

6

Setting up a global web-based database: is it feasible?

6.1 How to create a global database

In order to learn more about the possible genetic and environmental causes of CFA (and their interactions), comparable population data sets, collected over similar time periods and from many different geographical localities, need to be analysed. Storing and exchanging this data via a global database is desirable. However, there are a number of questions that need to be answered in order to create such a database. The potential answers to these questions will have to be examined and the remaining gaps in our knowledge on how to construct the database identified.

First, the most fundamental requirement in planning the database is to have access to people with training and skills in both biological/medical and computational/programming methods. Without such people on the design team, there will be a real risk that simple misunderstandings of terminology could lead to poor design and/or implementation of the database.

Second, the experience of some research teams working with global data on cancers in children (Dr Jim Kepner, Children's Oncology Group, Florida, USA) and in adults (Professor Bruce Armstrong, New South Wales Cancer Council, Australia), has confirmed the existence of two major problems:

- to have access to record linkage software, and
- to ensure the integrity of database queries.

Without adequate record-linkage software, important sources of historical, clinical and environmental exposure data may not be available. Geo-coding of current address data may not be possible so important spatial information on the distribution of cases can be lost. Even with the best designed databases, human misunderstandings and errors may lead to the unintentional extraction of data files that are not validly comparable between the participating research teams.

Any endeavour to create a global database will demand adherence to strict security measures ...

It is desirable to include DNA samples within the collection and to have access to the best gene-sequence and protein databases. Modern micro-array analysis of DNA creates large multidimensional data structures. However, the data analysis software supplied with commercial micro-array equipment does not usually offer the full range of statistical analysis tools required by research teams. It is highly desirable that computing software and hardware interfaces easily with the laboratory equipment. With such an interface, data files can be analysed by common statistical packages such as SPlus, SPSS, BMDP and SAS. Analysis using neural network software may also be useful in situations such as CFA research where the causes of the defects are thought to be multi-factorial in nature.

Once established, the database may be used for the registration of clinical trial participants (requiring randomization in real time) as well as for collaborative research on existing data. Existing data collections may need to be imported via a batch-loading process, with complex rules applied prior to importation in order to ensure the integrity of the data. Other issues of relevance to the design of the database include resources available for funding and maintenance, authentication of users, local versus central coding of data, identification of duplicate entries, and tracking participants over time and across centres.

In conclusion, any endeavour to create such a database will demand adherence to the strictest security measures possible in order to safeguard the highly sensitive personal data. If the data were to be exchanged via the Internet, the need for security and ease of access would have to be balanced by the need to ensure that the performance of the database allows quick access to data and information for all collaborators.

6.2 Linking bioinformatics to a proposed web site

In order to gain experience in the field of web site design and development and, in particular, in the linking of bioinformatics facility to a proposed web site, expertise was sought from those involved in the creation of facilities in two exemplar projects:

- Dr Olivier Cohen (*see Section 8, List of participants*) has been developing a bioinformatics platform dedicated to medical genetics in France to allow statistical information to be extracted from the database and to make it available to the medical and research communities through a web site.
- Dr Douglas Bratthal, (*see Section 8, List of participants*) is Director of the WHO Oral Health Country/Area Profile Programme (CAPP) created in January 1996, the purpose of which is the presentation of

a wide range of information on the Internet on dental and oral diseases.

6.2.1 *A web site based on an international human genetic database*

Dr Olivier Cohen has been developing a generically secured bioinformatics platform dedicated to medical genetics in France. The aim of this platform is to allow statistical information to be extracted from the database and to make it available to the medical and research communities through a web site. This information includes genetic nomenclatures and the patient data are described according to the familial genealogy. The database takes into account the different facets of the genetics, such as the clinical, chromosomal and gene aspects.

The aim is to make statistical information available to medical and research communities through a web site...

The web site provides users with information and anonymous data, related to genetic diseases, from the database. International nomenclatures are the basis for describing the chromosomal and gene-mutation features. The *London Dysmorphology Thesaurus* generates descriptions of clinical features from a list of syndromes and symptoms, according to a standard procedure.

The platform was initially dedicated to familial structural rearrangements of chromosomes that concern about one couple in 200. For a given chromosomal anomaly, the web site provides geneticists with assistance in diagnosis and genetic counselling. In real time, the user can get an ideogram of the rearranged chromosome according to international nomenclature (ISCN 1995), the assessment of the risk of imbalance at birth with a confidence interval, and specifically related papers. For research workers, interfaces exhibit the distribution of chromosomal breakpoints and genome regions observed at birth in trisomy or monosomy. These interfaces are interactive and allow the user to make contact in real time; 1000 contacts from about 50 different countries are currently users.

Impact at an individual level: The definition of individual risk factors is of a great importance. Indeed, each carrier has specific risks of imbalance at birth and miscarriages, varying from 0 % to approximately 80 %. Concurrently, different prenatal diagnostic strategies could be proposed, such as amniocentesis or chorionic villus sampling (CVS). The knowledge of the risk of imbalance at birth for each carrier allows for the proposal of a strategy based on objective reasons. Many factors have to be taken into account but, specifically considering the level of risk of imbalance at birth, the best strategy is that for which the risk of imbalance is greater than the iatrogenic risk.

Impact at population level: Since familial structural abnormalities of chromosomes are frequent, the impact of a risk assessment at population level is also very high. An initiative to conduct risk assessment could lead to better patient management with a consequent decrease of:

- handicaps prevalent in children (i.e. handicaps linked to the unbalanced chromosomal abnormalities);
- maternal morbidity (linked to repetitive reproduction failures or late pregnancy terminations);
- iatrogenic fetal loss (linked to CVS carried out when low risk of imbalance exists).

Giving more accurate information to the carriers should increase the understanding of their respective families – who are the actual targets of prevention. Indeed, it is only the patient who can inform his or her direct family members that each of them could be a carrier and can request a simple blood karyotype for screening. The importance of familial investigations is currently emphasised in genetic counselling consultations, but an educational initiative dedicated to the carriers and their families should increase targeted screening with a minimal cost-efficiency ratio.

Monogenic diseases: Using the MIM nomenclature, monogenic diseases can be considered. After registering the individual phenotypic expressions, the chromosomal status and the gene mutation description, the user can send data including the pedigree, through the Internet.

Increased security with smart cards: Each user needs a dedicated smart card to access genetic records. The smart card permits formal authentication by checking with a central electronic directory that certifies the user's identification and qualification.

Thematic networks: In order to improve collaborative networking, which is particularly useful in case of rare diseases, each user can ask for a thematic network to be opened. After agreeing on a charter of use, validated by legal experts, the user becomes the coordinator of that particular network. A diagnostic validation committee controls the quality of the records provided by users who have decided to share their data through the network.

6.2.2 Planning and managing the WHO Oral Health database

Considering the wealth of information available in the oral health field in the early 1990s, there was an urgent need to share it via the Internet in a standardized format that would make the data instantly available to thousands of users. The WHO Collaborating Centre at Lund/Malmö University in Sweden, that not only had wide experience in epidemiology

The main purpose of the CAPP web site is to present information on dental and oral diseases ...

and electronic communications, but was already an established Internet server, was approached for assistance in building, developing and implementing a pilot programme for such a database. The Centre responded positively and several proposals were presented and discussed. Finally, a model that included maximum input and participation from different sources was agreed upon and presented to other collaborating centres and organizations involved in the oral health project. A server to focus on periodontal diseases was simultaneously established at Niigata University, Japan.

Thus conceived, the WHO Oral Health Country/Area Profile Programme (CAPP) was created in January 1996; its main purpose being the presentation of information on the Internet on dental and oral diseases, including data for every country on the availability of its oral health services, dental education and manpower. To avoid a massive build-up of data sets over time, the programme structure was designed so that it would be easy to find and update data and images. Also, to facilitate access from computers with low or moderate capacity, a simple, “non-fancy” design was chosen for the web pages. Most importantly it was decided that, in order not to be dependent on outside expertise, all programming would be done by the WHO department involved.

The Home Page (<http://www.whocollab.od.mah.se/index.html>) of the CAPP has five sections:

- Main CAPP pages,
- About and Help,
- Links,
- Projects and Reports,
- Messages.

In the *Main CAPP pages* section, the user can select a country by either clicking on the list of countries arranged in alphabetical order or by selecting the list compiled on the basis of the WHO regions. Once a country is selected, a page specific for that country appears, showing the topics of the data available, such as:

- general information on the country, e.g. gross domestic product (GDP) and life expectancy,
- oral diseases including caries, tooth mortality and fluorosis,
- oral health manpower,
- dental education,
- oral health care system and services and, lastly,
- information relevant to oral health and care.

This means that each country has several pages of information under different topics. Another section includes *Methods and indices* – a much-used source of information for users who are planning and/or performing studies.

Topics such as oral mucous-membrane diseases or CFA, which are usually not routinely included, can also be presented when standardized study data are available. The ease with which the format and presentation of the data can be modified, and the programme instantly updated and made available to Internet users all over the world, is a major advantage of the CAPP. Furthermore, an active exchange of ideas, questions and comments from users is available through the e-mail service provided on CAPP's home page.

The presentation of CAPP on the Internet has made oral health information available globally. CAPP data can be updated on a daily basis. Information and further clarification is requested by many users of CAPP, including government bodies, universities, companies and private individuals. According to the log, the server was approached during its first year (1996) by more than 15 000 different computers (hosts) and from October 2000 to October 2001, more than a million requests for information from about 70 000 hosts were received. These figures clearly illustrate the growing need for this database and its usefulness all over the world. The number of individual requests recently exceeded 3000 per 24 hours – indicating the future scope and potential of CAPP.