Managing Data: Preservation and Access

IST400/600

Jian Qin

Summary of Duranti's article

Agent
Physical and Intellectual form
Action
Context

Authenticity
Mode
Form
State of transmission
Preservation and custody
Security

Reliability
Authority
Trustworthiness

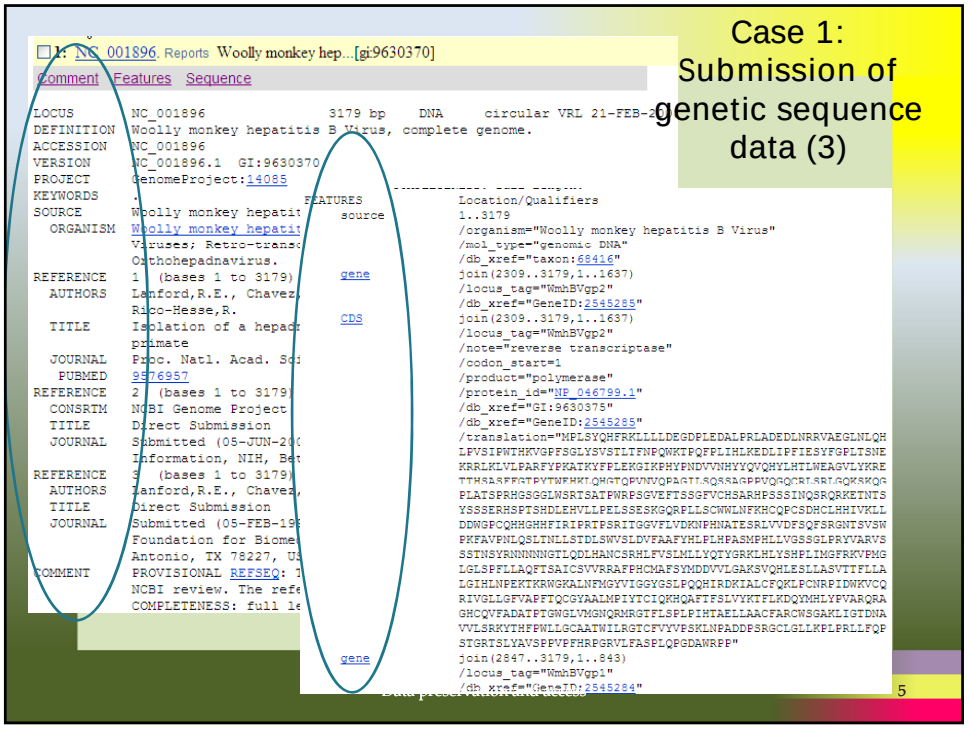
Case 1: Submission of genetic sequence data (1)

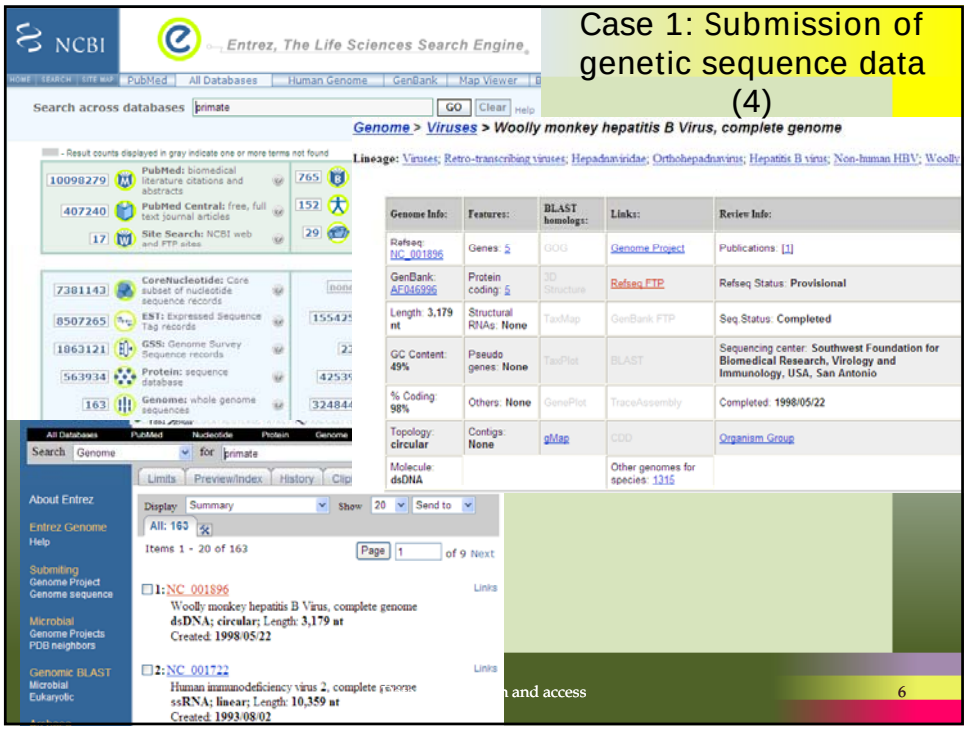
- Discipline: Biological sciences
- Needs for data:
  - Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication.
  - NCBI relies on data contributions to provide timely and accurate processing and biological review of new entries and updates to existing entries
- How do we translate these needs into data management needs?

Case 1: Submission of genetic sequence data (2)

- Need
  - Descriptive data elements for genetic sequence data
    - Submission interface for submitting and updating data sets
      - Tool development
    - Search interface for submitted sequence data
  - Backend management
    - Linking between relevant databases
    - Linking between data and publications
    - Data and database administration
    - Long-term preservation of data

Compiled based on http://www.ncbi.nlm.nih.gov/Genbank/index.html


Case 1: Submission of genetic sequence data (3)
1: NC_001896. Reports Woolly monkey hep...[gi:9630370]
Comment Features Sequence
LOCUS NC_001896 3179 bp DNA circular VRL 21-FEB-2
DEFINITION Woolly monkey hepatitis B Virus, complete genome.
ACCESSION NC_001896
VERSION NC_001896.1 GI:9630370
PROJECT GenomeProject:14085
KEYWORDS .
SOURCE Woolly monkey hepatit
ORGANISM Woolly monkey hepatit
Viruses; Retro-trans
Orthohepadnavirus.
REFERENCE 1 (bases 1 to 3179)
AUTHORS Lanford,R.E., Chavez,
Rico-Hesse,R.
TITLE Isolation of a hepad
primate
JOURNAL Proc. Natl. Acad. Sci
PUBMED 9576957
REFERENCE 2 (bases 1 to 3179)
CONSRTM NCBI Genome Project
TITLE Direct Submission
JOURNAL Submitted (05-JUN-20
Information, NIH, Bet
REFERENCE 3 (bases 1 to 3179)
AUTHORS Lanford,R.E., Chavez,
TITLE Direct Submission
JOURNAL Submitted (05-FEB-19
Foundation for Biome
Antonio, TX 78227, US
COMMENT PROVISIONAL REFSEQ: T
NCBI review. The refe
COMPLETENESS: full le
FEATURES Location/Qualifiers
source 1..3179
/organism="Woolly monkey hepatitis B Virus"
/mol_type="genomic DNA"
/db_xref="taxon:68416"
gene join(2309..3179,1..1637)
/locus_tag="WmhBVgp2"
/db_xref="GeneID:2545285"
CDS join(2309..3179,1..1637)
/locus_tag="WmhBVgp2"
/note="reverse transcriptase"
/codon_start=1
/product="polymerase"
/protein_id="NP_046799.1"
/db_xref="GI:9630375"
/db_xref="GeneID:2545285"
/translation="MPLSYQHFRKLLLLDEGDPLEDALPRLADEDLNRRVAEGLNLQH
LPVSIPWTHKVGPFSGLYSVSTLTFNPQWKTPQFPLIHLKEDLIPFIESYFGPLTSNE
KRRLKLVLPARFYPKATKYFPLEKGIKPHYPNDVVNHYYQVQHYLHTLWEAGVLYKRE
TTHSASFFGTPYTWFHKTQHGTQPVNVQPAGTLSQSSAGPPVQGQCRLSRTAQKSKQG
PLATSPRHGSGGLWSRTSATPWRPSGVEFTSSGFVCHSARHPSSSINQSRQRKETNTS
YSSSERHSPTSHDLEHVLLPELSSESKGQRPLLSCWWLNFKHCQPCSDHCLHHIVKLL
DDWGPCQHHGHHFIRIPRTPSRITGGVFLVDKNPHNATESRLVVDFSQFSRGNTSVSW
PKFAVPNLQSLTNLLSTDLSWVSLDVFAAFYHLPLHPASMPHLLVGSSGLPRYVARVS
SSTNSYRNNNNNGTLQDLHANCSRHLFVSLMLLYQTYGRKLHLYSHPLIMGFRKVPMG
LGLSPFLLAQFTSAICSVVRRAFPHCMAFSYMDDVVLGAKSVQHLESLLASVTTFLLA
LGIHLNPEKTKRWGKALNFMGYVIGGYGSLPQQHIRDKIALCFQKLPCNRPIDWKVCQ
RIVGLLGFVAPFTQCGYAALMPIYTCIQKHQAFTFSLVYKTFLKDQYMHLYPVARQRA
GHCQVFADATPTGWGLVMGNQRMRGTFLSPLPIHTAELLAACFARCWSGAKLIGTDNA
VVLSRKYTHFPWLLGCAATWILRGTCFVYVPSKLNPADDPSRGCLGLLKPLPRLLFQP
STGRISLYAVSPFVPFHRPGRVLFASPLQPGDAWRPF"
gene join(2847..3179,1..843)
/locus_tag="WmhBVgp1"
/db_xref="GeneID:2545284"
5

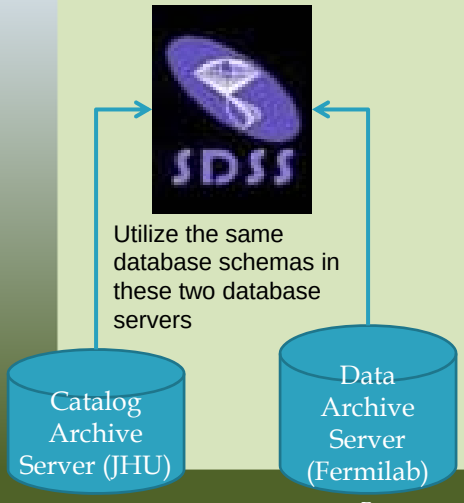

Case 1: Submission of genetic sequence data (4)
NCBI
Entrez, The Life Sciences Search Engine
PubMed All Databases Human Genome GenBank Map Viewer
Search across databases primate GO Clear Help
Genome > Viruses > Woolly monkey hepatitis B Virus, complete genome
Result counts displayed in gray indicate one or more terms not found
PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
CoreNucleotide: Core subset of nucleotide sequence records
EST: Expressed Sequence Tag records
GSS: Genome Survey Sequence records
Protein: sequence database
Genome: whole genome sequences
Lineage: Viruses; Retro-transcribing viruses; Hepadnaviridae; Orthohepadnavirus; Hepatitis B virus; Non-human HBV; Woolly
Genome Info: Features: BLAST homologs: Links: Review Info:
Refseq: NC_001896
GenBank: AF046996
Length: 3,179 nt
GC Content: 49%
% Coding: 98%
Topology: circular
Molecule: dsDNA
Genes: 5
Protein coding: 5
Structural RNAs: None
Pseudo genes: None
Others: None
Contigs: None
Genome Project
Refseq FTP
Refseq Status: Provisional
Seq Status: Completed
Sequencing center: Southwest Foundation for Biomedical Research, Virology and Immunology, USA, San Antonio
Completed: 1998/05/22
Organism Group
Other genomes for species: 1315
Search Genome for primate
Limits Preview/Index History Clip
About Entrez
Entrez Genome Help
Submitting Genome Project Genome sequence
Microbial Genome Projects PDB neighbors
Genomic BLAST Microbial Eukaryotic
Display Summary Show 20 Send to
All: 163
Items 1 - 20 of 163
Page 1 of 9 Next
1: NC_001896
Woolly monkey hepatitis B Virus, complete genome
dsDNA; circular; Length: 3,179 nt
Created: 1998/05/22
2: NC_001722
Human immunodeficiency virus 2, complete genome
ssRNA; linear; Length: 10,359 nt
Created: 1993/08/02
and access
6

Case 2: querying image data (1)

- Data case: SkyServer (http://cas.sdss.org/dr6/en/)
- Discipline: Astronomy
- Data size: 80GB containing 14 million objects and 50 thousand spectra
- Query type:
  - Point-and-click requests for
    - Images of the sky
    - Images of spectra
    - Tabular output s of the SDSS database

Compiled based on:
Szalay, A. et al. (2002). ftp://ftp.research.microsoft.com/pub/tr/tr-2001-104.pdf
Gray, J. et al. (2002). http://www.sdss.jhu.edu/ScienceArchive/pubs/msr-tr-2002-01.pdf

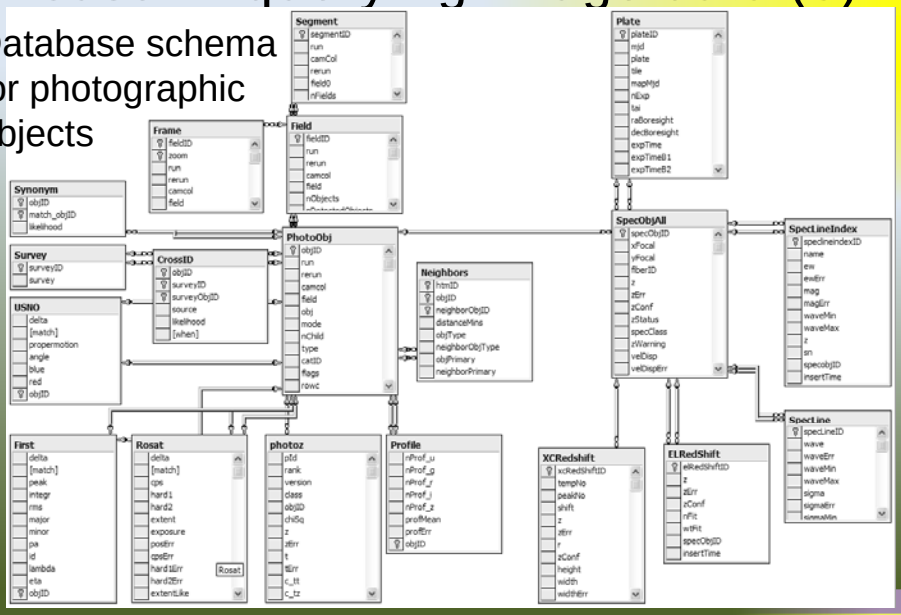
Case 2: querying image data (2)



- Data mining: Question answering
  - Defined 20 typical queries
  - Designed the SkyServer database to answer those questions
- Data objects:
  - Photographic objects
  - Spectroscopic objects

Case 2: querying image data (3)

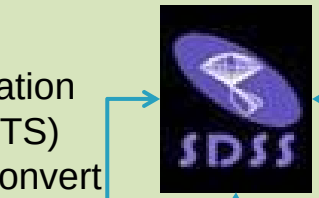
Database schema for photographic objects

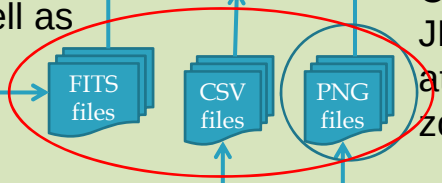


Case 2: querying image data (4)

SkyServer

Data Transformation Service (DTS) load and convert data as well as check for integrity





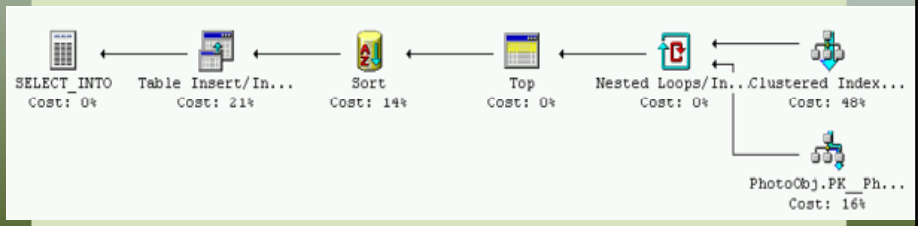
Converted to JPEG images at various zoom levels

The Main 2.5-meter Telescope

Case 2: querying image data (5)

Query: find all galaxies without saturated pixels within 1' of a given point

```
declare @saturated bigint;                          -- initialized "saturated" flag
set     @saturated = dbo.fPhotoFlags('saturated'); -- avoids SQL2K optimizer problem
select  G.objID, GN.distance                        -- return Galaxy Object ID and
into    ##results                                   -- angular distance (arc minutes)
from    Galaxy                          as G        -- join Galaxies with
 join   fGetNearbyObjEq(185,-0.5, 1) as GN          -- objects within 1' of ra=185 & dec=-.5
                  on G.objID = GN.objID             -- connects G and GN
where (G.flags & @saturated) = 0                    -- not saturated
order by distance                                   -- sorted nearest first
```



Case 2: query image data (6)

Query: Provide a list of moving objects consistent with an asteroid

```
select objID,                                       -- return object ID
       sqrt( power(rowv,2) + power(colv, 2) ) as velocity, -- velocity
       dbo.fGetUrlExpId(objID) as Url               -- url of image to examine it.
into  ##results
from   PhotoObj                                     -- check each object.
where (power(rowv,2) + power(colv, 2)) between 50 and 1000 -- square of velocity
  and rowv >= 0 and colv >=0                        -- negative values indicate error
```

What do the two cases tell us?

- A lot of data loading and conversion on a daily basis
- Preserving data is an ongoing process
- Preserving data needs metadata (organizing)
  - Who created the data, owns it, what it is about, use policy, etc.
- Short- and long-term access to data (access)
  - what is available, where the available data are located, and how to get them

Data preservation challenges

- Data formats
  - Vary in data types, e.g. vector and raster data types
  - Format conversions, e.g. from an old version to a newer one
- Data relations
  - e.g. there are data models, annotations, classification schemes, and symbolization files for a digital map
- Semantic issues
  - Naming datasets and attributes

Data access challenges

- Reliability
- Authenticity
- Leverage technology to make data access easier and more effective
  - Cross-database search
  - Integration applications