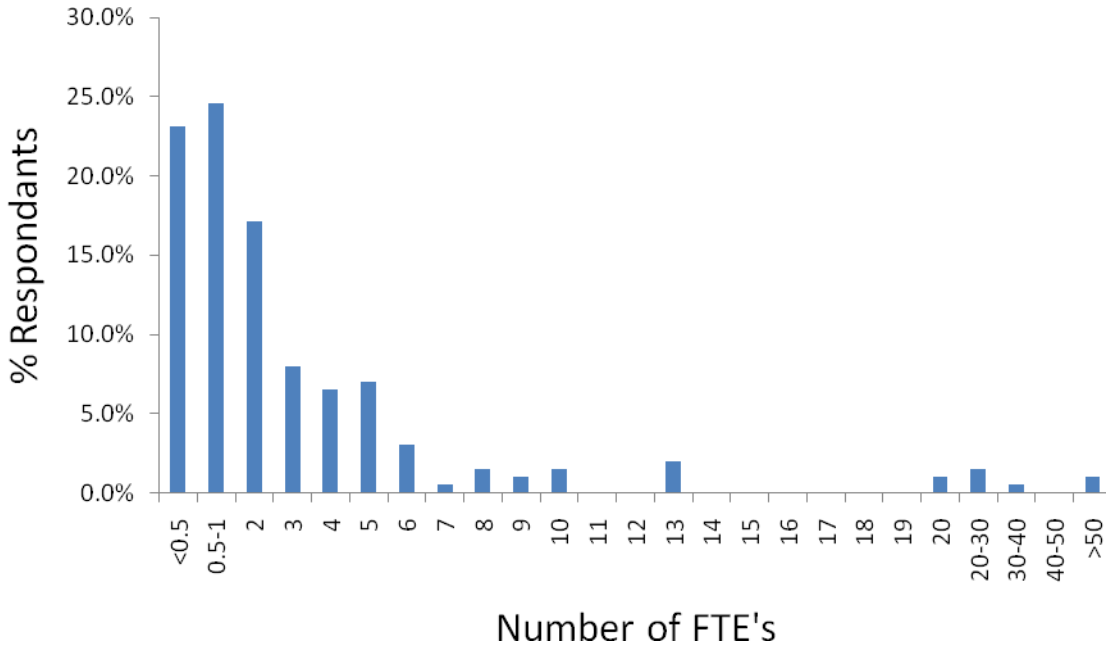

Figure 20. Plot of FTE's per respondent

Q66. If your institution has developed and hosts multiple databases please give the total

There were 138 responses to this question. Both for this and Q67 below it should be noted that the answers are on an individual basis, not an institutional one. Arguably it is the institutional distribution of multiple dbs and tools that are of most interest but further processing i.e. binning by institution would be necessary to generate these numbers.

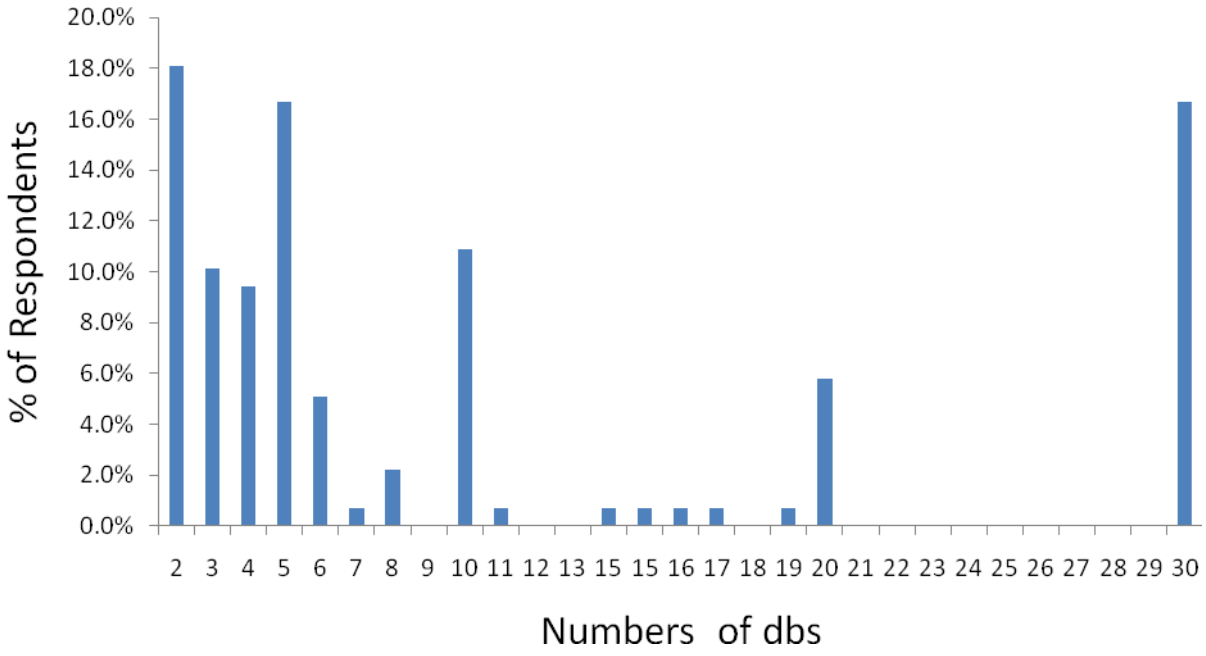

Figure 21. Plot of multiple databases per institution.

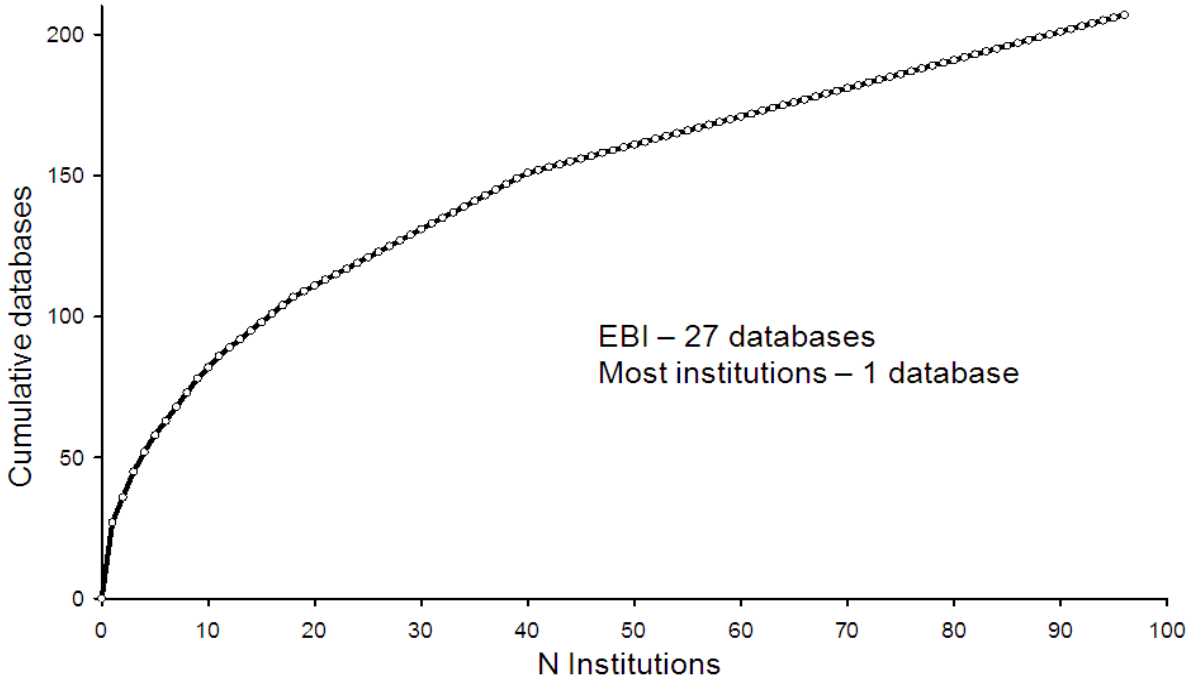

Figure 22. Plot of cumulative databases per institution.

Q67. If your institution has also developed and hosts web-based bioinformatics tools please give the total (or answer "none")

There were 155 responses to this question

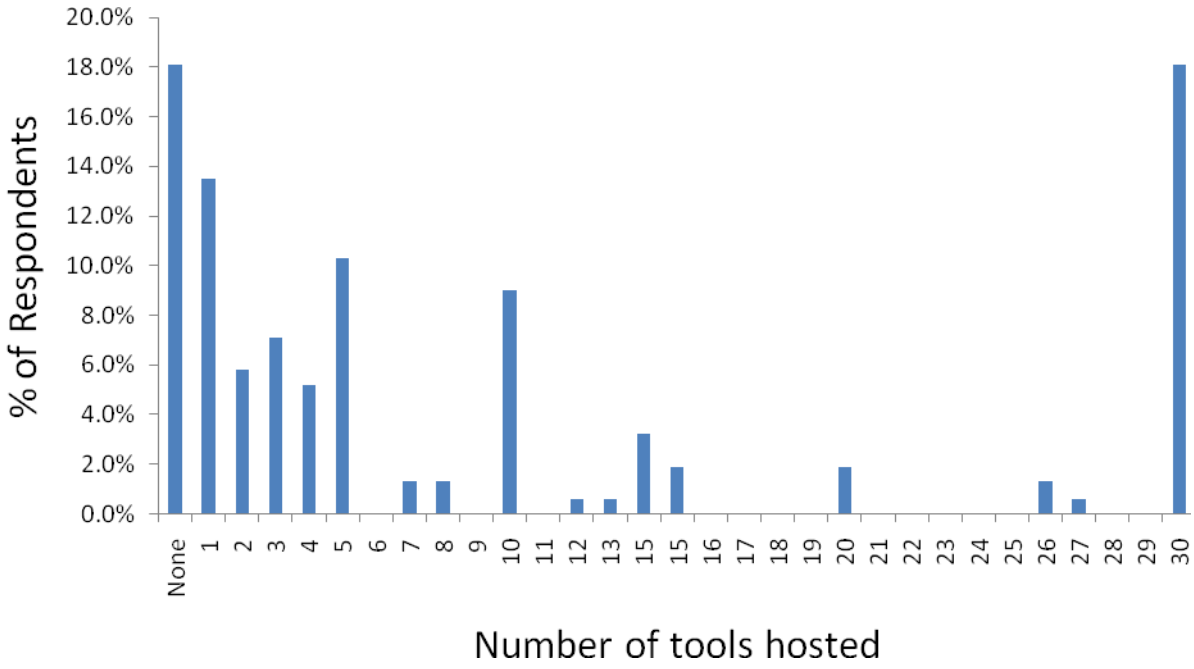

Figure 23. Plot of tools per respondent.

Q68. If your institution hosts multiple databases from 3 above please either give the additional URLs, or, if more than 5, then a home page where they can all be found (if you know any of these that have not been sent a survey link please forward this one - thanks!)

Answer Options	Response Frequency	Response Count
1	100.0%	124
2	39.5%	49
3	24.2%	30
4	13.7%	17
5	6.5%	8
answered question		124
skipped question		85

Q69. With what approximate frequency is the public version of your database updated?

Answer Options	Response Frequency	Response Count
The data is a complete compilation and does not need updating	1.0%	2
Daily	7.0%	14
Weekly	8.5%	17
Monthly	21.4%	43
6-monthly	23.9%	48
Yearly	13.9%	28
Other (please specify)	24.4%	49
answered question		201
skipped question		8

Q70. If your funding sustainability improved what would you consider for enhancement on a priority scale of 1-5 (lowest to highest)

Answer Options	1	2	3	4	5	Rating Average	Response Count
New features	3	14	28	73	60	3.97	178
Maintain current functionality and content	6	22	41	38	66	3.79	173
Expand the scope	11	26	33	50	49	3.59	169
Improve interface/usability	9	20	41	55	49	3.66	174
Utility enhancements	6	28	51	56	33	3.47	174
More interoperability features	9	21	50	55	30	3.46	165
New data sources	25	31	32	49	34	3.21	171
Increase update frequency	36	23	31	33	45	3.17	168
Links to more databases	17	40	53	48	9	2.95	167
Upgrade documentation	16	44	57	36	16	2.95	169
Upgrade hardware	43	52	33	21	13	2.44	162
Other (please specify and rate)							24
answered question							193
skipped question							16

Permissions and Comments

Q71. We would much appreciate if you could allow us the option to use some of your specific responses to illustrate particular points in presentations, reports or publications

Answer Options	Response Frequency	Response Count
Yes, you have my permission	41.3%	83
Yes, you have my permission but I would like to see the examples	51.7%	104
No, please do not use my responses	7.0%	14
answered question		201
skipped question		8

Q72. Please add any comments that you think are relevant to the future sustainability of European databases but not adequately captured in the questions above

A total of 29 free-text comments were made to this question

LIMITATIONS

There are two main reasons to review the limitations of the survey. The first is clearly that it is important to understand and interpret existing results and further analysis in the context of the inevitable imperfections perceived retrospectively. The second is to facilitate eventual re-use of all or parts of the survey, not only in the context of populating an eventual ELIXIR database registry but also for other applications. It should be pointed out that some limitations may have accrued from the fact the endeavour did not involve a questionnaire designer or any other type of survey expert. However, because the aggregate database domain experience of those who provided input to the questions (see acknowledgments) was profound, it remains a matter of speculation as to whether the engagement of such a professional would have produced a significantly more effective outcome.

Numbers and response rate

By employing more literature mining we could have probably pushed the number of circulated databases up to 600. However, to ensure both relevance and quality it would have been necessary to manually review all the URLs before launching the survey. This would have caused a significant delay. The circulation of unpublished resources is a moot point but it would be useful to be able to at least capture these in some way. Our final response rate of 38% (209/538) would have been higher if we had completed the URL check first so that we could have chosen not to circulate dead URLs, languished updates and databases that seemed eligible by their titles but upon inspection were revealed as either out of scope or unusable.

It is not possible to divine reasons for database administrators or PIs not responding unless we had resorted to telephoning non-responders, an option we chose not to take but could be considered perhaps on a sampling basis. Clearly some were daunted by the number and depth of the questions but there is no data to predict that a shorter survey would have significantly increased response rates.

Bias

While, as voluntary surveys go, our return rate was good, any survey with less than 100% response, not only on returns but completions on a per-question basis, is subject to bias. The results give an impression of a "pessimism bias" i.e. that less secure databases were more likely to respond. This is more useful for ELIXIR that an overly optimistic one. The bias issue can be at least partially addressed by comparing the pilot to the main survey for common answers i.e. to see if they are skewed by lower returns in the later, 38%, compared to 60% in the former. Fig. 22 below shows the result of such a comparison. Thus for one of the most basic questions the differences are small which suggests the lower returns from the main survey may not have strongly biased the results.

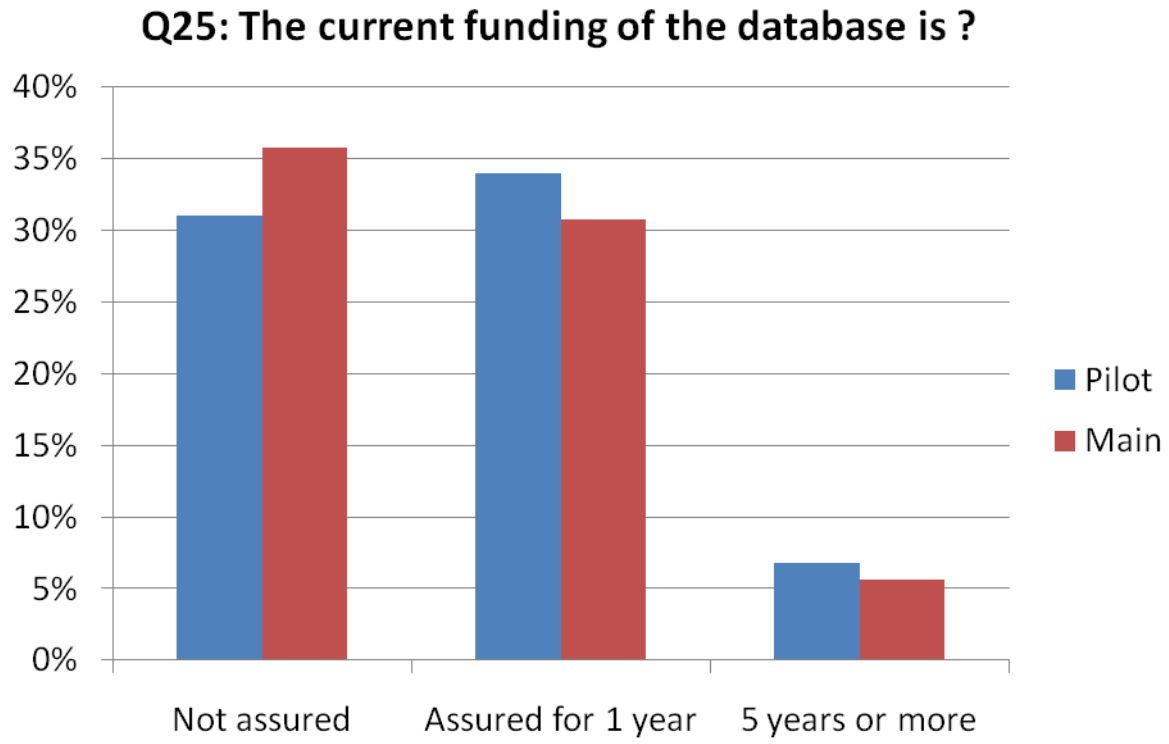

Figure 24. Responses for Q25 compared between the pilot and main survey.

Ambiguity

For the framing of questions it is inherently difficult to find the compromise between brevity that can increase ambiguity and verbosity that can fatigue or irritate respondents. The problem is exacerbated when a significant proportion of respondents are non-native English speakers.

ACKNOWLEDGEMENTS

Contributors to the development of the questionnaire included Janet Thornton, Graham Cameron, Andrew Lyall, Peter Stoehr, Rafael Najmanovich, Rodrigo Lopez, Peter Rice, Alex Bateman and Hamish McWilliam. Useful suggestions from members of the WP2 committee are also acknowledged. Dietrich Rebholz-Schuhmann and Antonio Jose Jimeno Yepes generated and analysed the Medline data. We would also like to express our thanks to the survey respondents who generously gave their time and thereby made a major contribution to this phase of the ELIXIR project.