PROPOSAL FOR AN INTERNET INTERFACE TO CALM HERBARIUM DATASETS

Alex R. Chapman

30th April 1998

SECOND DRAFT

Table of Contents

1. Introduction 1

1.1 Background 1

1.2 Prototype success 1

1.3 Vision for the future 2

2. Functionality 3

2.1 The Database Management System 3

2.2 The Firewall 3

2.3 The Browser 3

3. Security 4

3.1 Overview of KE Texhtml 4

3.2 Registration and Access Levels 8

3.3 Charges 9

4. Design Examples 11

5. Logical Site Map 10

6. Personnel roles 15

7. Appendices 15

Table of Figures

Figure 1. hypertext Transfer Protocol Daemon 5

Figure 2. KE Texhtml Process Model 7

Figure 3. FloraBase query form 11

Figure 4. FloraBase summary of records 12

Figure 5. FloraBase detailed record 13

Figure 6. FloraBase specimen data 14

1. Introduction

1.1 Background

The Science and Information Division's CALM Herbarium is custodian for a number of major corporate datasets. These include the:

- WAHERB specimen database of 435,000 herbarium collections which voucher plant science in the state
- WACENSUS plant names database, the authoritative listing of over 15,000 plant names, their currency and usage in Western Australia
- CALMLIB library catalogue, a data warehouse of CALM's three major science libraries, including the State's botanical library

The data contained in these systems records much of the work CALM undertakes in surveys, reserve management and flora conservation. It also provides an ongoing resource for CALM scientists and operations staff by providing both determinations of identity for existing specimens, and authoritative information on the status and relationships of plant names, in respect to the continuous upgrading of taxonomic knowledge.

It is the proposition in this paper that making such information more generally and easily available will not only aid the division's objectives of providing up-to-date and sound scientific information, but those of CALM concerning the conservation and management of the state's flora and the CALM estate. In return the CALM herbarium will rightfully gain acknowledgment as the key authority and information provider concerning the WA flora.

1.2 Prototype success

The CALM Herbarium has maintained the physical collection of specimen vouchers, nomenclatural information and botanical library for over seventy years and since 1985 this information has been entered into computer databases in order to increase the efficiency of collections maintenance and information retrieval.

In 1995, SID developed the prototype of the intranet web site CALMweb. In 1996 a prototype web interface to the herbarium's corporate databases commenced development as an alternative method for simplified database access for casual users. Due to the standardised graphical front-end, this tool quickly became the preferred method for integrating related datasets.

Detailed census data could be accessed from the specimen data or images, maps and descriptions could be accessed from the census. And the simplified query page let users perform powerful queries such as "show me all the conservation taxa originally published in Nuytsia" or "show me the Proteaceae taxa not yet formally published for which we have an image". This information was visible across CALM's expanding Wide Area Network to all staff with a web browser, and was widely used.

1.3 Vision for the future

The CALM Herbarium will be custodian for a part of CALM's NatureBase web site which will provide online access to a range of authoritative flora information products.

The site, to be called FloraBase, will enable a registered user to query on a range of plant name attributes and retrieve a sorted list of species names and related general information. Users with higher levels of privilege will then be able to access more detailed information components regarding a particular species, such as maps, images and family, genus or species descriptions. Users with the highest privilege will also have access to the specimen database records from which the summary components are largely derived.

FloraBase will also contain links to further documentation about the site, pages concerning the availability of publications and CDs of related information (such as the Census, the WA Flora Catalogue or Nuytsia), or in the future, interactive identification aids distributed via the Internet.

By registering users, the Herbarium will be able to effectively monitor and control access to the various components of FloraBase, and tailored to the needs of researchers, the general public and consultants. Where appropriate, access will be limited to a fixed period, and an up-front charge for use of the content can be made. Registration will also include a signed contractual agreement between parties to ensure that the content is not misused or otherwise published in another form.

In this way the CALM Herbarium can protect its intellectual property and data ownership rights, while making the information available to bona fide users and recovering some of the considerable cost of maintaining the information base.

2. Functionality

As discussed above this system has been in use over the CALM intranet for nearly two years and has been refined into a highly usable and efficient data retrieval system, with additional data elements frequently added into the system.

There are three main components for the secure delivery of this data via CALM's NatureBase web site.

2.1 The Database Management System

TEXPRESS from Knowledge Engineering is the object-oriented database management system in use at the CALM Herbarium. It is the core tool in which all the corporate herbarium databases are developed and maintained and it also provides the web interface technology in which FloraBase has been created.

The TEXPRESS database web interface provides a path for external internet queries to pass through the departmental firewall securely via the use of a trusted proxy user. It is this known user that performs the query and returns the result to the internet user, so that the internet user never breaches the firewall themselves. The proxy user has clearly defined access rights and is not able to perform any other function inside the firewall other than perform the requested query on a particular database on a nominated machine, so departmental security is maintained. This method of access has been detailed previously to CALM's Information Systems Section manager and consultant system administrator and been accepted as one of the most secure systems available.

2.2 The Firewall

Firewall software provides an effective barrier to internet traffic entering the Departmental domain of computers. CALM's Internet firewall software is Firewall 1, one of the most secure and respected firewall products on the market.

2.3 The Browser

The internet user must have a browser capable of handling dialog boxes, forms and tables in order to access FloraBase. The database web interface will be compatible with versions 3, 4 and upwards of Netscape Navigator and version 4 and upwards of MS Internet Explorer, the two major browsers currently in the marketplace.

3. Security

The TEXPRESS database web interface provides a path for external internet queries to pass through the departmental firewall securely via the use of a trusted proxy user. It is this known user that performs the query and returns the result to the internet user, so that the internet user never breaches the firewall themselves. The proxy user has clearly defined access rights and is not able to perform any other function inside the firewall other than perform the requested query on a particular database on a nominated machine, so departmental security is maintained. This method of access has been detailed previously to CALM's Information Systems Section manager and consultant system administrator and been accepted as one of the most secure systems available.

The firewall or web server software will provide the functionality of challenging users of FloraBase for a valid username/password combination and if the user is registered and the password correctly entered, a process is spawned for the user at their assigned level of access and they are passed on to the appropriate FloraBase query page.

The following is an overview of the database web interface as presented in the TEXPRESS Version 5 manual.

3.1 Overview of KE Texhtml

KE Texhtml is designed to support an HTTP server machine to provide HTML documents to Netscape Internet clients.

A Netscape client selects a *URL* to connect to a server machine and retrieve a resource from that machine. Although the resource can be in any format, Netscape is designed to support HTML documents and correctly interpret the markers contained within them.

HTML documents can contain a variety of formatting markers which make the presentation of the document more visually appealing. They can also contain **anchors** which are hypertext links to positions within the same document or to other documents on the Internet. References to other documents are made by including their URL in the anchor. In this way, users need only identify a starting URL after which they can navigate through a series of documents, perhaps on a series of different server machines on the Internet, by use of the embedded URLs. This gives rise to the concept of the **World Wide Web** (WWW) where files spread across server machines throughout the world are linked together in a variety of ways, like a spider web.

The HTTP server daemon, *httpd*, provides a means of delivering static HTML documents to Netscape clients upon request. However, it is also extensible, such that a request can result in the execution of a program on the server machine which will generate an HTML document.

The figure below gives an overview of this process.

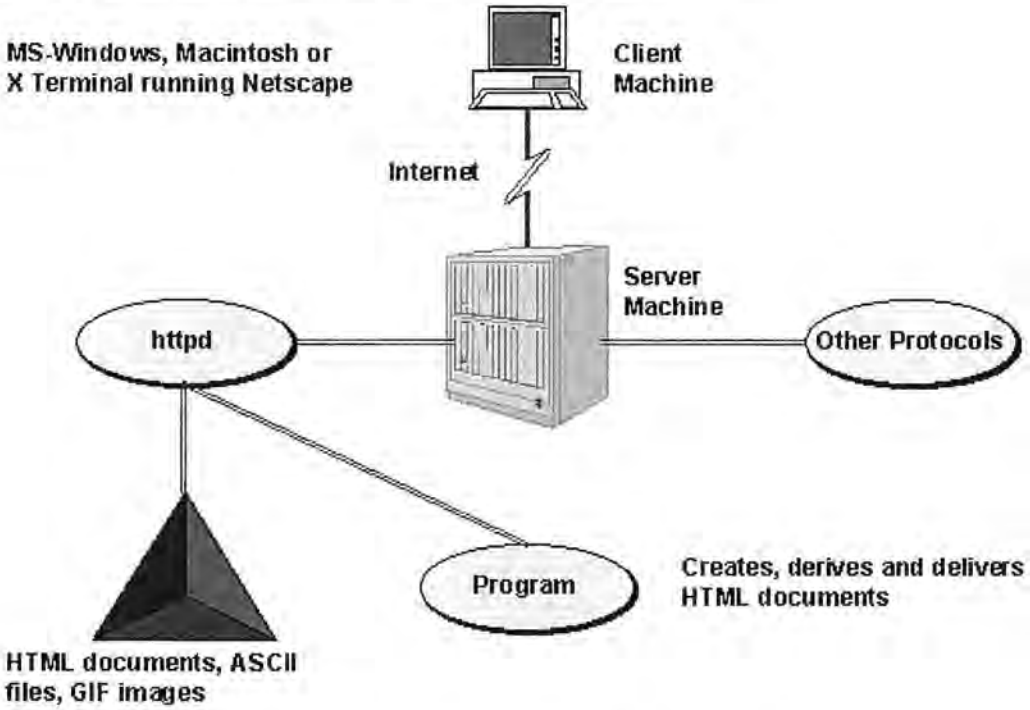


Figure 1. hypertext Transfer Protocol Daemon

KE Texhtml has been designed to be invoked by *httpd*. When a URL with a file name of **/cgi-bin/texhtml** is requested, the request is transmitted across the Internet to the server machine's *httpd* which then invokes *texhtml*. Depending on the associated arguments, one of two actions is taken:

1. If invoked with one argument, KE Texhtml determines the appropriate host machine at the server site from its configuration files and generates HTML document (or downloads an existing document) from that host. This document is generally in the shape of a query form for a database or databases.

2. If invoked with more than one argument, KE Texhtml uses its first argument to determine, from its configuration files, the host on which the database exists, the template *Texql* statement to be performed and the template HTML document to be produced. It then substitutes the remaining arguments into the *Texql* query before it is executed.

KE Texhtml consist of three processes. *texhtml* is invoked by *httpd* for every Internet request. For each of the above actions, it connects to one of possibly several resident *texhtmlserver* processes on one of possibly several machines. This process is a KE Texpress C-API based application and so on start-up, it in turn connects to a *texserver* process.

The separation of these processes enables the construction of a *fire wall* providing secure access to machines containing databases and restricting Internet access to one machine which contains no privileged data. This fire wall security approach is shown in the figure below.

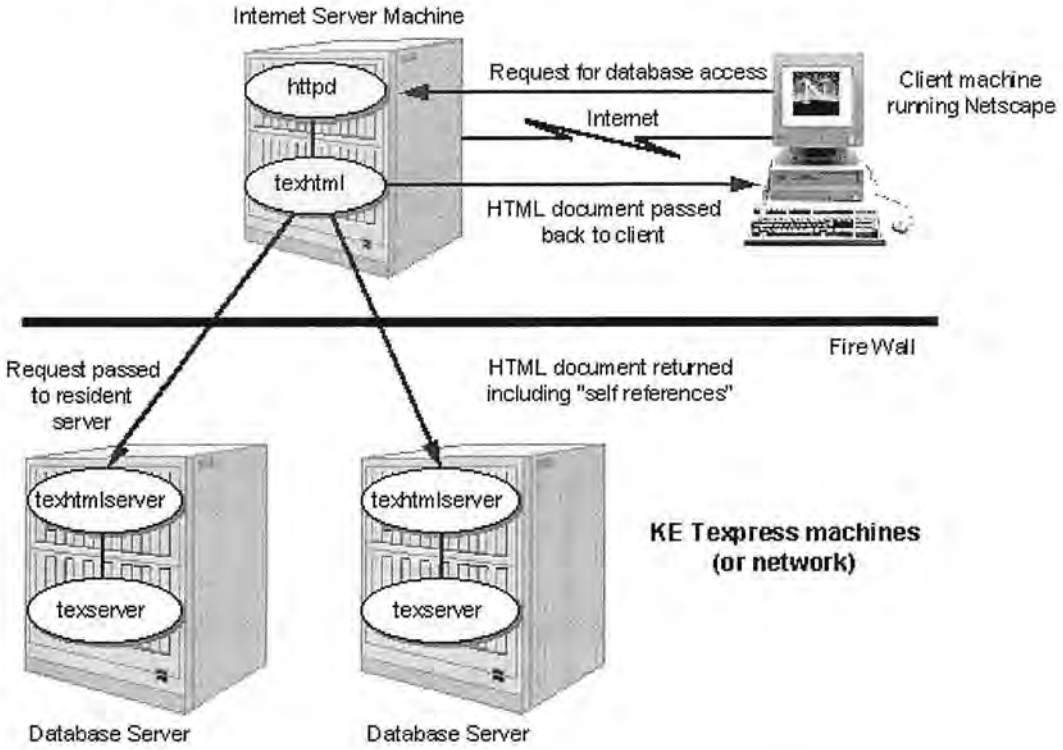


Figure 2. KE Texhtml Process Model

3.2 Registration and Access Levels

The Herbarium will be able to effectively monitor and control access to the various components of FloraBase by registering users for a fixed period. In this way the CALM Herbarium can protect its intellectual property and data ownership rights, while making the information available to bon fide users and, where appropriate, an up-front charge for use of the content can be made, thereby recovering some of the considerable cost of maintaining the information base.

Registration application will require an online form to be filled in with contact details, organisation and purpose for access. On receipt of the application the applicant will be assessed and a level of access assigned. An appropriate contractual agreement will be printed and despatched by mail for signing. On receipt of the signed agreement the applicant will be assigned a username and password, the NatureBase system administrator notified of the additional record to be added to the access table in Firewall 1, and the user notified of their access rights.

It is envisaged that there will be five levels of access required (see Table 1).

Table 1. Proposed FloraBase Access Levels

Access Level	Description of content accessible at this level	Example of the type of user assigned this access level
Level 1	FloraBase summary screen with name-related information and an indication of other content available (see Figure 4) and all of CALMLIB.	General user interested in information about plant names, their publication & currency; library clients.
Level 2	The above plus: FloraBase maps, images and descriptions, either separately from the buttons on the summary screen, or compiled into a single detailed FloraBase page (see Figure 5.)	General user interested in a range of summary information about a species or group of species, such as operations staff needing a printout of up-to-date data to aid field identification.
Level 3	All the above plus: access to the specimen database, however, detailed locality information regarding conservation taxa will be masked (Figure 6).	General CALM staff and other trusted users with a need for general purpose access to specimen-based data, such as other Australian herbaria staff.
Level 4	All the above plus: access to the specimen database, including detailed locality information regarding conservation taxa.	CALM Herbarium staff and other trusted users with a justifiable need for detailed information on conservation taxa, eg. CALM regional ecology officers.
Level 5	All database web interface forms and reports.	CALM Herbarium staff and other trusted users with a specific research need for access at this level.

Table 2. Functional Summary of FloraBase Access Levels

		NAMES		DESCRIPTIONS		SPECIMENS		LIBRARY
Level	Password	Query	Buttons	Query	Buttons	Query	Buttons	Query
DAMPIER	No	Yes	View		View		View	Yes
DRUMMOND	Yes	Yes	View		Active (no S)		View	Yes
PREISS	Yes	Yes	View		Active		Active (-CT locs)	Yes
FITZGERALD	Yes	Yes	View		Active		Active (+CT locs)	Yes
GARDNER	Yes	Yes	View	Yes	Active	Yes	Active	Yes

3.3 Charges

The CALM Herbarium considers that the FloraBase initiative is a major advancement in the timely presentation of accurate and authoritative information on the state's flora. As such it will be of great interest, and more importantly, great use, to a wide range of people working with our flora.

The range of users is generally indicated in Table 1, however, they fall into roughly three categories.

1. **Scientific users**: CALM research scientists and operations staff with a need to access detailed information for scientific, conservation or management purposes, or researchers from other state or international herbaria or universities.
2. **General users**: Members of the public, students, or amateur botanists such as members of the Wildflower Society or the CALM Regional Information Network of herbaria with a general interest in the state's flora.
3. **Commercial users**: Consultants, ecological survey firms, landscape architects and wildflower pickers who stand to benefit from easy access to quality data.

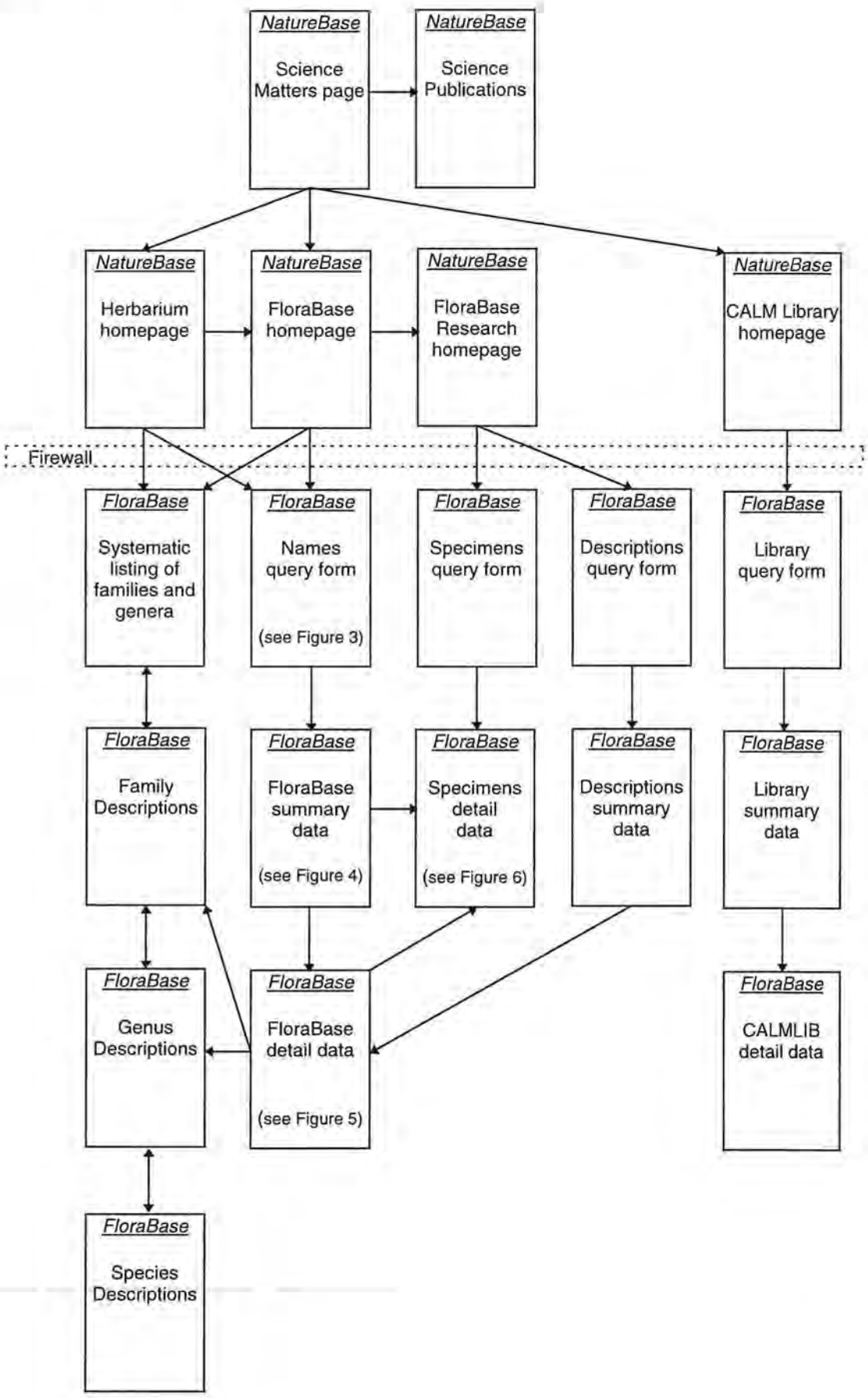
While use of the data from the site will be controlled by contractual arrangements, associating a registration charge for some categories of user will also help control the flow of information.

General users who want access to more than the first level of information (see Table 1) should be expected to pay a moderate registration charge, so that they associate a value with the information provided and to ensure the site does not present unfair competition to products in other media formats.

Commercial users stand to benefit greatly from access to FloraBase largely due to decreased time in preparing species lists, checking spelling of scientific names, identification of species and tracking down species name changes. These users should pay a substantial registration fee.

We do not expect to charge CALM staff for access, and there is a long-standing agreement amongst herbaria and universities for the free exchange of data for research purposes.

4. Logical Site Map



5. Design Examples

The following figures are presented to help indicate the look, feel and content of representative portions of FloraBase.

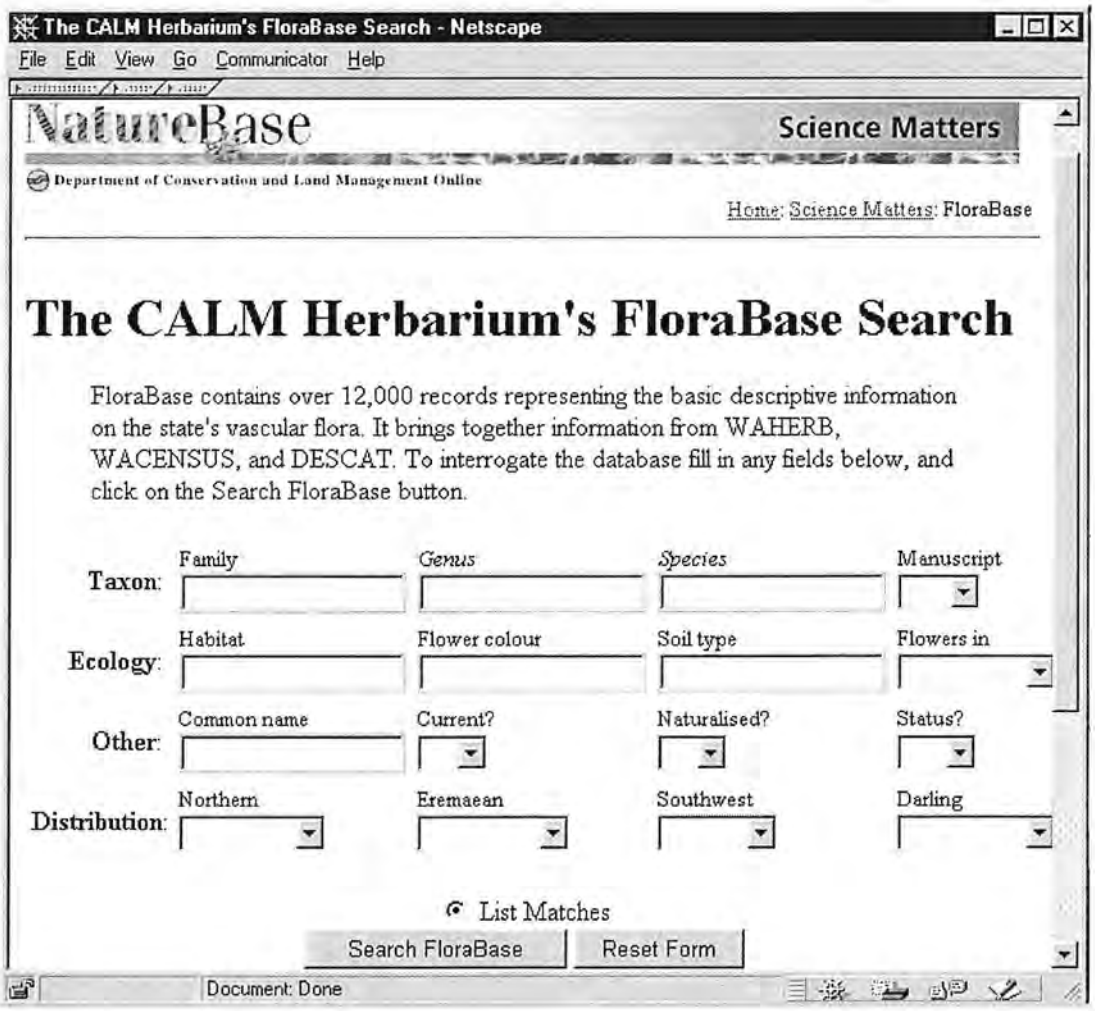


Figure 3. FloraBase query form

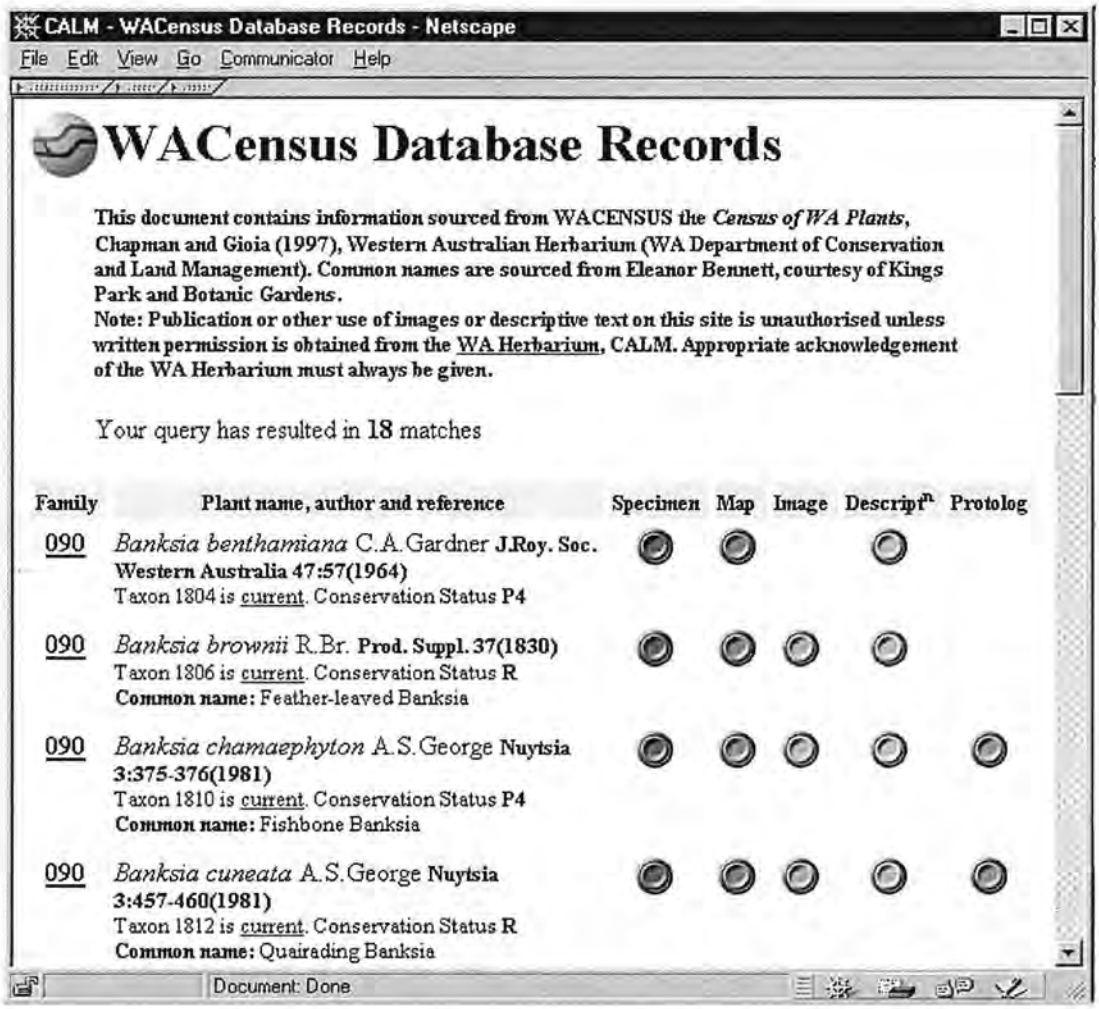


Figure 4. FloraBase summary of records

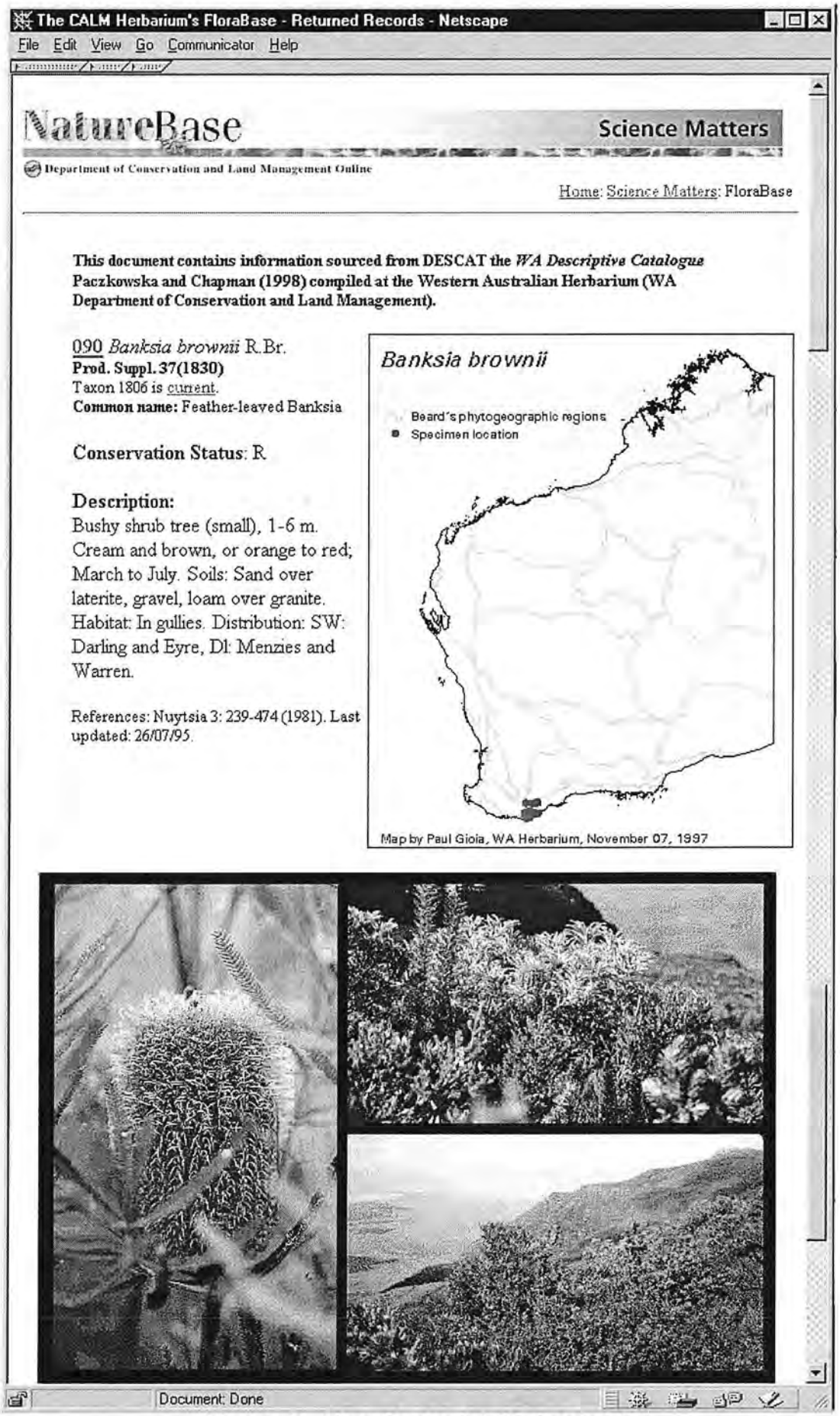


Figure 5. FloraBase detailed record

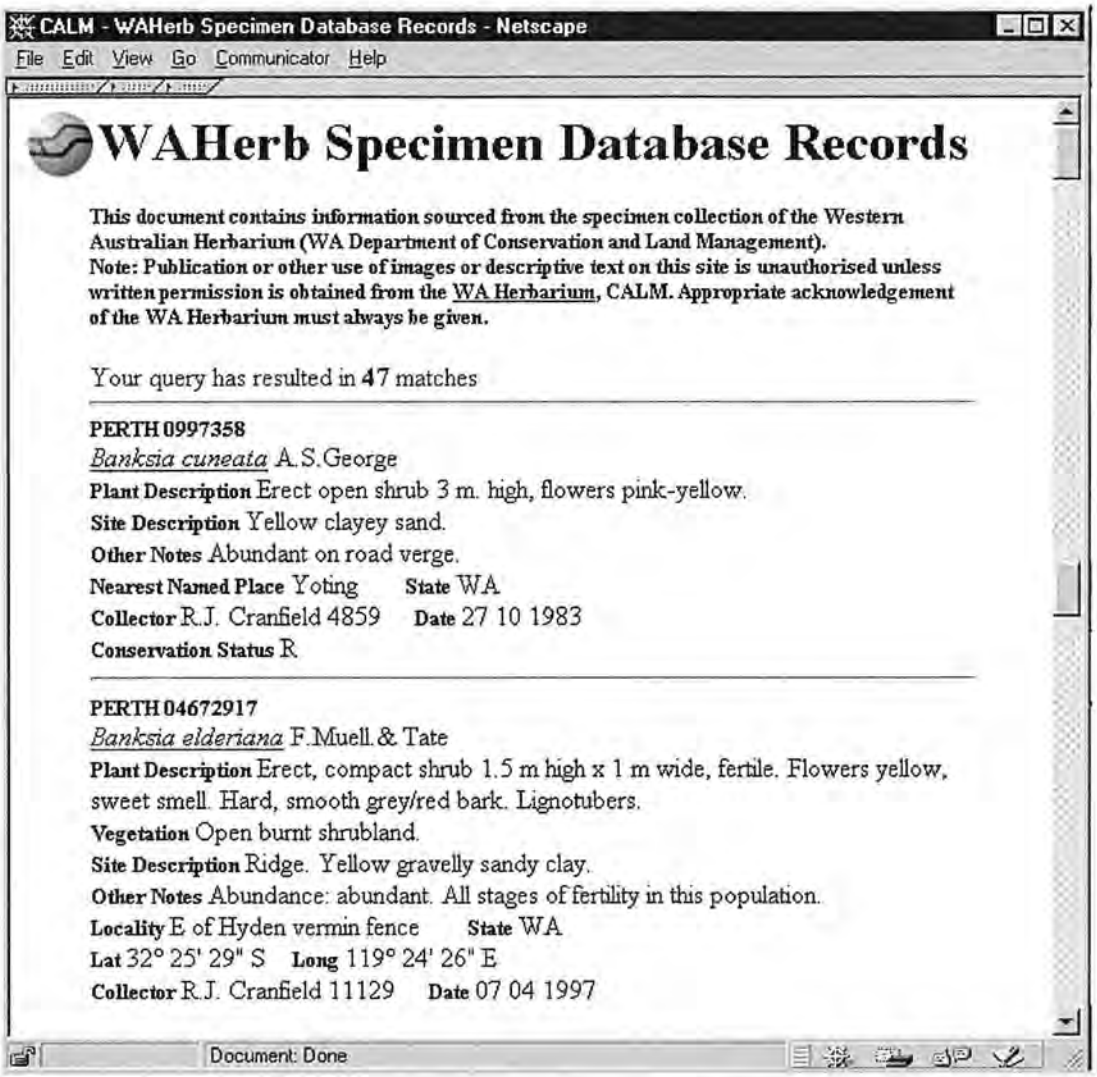


Figure 6. FloraBase specimen data
(note difference in handling conservation taxa records)

6. Personnel roles

Personnel roles involved in maintaining FloraBase online include:

- Database operators for basic data maintenance
- Database manager to ensure database functionality and data integration is maintained
- Registration manager as the point of contact for email registration applications, vetting, account creation and general correspondence
- Web designer to maintain and improve the look and feel of the web site
- System administrator of the FloraBase database server, to ensure functionality and security are maintained
- System administrator of the NatureBase web server, to ensure accounts are created and managed correctly by the Firewall and Web Server software and that TCP/IP communications are maintained across the Firewall.

7. Appendices

1. KE Texhtml Internet Services Guide

FloraBase - Access and Charges

Access Level	Description of content accessible at this level	Example of the type of user assigned this access level
Level 1	FloraBase summary screen with name-related information and an indication of other content available and all of CALMLIB.	General user interested in information about plant names, their publication & currency; library clients.
Level 2	The above plus: FloraBase maps, images and descriptions, either separately from the buttons on the summary screen, or compiled into a single FloraBase page	General user interested in a range of summary information about a species or group of species, eg. operations staff needing a printout of up-to-date data to aid field identification.
Level 3	All the above plus: access to the specimen database, however, detailed locality information regarding conservation taxa will be masked)	General CALM staff and other trusted users with a need for general purpose access to specimen-based data, such as other Australian herbaria staff.
Level 4	All the above plus: access to the specimen database, including detailed locality information regarding conservation taxa.	CALM Herbarium staff and other trusted users with a justifiable need for detailed information on conservation taxa, eg. CALM regional ecologist.
Level 5	All database web interface query forms and reports.	CALM Herbarium staff & other trusted users with a specific research need for toplevel access

	REQUIRED ACCESS LEVEL			
USER CATEGORY	**CALM USERS**		**INTERNET USERS**	
	LEVEL	**CHARGE**	**LEVEL**	**CHARGE**
General user				
Library user				
Herbarium volunteer				
Herbarium visitor				
Herbarium technical staff				
Herbarium scientist				
CALMScience scientist				
" technical staff				
" consultants				
" admin. officer				
Field operations staff				
Regional ecologist				
Other CALM divisions				
CALM business units				
University student				
University staff				
Wildflower Society member				
Wilderness Society member				
Wildflower pickers				
Mining company				
Bioprospecting company				
Environmental consultants				