Programmatic Web Access

657.019 Scripting for Biotechnologists (WS 2018/19)

Goals

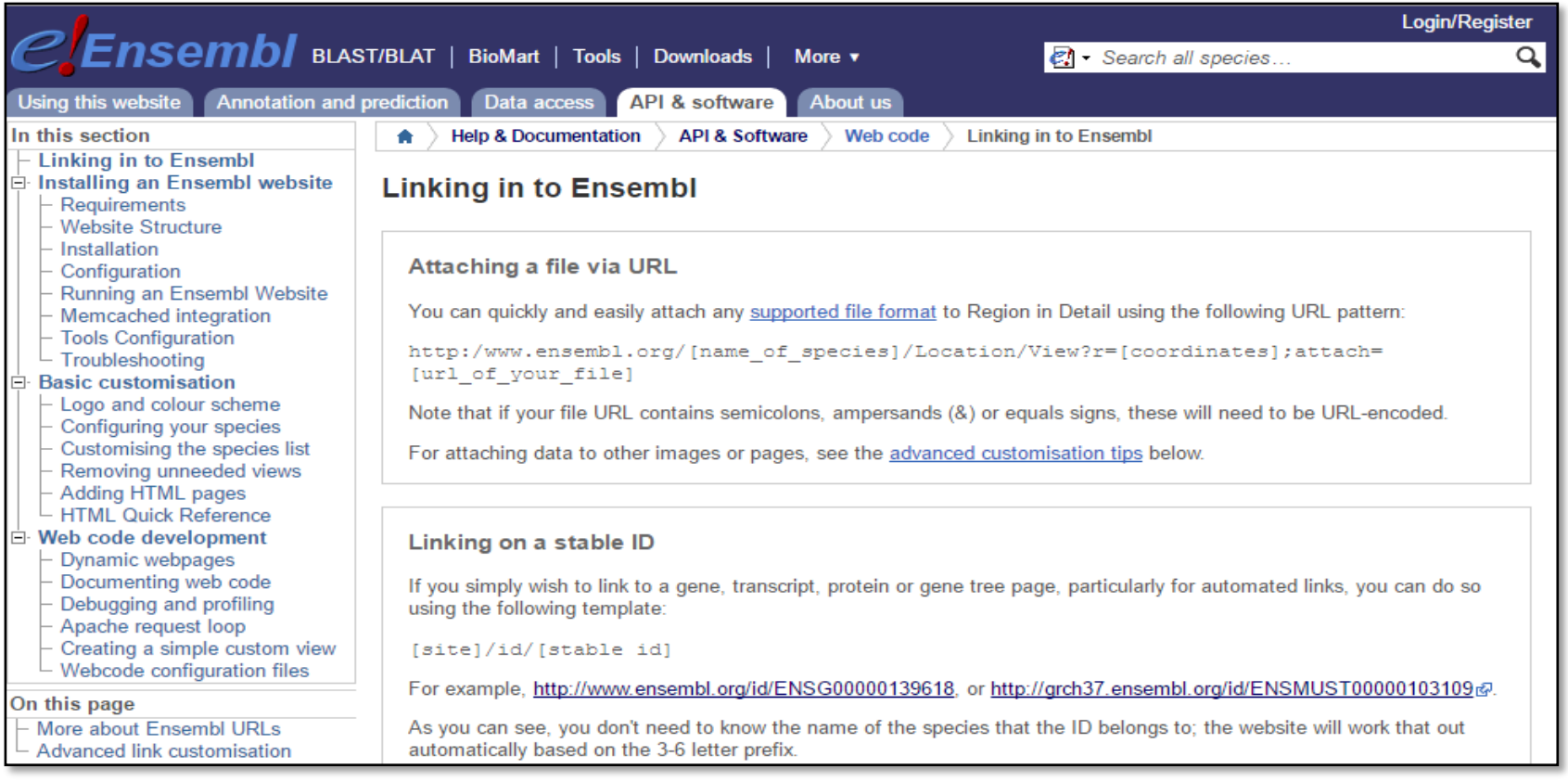
- Find Web addresses for programmatic data access to bioinformatics sites
  - Called API (Application Programming Interface) URLs
  - Just means: there is some syntax to follow in the URL
  - As opposed to access through a Web browser
- Use these URLs from the command line to retrieve data
- Process the downloaded data with Perl and shell scripts
  - Build an analysis pipeline

Bioinformatics APIs

- Ensembl
  - Genomes
- UniProt
  - Proteins
- NCBI
  - Assorted databases

Ensembl (genomes)

- http://www.ensembl.org/info/docs/webcode/linking.html



Ensembl views

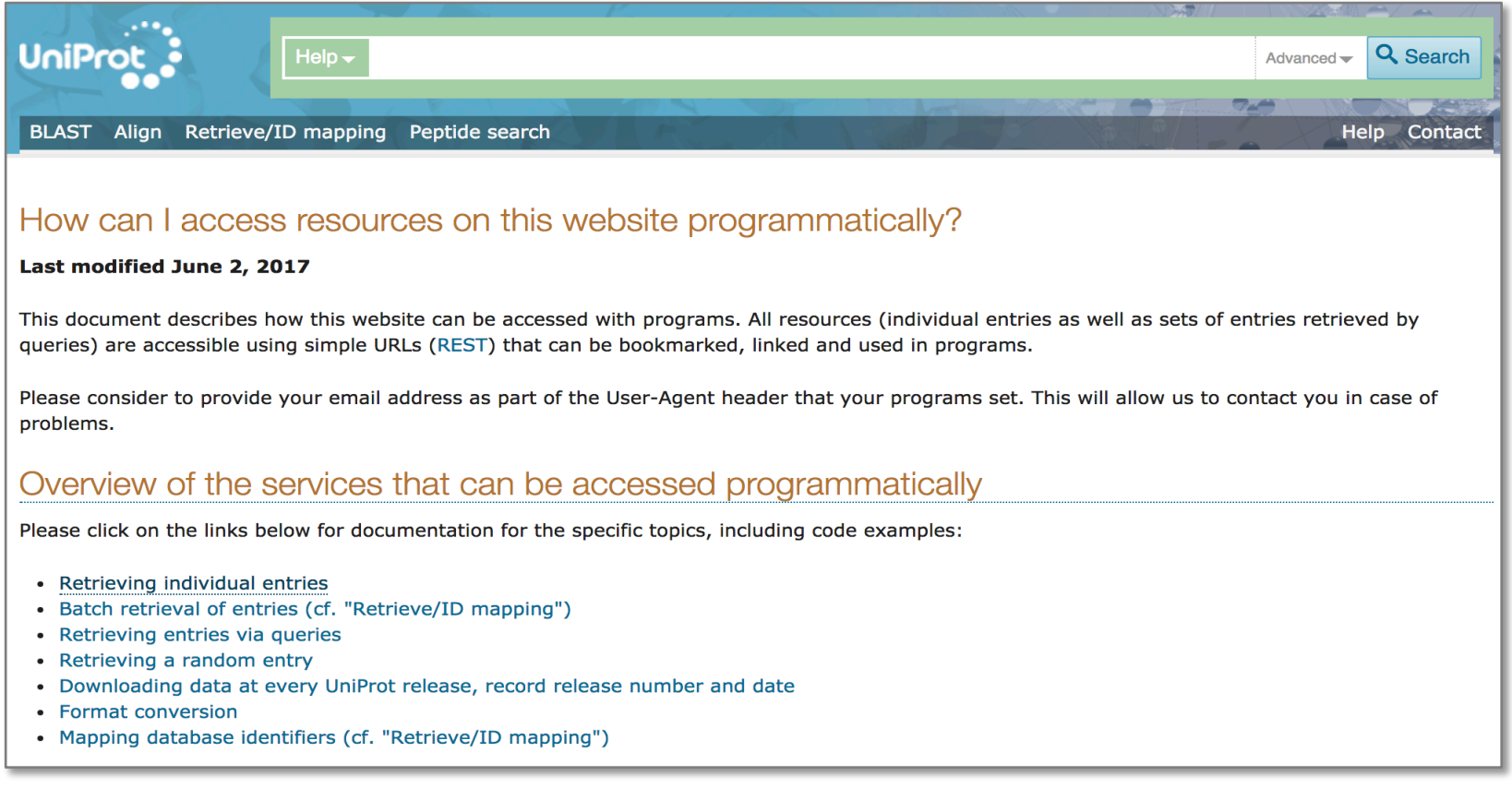
- Simple ID linking
  - http://www.ensembl.org/id/ENSG00000139618
- Location-based views
  - http://www.ensembl.org/Homo_sapiens/Location/View?g=ENSG00000012048
  - http://www.ensembl.org/Homo_sapiens/Location/View?r=17:38449840-38530994
- Feature-based views
  - http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000012048

Ensembl stable IDs

- ENS(species)(feature type)(identifier)
  - ENS: it's an ensembl ID
  - (species): 3-letter species code, none for humans
  - (object type): E=exon, FM=Ensembl protein family, G=gene, GT=gene tree, P=protein, R=regulatory feature, T=transcript
  - (identifier): 11-digit number
- Examples:
  - ENSG00000139618: human gene
  - ENSMUST00000103109: mouse transcript
  - ENSDARP00000006884: zebra fish protein
- More info:
  - http://www.ensembl.org/info/genome/stable_ids/index.html

UniProt (proteins)

- http://www.uniprot.org/help/api

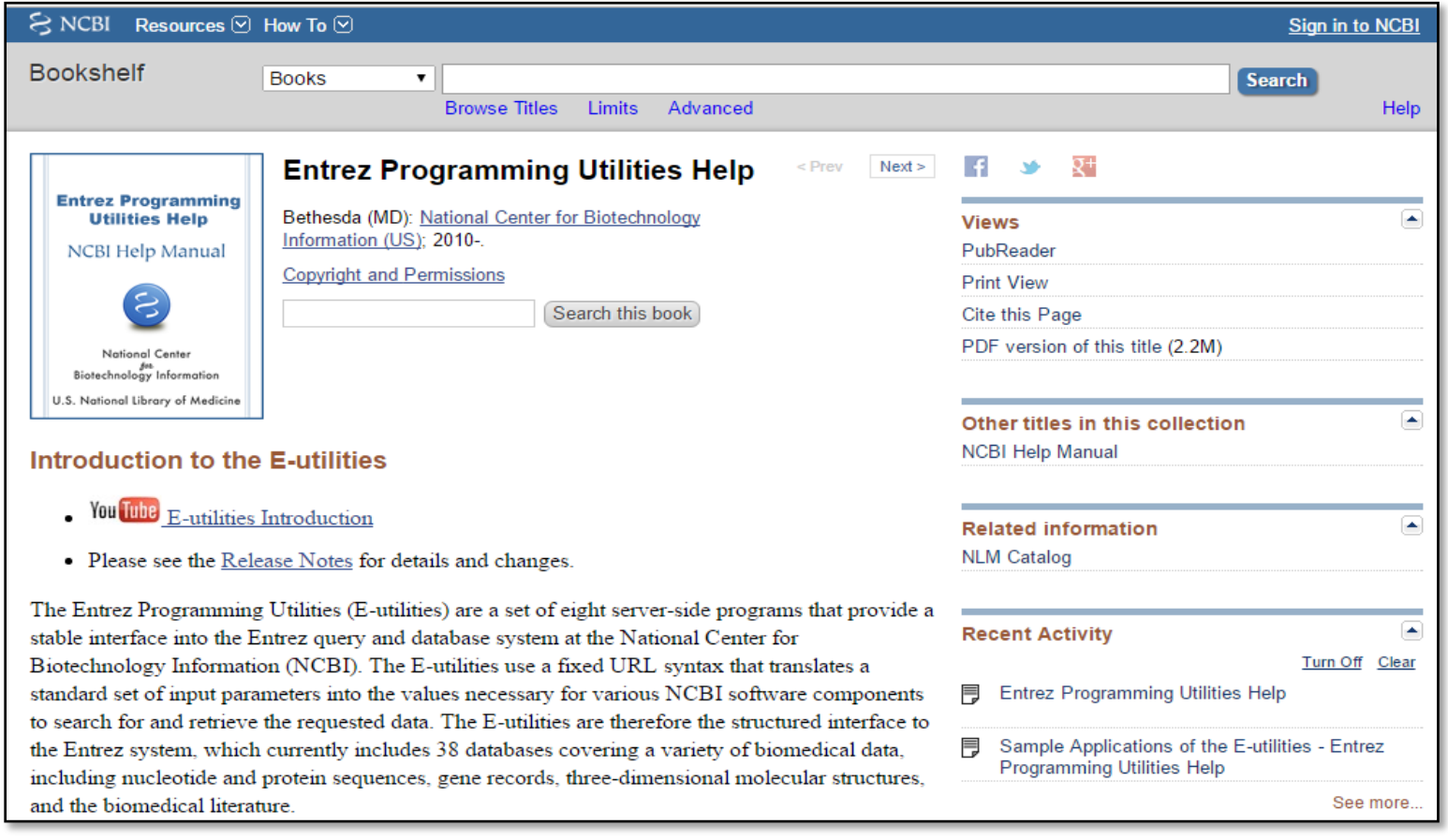


UniProt entry retrieval

- Basic Web page
  - http://www.uniprot.org/uniprot/P12345
- Specific formats
  - http://www.uniprot.org/uniprot/P12345.txt
  - http://www.uniprot.org/uniprot/P12345.xml
  - http://www.uniprot.org/uniprot/P12345.fasta
  - http://www.uniprot.org/uniprot/P12345.gff
  - http://www.uniprot.org/uniprot/P12345.rdf
- Extension indicates the desired format

NCBI (assorted)

- http://eutils.ncbi.nlm.nih.gov



NCBI EFetch

- Returns formatted data records for a list of input UIDs
- Fetch the first 100 bases of the plus strand of GI 21614549 in FASTA format:
  - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=21614549&strand=1&seq_start=1&seq_stop=100&rettype=fasta&retmode=text
- Fetch full XML record for Gene ID 2:
  - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=gene&id=2&retmode=xml
- db and id are required parameters
- rettype and retmode indicate format

NCBI databases

- https://www.ncbi.nlm.nih.gov/books/NBK25497/table/chapter2.T._entrez_unique_identifiers_ui/?report=object only

- E-Utility Database Names are different from Entrez Database

Table 1

– Entrez Unique Identifiers (UIDs) for selected databases

Entrez Database	UID common name	E-utility Database Name
BioProject	BioProject ID	bioproject
BioSample	BioSample ID	biosample
Biosystems	BSID	biosystems
Books	Book ID	books
Conserved Domains	PSSM-ID	cdd
dbGaP	dbGaP ID	gap
dbVar	dbVar ID	dbvar
Epigenomics	Epigenomics ID	epigenomics
EST	GI number	nucest
Gene	Gene ID	gene
Genome	Genome ID	genome
GEO Datasets	GDS ID	gds
GEO Profiles	GEO ID	geoprofiles
GSS	GI number	nucgss
HomoloGene	HomoloGene ID	homologene

Efetch return formats

- https://www.ncbi.nlm.nih.gov/books/NBK25499/table/chapter4.T._valid_values_of__retmode_and/?report=objectonly

- Valid values of **rettype** and **retmode** vary by database

– Valid values of &retmode and &rettype for EFetch (null = empty string)

Record Type	&rettype	&retmode
All Databases		
Document summary	docsum	xml, *default*
List of UIDs in XML	uilist	xml
List of UIDs in plain text	uilist	text
db = bioproject		
Full record XML	xml, *default*	xml, *default*
db = biosample		
Full record XML	full, *default*	xml, *default*
Full record text	full, *default*	text
db = biosystems		
Full record XML	xml, *default*	xml, *default*
db = gds		
Summary	summary, *default*	text, *default*
db = gene		
text ASN.1	*null*	asn.1, *default*
XML	*null*	xml
Gene table	gene_table	text
db = homologene		
text ASN.1	*null*	asn.1, *default*
XML	*null*	xml

`wget`: command line download

- Without a Web browser
  - Need only a terminal
  - Can be used as a command within a script for batch download
- Basic syntax: **`wget <URL>`**
- Safer to surround URL with single quotes to escape special characters within URL
- On Macs, a similar program **`curl`** is available

wget: most useful options

- Downloaded (Output) file name
  - **wget -O outfile.html 'http://site/path?param=value'**
  - Not lowercase **-o** (log file)
- Read URLs from (input) file
  - **wget -i url_list.txt**
  - One URL per line in the URL list file
- Save downloaded files to a directory (Prefix)
  - **wget -P downloadDir http://...**
  - Not **-p** (page requisites: download all images, etc. needed to display HTML page)

More on Perl one-liners

- Separate multiple actions
  - `perl -ane 's/junk_text_regex//g; print if /good_stuff_regex/'`
  - Order of actions matters

- Not to match a regex (inverse match, negation)
  - `perl -ane 'print unless /regex/'`

Summary

- We point to things on the Web using Uniform Resource Locators (URLs)
- We can systematically access information from bioinformatics sites with defined URL syntax
- Each bioinformatics site has its own URL syntax to be followed
- Automated downloading can be combined with shell scripting and regex matches to build powerful analysis tools!