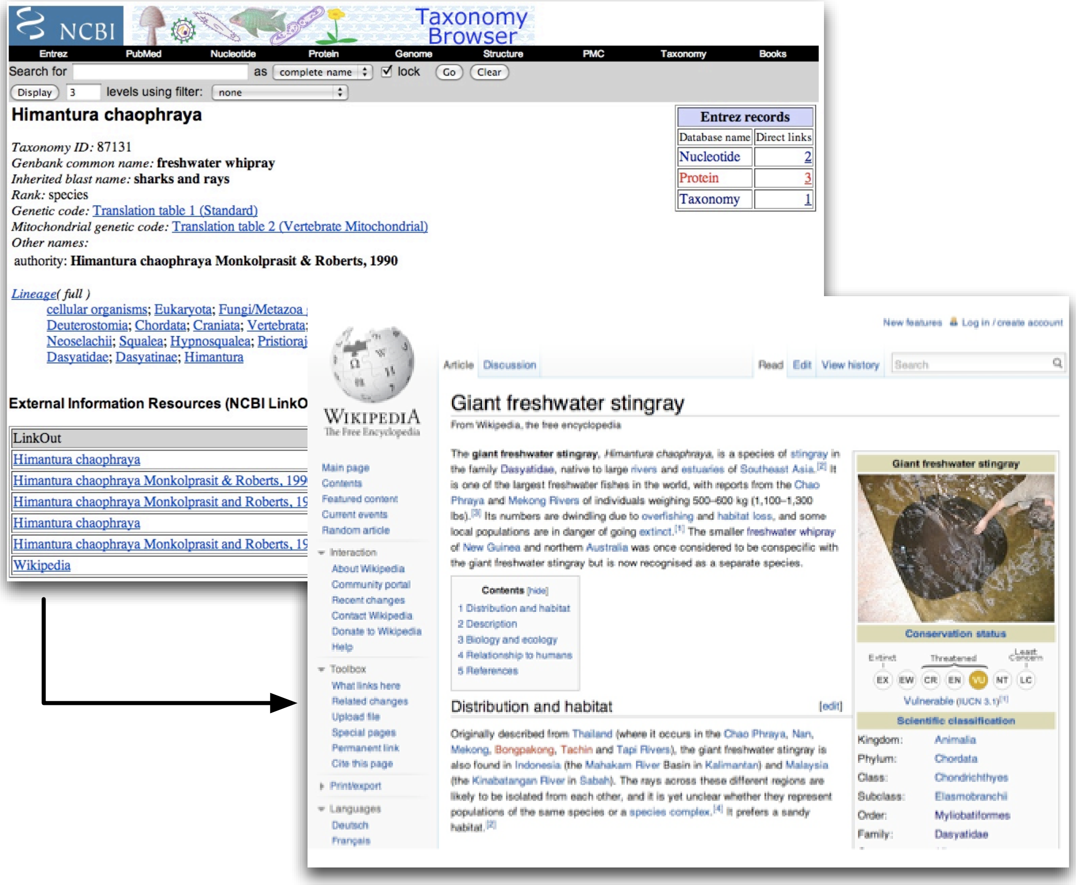
Phyloinformatics in the age of Wikipedia

Roderic D M Page
DEEB, FBLS
University of Glasgow, Glasgow G12 8QQ, UK
r.page@bio.gla.ac.uk

One of the great challenges of phyloinformatics is linking together information on phylogenies, taxa, genomes, specimens, and publications. One approach to linking disparate data is to use shared identifiers. For example, if a bibliographic database and a nomenclature database both use the same identifier for a publication (such as a DOI), then we can easily link the two pieces of information together using that identifier. An obstacle to this approach is the lack identifiers, or failure to reuse existing identifiers (Page, 2008). For example, sequences in GenBank may lack bibliographic identifiers, even if the paper in which the sequences were published has an identifier (Miller et al., 2009). This low "link density" is a major obstacle to linked data approaches to integrating biodiversity data. Link density can be increased either by reusing existing identifiers, or by creating maps between existing identifiers.

This talk describes a mapping between the NCBI taxonomy database and Wikipedia. These two databases were chosen because the NCBI taxonomy contains all the taxa for which sequences are publicly available, and for many taxa Wikipedia is the first site returned in a Google search on that taxon's scientific name (Page, 2010). The bulk of the mapping was created automatically by matching strings in the two databases, either directly, or via synonyms (NCBI) or redirect pages (Wikipedia). To support manual editing and correction the mapping was loaded into a wiki available at http://iphylo.org/linkout. Implemented using semantic wiki (http://semantic-mediawiki.org/), the wiki provides RDF export of the mapping. NCBI taxonomy ids are linked to the equivalent URI in the Uniprot database (http://www.uniprot.org), and Wikipedia pages are mapped to the equivalent URI in Dbpedia (http://dbpedia.org). Both Uniprot and Dbpedia provide data in RDF, which when combined with the mapping from http://iphylo.org/Linkout enable a wide range of queries. For example, we could query Wikipedia for images of taxa in GenBank, or query GenBank for sequences from taxa in the International Union for Conservation of Nature Red List of Threatened Species (http://www.iucnredlist.org/).

To date a total of 52,956 NCBI taxa have been mapped to the corresponding taxa in Wikipedia. The mapping for any individual NCBI taxon can be found by appending the taxonomy id to the URL "http://iphylo.org/linkout/Ncbi:", for example http://iphylo.org/linkout/Ncbi:87131. The mapping has also been uploaded NCBI's Linkout, so that the link to Wikipedia is shown on the NCBI page for the taxon (Fig. 1).



Himantura chaophraya

Taxonomy ID: 87131
Genbank common name: **freshwater whipray**
Inherited blast name: **sharks and rays**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
authority: **Himantura chaophraya Monkolprasit & Roberts, 1990**

Lineage(full)
cellular organisms; Eukaryota; Fungi/Metazoa
Deuterostomia; Chordata; Craniata; Vertebrata;
Neoselachii; Squalea; Hypnosqualea; Pristioraj
Dasyatidae; Dasyatinae; Himantura

Entrez records	
Database name	Direct links
Nucleotide	2
Protein	3
Taxonomy	1

External Information Resources (NCBI LinkO

LinkOut
Himantura chaophraya
Himantura chaophraya Monkolprasit & Roberts, 199
Himantura chaophraya Monkolprasit and Roberts, 19
Himantura chaophraya
Himantura chaophraya Monkolprasit and Roberts, 19
Wikipedia

Giant freshwater stingray

From Wikipedia, the free encyclopedia

The **giant freshwater stingray**, *Himantura chaophraya*, is a species of stingray in the family Dasyatidae, native to large rivers and estuaries of Southeast Asia.[2] It is one of the largest freshwater fishes in the world, with reports from the Chao Phraya and Mekong Rivers of individuals weighing 500–600 kg (1,100–1,300 lbs).[3] Its numbers are dwindling due to overfishing and habitat loss, and some local populations are in danger of going extinct.[1] The smaller freshwater whipray of New Guinea and northern Australia was once considered to be conspecific with the giant freshwater stingray but is now recognised as a separate species.

Contents [hide]
1 Distribution and habitat
2 Description
3 Biology and ecology
4 Relationship to humans
5 References

Distribution and habitat [edit]

Originally described from Thailand (where it occurs in the Chao Phraya, Nan, Mekong, Bongpakong, Tachin and Tapi Rivers), the giant freshwater stingray is also found in Indonesia (the Mahakam River Basin in Kalimantan) and Malaysia (the Kinabatangan River in Sabah). The rays across these different regions are likely to be isolated from each other, and it is yet unclear whether they represent populations of the same species or a species complex.[4] It prefers a sandy habitat.[2]

Kingdom:	Animalia
Phylum:	Chordata
Class:	Chondrichthyes
Subclass:	Elasmobranchii
Order:	Myliobatiformes
Family:	Dasyatidae

Fig. 1. The NCBI page for *Himantura chaophraya* (taxonomy id 87131) linked to the Wikipedia page for the Giant freshwater stingray.

Acknowledgements

My thanks to Scott Federhen and Kathy Kwan at NCBI for all their assistance in getting the mapping into NCBI Linkout.

References

Miller, Holly; Catherine Norton; Indra Neil Sarkar. (2009). GenBank and PubMed: How connected are they? BMC Research Notes 2009, 2:101 doi:10.1186/1756-0500-2-101

Page, R. D. M. (2008). Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9:345-354. doi:10.1093/bib/bbn022

Page, R. D. M. (2010). Wikipedia as an encyclopaedia of life. Nature Precedings hdl:10101/npre.2010.4242.1