

# 3

## THE ENHSIN PILOT NETWORK ANTON GÜNTSCH

### INTRODUCTION

The *ENHSIN* pilot network<sup>1</sup> is a simple XML<sup>2</sup>-based prototype for a common access system for distributed heterogeneous biological collection databases at specimen level. It has been designed and implemented in a relatively short time as a demonstrator for project participants, members of projects with a similar scope such as the Species Analyst project (Viegas, 1998) and *REMIB* (Lara, 2000), and decision makers.

### SYSTEM ARCHITECTURE

The *ENHSIN* pilot system consists of four major components (see Fig. 1).

**DATA SOURCES:** biological collection databases that may differ widely in their internal structure but which do provide data in accordance with a unified common data structure. This structure is currently given by a simple relation (e.g. a table or a view).

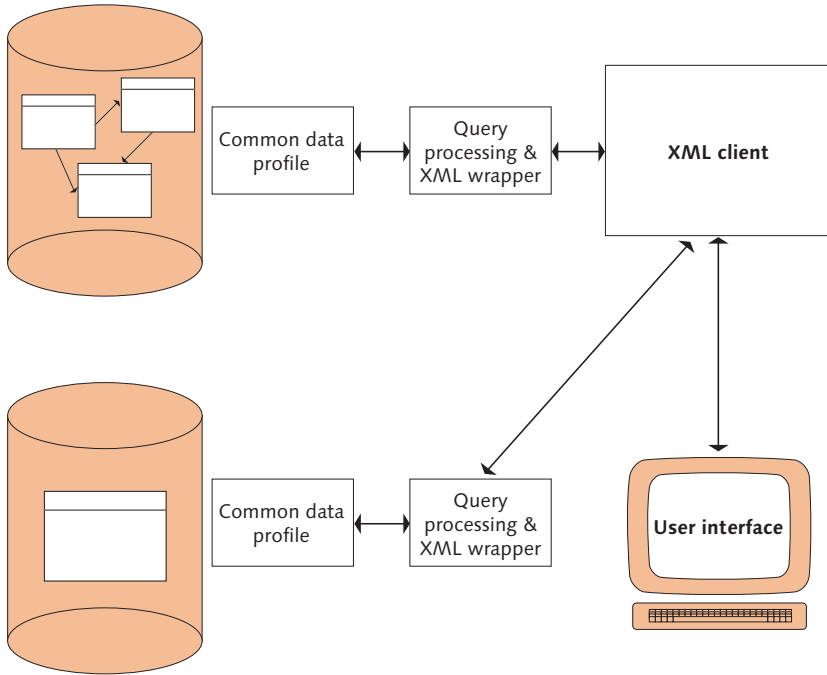
**USER INTERFACE:** a query form containing fields for genus, species epithet, collector's name, collection date, and country (Fig. 2). In addition, the maximum number of records to be returned per data source can be specified. By choosing 'fuzzy retrieval' the user instructs the system to search in less structured fields (see element set description), which usually produces a larger

---

<sup>1</sup> [www.bgbm.org/BioDivInf/projects/ENHSIN/XMLClient.htm](http://www.bgbm.org/BioDivInf/projects/ENHSIN/XMLClient.htm)

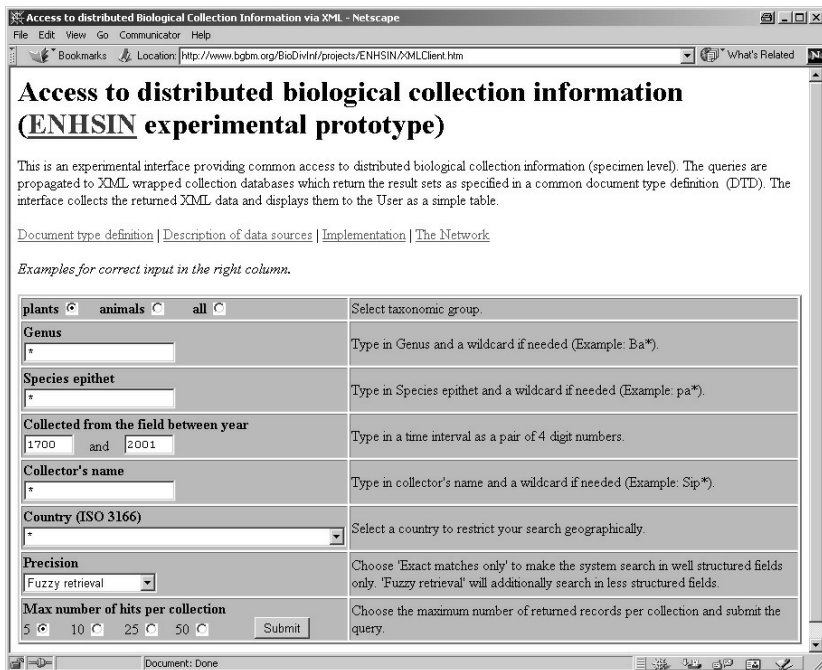
<sup>2</sup> [www.w3.org/XML/](http://www.w3.org/XML/)

Figure 1. System architecture



34

Figure 2. ENHSIN pilot user interface



but less precise query result. Additionally, the interface allows for selecting a taxonomic group (animals, plants, and all) to reduce the number of databases to be queried.

**CENTRAL XML CLIENT:** an Active Server Page<sup>3</sup> (ASP), which receives a query from the user interface, propagates it to the data sources, receives and parses the XML coded answers, and returns them as an html table. It also produces site maps by linking to the PARC Map Viewer<sup>4</sup> if unit data contain appropriate geographic coordinates. The map viewer, a free of charge service to create maps by simply accessing its interface with a parameterized url, is inactive at present. Parc is currently seeking for a way to relaunch the service on a hardware and software platform, which fits better into the company's technical infrastructure.

By using a browser with at least basic XML processing functionality (e.g. Microsoft Internet Explorer 5.0) the raw XML data can be loaded and viewed. Otherwise (e.g. using Netscape 4.7) it is possible to view the XML code by displaying the page source code.

**XML WRAPPER:** an Active Server Page, which is installed on the web server of each data source's site to process queries from the central client. It constructs an equivalent sql statement, receives the matching records from the data source, and sends them back to the client as an XML document. The XML document follows the grammar defined in a public document type definition (DTD).

35

## THE PILOT NETWORK

The *ENHSIN* pilot network currently provides access to seven collection information sources, ranging from small desktop applications to large-scale collection information servers:

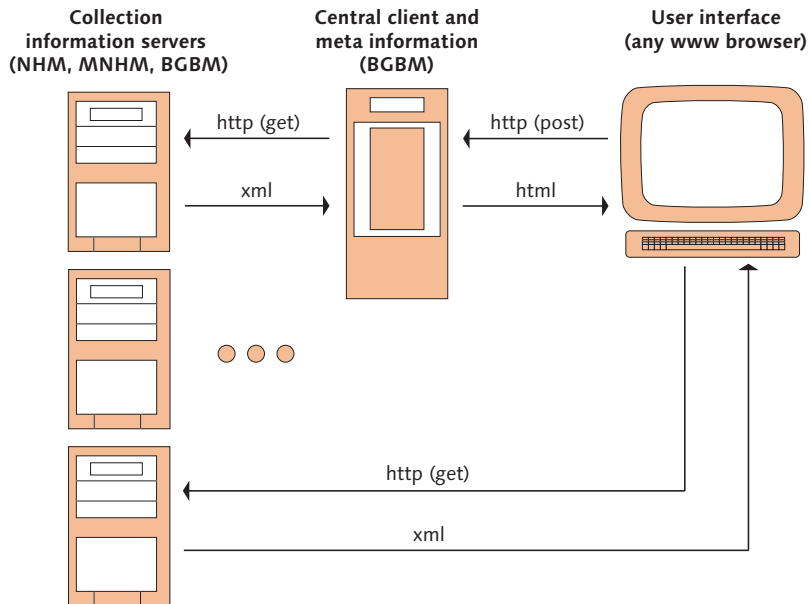
- ◆ Homoptera collection, Zoologisches Forschungsinstitut and Museum Alexander Koenig, Bonn.
- ◆ Lichen collection, Botanic Garden and Botanical Museum, Berlin-Dahlem
- ◆ Fruit and seed collection, Botanic Garden and Botanical Museum, Berlin-Dahlem.
- ◆ Collection nationale de poissons, France, Muséum National d'Histoire Naturelle, Paris.
- ◆ Fish collection (partial data), The Natural History Museum, London.
- ◆ Herbar national, Muséum National d'Histoire Naturelle, Paris.
- ◆ Freshwater mollusks of Salzburg, Austria.

---

<sup>3</sup> <http://msdn.microsoft.com/404/default.asp>

<sup>4</sup> <http://pubweb.parc.xerox.com>

Figure 3. The pilot network – information flow



All databases communicate with the central client software on the basis of the hypertext transfer protocol (HTTP) and XML (Extensible Markup Language), regardless of whether they are installed on a remote website or in the local area network of the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM), which is currently running the *ENHSIN* user interface on its World Wide Web server (Fig. 3).

An additional 'meta information' database has been developed, installed, and connected to the central web client software, which holds high-level collection descriptions and technical specifications needed to establish connections to collection resources. The metadata are used to display collection information even if a provider database is temporarily offline and registers new data sources once they have been properly installed. Additionally, they can be used to index collections such that queries are distributed to relevant data sources only. At present, this index only contains the two attributes 'plant collection' and 'animal collection'.

## THE ENHSIN ELEMENT SET AND XML INTERCHANGE FORMAT

The *ENHSIN* element set currently consists of 34 elements, covering both unit-level data items and meta-level elements describing the entire set of collection objects (see Güntsch & Berendsohn, 2001. and Güntsch, 2000b for a detailed list). Based on this element set, an XML interchange format has been developed, which is specified with an XML document type definition

(Güntsche, 2000a). Both element set and the document type definition (DTD) have been designed in a way that makes it easy for collection information providers to take part in the network by offering highly structured or semi-structured elements for the same content (Güntsche & Berendsohn, 2001).

A well-structured collection site record in the database can be represented with the following XML document for example:

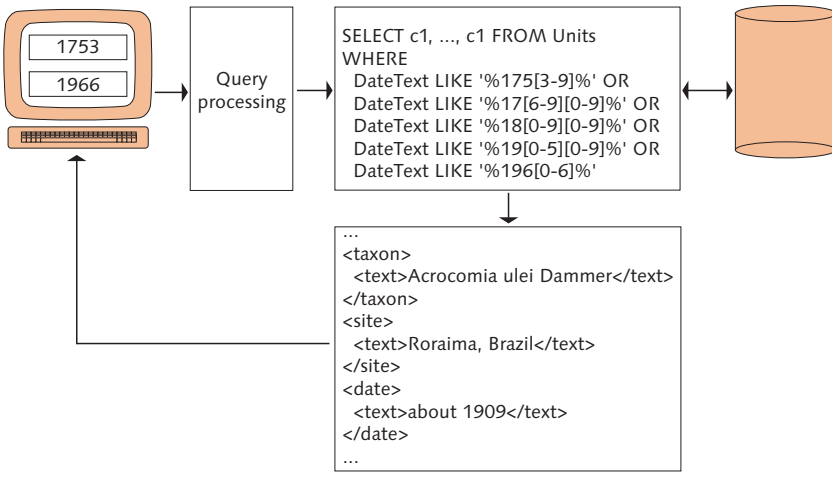
```
...
<site>
  <country>Afghanistan</country>
  <place>Tangi Gharuh</place>
  <lat>34.16</lat>
  <long>69.48</long>
</site>
<date>
  <day>21</day>
  <month>8</month>
  <year>1952</year>
</date>
...
```

Within the same framework a nearly unstructured collection site description can be represented and transmitted like this:

```
...
<site>
<text>
BRASIL, Rio Grande do Sul: Parque Nacional dos Aparados da Serra,
Itaimbezinho (mun. Cambará do Sul). Nas casca de árvore da mata:
junto ao canyon. Alt. 1200 m.
</text>
</site>
<date>
  <text>16 Apr. 1993</text>
</date>
...
```

This tolerant data specification for providers of collection information (see Berendsohn, this volume) is at the same time a challenge for the programming of applications and web services that are able to query and present structured and unstructured data through a single user interface. The *ENHSIN* pilot system exemplifies this approach by analysing unstructured gathering dates and species names that are provided as a single string. For example, a user query for objects collected between two years is dynamically converted into an SQL statement searching for an equivalent regular expression in a free text date string if further structured date elements are not provided (Fig. 4). Other free text search functions such as extraction of geographic co-ordinates or thesaurus-based search for place names may be implemented in the future.

Figure 4. Searching for 'fuzzy' dates



## SOFTWARE INSTALLATION FOR DATA PROVIDERS

The problem of software installation and configuration for data providers is considered a major obstacle for the success of a European biological collection information network. Therefore, keeping the installation procedure for the wrapper software as simple as possible was an important software design aim.

To link a collection database to the central network, the provider has to create a view on this database which implements a subset of the *ENHSIN* element set. This subset must at least contain the elements marked as mandatory. If the data-base management system does not allow creating views, a query or temporary table can be used alternatively. Once the view has been created, the wrapper software has to be installed. For this, an active server page (ASP) has to be copied into the script directory of the provider's web server. Two parameters (name of the view and name of the local database user) have to be modified.

Finally, the new information source has to be registered. A simple (one-page) questionnaire has to be completed providing information at collection level such as collection name, institution name, content description, and IPR statements. This questionnaire is electronically sent to the central system maintenance to register the service.

An instruction document, which guides data providers through the complete installation procedure avoiding technical terminology as much as possible has been developed and refined in co-operation with data providers. The entire installation procedure takes usually between two hours and a day depending on the complexity of the view to be created. The experiences with connecting new data sources to the *ENHSIN* pilot system showed that the main problems consist in the correct spellings of element names when creating the view and setting the data source name properly. Therefore, the installation could be further made easier by providing software tools to support these steps.

## OUTLOOK

Discussions with project participants and related projects indicated that the *ENHSIN* pilot network is considered to be a successful initial system. To further evolve into a large-scale European collection information network, however, several issues have to be addressed:

**OPERATING SYSTEM:** Currently, the XML wrapper software is implemented with Microsoft Active Server Pages which forces data providers to use the Microsoft Internet Information Server (IIS) as a web server. Since the majority of installed web services are based on non-Microsoft technologies (such as Apache web server), a future version of the wrapper software should be implemented with a generic scripting language such as Perl or PHP.

**PARALLEL ACCESS:** The central client software is currently retrieving the results of a query sequentially from the participating databases. This means that the overall loading time of the returned XML documents is given by the sum of the individual loading times. If a future system will be connected to a much higher number of information providers this fact would significantly slow down the system's performance. Therefore, parallel processes should replace the sequential loading mechanism.

**QUERY LANGUAGE:** At present, queries are transmitted from the central client to the xml wrappers with simple parameterised URLs. To achieve more flexible query mechanisms, a standardised query language should be developed on the basis of XML. The language should be able to express arbitrary simple Boolean queries plus a number of useful string comparison and 'fuzzy' search operators.

**CONTENT:** The project tried to define the element set as small and simple as possible to encourage participation in the network and to be able to develop client software in a reasonable time. Nevertheless, the architecture is capable of processing much wider element set definitions, such as the one currently developed by the *CODATA/TDWG* Working Group on Biological Collection Data Access (see Berendsohn, this volume), in which *ENHSIN* plays an important role.

---

### References

Güntsch, A. & Berendsohn W.G., *in press*. Maximise Common Denominators: Towards an International Data Access Profile for Biological Collection Information. *Proceedings of the 17th International CODATA Conference*, Baveno, October 2000.  
[www.codata.org/conf2000/html/prog.pdf](http://www.codata.org/conf2000/html/prog.pdf)

Güntsch, A 2000a [Jul]. *ENHSIN* Pilot Network DTD, Berlin.  
[www.bgbm.org/BioDivInf/Projects/ENHSIN/PilotCollectionDTD.htm](http://www.bgbm.org/BioDivInf/Projects/ENHSIN/PilotCollectionDTD.htm)

Güntsch, A 2000b [Dec]. The *ENHSIN* Pilot – Implementation Issues, Berlin.  
[www.bgbm.fu-berlin.de/BioDivInf/projects/ENHSIN/PilotImplementation.htm](http://www.bgbm.fu-berlin.de/BioDivInf/projects/ENHSIN/PilotImplementation.htm)

Güntsch, A. & Berendsohn W.G., 2001. Wrapping up collections: the *ENHSIN* pilot access system. *17th Annual Meeting of the Taxonomic Databases Working Group* (TDWG 2001), Sydney November 2001, Abstract Volume.  
[plantnet.rbgsyd.gov.au/bioforum/TDWG\\_program/tdwg\\_abstracts.html](http://plantnet.rbgsyd.gov.au/bioforum/TDWG_program/tdwg_abstracts.html)

Lara, L. 2000. Informatics tools for the management of biodiversity information. *16th Annual Meeting of the Taxonomic Databases Working Group* (TDWG 2000), Frankfurt November 2000, Abstract Volume.

Vieglas, D. 1998. Integrating disparate biodiversity resources using the information retrieval standard z39.50. *TDWG 1999 Abstracts*, Cambridge, USA.  
[www.tdwg.org/rep1999.html#dave](http://www.tdwg.org/rep1999.html#dave)