Indexing for Life

D7.1 - Taxon-centric portal at EBI: Extension of taxon-centric portal at EBI to support the cross-mapped CoL Taxonomy.

Work package 7

Guy Cochrane

15 November 2013

Capacities Programme of Framework 7: EC e-Infrastructure Programme – Virtual Research Communities - INFRA-2010-2

Grant Agreement No:	261555
Project Co-ordinator:	Dr Alastair Culham
Project Homepage:	http://www.i4Life.eu
Duration of Project:	36 months
Start Date:	November 2010
End Date:	November 2013

D7.1 Taxon-centric portal at EBI: Extension of taxon-centric portal at EBI to support the cross-mapped CoL Taxonomy.

Introduction

The Taxon Portal (TP; http://www.ebi.ac.uk/ena/data/warehouse/search?portal=taxon) provides a single direct taxonomic entry point into globally comprehensive nucleic acid sequence data. Supporting the Catalogue of Life (CoL) taxonomy and the *de facto* standard molecular biological taxonomic classification (known as the 'NCBI Taxonomy'), the TP provides an important and previously unavailable interface to sequence data presented from the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/). The TP offers both interactive web and rich web service interfaces that, through their provision of search, taxonomic navigation and connectivity with the CoL, represent an important discovery and analysis tool that integrates broad molecular data resources into CoL.

Scope of deliverable

Our approach to i4Life workpackage 7 has been to develop the necessary software and to make frequent deployments as soon as any useful components can be made available to users. D7.1 represents the culmination of this process and the maturation to full production status of the complete TP, at this point comprising a rich set of functionalities comprehensively integrated into the ENA service environment.

Interactive web pages

The core unit of presentation of TP is a taxon page, offering descriptive information relating to the taxon indicated and tab-selectable summary information and connectivity into sequence data, taxonomic navigation and translation tables appropriate for coding sequence analysis for the taxon (see figure 1 for example).

A typical direct use of the TP would involve a user searching a taxonomic name from the entry page (http://www.ebi.ac.uk/ena/data/warehouse/search?portal=taxon) against a chosen taxonomy. In cases where multiple taxa result from the search (in cases, for example where the user has elected to include taxa subordinate to the query taxon), a results summary is shown allowing selection of the appropriate taxon page. In cases where a search results in a single hit, the taxon page is shown directly.

Many users approach TP indirectly. The ENA Advanced Search (http://www.ebi.ac.uk/ena/data/warehouse/search) provides links into the TP entry page and taxonomic search filters are provided against all taxonomically annotated records (see figure 2). Links to appropriate TP pages are shown on all ENA records for which taxonomic annotation is available (i.e. all raw data, sequence and sample records). Finally, links are also provided from a number of resources external to ENA using the systematic TP URL structure described at http://www.ebi.ac.uk/ena/about/browser#taxonomy_portal_options. Further usage information is provided at http://www.ebi.ac.uk/ena/about/taxon-portal-web-interface.

Figure 1 – Screenshots of a sample TP taxon page, http://www.ebi.ac.uk/ena/data/view/Taxon:9606, showing a typical view with link to CoL (a) and close-up views of the 'Portal' (b) 'Navigation' (c) and 'Genetic Code' (d) tabs.

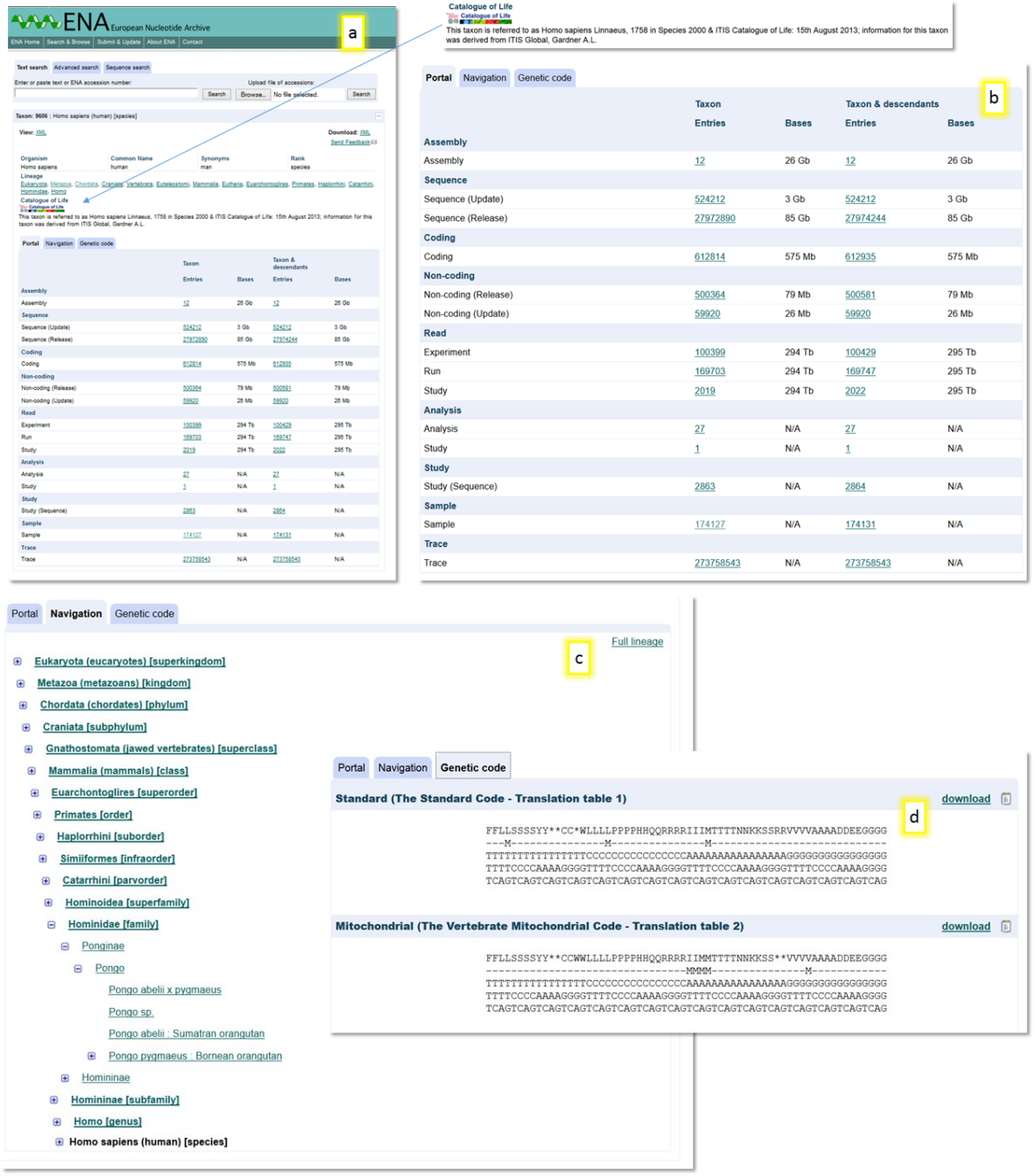
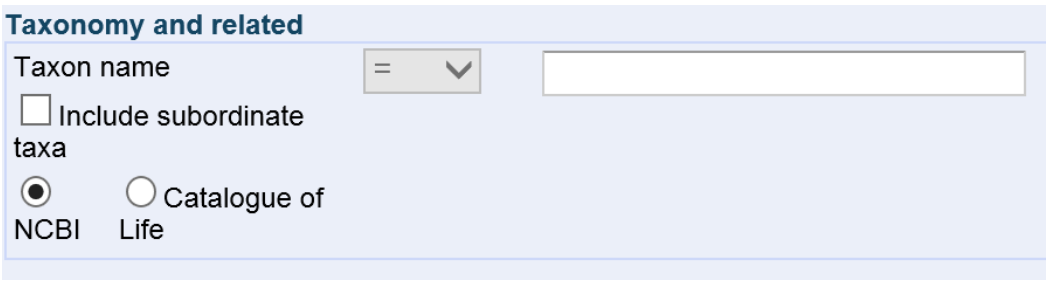


Figure 2 – Screenshot of the TP search module that appears in all Advanced Search pages from which ENA records with taxonomic annotation may be accessed.



RESTful web service interface

The TP offers powerful functionality through a RESTful programmatic interface. Search filters support hierarchy-aware[1] access to NCBI Taxonomy and CoL and a scientific name search (see table 1). Output formats are native ENA XML (e.g. http://www.ebi.ac.uk/ena/data/view/Taxon:2759&display=xml) and Darwin Core XML (e.g. http://www.ebi.ac.uk/ena/data/view/Taxon:2759&display=dwc). Full documentation is provided at http://www.ebi.ac.uk/ena/about/browser#data_warehouse.

Table 2 – RESTful query taxonomic filter functions available from TP.

Function	Description	Parameters	Example
tax_eq	All records that match the given NCBI taxonomy identifier	NCBI taxonomy identifier	tax_eq(9606)
tax_tree	All records that match the given NCBI taxonomy identifier or are descendants of it	NCBI taxonomy identifier	tax_tree(2759)
tax_name	All records that match the given NCBI scientific name	NCBI scientific name	tax_name(Homo%20sapiens)
col_tax_eq	All records that match the given CoL taxonomy identifier	CoL taxonomy identifier	col_tax_eq(6850099)
col_tax_tree	All records that match the given CoL taxonomy identifier or are descendants of it	CoL taxonomy identifier	col_tax_tree(6850295)

Taxonomy

TP supports both the Catalogue of Life (CoL) taxonomy and the *de facto* standard molecular biological taxonomic classification (known as the 'NCBI Taxonomy'). While the NCBI Taxonomy remains the underlying organising classification for ENA content, through a regular application of the i4Life cross-mapper, searches including CoL taxonomic names are supported. Such integration provides connectivity between not only CoL and sequence data but also between other data organised or made available under CoL. Examples include the data resources of the i4Life global biodiversity programmes (GBIF, CBoL, EOL and IUCN) and many beyond. While current TP functionality covers unrestricted access public data from the nucleic acid sequence domain, the near-ubiquitous use of the NCBI Taxonomy across the molecular biology data resources of EMBL-EBI and beyond will support future enhancements to the portal that provide, in response to taxonomic queries, direct summary information and onward links to data covering such domains as proteins, structures, metabolites and systems biology.

Usage

Because the TP has been available for some time in *beta*, we are able to present usage statistics: On average in 2013, excluding robots, but including programmatic and web calls, we have received 15,000 monthly visitors in 30,000 visits and have served 1,000,000 hits.

[1] In the TP system, the NCBI Taxonomy provides the taxonomic hierarchy for both NCBI Taxonomy and CoL queries, but mapped names enable navigation of the hierarchy in either taxonomy's namespace.

Technology
The TP comprises a number of technical components:

1. Persistent database tables (one for each of the ENA data classes): These tables (operated under the Vertica analytical database technology) capture counts of entries per taxon and sub-taxon to provide fast access to this summary information. ENA accession numbers and NCBI taxonomic identifiers are extracted from records for each data class and included in the TP. The full taxonomic tree is indexed in memory and accession number counts are associated with individual taxa. Finally, taxonomic lineages are used to calculate the number of entries associated not only with individual taxa but also with subtaxa.

2. Search structures: To provide rapid access to entries by tax_id or by sub_tree_tax_id, two columns are added to the data class-specific tables. First, tax_id column contains the NCBI tax_id and is used for queries focusing on entry retrieval for a single taxon only. Second, the sub_tree_tax_id column contains the taxonomic linage using an efficient encoding where <N integer>.<M integer>, etc. refers to the Mth child of the Nth parent (e.g. 45.34). This provides a dense encoding of taxonomic lineage and allows us to search sub-tree taxa using a LIKE query. A separate table is created to contain mappings from NCBI tax_id to the encoded lineages. In addition the tables allow users to combine taxonomic searches with other search criteria.

3. Web layer: The TP is presented as part of the ENA Browser to allow users to search, view and download content associated with specific taxa and subtaxa and to examine counts of associated entries. Data downloads are supported in a variety of formats including XML, flat files and fasta for sequence data. The ENA Browser HTML and REST views are provided by redundant Tomcat web nodes operating from the two EMBL-EBI London data centers.

4. ENA Advanced Search: This layer allows users to perform interactive and programmatic searches that combine taxonomy lookups with other data class-specific search criteria and the ability for users to download either the resultant data objects or customisable tabular reports upon these objects.

5. CoL support: CoL is supported in the ENA taxonomy portal and Advanced Search. CoL identifiers are injected at the time of indexing and used to display CoL taxonomy information in the TP. CoL and the i4Life mapping product are stored in an oracle relational database dedicated to this use. Advanced Search allows users to filter by CoL taxonomy or to retrieve CoL taxonomy identifiers in tabular reports.

6. An in-memory cache layer: An in-memory cache service holds summary ENA record counts (see point 1.) in memory for fastest possible performance.

7. A data layer: The data accessed through the taxonomy portal are stored using a combination of Oracle and disk-based approaches and are served from the EMBL-EBI's two London data centers for high availability.

Benefits to ENA

The impact of the work in this deliverable upon the usability of the ENA data resource has been substantial. Usage statistics (see above) already show that a significant number of users are exposed to the TP. As time proceeds and the new functionalities become increasingly known, we expect the scientific outputs to be significant for both our existing user base and new user communities.

Through TP, our body of existing users, including evolutionary biologists, phylogeneticists, taxonomists, systematists and molecular ecologists, are now provided with a way in which they can interrogate and explore ENA content under a simple but powerful interface. Having treated taxonomic name as a simple attribute for many years in the ENA search and browser interfaces, hierarchy awareness (providing the ability to browse and group according to parent and subordinate taxa) and support for synonym dictionaries (allowing users to query based on common and prior names for taxa) offer the user not only a smoother and simpler experience, but also the capacity to work with ENA in ways that were previously not possible.

Combining the taxonomic functionalities developed for TP with gene name dictionary work also supported by i4Life (M7.1a and M7.1b), the Marker Portal (http://www.ebi.ac.uk/ena/data/warehouse/search?portal=marker), an interface to support marker selection and analysis, allows users to slice ENA marker locus sequences by taxonomy and targeted marker gene. We have already received very positive feedback from users about this service. Further details on the Marker Portal are available from http://www.ebi.ac.uk/ena/about/marker-portal-web-interface.

Through the inclusion of support for CoL in the TP, ENA has become able to face new user communities. These users, who come from fields often distant from molecular biology, are expected to include those working in biodiversity, conservation, ecology, marine science, environmental science, industry, agriculture and government. We hope to seed the use of the TP (and other EMBL-EBI services) within conservation biology communities through the publication of a paper relating to the use of molecular data in the prioritisation of species for conservation programmes, co-authored by a number of i4Life partners (see M2.1a).