Bulletin UASVM Horticulture, 68(1)/2011
Print ISSN 1843-5254; Electronic ISSN 1843-5394

Local Using of Integrated Taxonomic Information System (ITIS)

Lorentz JÄNTSCHI, Radu E. SESTRAŞ

University of Agricultural Sciences and Veterinary Medicine, 3-5 Manastur Street, Cluj-Napoca 400372, Romania; lori@academicdirect.org

Abstract. In order to use a taxonomy system for queries in the project entitled Evaluation in the spontaneous flora of phytopharmaceutical action in relation with structure, composition or genotype using metaheuristic algorithms and evolutionary programming" (Postdoctoral fellow at University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca supported by POSDRU/89/1.5/S/62371) a study of IT IS (Integrated Taxonomic Information System) database usage in a local network were conducted. Installation, and query construction on the ITIS database are discussed.

Keywords: taxonomy systems; structured query language; topology of taxonomic systems

INTRODUCTION

ITIS is a perpetual project like PubMed (NIH PubMed, www), PubChem (PubChem Website, www; Bolton *et al.*, DOI; Wang *et al.*, 2010), Genome (NCBI Genome, www), or other like projects, which is coordinated under tutee of Smithsonian Institution (Washington, DC, USA) and formed as a partnership of federal US agencies. Project aim is to provide scientifically credible taxonomic information. Two technical work groups - the Database Work Group (DWG) and the Taxonomy Work Group (TWG) have specific responsibilities in the project.

It is updated regularly (once a month with about 3000 records adding/changing; it contains over 500000 taxonomic units; full ITIS database is available for downloading in various database formats (Informix, MS Sql, MySQL) from http://www.itis.gov/downloads/.

Starting with the instructions available on the ITIS website (ITIS Project, www), the aim of the present study was to install, prepare to use and finally to make a home-made software to query the ITIS database.

MATERIAL AND METHOD

Following steps were done in order to install the ITIS database on a local database server:

- Downloading of Full ITIS Data Set (MySQL by table) "itisMySQLMMDDYY.TAR.gz" where MM is the two digits month, DD is the two digits day and YY is the two digits year (26 July 2011 version was downloaded);
- Unzipping and copying to a MySQL database server machine and thereafter executing of "mysql -uroot -p --enable-local-infile < dropcreateloaditis.sql" will provide the full ITIS database on a local database server;

RESULTS

In order to prepare the ITIS database for usage, the understanding of the database structure and of the manner of information structure is required. Following results were obtained

seeking for the information structure:

- Organizing of the information in ITIS database
  - "taxonomic_units" table contains names (concatenating of `unit_name1`, `unit_name2`, `unit_ind3`, `unit_name3`, `unit_ind4`, `unit_name4` give the full qualified name of an entity) and hierarchy (`tsn` is the unique code identifier of the entity; `parent_tsn` code identifier give a link to its parent group) of living organisms or groups of organisms; other fields (such as `name_usage`, `unaccept_reason`, `credibility_rtng`, `completeness_rtng`, `currency_rating`, `initial_time_stamp`, `taxon_author_id`, `update_date`) are record specific information linking the entity with other information stored in the database in other tables;
  - some information in the table may be found in duplicate; ex. "Bacteria" (`unit_name1`='Bacteria') has two entries (records), one valid (`tsn`=202421, `name_usage`='valid') and one invalid (`tsn`=50, `name_usage`='invalid')
  - "kingdoms" table have six records (`kingdom_id`, `kingdom_name`, `update_date`) as follows: ('1', 'Monera', '1996-03-26'), ('2', 'Protozoa', '2004-06-04'), ('3', 'Plantae', '1996-03-26'), ('4', 'Fungi', '1996-03-26'), ('5', 'Animalia', '1996-03-26'), ('6', 'Chromista', 2004-06-04');
  - hierarchy in the phylogeny depends on kingdom;
    - it should be noted that the levels in the hierarchy did not means that every entity from a lower group has a link to the immediate upper group; it means only that it exists at least one entity from a lower group which points to the immediate upper group;
    - three kingdoms - namely "Plantae", "Fungi" and "Chromista" - are hierarchized on 22 levels (Kingdom ↳ Subkingdom ↳ Division ↳ Subdivision ↳ Class ↳ Subclass ↳ Order ↳ Suborder ↳ Family ↳ Subfamily ↳ Tribe ↳ Subtribe ↳ Genus ↳ Subgenus ↳ Section ↳ Subsection ↳ Species ↳ Subspecies ↳ Variety ↳ Subvariety ↳ Form ↳ Subform);
    - other two kingdoms - namely "Monera" and "Protozoa" - has one level less (21 levels), but with different level names (Kingdom ↳ Subkingdom ↳ Phylum ↳ Subphylum ↳ Superclass ↳ Class ↳ Subclass ↳ Infraclass ↳ Superorder ↳ Order ↳ Suborder ↳ InfraOrder ↳ Superfamily ↳ Family ↳ Subfamily ↳ Tribe ↳ Subtribe ↳ Genus ↳ Subgenus ↳ Species ↳ Subspecies);
    - the last but not the least kingdom - namely "Animalia" has a hierarchy derived from "Monera" and "Protozoa" (or viceversa); thus the 21 levels from "Monera" and "Protozoa" are continued (from ↳ Subspecies) in "Animalia" with ↳ Variety ↳ Form ↳ Race ↳ Stirp ↳ Morph ↳ Aberration.
- A simple search in "taxonomic_units" table should be conducted like the following SQL phrase which query about the green alga known as Prototheca that lacks chlorophyll: SELECT * FROM `taxonomic_units` WHERE `unit_name1` LIKE 'Prototheca%' OR `unit_name2` LIKE 'Prototheca%' OR `unit_name3` LIKE 'Prototheca%' OR `unit_name4` LIKE 'Prototheca%' when following record are retrieved:

tsn	…	unit_ name1	…	unit_ name2	…	name_ usage	…	initial_ time_stamp	parent_ Tsn	…	kingdom_ id	rank_ id	update_ date	…
9828		Petalomonas		prototheca		accepted		6/13/1996 14:51	9824			3	220 7/10/1996	

- Searches for "Petalomonas prototheca" descendants in the phylogeny tree should be conducted like this: SELECT * FROM `taxonomic_units` WHERE `parent_tsn` = 9828; this query retrieved an empty set; Searches for "Petalomonas prototheca" fellow entities in the phylogeny tree should be conducted like this: SELECT * FROM `taxonomic_units`

WHERE `parent_tsn` = 9824; this query retrieved 22 records (entities) namely ("tsn"+": " + "unit_name1"+" "+" unit_name2"): 9825: Petalomonas abscissa; 9826: Petalomonas mediocanellata; 9827: Petalomonas phacoides; 9828: Petalomonas prototheca; 9829: Petalomonas steini; 180835: Petalomonas alata; 180836: Petalomonas angusta; 180837: Petalomonas applanata; 180838: Petalomonas asymmetrica; 180839: Petalomonas gigas; 180840: Petalomonas inflexa; 180841: Petalomonas involuta; 180842: Petalomonas klinostoma; 180843: Petalomonas minuta; 180844: Petalomonas mira; 180845: Petalomonas platyrhyncha; 180846: Petalomonas praegnans; 180847: Petalomonas pusilla; 180848: Petalomonas quadrilineata; 180849: Petalomonas sexlobata; 180850: Petalomonas tricarinata; 180851: Petalomonas ventritracta.

÷ A more complex query may retrieve in one interrogation the record of the Prototheca's parent: SELECT * FROM `taxonomic_units` WHERE `tsn`= (SELECT `parent_tsn` FROM `taxonomic_units` WHERE `unit_name1` LIKE 'Prototheca%' OR `unit_name2` LIKE 'Prototheca%' OR `unit_name3` LIKE 'Prototheca%' OR `unit_name4` LIKE 'Prototheca%')

÷ If a search for the level in the phylogeny tree is desired for Prototheca, then the query should be made as follows by using the "hierarchy" table if the ITIS database: SELECT * FROM `hierarchy` WHERE `hierarchy_string` LIKE '%-9828' when following string are retrieved: 202422-9601-9602-9816-9817-9824-9828;

÷ If the result of the Prototheca hierarchy is desired in one query, then the search should be conducted as follows: SELECT `hierarchy_string` FROM `hierarchy` WHERE `hierarchy_string` LIKE CONCAT('%-',(SELECT `tsn` FROM `taxonomic_units` WHERE `unit_name1` LIKE 'Prototheca%' OR `unit_name2` LIKE 'Prototheca%' OR `unit_name3` LIKE 'Prototheca%' OR `unit_name4` LIKE 'Prototheca%')) when again the string 202422-9601-9602-9816-9817-9824-9828 are retrieved.

APPLICATION:
A simple procedure to retrieve the phylogeny from ITIS database

After the stage of the minimal understanding of the database structure, the next step was made: creation of home-made software to do queries on ITIS database. A procedure does the queries, and a main program interfaces with the user. The procedure is written in PHP language and it assumes that it exists a calling program (implemented in PHP too), and it opens a connection to a MySQL database server able to query the ITIS database. The procedure uses three queries (namely $q1, $q2 and $q3 in the implementation of the procedure) in order to retrieve hierarchy string ($q1 query), names of the levels of the phylogeny ($q2 query) and types of the levels in the phylogeny ($q3 query). It is assumed that the query string is one word only (Fig. 1).

```
function name_to_phylogeny($name){$tree=array();
    $c=mysql_connect(server,user,password);$q=mysql_query("USE `ITIS`");
    $q1="SELECT `hierarchy_string` FROM `hierarchy` WHERE `hierarchy_string` LIKE";
    $q1.=CONCAT('%-',(SELECT `tsn` FROM `taxonomic_units` WHERE `unit_name1` LIKE";
    $q1.=" '".$name."%' OR `unit_name2` LIKE '".$name."%' OR `unit_name3` LIKE";
    $q1.=" '".$name."%' OR `unit_name4` LIKE '".$name."%'))"; $q=mysql_query($q1);
    while($r=mysql_fetch_row($q))$entities[]=$r[0];mysql_free_result($q);
    foreach($entities as $entity){$result=array();
        $phylogeny=explode("-",$entity);
        foreach($phylogeny as $level){
            $q2="SELECT `unit_ind1`,`unit_name1`,`unit_ind2`,`unit_name2`,";
            $q2.="`unit_ind3`,`unit_name3`,`unit_ind4`,`unit_name4`,";
            $q2.=`rank_id`,`kingdom_id`";
```

```
            $q2.=" FROM `taxonomic_units` WHERE `tsn` = '".$level."'";
            $q=mysql_query($q2);$r=mysql_fetch_row($q);mysql_free_result($q);
            for($i=0;$i<8;$i++)if(!$r[$i])unset($r[$i]);
            $q3="SELECT `rank_name` FROM `taxon_unit_types`";
            $q3.=" WHERE `kingdom_id`='".$r[9]."' AND `rank_id`='".$r[8]."'";
            $q=mysql_query($q3);$s=mysql_fetch_row($q);mysql_free_result($q);
            unset($r[8]);unset($r[9]);$result[$s[0]]=implode(" ",$r);
        }$tree[]=$result;
    }mysql_close($c);return($tree);
}
```

Fig. 1. Name to phylogeny function (PHP implementation)

If the name of the searched entity comes via a GET form submission method and is accessible via $_GET["name"] environment variable, then the name_to_phylogeny function should be called as in Fig. 2 below:

```
$tree=name_to_phylogeny($_GET["name"]);
foreach($tree as $result){$t="";
    foreach($result as $k => $v){
        echo($t.$k.": ".$v."\r\n");$t.="\t";
    }
}
```

Fig. 2. Call of the name_to_phylogeny function to do queries (PHP implementation)

If the query name is "Prototheca" coming via a HTTP submission form ($_GET["name"]=="Prototheca") then obtained result is as given in Fig. 3.

```
Kingdom: Plantae
    Division: Euglenophycota
        Class: Euglenophyceae
            Order: Sphenomonadales
                Family: Sphenomonaceae
                    Genus: Petalomonas
                        Species: Petalomonas prototheca
```

Fig. 3. Result of the name_to_phylogeny("Prototheca") query on ITIS database

DISCUSSION

The program implementing queries on ITIS database is online available for general use purpose: http://l.academicdirect.org/Horticulture/GAs/62371/ [Link].

There are differences between ITIS classification (ITIS Project, www) and USDA Plants (USDA Plants, www) classification; for example `Supradivision` level in USDA Plants classification is not present in ITIS hierarchy.

Regarding the Prototheca species (correct naming according to ITIS standardization: Petalomonas prototheca species) a query on NCBI Taxonomy and/or Genome project (NCBI Taxonomy, www; NCBI Genome, www; Sayers *et al*., 2009; Benson *et al*., 2009) retrieve a slightly different classification. Thus, Fig. 4 gives these differences:

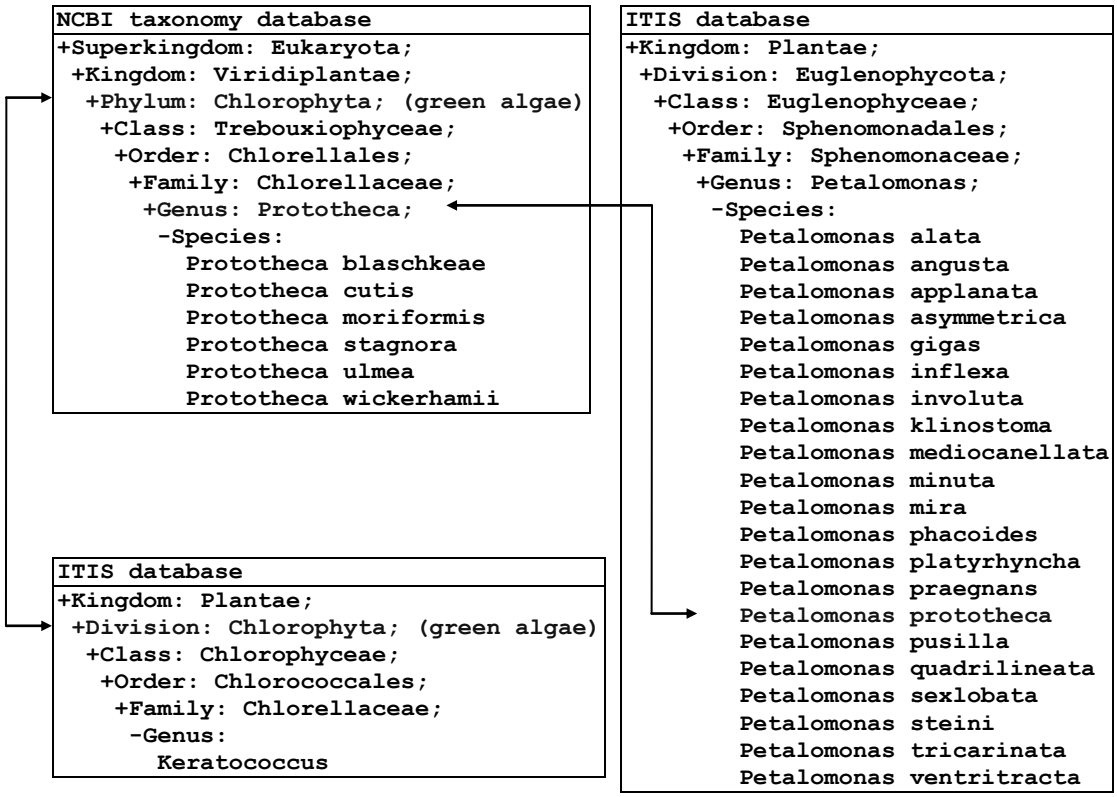

Fig. 4. Prototheca entity in NCBI and ITIS databases

The above table reveals that both projects (Genome and ITIS) are still at the beginning of the systematization of the organisms and creating of a unique taxonomy reference will require a lot of work to do.

Acknowledgments. The study was supported by POSDRU/89/1.5/S/62371 through a postdoctoral fellowship for L. Jäntschi.

REFERENCES

1. Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and E. W. Sayers (2009). GenBank. Nucleic Acids Research 37:D26-31 [Epub 2008 Oct 21].

2. Bolton, E., Y. Wang, P. A.Thiessen and S. H. Bryant (DOI). PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 In: Annual Reports in Computational Chemistry 4:27p. [Epub 2008 Dec 23, free author manuscript].

3. ITIS Project. (www). ITIS: Integrated Taxonomic Information System (project, website): http://www.itis.gov/ [retrieved on August 6, 2011].

4. NCBI Genome. www. NCBI Genome: Information by genome sequence (project, website): http://www.ncbi.nlm.nih.gov/sites/genome [retrieved on August 6, 2011].

5. NCBI Taxonomy. www. NCBI Taxonomy Project (website): http://www.ncbi.nlm.nih.gov/ Taxonomy/ [retrieved on August 6, 2011].

6. NIH PubMed. www. PubMed.gov: US National Library of Medicine & National Institutes of Health (project, website): http://www.ncbi.nlm.nih.gov/pubmed [retrieved on August 6, 2011].

7. PubChem Website. (www). PubChem: A National Center for Biotechnology Information Project (website): http://pubchem.ncbi.nlm.nih.gov/ [retrieved on August 6, 2011].

8. Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt,

G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko and J. Ye (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37:D5-15 [Epub 2008 Oct 21].

9. USDA Plants. (www). USDA Plants: United States Department of Agriculture Natural Resources Conservation Service Plants Database (project, website): http://plants.usda.gov/ [retrieved on August 6, 2011].

10. Wang, Y., E. Bolton, S. Dracheva, K. Karapetyan, B. A. Shoemaker, T. O. Suzek, J. Wang, J. Xiao, J. Zhang and S. H. Bryant (2010). An overview of the PubChem BioAssay resource Nucleic Acids Research 38:D255-66. [Epub 2009 Nov 19].