The ELIXIR Strategy for Data Resources
Draft Report from Workpackage 2
The ELIXIR Preparatory Phase

1. Executive Summary

The creation of a robust information infrastructure for biological research is the core motivation for the ELIXIR project. Data resources are the foundation of that infrastructure. Workpackage 2 of the project documents the scientific need for a coordinated European infrastructure for the data.

Through a survey of existing data resources, it has identified about 500 existing biological databases in Europe and has collected detailed information from over 200 databases offered from about 100 institutions. This uncovers a palette of databases ranging from large (say 20 staff) projects for major core databases through to specialist collections which are the part time effort of individual researchers. The total European effort currently involves at least 350 staff Europe-wide with reported annual direct costs of about €30 million. However, even conservative estimates would put the total expenditure (including indirect costs) at more than €50 million. A usage community of several hundred thousand scientists creates some 60 million web hits per month exploiting the information.

Based on these data and discussions, the workpackage committee has identified the key scientific needs focussed around information intensive, high-throughput biomolecular research, outlining an essential core of information resources covering genomes, genes, gene products and other biologically active molecules such as drugs and metabolites. It also recognises the need to provide information on the interactions and processes in which these molecules are involved and to connect the molecular information to the biology of cells, organs and organisms.

The usage of this shared record of science is testament to its value, and ensuring that the data resources to be supported under ELIXIR can be shared without restriction was seen as crucial. ELIXIR should give integrated access to the information it provides and tailor its services to meet the needs of both general and specialist users – e.g., in drug discovery.

Existing funders of many key databases expect ELIXIR support to relieve them of that burden, and have no long term expectation of continuing their support. This means that it is crucial that ELIXIR succeeds as a source of funding.

Whilst there are some resources whose inclusion in ELIXIR is uncontroversial, candidates for support are numerous, and, in a world of limited funds, clear procedures for assessing the case for inclusion must be developed. The overall funding of the information infrastructure will almost certainly come from a mixture of ELIXIR (pan-European) and national funding, and the balance will be determined by the member states and in discussion with the proposed ELIXIR Scientific Committee (see Workpackage 5). ELIXIR must work to ensure support for essential national initiatives alongside its European endeavour.

The benefit of ELIXIR should be maximised by ensuring that it interoperates with the information infrastructures of other scientific areas: e.g., medicine, chemistry. As part of the ESFRI initiative to create support for European research infrastructure, this document is focussed on Europe, however, ELIXIR must be an integral part of the global infrastructure for biological information. Its societal benefit will depend on translational research to develop applications of the science. Health and medicine are the obvious, but not the only application areas. ELIXIR recognises that realising that much of the potential benefit will depend on commercial developments and will strive to stimulate such exploitation. In serving the commercial sector, ELIXIR will also seek support from industry, but will not adopt licensing schemes which constrain the exploitation of the data.

Key recommendations are:

1. ELIXIR should ensure the existence of the information infrastructure in Europe necessary to support world class life science research in the biomolecular domain, and in the relationship of biomolecular information to more holistic biology.
2. Information resources on all aspects of biologically active molecules will be required.
3. In the future, core resources of widespread utility may have no source of support other than through ELIXIR, making it essential that ELIXIR create mechanisms for their support.
4. Non-core resources are also crucial, and the success of ELIXIR will depend on them being supported. This support may or may not come from ELIXIR. ELIXIR must work with them to ensure that there are sources of funding irrespective of whether that funding is through ELIXIR.
5. Unrestricted access to the information and freedom to exploit it is a key principle of ELIXIR which must not be threatened. This does not preclude seeking funding from commercial organisations who exploit the infrastructure.
6. ELIXIR should strive to ensure interoperability with neighbouring scientific domains such as medicine, epidemiology and chemistry.
7. While maintaining a biomolecular focus, ELIXIR should exhibit some pragmatism in order to fulfil its mission, including activities on the basis of their ability to contribute to the mission rather than rigid principles of eligibility.
8. ELIXIR should establish processes to assess the suitability of activities for inclusion in ELIXIR. This scrutiny should be routinely applied both to existing projects and to proposed new projects.
9. ELIXIR should engage in activities designed to stimulate the application of its resources to societal benefit. Applications in health and medicine are an obvious, but not exclusive priority.
10. ELIXIR should strive to provide integrated views of data which enable their exploitation as a whole, and also views which are targeted to the needs of key user communities, e.g. in drug discovery.

2. Introduction

Modern life science research generates huge quantities of data. These data inform all aspects of biological and biomedical research and are valuable at all stages in the scientific process. At the one extreme, undirected exploration of the information can throw up exciting hypotheses. At the other, the data necessary to test a specific hypothesis may be available without the need for any new experiment. It is widely accepted in biology that open sharing of data[1] optimises the benefits to science and ultimately to society. A diversity of data is now available in publicly available collections. These are provided by institutions throughout Europe with or without tools to explore them. The activities are supported by institutional, grant and infrastructural funding from national and European governmental, charitable and commercial sources. Although ELIXIR is concerned with

[1] Concerns about the confidentiality of personal data and the rights of individuals to exert appropriate control over the use of their personal data must be respected.

shared public collections, some related data are made available under commercial license schemes, and others under hybrid schemes, which provide free access for public science but charge for commercial exploitation. The original data come from many sources in all sectors of science. Historically data collected by individual scientists were collected in composite public databases for "secondary" usage. They still are, but nowadays there are also major public data collection projects (e.g., for genomes) which have the explicit goal of generating data of generic value.

Taken together the interconnected collections of data form an "information space" which has become a key part of the life-science infrastructure. Over three decades this information infrastructure has evolved in response to ever-changing scientific need. The overall strategy, structure and support mechanisms for the infrastructure have exploited the funding methods of pre-information-intensive science, which have adapted, with differing degrees of success, to the new need. A major ELIXIR goal in this preparatory phase is to engage stakeholders in a rational examination of life-science information infrastructure in Europe and produce a plan which maximises its utility and gives it a secure future.

Other workpackages in ELIXIR address key tasks essential to these goals ranging through technical, legal, financial and many other matters. However the data resources workpackage is at the heart of the project.

Translating science into societal benefit

While the justification for ELIXIR is often phrased in terms of basic science – improved understanding of living systems - it is worth stressing that the science is readily translated into applications of immediate and substantial societal benefit. All endeavour which touches living systems will exploit the knowledge made available under ELIXIR. Examples include:

Health and medicine — Advances in medicine, such as better drugs, personalised treatments and understanding of genetic risk, are obvious benefits from the investment in the biological information infrastructure.

Personal care — already the producers of personal care products such as shampoos and skin creams utilise the methods and data of bioinformatics to study the action of their products.

Agriculture — the development of healthy, high-yield farm animals and crops sees great benefit from understanding and potentially manipulating their genetic make-up.

Food science — creating nutritious food depends on understanding the molecular effects of its ingredients.

Brewing and fermentation — the quality, cost effectiveness and environmental impact of such industries can all be improved by the selection and manipulation of the micro-organisms used.

Biofuels — optimising the production of biofuels from plants will involve understanding their biomolecular processes

Forestry — trees which can produce high quality timber and withstand harsh environments are already being selected on the basis of their genes.

Fishery — the now-widespread farming of fish poses problems of disease control, and identifying or developing strains which are genetically disease resistant is seen as preferable to extensive use of drugs to control disease.

Environment — understanding the biosphere at the molecular level is crucial to protecting it, and clean-up methods which exploit micro-organisms rather than chemicals are often less insulting to the environment.

3. The process

The ELIXIR preparatory phase project created a "Data Resources Committee" which has, through a series of meetings, worked to prepare this report. The composition of the Committee was as follows:

Graham Cameron	EMBL-EBI Hinxton, UK (Chair)
Janet Thornton	EMBL-EBI Hinxton, UK (Co-chair)
Jean Weissenbach	Genoscope, France
Rob Cooke	GlaxoSmithKline Plc, UK
Gianni Cesarini	University of Rome, Italy
Des Higgins	University College Dublin, Ireland
Torsten Schwede	Swiss Institute of Bioinformatics,
Dawn Field	Centre for Ecology and Hydrology, UK
Hans-Werner Mewes	Helmholtz Association, Germany

In addition the following scientists attended individual of the meetings:

Dr Ron Appel	Swiss Institute of Bioinformatics,
Professor Marie-Paul Lefranc	CNRS, Montpellier 2
Chris Southan	EMBL-EBI Hinxton, UK

4. The current European information landscape

The journal Nucleic Acids Research publishes an annual database issue describing biological databases. The associated online archive lists more than a thousand databases which have been described over the years. (Galperin 2008) A previous analysis of that archive (2005) indicated that about a third of these originated in Europe, and that they covered the entire biomolecular spectrum. (See figure 1.)

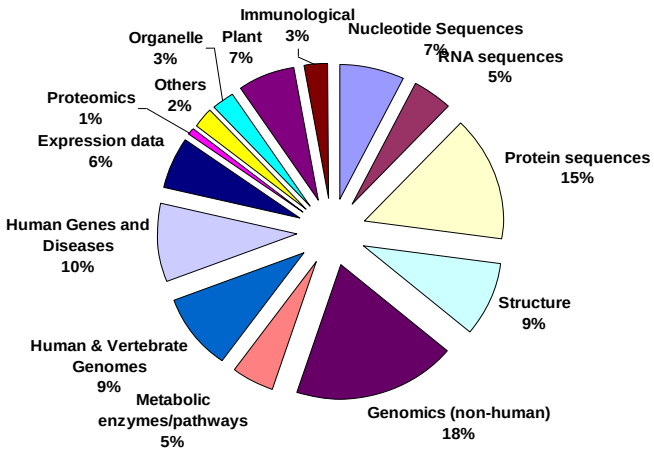


Figure 1. Subject matter of biological databases (2005)

Under the ELIXIR preparatory phase we carried out a survey of biological database providers to establish the state-of-the-art. This was done electronically, and input was solicited from 531 European databases identified from the Nucleic Acids Research archive, supplemented by others found by web search, word of mouth, and knock-on suggestions from the respondents in the initial sample. We received 208 responses from 97 different institutions. This is a pleasing response rate by comparison with most surveys, and it is probably safe to assume that nearly all significant databases are represented, but it would be foolish to assume that we have the complete picture. Of the 323 that did not respond, individual investigation indicated that at least 54 had not been updated since 2005. Clearly some are no longer active. The survey was carried out using the tool "Survey Monkey" (www.surveymonkey.com). The survey report is attached as Annex 1.

A few specialist institutions, like the EBI and SIB (the Swiss Institute of Bioinformatics) reported on a fair number of resources being offered, however, most institutions responding offer only one or two databases. Figure 2 shows cumulative databases per institution. Note that, we will, by and large avoid identifying particular databases and institutions by name as the survey preamble assured them that we would base our analysis on aggregate results.

Although a large number of European nations are involved in database provision (see figure 3), a few are strikingly active. This is typically a result of particular institutions such as the EBI and the Sanger Institute in the UK, the Martinsried Institute for Protein Sequences and the Technische Universität Braunschweig in Germany. In Switzerland, the Swiss Institute of Bioinformatics accounts for almost all the activity.

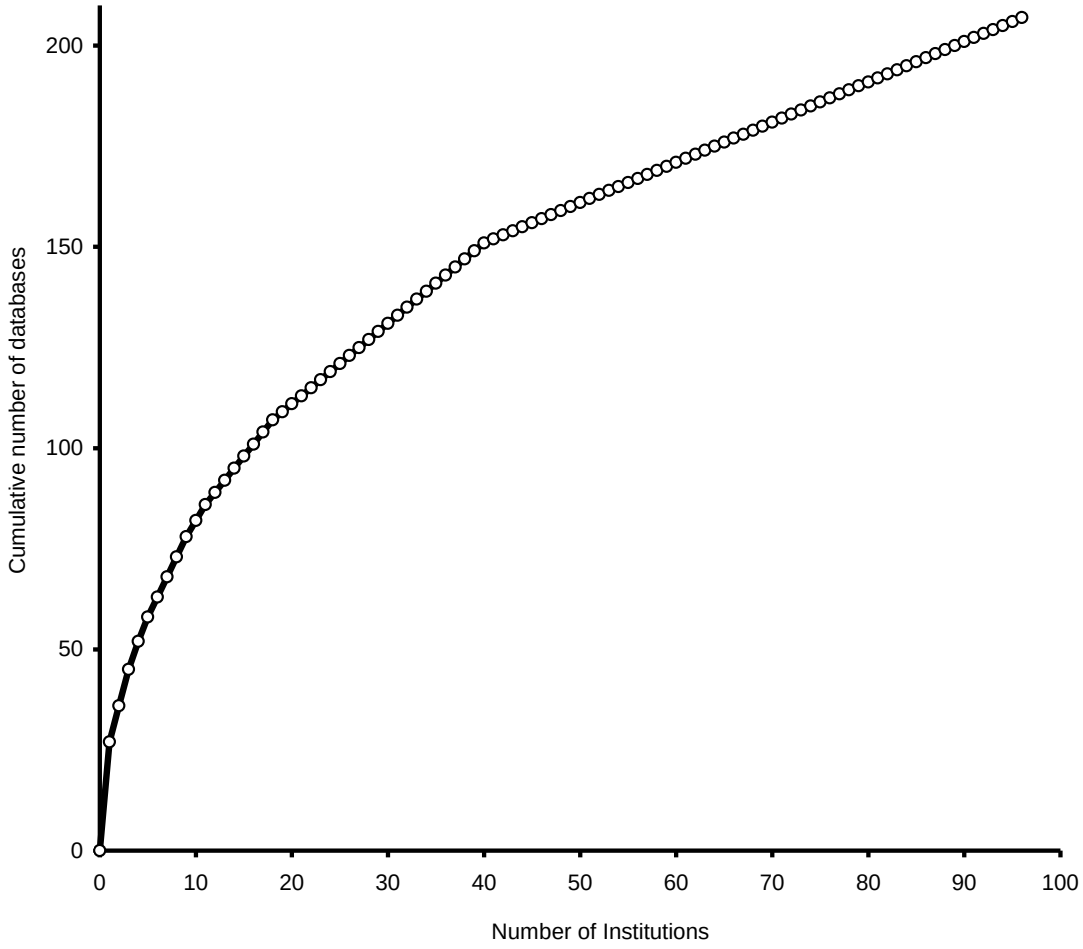


Figure 2. Cumulative databases by institution

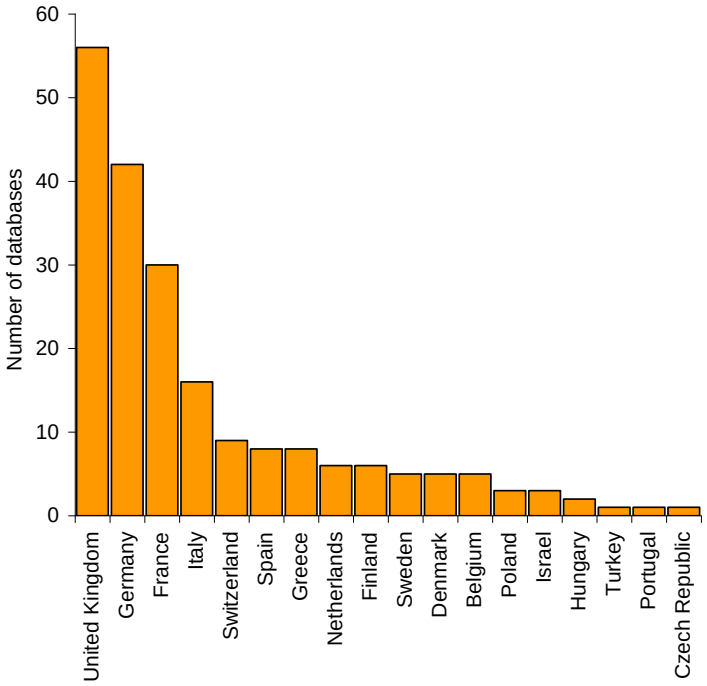


Figure 3. Number of databases by country

The subject matter of the databases polled is extremely diverse. Respondents were asked to assign keywords describing the subject matter of their databases, and most assigned multiple keywords. Table 1 shows the frequencies of all keywords which appeared more than ten times. In all, over two thousand unique keywords were assigned.

Keyword	Total	Keyword	Total
Eukaryotic	82	Enzymes	35
Protein sequence	79	Experimentally determined protein function	34
DNA sequence	75	Mus musculus	34
Gene names	70	Vertebrate	34
Publications	68	Predicted DNA features	32
Genomic sequence	64	Disease association data	31
Species specific	61	Predicted RNA features	27
Protein domains	60	Transcript expression data	27
Homo sapiens	56	Phylogenetic group specific	26
Predicted protein features	56	Drosophila melanogaster	24
Ontologies	52	Experimentally determined DNA features	24
Predicted protein function	46	Images	22
Protein 3D molecular structures	45	Saccharomyces cerevisiae	22
Mammalian	44	Genotyping	20
Prokaryotic	44	Caenorhabditis elegans	19
Transcribed sequence	44	Experimentally determined RNA features	18
Protein-protein interactions	41	Protein expression data	16
RNA sequence	39	Small molecule chemical structures	15
Sequence polymorphisms	39	Clinical data	13
Protein family specific	38	Locus specific	13
Experimentally determined protein features	36	Mass-spectrometry data	10

Table 1. Keywords and their frequency

4.1. Approach to analysis

Many of the questions in the survey asked for responses which collected “binned” data. E.g., question 7 is:

“Please estimate approximate total number of entries:”

And invites responses:

<1K	1-5K	5-10K	…..	100-200 mill	>200 mill

In calculations the median of the bin was used for all responses in the bin except for the open-ended top bin where the lower bound was used.

This means that for the example above the values used were:

500	3000	7500	…..	150 mill	200 mill

All such “binned” questions were treated in this way.

Perhaps the most useful way of looking at the data is as “cumulative charts”. An example might be the total number of web hits as a function of the number of databases, sorting the databases in decreasing number of hits. This method is shown in figure 4, which indicates that collectively, all the databases surveyed account for about 60 million hits per month, with the first accounting for about 12 million and the first 10 or twenty databases accounting for almost all of the total.

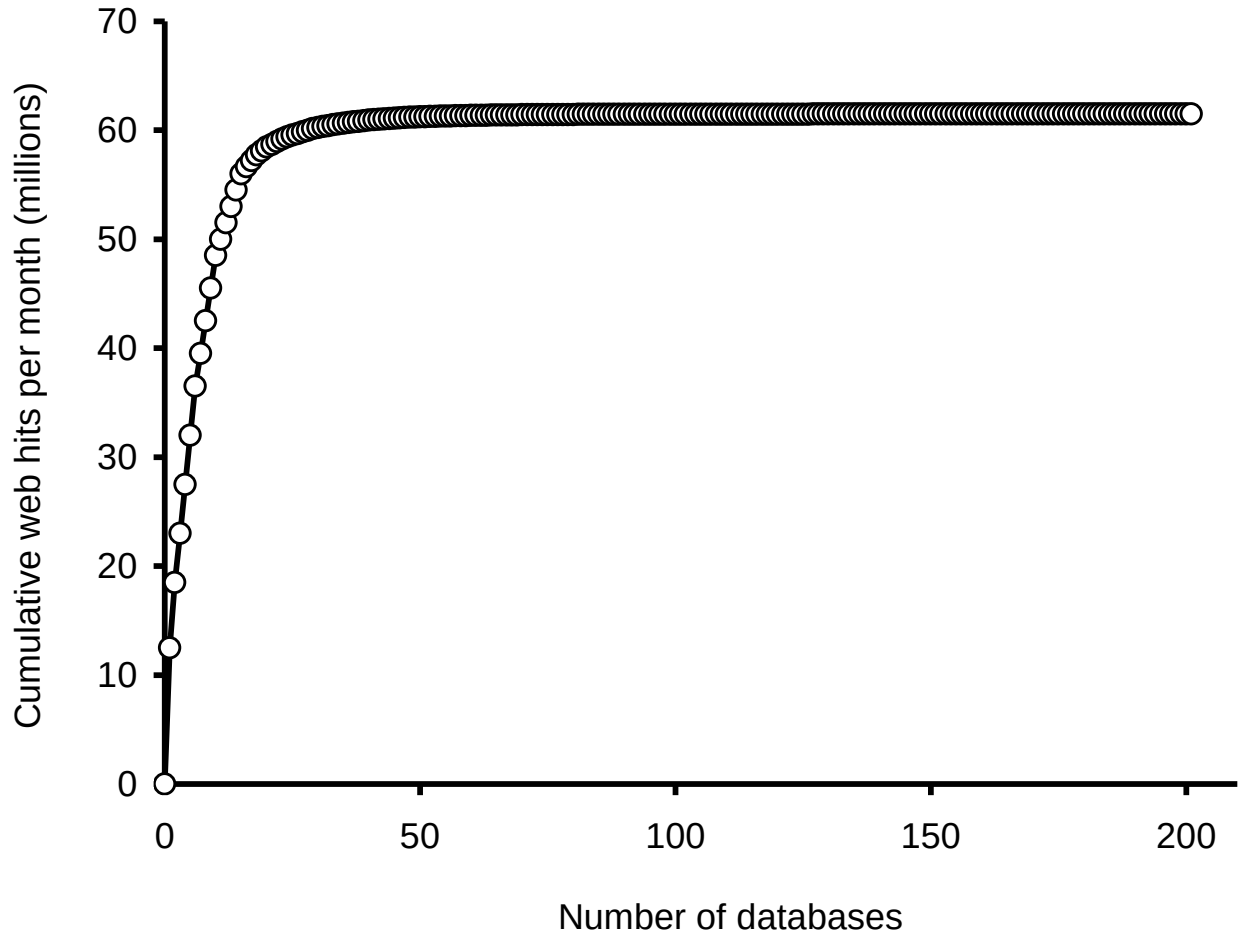


Figure 4. Cumulative hits per month (millions) by database

A few institutions are dedicated to information provision and hold several big databases. The corresponding analysis of cumulative hits per institution shows an even steeper initial rise. Figure 5 shows this analysis. In the subsequent figures we will present the

information by database rather than by institution. It should be borne in mind that, in most cases, the effect of analysis by institution would be similar to that shown in figure 5.

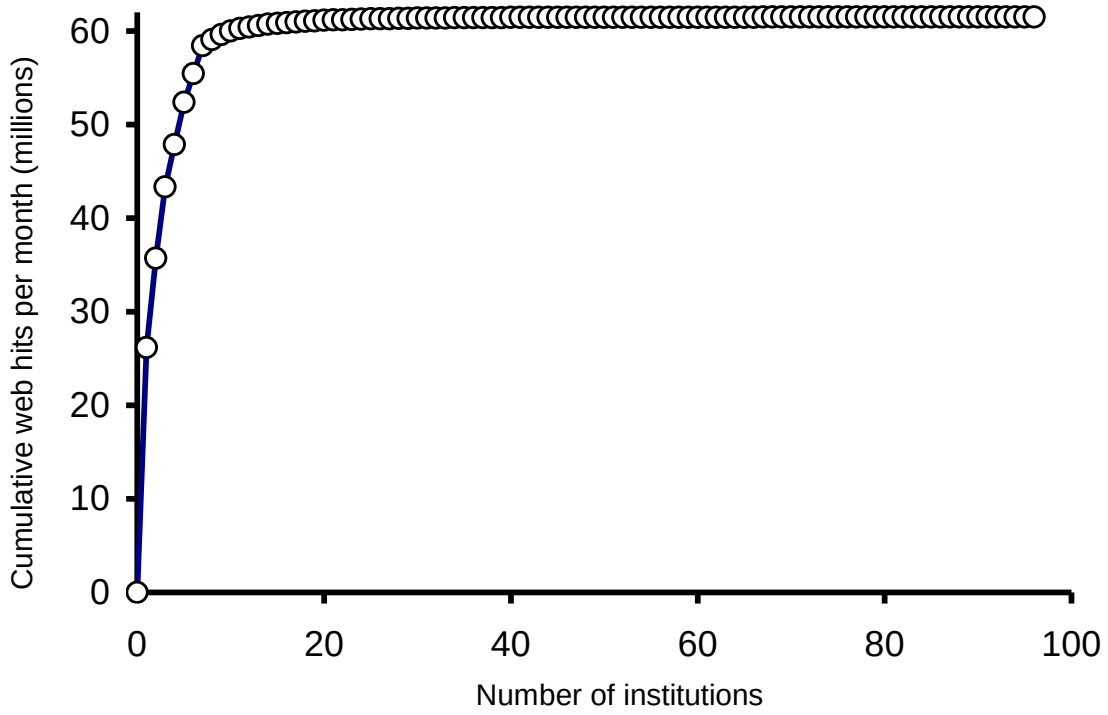


Figure 5. Cumulative hits per month (millions) by institution

4.2. Caveats

As noted above, we do not here identify databases and institutions by name. Aside from our assurance to them that we would not do so, there are good reasons for this. As always with such surveys, questions can be interpreted differently and produce anomalous results. For example, one project whose purpose was to create a database included all the laboratory work necessary to that database in the costs – not what we intended. That anomaly was corrected after a phone call, but others persist. It would be unreasonable to expose them without discussion with individual database providers to allow them to recheck the data. This is a labour intensive task which has not been completed.

It is our intention to reuse the framework of the survey to create an online database registry, where the providers can decide for themselves what to make public.

Despite these concerns, we are very confident indeed that the patterns shown by the aggregate analysis presented here will persist as we revise the methodology, and do represent a realistic overview of the data provision landscape.

4.3. How much data?

Table 2 summarises the reported sizes of the databases in gigabytes. We can see that there are a very large number of rather small databases and only a handful pushing up to the terabyte range. Whether this will remain the case as we start to see the full effect of modern sequencing methods is an open question.

Size	NDB
0 to 0.5 gigabytes	47
0.5 to 1 gigabytes	32
1 to 2 gigabytes	21
2 to 4 gigabytes	13
4 to 6 gigabytes	6
6 to 8 gigabytes	5
8 to 10 gigabytes	7
10 to 20 gigabytes	8
20 to 50 gigabytes	17
50 to 100 gigabytes	14
100 to 200 gigabytes	6
200 to 500 gigabytes	8
500 to 1000 gigabytes	6
1000 to 2000 gigabytes	2
2000 to 3000 gigabytes	2
3000 to 4000 gigabytes	0
4000 to 5000 gigabytes	5

Table 2. Reported sizes of databases

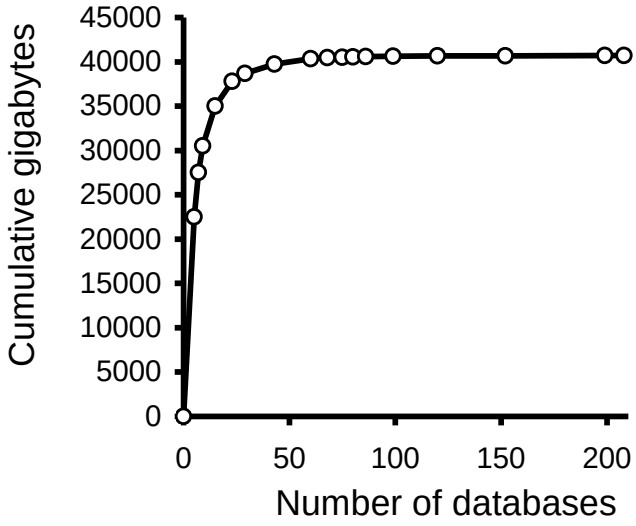


Figure 6. Cumulative gigabytes by database

It should be stressed that the sheer volume of data reflects neither the value to science, nor the effort required to amass them. The effort required to collect molecular structure data, for example, is enormous in proportion to the amount of data collected, but the value to science more than merits the effort.

If we plot these data cumulatively, as in our previous graphs, we see the familiar pattern of a few databases accounting for almost all of the data. (Figure 6) Note that, in this chart we have one point per "bin" from the questionnaire, *not* one point per database. For example, the first point represents the first five databases.

A similar picture is shown if we count the number of entries (individual records) in the database. Figure 7 illustrates the cumulative percentage accounted for by decreasing database size. The overall total number of records is a staggering 2.15 billion.

So overall, about 2 billion records sum to about forty terabytes of data in total from some 200 or so databases, with half a dozen databases accounting for most of the data.

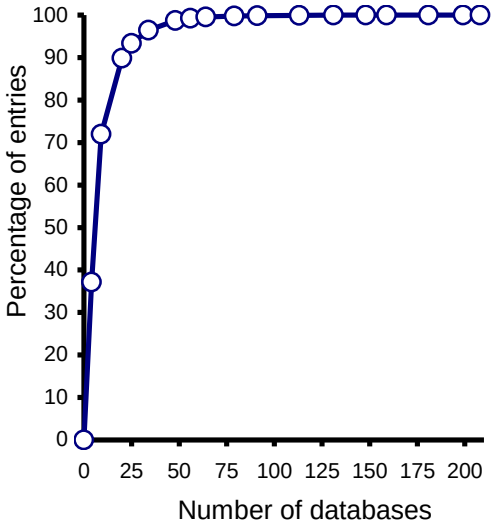


Figure 7. Cumulative entries by database

4.4. Making the data available

Unsurprisingly, all of the databases give some sort of web access to their data. More interesting, however, is the trend to give programmatic access to the information. Already 30% of databases have some kind of applications programming interface, and a further 20% have plans to do so within a year.

In general there are no restrictions on the use of the data in the databases. Although only 113 databases allow the data to be

downloaded in their entirety, 67 of the remainder do not offer this only for practical reasons, with only 28 saying that there are restrictions on the use of the data.

4.5. How much usage?

Many metrics of usage exist, but most commonly the number of web hits is reported. The survey collected the number of web hits per month for the various databases. This is plotted cumulatively in figure 8 (same data as figure 4 in our initial description of the method of analysis).

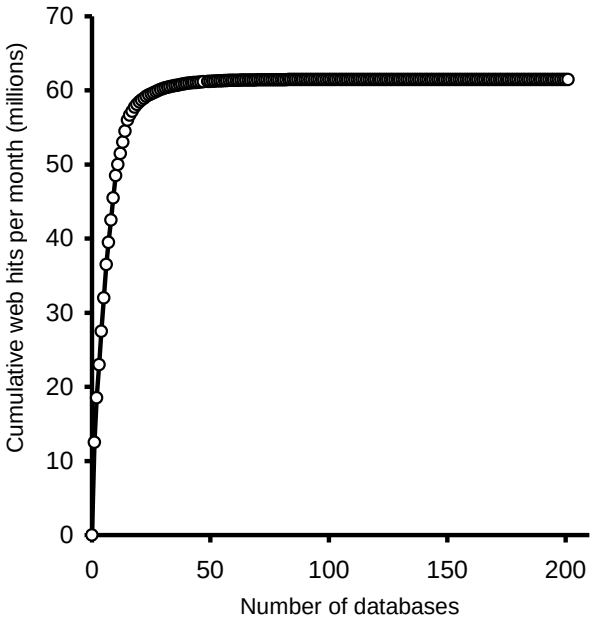


Figure 8. Cumulative web hits by database

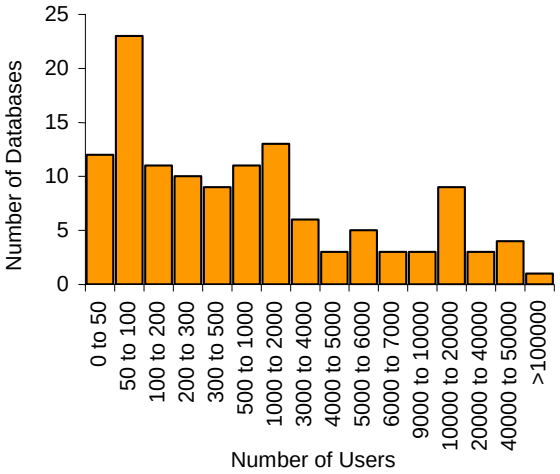


Figure 9. Number of unique users reported

Web hits in themselves, however, are not necessarily a good measure of usefulness to science – the numbers are influenced by aspects of design which have no bearing on the amount of useful information transmitted. A measure often favoured that of "unique users", plotted in figure 9. This too has its weaknesses, all that we collect are unique IP addresses, which are heavily confounded by details of the implementation at the user site. Some sites connect (literally) hundreds of users who appear as one – other implementations can make a single user appear as more than one. Figure 9 shows the number of databases reporting in the various "bins" on the question asking how many unique user they had. (Note that the horizontal scale has no meaning other that that of the category labels. It is not linear, logarithmic or the like.) The same people will of course use more than one database, thus these figures do not allow us to estimate the total number of users in Europe. It is worth noting that the EBI web site sees many hundreds of thousands of unique users in a given year. A few hundred thousand are certainly located in Europe, and they will certainly also use resources other than those of the EBI.

In all, some 60 million web hits per year are testament to the usefulness of these resources to several hundred thousand users.

4.6. Citations

Although database usage provides the most obvious measures of impact, it is worth looking at the rate of citation of the papers which describe the databases. In their thirty year history, the entire population of databases surveyed claim a total of over forty thousand citations. Figure 10 plots the cumulative citations for databases, which shows that citations are not so dominated by a small number of databases as other metrics. There is a modest positive correlation between the number of citations and usage (+0.36[1]), and between citations and the annual cost of the databases (+0.30), but no correlation between database size and citations. The most obvious correlation is between database age and citations (+0.58). This latter correlation is not purely an indication of a greater accumulation of citations over time. Even if we take the number of citations per year of existence for each database, there is still a correlation with age (+0.27). That is, the average rate of citation of older databases is greater. Detailed inspection of the data does indicate that citations are an interesting metric and worth considering separately. Anomalies include:

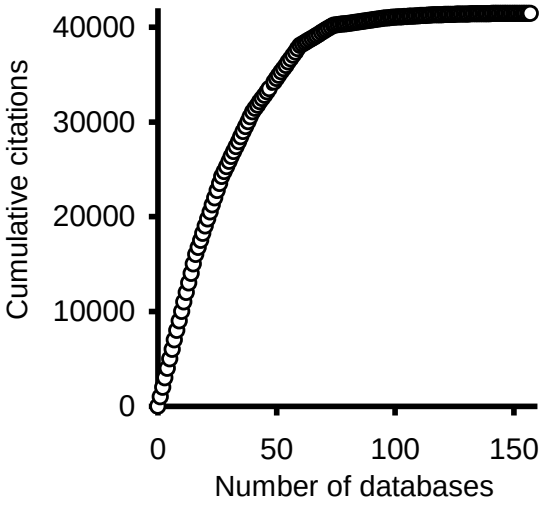


Figure 10. Cumulative citations by databases

- Only three of the top ten databases in size and cost appear in the top ten by citations.
- Three databases in the bottom quartile in cost appear in the top ten by citations.

However:

- None of the databases in the top ten by cost or usage appear in the bottom half by citation.

4.7. What does it cost?

Figure 11 shows the composite of funding from 152 databases which reported some public funding within Europe. As one might expect, most live in some kind of a mixed economy.

Of the remaining 56 databases, five are funded entirely from non-European funds, and four are entirely commercial. This means that 47 report no costs whatsoever! Table 3 presents an interesting breakdown of commercial income. Of 31 databases who report that they make charges to commercial users, only 10 make any commercial income, while of the 192 who report that they are free to all, six in fact report some commercial income.

The latter is not so surprising, as there are other ways of making commercial revenue apart from directly charging for use.

The reported total investment to date of the databases is just over €300 million from only 142 databases which supplied the information. Figure 12 shows the cumulative costs starting with the most expensive databases. Again we see the familiar pattern of a small number of large databases consuming most of the resources.

This pattern is repeated for the €36 million per year spent on the running of the databases, as shown in figure 13. (151 databases provided this information).

[1] Correlations are all Pearson product moments. Various aspects of the data – binned responses, zero bounding, etc. – create data bizarre distributions which prohibit any probabilistic assertions.

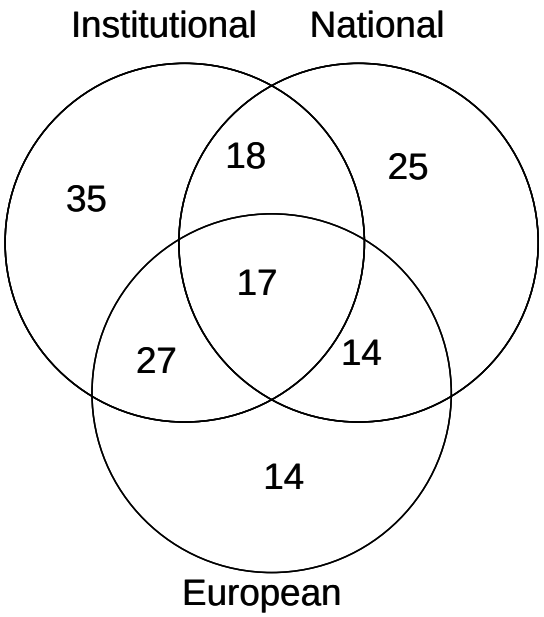


Figure 11. Sources of public funding

	Has no commercial income	Has commercial Income	Total
Academic but charges commercial users	21	10	31
Free to all	171	6	177
Total	192	16	208

Table 3. breakdown of commercial income.

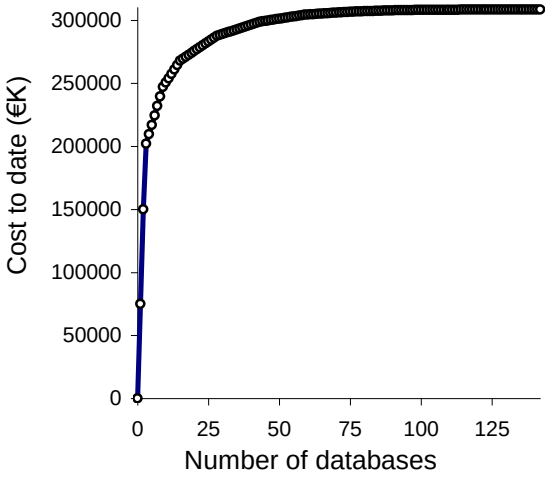


Figure 12. Cumulative cost to date of databases

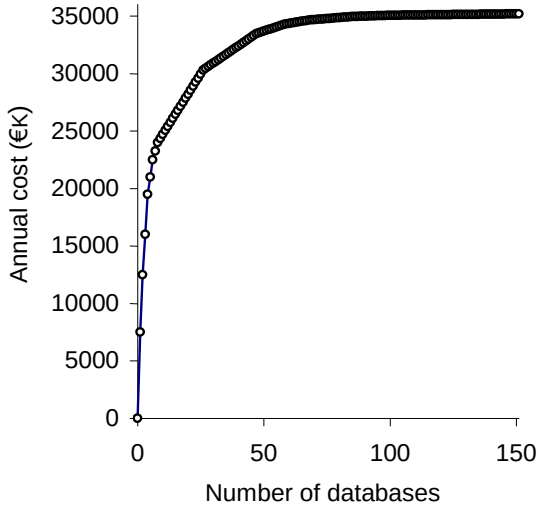
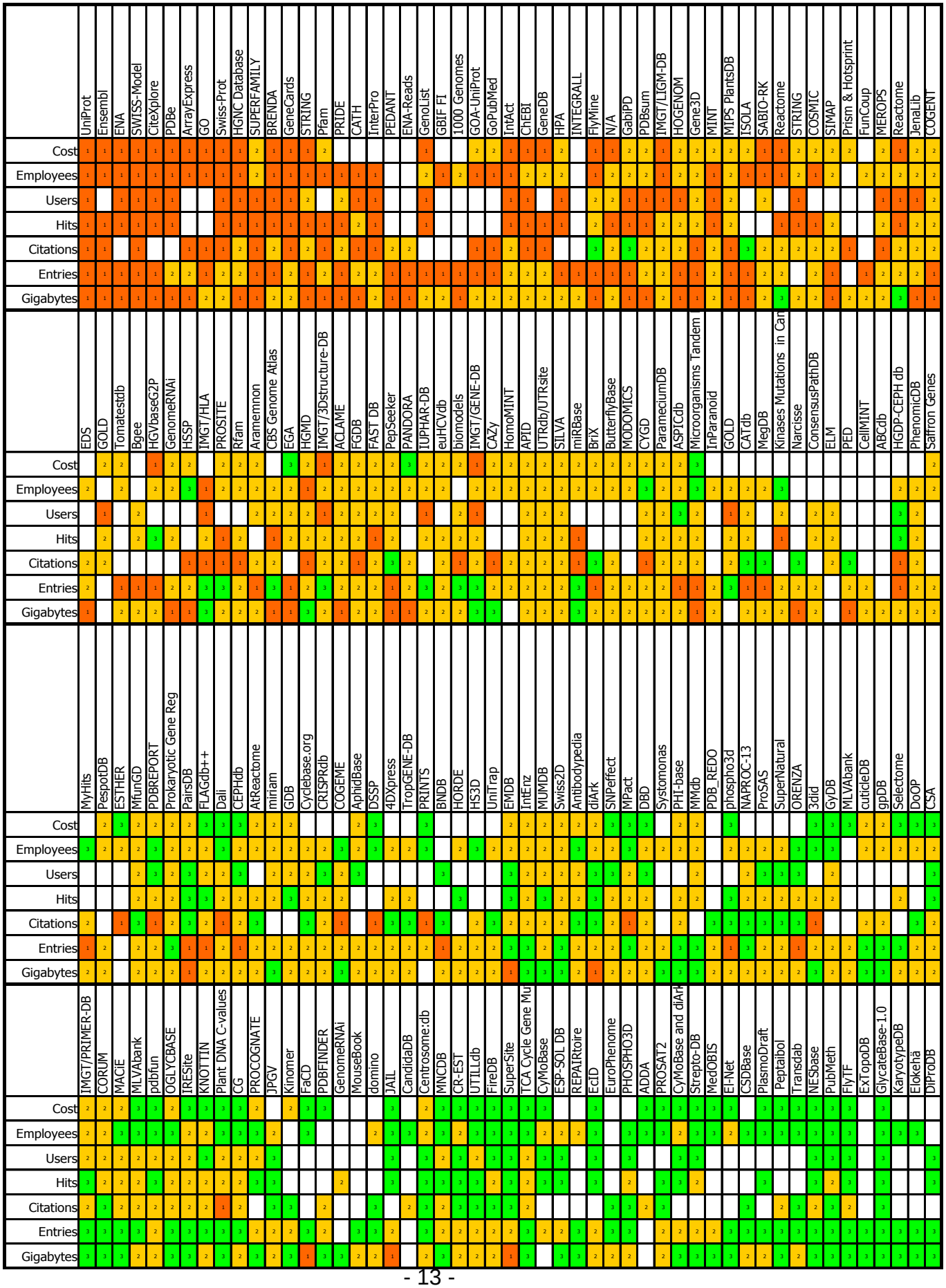


Figure 13. Cumulative annual cost of databases

In looking at each of the database metrics individually, it is hard to get a good overview of the composite picture. Figure 14 attempts to allow one to browse the data in a way which gives insight into the relationships between the main key data collected. For several measures, each database has been cast in one of three categories: quartile 1, quartiles 2 & 3, and quartile 4 – with 1 being the highest and 4 the lowest. The categories are coloured red, yellow and green respectively. The main message is that the various indicators are well correlated, with the more expensive databases having more users, and more data. This provides some reassurance that the current investments are reasonably rational. (Note, the "binning" of responses means that the ranking was coarse grain and incomplete. Boundaries between quartiles were moved to the nearest "bin boundary" so the quartile divisions are only approximate.)

Figure 14. Ranking of databases on key indicators. Red=Top quartile, Yellow=Middle two quartiles, Green=bottom quartile, White=no data



4.8. Conclusions from the survey

Annex 1 presents a full report of the survey, however, here we can outline the main conclusions.

- We identified about 500 public biological databases in Europe, of which 200 or so responded to our survey. Although we only captured information from 40% of the databases, we hope to have uncovered much more than 40% of the total data and total costs – the non-respondents are typically smaller databases.
- About 100 institutions are responsible for the 200 or so responding databases.
- A few specialist institutions offer many databases, but most are conventional academic centres offering one database.
- Only a tiny fraction operate on a commercial basis.
- These databases cover the entire spectrum of biomolecular information.
- They see about 60 million web hits per month from a user community of several hundred thousand.
- The total amount of data served is only about 40 terabytes – this is inconsistent with some recent assertions about data volume and may indicate that there are more underlying data which are not served directly to users.
- The staff of the databases ranges from a small fraction of an individual's time through to a few large databases with 20 or more staff. In total there are about 350 full time staff accounted for in the responses to the survey.
- The estimated total development effort to date is about 1000 person years. (This figure in fact seems low by comparison with the estimated total spend of €300 million.)
- The annual direct cost reported is about €35 million. It is easy to imagine, say, €15 million more associated with serving these databases.
- The reported total investment to date is about €300 million.

As always with surveys, there are inescapable anxieties about how complete or representative our sample is. It is clear that the totals we calculate are lower than the real figures. Our initial list of 500 will certainly have missed some; only about 200 of those replied, and of those 200, 47 reported no costs. The magnitude of this under-reporting is extremely hard to estimate. We feel that a factor of two is not unlikely. ELIXIR will attempt to refine these data over time.

5. The scope

Based on the results of the survey and their general knowledge of the current data resources, the workpackage committee sought to enunciate the scope of the current resources and how this scope might be developed into the future.

ELIXIR espouses a biomolecular focus, and is cautious of over-expansion beyond that focus. In particular the medical relevance of biomolecular data should, where possible, be reflected by connecting to medical data resources rather than expanding the scope of ELIXIR to include medicine.

Biologically active small molecules do bring some chemical information within scope, it being crucially important to have public domain resources available in this area. Shared ontologies will be crucial to this, and should receive appropriate attention within ELIXIR.

It should be stressed that, whilst retaining focus, the need for ELIXIR extends to the unification of the entire biomedical electronic knowledge space. Achievement of this ambition depends on connections to robust information infrastructure in related disciplines, and their support is also crucial. Two ELIXIR workpackages explore the relationship to important neighbouring disciplines:

- WP9 Interdisciplinary Interactions between biological information and Medical/Health and Nutrition Information
- WP10 Interdisciplinary Interactions with Chemical, Plant, Environment & Agriculture Databases

Other projects of the ESFRI roadmap connect well to ELIXIR and in many cases regard ELIXIR as a necessary underpinning to their work. Working groups are being established to develop these interfaces.

ELIXIR's emphasis of its biomolecular "focus" is not intended to exclude the support of any data resources which are not molecular. Our goal is to work together to ensure that the information needs of life-science research are satisfied, and the selection of resources to be supported will be based on that goal.

5.1. The historical development of data resources

Key shared collections of biomolecular information began with the Protein Structure Database (PDB) which was created in 1971 to make available three dimensional structures of bio-macromolecules. It has since grown to a worldwide effort which makes available over fifty thousand molecular structures (June 2009). By the early eighties, DNA sequencing had become commonplace, and the first public nucleotide sequence databases were established: The EMBL Data Library in Europe and GenBank in the USA. This too is now an international effort involving Europe, Japan and the USA. Around the same time protein sequence databases were being established, and these are now unified in the global UniProt project.

In terms of sheer volume, data from genome projects began to overshadow all other data types in the nineties and into the new millennium. This has been accompanied by the advent of new high throughput methods to study diverse aspects of the molecular basis of living systems. These include methods to study gene expression, proteomics, genetic variation and many others.

A much-rehearsed (and of course oversimplified) gene-centric "central dogma" has emerged which goes:

Genomes contain genes

Genes produce transcripts

Transcripts translate to protein sequences

Protein sequences form complex 3D structures.

This central dogma forms an implicit organising framework for much biomolecular data. The not-always-enunciated goal is to understand the molecules and molecular processes of living cells. Metabolites and other "small molecules" encountered in the environment or perhaps introduced intentionally as drugs form an integral part of the biomolecular ecosystem. Information on these molecules is necessarily within the scope of ELIXIR.

Beyond molecular information

ELIXIR's biomolecular focus in principle includes information about biologically active molecules, their activity, and the complex processes in which they are involved. As we gain more information about the molecular components of cells and their interactions, a better understanding of cellular processes will emerge and influence cellular biology research, which itself is moving into high-throughput mode. The ambition of understanding even cellular processes is, however, not enough. There is compelling motivation to connect this knowledge to an understanding of biology at the organ and organism level, addressing obvious questions like:

- What are the molecular correlates of disease?
- Can we intervene with drug molecules and influence the phenotype?

That is, there is a clear medical desire to exploit this knowledge to benefit human wellbeing. The benefits are, however, not restricted to medicine – understanding, say crop plants, could result in insight enabling the production of strains with higher disease resistance, greater yield or with a better tolerance of harsh environments. Further examples are discussed below.

Although ELIXIR's focus is on biomolecular information, the value of that information can only be realised if relevant phenotypic information is available.

However, even if we are fortunate enough to have robust biomolecular and phenotypic data, we still cannot make sense of that information without further contextual information. This includes "sample information" - where the molecules being assayed came from (maybe diseased tissue in a given individual) and other environmental aspects (for example: Was a particular drug being used?)

Thus, although the focus of the ELIXIR infrastructure is biomolecular, this does not mean that it will deal only in molecular information. Optimising the scientific benefits from the information will depend on connecting it to larger scale life-science. It is the hope that cooperation with the information infrastructure of related domains, such as medicine, will allow "joined up" science, but it is also clear that ELIXIR resources themselves will often have to reach beyond the molecular level.

We see thus ELIXIR's scope as including information on biologically active molecules and their behaviour, and the information necessary to understand the relationship of those molecules and molecular processes to more holistic biology.

5.2. Key modules of the information infrastructure

Our survey indicates heavy reliance on key resources which are central to the biomolecular focus.

Below, we summarise some of these core information building blocks which ELIXIR must support. Of course, for many of these, there are already established databases providing service today. However, in discussing them, this report does *not* endorse existing operations. Inclusion in this list indicates that the *functionality* is required. It must be provided either by an existing source or by some replacement of that resource – discontinuation without replacement is not an option.

Molecular components

The fundamental information on genomes, genes and gene products (our central dogma) is extensively documented in electronic form. This results in a few databases whose main function is to describe the molecular components of living systems. They include:

DNA and RNA (nucleotide) sequences have been determined from approaching half a million species, and over 150 million sequence records are stored in the shared global archive (June 2009). These include data sets generated by individual scientists and complete genomes from genome projects. These sequence data have "annotation" attached to them describing their source and the biological properties of particular sequence regions. The most important annotation is that of genes - sequence records contain assertions about the location, structure and products of genes. Although the initial elucidation of gene function involved painstaking laboratory research, nowadays genes can receive an initial annotation by computer programs which exploit existing knowledge to identify and characterise genes on the basis of similarities and known patterns.

To carry out biological research, access to the entire archive of nucleotide sequences and their annotation is essential.

Protein sequences nowadays are almost all determined by translation of the gene sequences, and today there are more than nine million (June 2009) protein sequences in the public domain. These too are accompanied by detailed annotation, and sensibly, the most detailed documentation of the biological function of proteins is recorded in the annotation of protein sequences. These data too are an indispensible component of the biological information infrastructure.

Characteristics of protein sequences – Evolutionary relationships between genes of similar function are reflected in sequence similarities. Thus biological function can often be inferred by looking for similarities and for sequence patterns which occur in proteins of known function. Information collections which document motifs and domains typical of certain functions have been built, and are used as a tool to characterise novel proteins. These collections are extremely heavily used and have become a crucial part of the infrastructure.

Protein structures determined by x-ray crystallography and by nuclear magnetic resonance (NMR) give perhaps the most detailed information possible about biological macromolecules. Structure and function are inexorably linked. Indeed one could view the sequence based function predictions as a proxy for the far more demanding business of determining three dimensional structures in the search for function. Structure determination is difficult, and is not amenable to the same levels of automation as sequencing. Despite this, over 50 thousand structures have been determined, and automation of some parts of the process has caused a substantial acceleration in the production of new structures.

Small molecules – The information so far discussed all broadly relates to molecules directly encoded in the genome, and the "central dogma" which we have iterated helps to form an organisational framework and define natural lines of cleavage between the component information collections. Further information collections on biologically active molecules not encoded by the genome – chemicals – will also be necessary, for example to understand the action of drugs, metabolites and nutrients.

Molecular "behaviour"

The core information collections on molecular components can be distinguished, albeit imperfectly, from information on the "behaviour" of molecules. This includes where and when they can be found, and what interactions and processes they are involved in. Increasingly sophisticated high-throughput methods give ever more fine-grain data on what molecules are present in particular samples and their interactions and roles.

This information is qualitatively different from our parts-list of components. For example, the presence of a particular protein or metabolite can be heavily influenced by aspects of the cellular micro-environment, some of which may be under experimental control and some not. The data therefore are "snapshots" of a changing system, unlike the relatively universal information on molecular sequence or structure. This poses challenges in deciding what is worth storing, and in interpreting available data.

Transcription is the process whereby RNA sequences corresponding to genes are created from the genes. (The translation process creates proteins from these RNA transcripts.) Within biological systems there is detailed control over which gene products are created under different conditions. Some products are required in some tissues but not in others, some are required at particular developmental stages, some in response to the presence of particular drugs or nutrients and so on. Modern methods, initially based

on microarrays, now on rapid sequencing, allow whole-scale analysis of the level of transcription of genes. This can yield key insights, for example, into the mechanisms of cellular processes, disease states and responses to drugs. Shared collections of transcription information already exist and these can be used to generate 'molecular atlases' of which genes are expressed where and when in an organism. Such atlases will be an essential component needed to understand biology.

Proteomics methods, based on mass spectrometry and 2D gel electrophoresis, allow the detection of protein molecules in cells, and are becoming more precise and facile. This caused an acceleration in the acquisition of proteomics data of enough value to generate shared collections, which will be a further component of the ELIXIR infrastructure.

Molecular interaction data — Interactions between the molecular entities must be documented. This includes biomolecular interactions and interactions between chemicals and biological systems including drugs effects and side effects.

Pathways — Information resources which document the composite of interactions, processes and influences which build pathways will form an important part of the infrastructure.

Models — The above examples of core information illustrate but do not exhaust the requirements. In the era of systems biology, resources which document the molecular processes have emerged. These include databases of:

- molecular models,
- mathematical models of biological processes,
- simulations of molecules and processes, which attempt to describe biological processes in a quantitative and dynamic way.

Core data and more data

Core information resources, such as those described above, aim to provide complete collections of generic value to life science. They coexist with a broad range of databases with diverse motivations, often specialising in a particular scientific topic. Core databases tend to have a *de facto* stamp of approval from the community and from key stakeholders such as funders and journals. Non-core databases exist for a variety of (not mutually-exclusive) motivations which include:

Investigator-led databases are typically the product of research groups (though they may well be served to external users). Their content reflects the research interests of their provider (E.g., documenting catalytic sites).

Organism specific databases store information about a particular organism or class of organisms in a depth which would not be possible in the general purpose databases.

Specialist databases handle data whose structure cannot easily be represented in the more general database (say immunoglobulins).

Derivative/Summarising databases combine and organise data from a range of other databases, such as a non-redundant set of coding sequences.

Support databases are built to support the operation of the core databases or to be used in conjunction with them to increase their value. For example they may provide controlled vocabularies for a range of core databases (say organism names).

These non-core databases must be supported and made visible under ELIXIR. Their support may include local (national) funds as well as funds targeted to the scientific topic which they address. The extent of ELIXIR core support depends on as yet uncertain decisions from the member states. They may prefer a model where funds for local

databases are provided directly from national sources rather than channelled indirectly through the combined ELIXIR pool of funds. Workpackages four and five are dedicated to the organisational, legal and financial aspects of ELIXIR and will explore these possibilities.

Research literature

Databases and literature have been tightly connected ever since the first biomolecular databases appeared. Workpackage 8 explicitly considers the literature and its interactions with the data resources. Data records appearing in the databases cite the relevant literature, and, for many kinds of data, the literature quotes "accession numbers" or other identifiers in the databases. For decades researchers have searched scientific literature electronically through collections such as Medline, which, until relatively recently have included electronic abstracts but not the full text of articles. Despite the connections between data and literature, historically they have been served to the community rather independently. This is changing. Nowadays most new literature is available in electronic form, and the business model for scientific publishing is adapting in ways which diminish restrictions on the access to the literature.

In addition, the imposition of defined structures on electronic articles and the ability to add value through data mining combine to make the literature more database-like. All of this promotes tighter connections between the literature and the databases, and invites the development of services which combine access to the two. Whether literature resources are seen as part of the ELIXIR infrastructure is an open question, but, to be competitive, ELIXIR must offer combined access to literature and data. It seems likely that the best way to achieve this will be to include it in the scope of ELIXIR.

Joined up information

Identification of the key kinds of information necessary to the ELIXIR goals is only a part of the task. It poses some challenges, and, even if we make astute initial choices, science will advance and revisions will be necessary. In section 6 (page 20) we discuss methods for priority setting. However, in some senses, agreeing the palette of information to be included is the easy part of the problem. Serving this information to the user community in ways which optimise its utility is extremely challenging. (Workpackage 3 explores the needs of the users in some depth.) The challenges are manifold, but perhaps the most significant are:

Accurate rapid searching. How do you take a user to exactly the information of interest when the entire archive reflects all the repetition, similarity and confusion of science? Typical searches return hits to multiple database entries often containing obsolete or fragmentary information, even where there is also a clear, complete and definitive entry.

Integrating the information. In the discussion above we describe components of the information as discrete entities. In fact the scientist is thinking in terms of biological processes, and wants to see all of the information relevant to the current topic of concern. Looking at a particular gene, for example, one would wish to explore its products and variants, their reactions and pathways, expression information, drugs which act on its products, relevant literature, all with ease.

Presentation of the information in ways which take the scientist to the core facts is challenging in such a multidimensional space, and skilled programming is necessary.

Workpackage 12 is dedicated to the tools and methods of bioinformatics. Much of its subject matter is the tools which *analyse* the data, e.g. sequence alignment tools. There is something of a distinction between these and tools used to *navigate* the data, e.g. database search tools. The boundary is, of course, blurred. However it is impossible to

consider the databases, the subject matter of this workpackage, in isolation of the navigation capabilities offered. Intelligent search tools are a clear prerequisite to enable scientists to home in on the topics of interest. Alongside these tools, there is a requirement for integrative *views* of the data to facilitate both the searching and the presentation of the data.

For example, a researcher from a pharmaceutical company studying drug targets may want to look at human genes presented with a composite of all of the relevant gene, product, variation and other information. The experience should be that of browsing a database of drug targets, even although the information browsed is actually stored in a range of databases and assembled by software into the required view. Although the information engineers who assemble the data collections may have a mindset which focuses on "instantiated" data resources, these user views are essential components of the system, which must appear every bit as real to the user as the actual databases.

6. Changing needs and mechanisms for prioritisation

Many components of the required information infrastructure are uncontroversial, and above we have outlined some which seem clear candidates for inclusion. However, our vision is constrained by the perspectives and priorities of today – science will advance, and the needs will change. We must have transparent processes for reviewing the scope of the infrastructure and adjusting it to suit evolving scientific requirements. Funds, no matter how generous, will be limited, and priorities must be set. At the strategic level, the governance structure described in workpackage four will play a large role in steering the project and ensuring accountability. However it will be part of the routine business of ELIXIR to consider and amend the palette of information included on the basis of a robust and transparent analysis. Resources to be supported under ELIXIR must be required to make a strong case for that support. This is not only true for the initiation of support, but also for its continuation. Existing projects must be rigorously reviewed on a regular basis to ensure that they remain valuable. This scrutiny will be based on scientific justification and, where possible, quantifiable metrics. ELIXIR's leaders must establish procedures to do this.

A part of the motivation for ELIXIR is to ensure continuity within the information infrastructure. This need for continuity will influence the nature of the review procedure and the implementation of its recommendations. It is likely that there will be a significant "rolling commitment" to key resources, meaning that subject to satisfactory review, they typically have several years future security, so that, should there be a decision to discontinue or replace them, a transition period, for example to a new supplier, can be allowed.

6.1. Criteria for ELIXIR support of data resources

The committee considered that, in order to be supported under ELIXIR, a data resource will need to provide justification based on some, if not all, of the following criteria:

The scientific need

The case for a data resource must demonstrate that there is a genuine scientific demand. Demand can take several forms. Some are listed below:

- User demand — Actual and/or anticipated user communities.
- Data generator demand — New scientific projects to generate public data collections often require a custodian for their data. The mission of ELIXIR is to serve whole communities, not to provide support for individual projects. However ELIXIR may be a natural home for the data if they are of general utility. ELIXIR should encourage data

generators to budget for the storage and sharing of their data rather than simply assuming that core ELIXIR funds will be available.

- Funding agency demand — Funding agencies sometimes recognise the need to provide for custodianship of data from their projects. ELIXIR should view very positively such requests. However, their appropriateness to ELIXIR must be measured against the other criteria outlined here. The fact that a project is fundable is not in itself justification for embarking upon it.
- Journal demand — For many kinds of data it makes sense for journals and databases to work together. This can be initiated by either side: databases may evoke the support of journals to ensure that data associated with publications are deposited, or journals may ask databases to act as sources for data associated with publications. This connection to the scientific literature is valuable, but, again it should not overshadow the need to demonstrate scientific appropriateness.
- Standardisation and connectivity — Clearly any service must support community-accepted standards, but, furthermore, part of the justification for a project may be its contribution to standards. For example, a database may have a role as an "external authority", say in standardising nomenclature.

Context and appropriateness

Information resources included in ELIXIR must be appropriate to its focus and mission. Issues to be borne in mind when considering appropriateness include:

- Does the resource really fit into the ELIXIR's scientific focus, or at least contribute to the overall ELIXIR mission?
- Is the resource made available on a basis which is acceptable to ELIXIR? ELIXIR is about data sharing, and thus some kinds of restrictions on data use and reuse could disqualify a resource from inclusion. For example, a resource made available under stringent commercial license arrangement is unlikely to receive ELIXIR support.
- Does the scientific community see ELIXIR as the right home for the resource?
- If there are peer databases world-wide who share the task, data exchange with them must be agreed.
- Are other organisations outside of ELIXIR willing and able to carry out the task in a way which will satisfy the scientific community, thus making it unnecessary for ELIXIR to include it?
- Are there strong strategic reasons for the project, and for ELIXIR involvement? E.g., Is a European presence necessary in a global endeavour?
- If ELIXIR involvement would involve direct competition, will the proposed resource be strong with respect to the competitors?

Database statistics

For existing resources, the case for support can be accompanied by quantitative information about the resource, its costs and its usage. The ELIXIR Data Provider Survey, discussed in section 4 (page 4), has generated some of this information for existing data resources in Europe. Statistics could include information on:

Size and complexity

Measures of database size both in scientifically meaningful terms (e.g., number of structures) and in sheer gigabytes. This can be accompanied by growth rates and some measures of complexity (e.g., the number of tables in the database schema).

Data acquisition rates

The flow of the data into the databases can be quantified. This includes both data submitted directly from scientists and that acquired by exchange with other databases.

Usage Statistics

There are many possible metrics of usage, and weakness in them abound. Web hits, number of unique users, and amount of data transferred to users are obvious examples.

Cost and value for money

Costs of running resources must be justified by the scientific case and usage information described above.

7. Principles of data sharing

The very existence of the ELIXIR project reflects the fact that data sharing has become the norm in the biomolecular domain[1]. At an early meeting, the committee expressed the opinion that ELIXIR should "espouse the strongest possible public domain principles". Data supported by the infrastructure should be downloadable in their entirety and subject to no restrictions in use and reuse. Insistence on acknowledgement of the data source is considered acceptable, as is prohibiting the distribution of a modified data collection in a form which could be confused with the original.

The time at which it is reasonable to expect a research project to release its data reflects the practice of particular communities and project motives. However it was felt that any data discussed in a publication should be made available before or at the time of publication. In traditional science, it is normal and reasonable for a scientific group to retain their data until they have completed and published their analyses.

Nowadays, of course, some high throughput projects are funded explicitly in order to create datasets of general value – genome projects being the best example. It was felt that, where the core justification for funding was to create data of value to the community, there is an obligation to release the data as soon as they are of potential value to others. This does not mean that the databases should be swamped with preliminary data or unreduced raw data, but it does disallow the withholding of data from this kind of project to allow privileged exploitation before release.

Clearly where biological data can be identified to individuals, appropriate access restrictions associated with confidentiality, consent and ethics must be applied. This, however, should not be confounded with protectionism, and aggregate data which breaches no such constraints can often be made available.

It is ELIXIR's goal to ensure that Europe has an information infrastructure for biological research which is:

- Complete
- High-quality
- Of optimal usefulness to science
- Good value for money

The time honoured and cost-effective way of amassing and ensuring the completeness and quality of scientific information is by utilising the goodwill and expertise of the scientific community. The data resources supplied under ELIXIR are populated with information generated by scientists, and made available by those scientists to ELIXIR free of charge and free of restriction. ELIXIR does not own the information it makes available, it is the custodian on behalf of the scientific community. It therefore has no moral right to charge for the right to use the information, and attempts to do so would certainly damage the goodwill on whose basis the information is deposited.

[1] Field *et al*, 'Omics data sharing, Science (2009) (in press)

Optimising the usefulness of the information depends strongly on not restricting its use. The ability to analyse entire datasets without constraint, to follow links to connected data without concern about ownership, and to rework and recombine different data is the essence of bioinformatics. Any restrictions which inhibit this creativity stifle science.

Some data providers assert that their data are "in the public domain" when in fact they mean that a user is at liberty to view the data through a portal which they provide, constrained by the capabilities of the portal. This is not consistent with ELIXIR's policy, which expects data to be exploitable in their entirety.

ELIXIR's foundation is in basic research, and it is very strongly of the view that the information it provides should be freely available to public research at the point of use. Any other system is bad for science, will breed ill-will and will create administrative overhead which simply makes the whole exercise simultaneously less effective and more expensive.

What of our industrial users?

It can be argued that:

- Commercial operations are accustomed to paying for services, and it is reasonable that they should contribute to the services of ELIXIR.
- Service providers could earn some money from industry to lessen the burden on the public purse.
- The ability to attract commercial funding demonstrates to government bodies that ELIXIR is of economic benefit, thereby adding weight to the arguments for its public support.

These arguments are completely accepted by ELIXIR. ELIXIR should seek funding from the commercial sector, and we have every confidence that it will do so successfully. However, ELIXIR should *not* raise commercial revenue through mechanisms which require licensing of the right to use the data or impose restrictions on the use of the data.

The historic evidence is that industry is willing to contribute to the information infrastructure in recognition of the benefits it receives, even in the absence of any negative consequences for not doing so.

8. ELIXIR and the future

8.1. Trends

The overall financial picture for ELIXIR is examined in Workpackage 5. Here we consider only the costs of the actual data resources. Talk of 200 or more databases and a total expenditure of €300 million to date might at first sight seem alarming, but by comparison with the total European life-science research expenditure, it is not huge. However, if we plot the population of databases in our sample over time, we see an alarmingly steep curve. (Figure 15), more so if we extrapolate the trend out a few years (Figure 16). However, this growth is not reflected in exponentially growing costs. The truth is that small biomolecular databases come and go over time, but the key core databases are enduring. The average age of the databases responding to

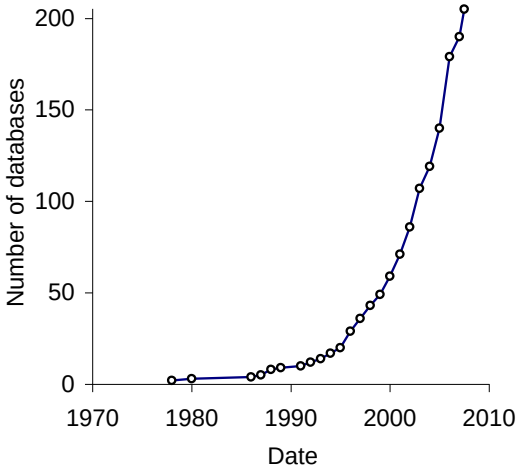


Figure 15. Growth in the number of databases

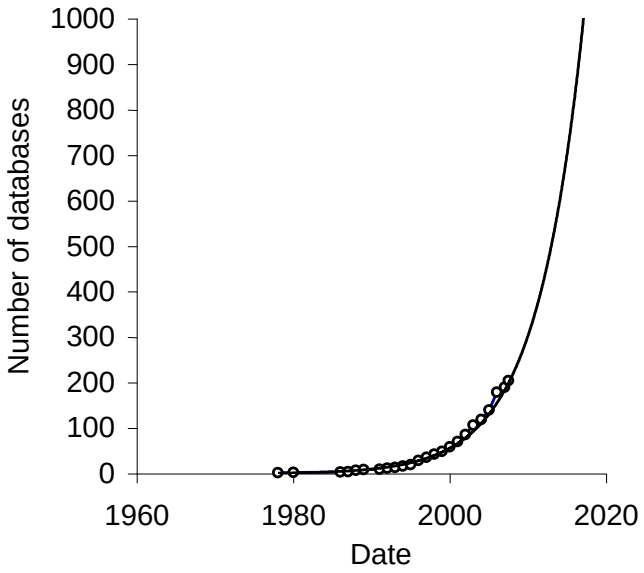


Figure 16. Trend in the number of databases

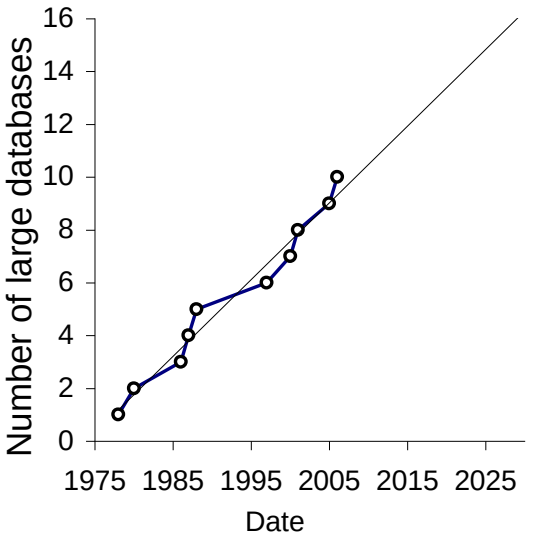


Figure 17. Trend in the number of "large" databases

our survey was just over three years (when the data were collected in 2008). However, if we look only at the data for the ten most costly databases (on annual spend) we discover that their average age is over fifteen years. These ten account for seventy percent of the total annual expenditure reported in the survey. The growth in that population is shown in figure 17. It exhibits a fairly linear growth and is much less alarming.

The most significant costs of the information infrastructure for life science are in the enduring core data resources, and the evidence is that, with the progress of science, we seem to create a new significant data resource every one or two years. Within the ELIXIR database provider survey, we collected no trend information within databases. It was essentially a snapshot of the situation in 2008.

8.2. What data resources must ELIXIR support?

While there are some subtleties about which resources to include in ELIXIR, the current requirements listed in section 5.2 above (page 16) are mostly uncontroversial. What ELIXIR eventually supports will depend on the decisions of the member states about the financial model. It should be stressed that a decision that a resource should not be supported through ELIXIR does not mean it should not be supported at all. Of course, in the real world of limited funds, some proposed resources will not be funded. However, there will be essential resources which will be funded from non-ELIXIR sources, either because they fall outside ELIXIR's scientific scope, or because the member states would prefer to fund them from national sources rather than channel the funds through ELIXIR.

Thus, in this workpackage although we have identified components which need support and can speculate about coming requirements, the strategic decisions about support in or outside ELIXIR may have to be taken down-stream.

It is however worth noting that some core projects will be under genuine threat without ELIXIR support. Agencies currently funding many of the key databases expect these costs to be borne by ELIXIR in the future, and have no plans for long term continuation of their own support. The continuation of these databases, be it at their current location or elsewhere, is under genuine threat should ELIXIR fail to provide for them.

It is clear that a palette of core and associated resources not unlike that described in section 5.2 must be supported. It is also clear that our ability to anticipate what will be required in future will be imperfect. Some likely themes can be identified. They might

include images of cellular phenotypes, more information on metabolism, more connectivity to health and disease information, to name a few, However, it is clear that ELIXIR must have robust processes for considering cases for support. The ideas outlined in section 6 (page 20) should contribute to this process.

9. Recommendations

The discussions above can be summarised in the following set of recommendations from the committee:

1. ELIXIR should ensure the existence of the information infrastructure in Europe necessary to support world class life science research in the biomolecular domain, and in the relationship of biomolecular information to more holistic biology.
2. Information resources on all aspects of biologically active molecules will be required.
3. In the future, core resources of widespread utility may have no source of support other than through ELIXIR, making it essential that ELIXIR create mechanisms for their support.
4. Non-core resources are also crucial, and the success of ELIXIR will depend on them being supported. This support may or may not come from ELIXIR. ELIXIR must work with them to ensure that there are sources of funding irrespective of whether that funding is through ELIXIR.
5. Unrestricted access to the information and freedom to exploit it is a key principle of ELIXIR which must not be threatened. This does not preclude seeking funding from commercial organisations who exploit the infrastructure.
6. ELIXIR should strive to ensure interoperability with neighbouring scientific domains such as medicine, epidemiology and chemistry.
7. While maintaining a biomolecular focus, ELIXIR should exhibit some pragmatism in order to fulfil its mission, including activities on the basis of their ability to contribute to the mission rather than rigid principles of eligibility.
8. ELIXIR should establish processes to assess the suitability of activities for inclusion in ELIXIR. This scrutiny should be routinely applied both to existing projects and to proposed new projects.
9. ELIXIR should engage in activities designed to stimulate the application of its resources to societal benefit. Applications in health and medicine are an obvious, but not exclusive priority.
10. ELIXIR should strive to provide integrated views of data which enable their exploitation as a whole, and also views which are targeted to the needs of key user communities, e.g. in drug discovery.