From Web APIs for Biology (WABI) to Semantic Web API for Biology (SABI)

Center for Information Biology and DNA
Data Bank of Japan (**DDBJ**)
National Institute of Genetics

Hideaki SUGAWARA

Motivation to develop
DDBJ Web API in 2002

It is laborious and sometimes impossible for us to write a computer program that finds useful data resources, sends a proper query and processes the output.

⇩

To improve the problem solving environment for bioinformaticians, we implemented a SOAP (Simple Object Access Protocol) server* and Web services that provide a programmatic interface.

* [Reference] Sugawara H, Miyazaki S. Biological SOAP servers and web services provided by the public sequence data bank. Nucleic Acids Res. 2003 Jul 1;31(13):3836-9

DDBJ : DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, through close collaboration with EBI in Europe and NCBI in USA.

Hot Topics

Feb. 1, 2008	Releasing millions of entries from DDBJ: What has the new computer system in a year brought with? NEW
Feb. 1, 2008	Update of the databases related to the H-Invitational NEW
Dec. 27, 2007	(Mar. 2008) 7th Japan-Korea-China Bioinformatics Training Course NEW
Aug. 1, 2007	[Caution] Revision of DDBJ flat file format

Maintenance

Jan. 22, 2008	ARSA database search (DDBJ, DAD) temporary unavailable

Sequence Data Submission

- Submit my sequences
  Orientation for the data submission
- Update my entries
  Guidance for the update of the entry

FTP/Web API

- FTP (ftp.ddbj.nig.ac.jp)
  Download data files
- Web API
  Programmatic interfaces of DDBJ Web services

Web API for Bioinformatics (WABI)

-expansion of "XML Central of DDBJ"-

What's new

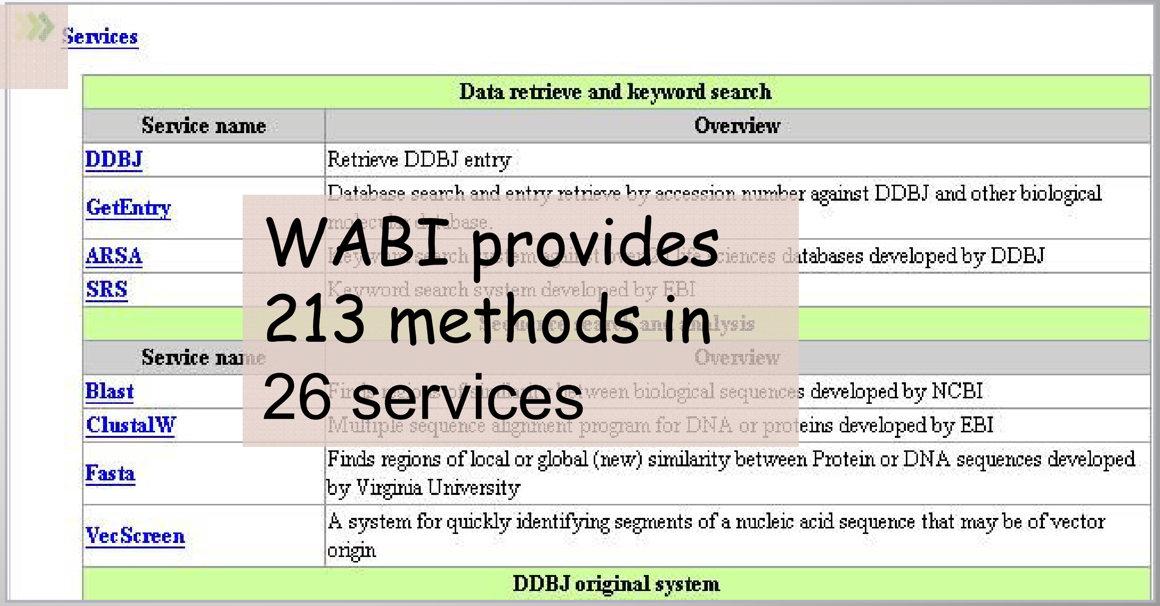
Feb 06 2008 ARSA Service was updated.

Oct 11 2007 GetEntry Service was updated.(19 databases were added for entry search.)

Oct 02 2007 XML Central of DDBJ revised the web page.

Aug 24 2007 XML Central of DDBJ released REST service.

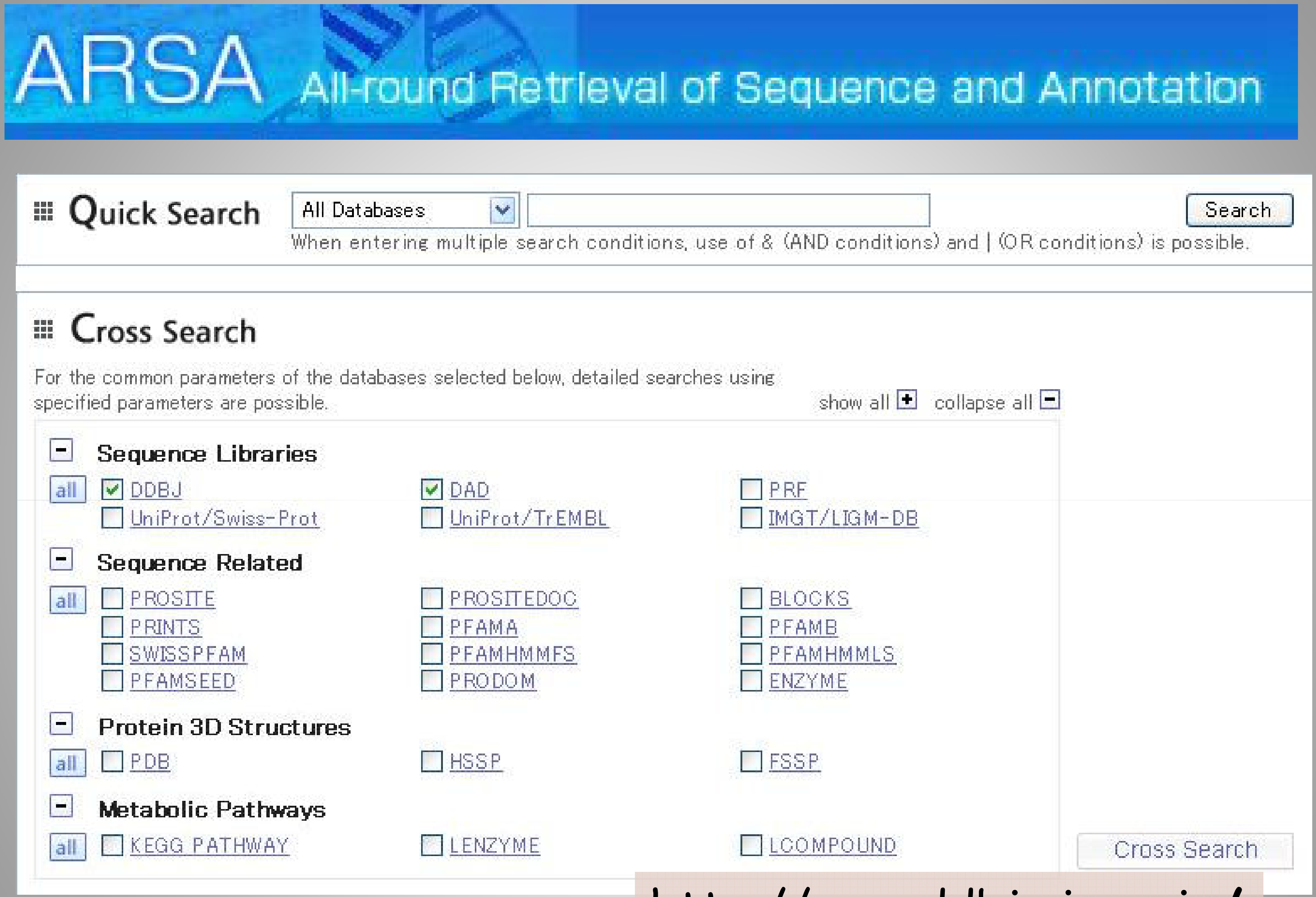
http://xml.nig.ac.jp/



Services

Data retrieve and keyword search	
Service name	**Overview**
DDBJ	Retrieve DDBJ entry
GetEntry	Database search and entry retrieve by accession number against DDBJ and other biological molecular database.
ARSA	Keyword search system [illegible] sciences databases developed by DDBJ
SRS	Keyword search system developed by EBI
Sequence search and analysis	
Service name	**Overview**
Blast	Finds regions of similarity between biological sequences developed by NCBI
ClustalW	Multiple sequence alignment program for DNA or proteins developed by EBI
Fasta	Finds regions of local or global (new) similarity between Protein or DNA sequences developed by Virginia University
VecScreen	A system for quickly identifying segments of a nucleic acid sequence that may be of vector origin
DDBJ original system	

WABI provides
213 methods in
26 services



ARSA All-round Retrieval of Sequence and Annotation

Quick Search

All Databases ▾ Search

When entering multiple search conditions, use of & (AND conditions) and | (OR conditions) is possible.

Cross Search

For the common parameters of the databases selected below, detailed searches using specified parameters are possible.

show all ⊞ collapse all ⊟

⊟ Sequence Libraries

all
- ☑ DDBJ
- ☐ UniProt/Swiss-Prot
- ☑ DAD
- ☐ UniProt/TrEMBL
- ☐ PRF
- ☐ IMGT/LIGM-DB

⊟ Sequence Related

all
- ☐ PROSITE
- ☐ PRINTS
- ☐ SWISSPFAM
- ☐ PFAMSEED
- ☐ PROSITEDOC
- ☐ PFAMA
- ☐ PFAMHMMFS
- ☐ PRODOM
- ☐ BLOCKS
- ☐ PFAMB
- ☐ PFAMHMMLS
- ☐ ENZYME

⊟ Protein 3D Structures

all
- ☐ PDB
- ☐ HSSP
- ☐ FSSP

⊟ Metabolic Pathways

all
- ☐ KEGG PATHWAY
- ☐ LENZYME
- ☐ LCOMPOUND

Cross Search

http://arsa.ddbj.nig.ac.jp/

ARSA WABI

- The system provides not only free keyword search but also field search by using XML path against some bio-databases.

e.g,

```
/ENTRY/DDBJ/definition = '16S ribosomal RNA'
```

```
/ENTRY/DDBJ/length > 10000 AND
 /ENTRY/DDBJ/length < 20000
```

```
(/ENTRY/DDBJ/feature-table/feature{/f_key=='rRNA'
AND  /f_quals/qualifier{/q_name=='product' AND
/q_value='16S ribosomal RNA'}})
```

Categorization of WABI services

Service	Description
getentry, SRS, ARSA	Entry retrieval and keyword search of more than 20 databases
BLAST, FASTA, ClustalW, VecScreen, Mapping	Homology search and multiple alignment
GIB, GIB-V, GIB-IS, GIB-ENV, GTPS, GTOP	Retrieval of the secondary databases developed by DDBJ
Ensembl, NCBIGenomeAnnotation, RefSeq, GO, OMIM	Retrieval of useful databases developed by other sites
TxSearch	An unified taxonomy by the International Nucleotide Sequence Database Collaboration (INSDC=DDBJ/EMBL/GenBank)

Mapping of WABI to BioMOBY's service ontology

Service > Bioinformatics > **Database** > **Searching** > GettingNucleotides

GetEntry (DDBJ), RefSeq (NM_, NC_ prefix etc)

Service > Bioinformatics > **Database** > **Searching** > GettingAminoacids

GetEntry (DAD, Uniprot-TrEMBL, Uniprot-Swiss-prot, PRF, PDB)
RefSeq (NP_ prefix etc)

Service > Bioinformatics > **Database** > **Searching** > GettingStructures

PDB

Service > Bioinformatics > **Database** > **Searching** > GettingText

TxSearch , GIB, GTPS, OMIM, Ensembl etc

Service > Bioinformatics > **Database** > **Searching**

ARSA, SRS, OMIM

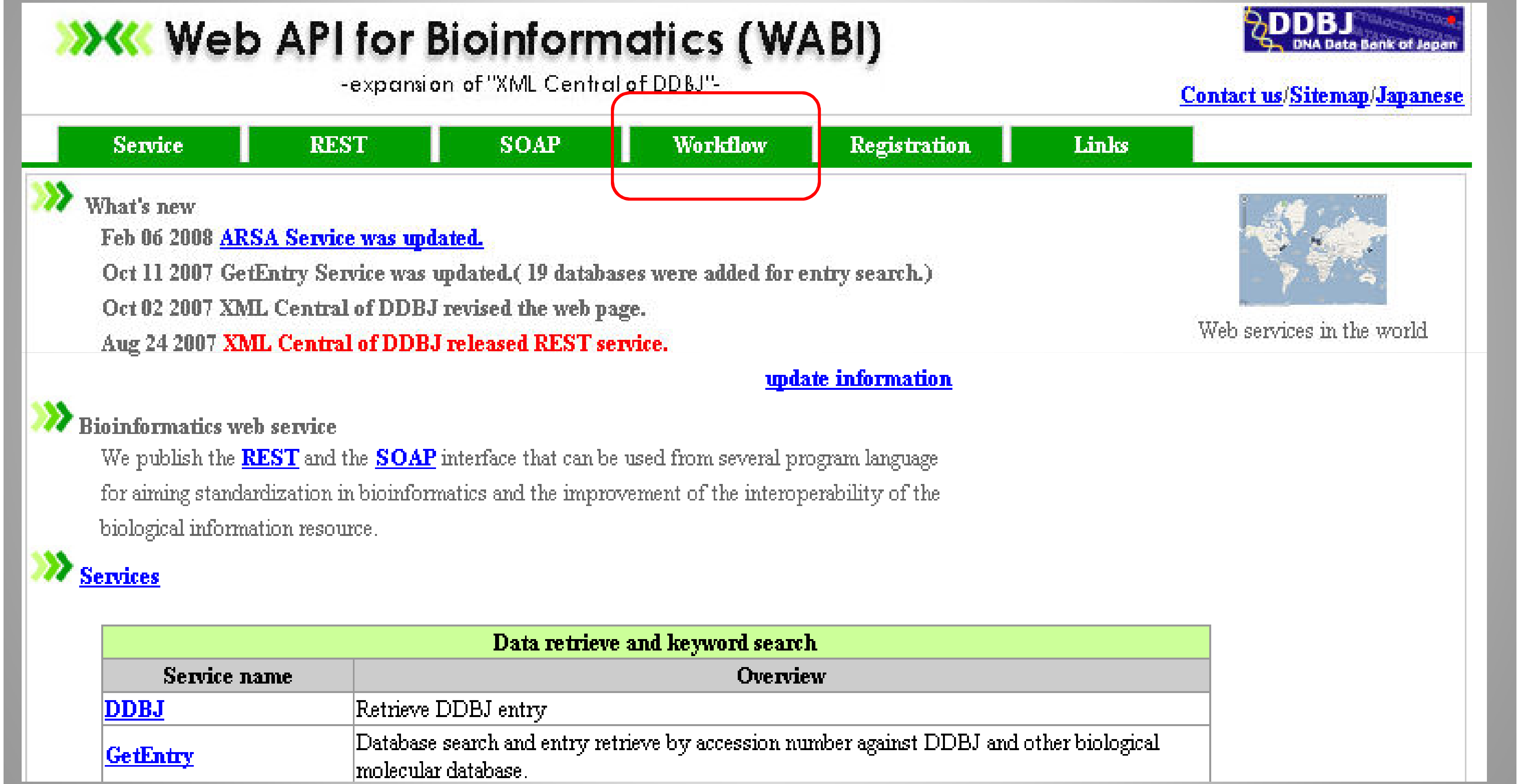
Service > **Analysis** > **Alignment** > Pairwise_Sequence_Comparison > Pairwise_Local_Aligning

BLAST, FASTA, VecScreen, Mapping

Service > **Analysis** > **Alignment** > Multiple_Sequence_Comparison

ClustalW

WABI workflows



Web API for Bioinformatics (WABI)

-expansion of "XML Central of DDBJ"-

What's new

Feb 06 2008 ARSA Service was updated.

Oct 11 2007 GetEntry Service was updated.(19 databases were added for entry search.)

Oct 02 2007 XML Central of DDBJ revised the web page.

Aug 24 2007 XML Central of DDBJ released REST service.

Web services in the world

update information

Bioinformatics web service

We publish the REST and the SOAP interface that can be used from several program language for aiming standardization in bioinformatics and the improvement of the interoperability of the biological information resource.

Services

Data retrieve and keyword search	
Service name	**Overview**
DDBJ	Retrieve DDBJ entry
GetEntry	Database search and entry retrieve by accession number against DDBJ and other biological molecular database.

SNPs workflow

It is to classify all of the human genes from the viewpoint of SNPs.

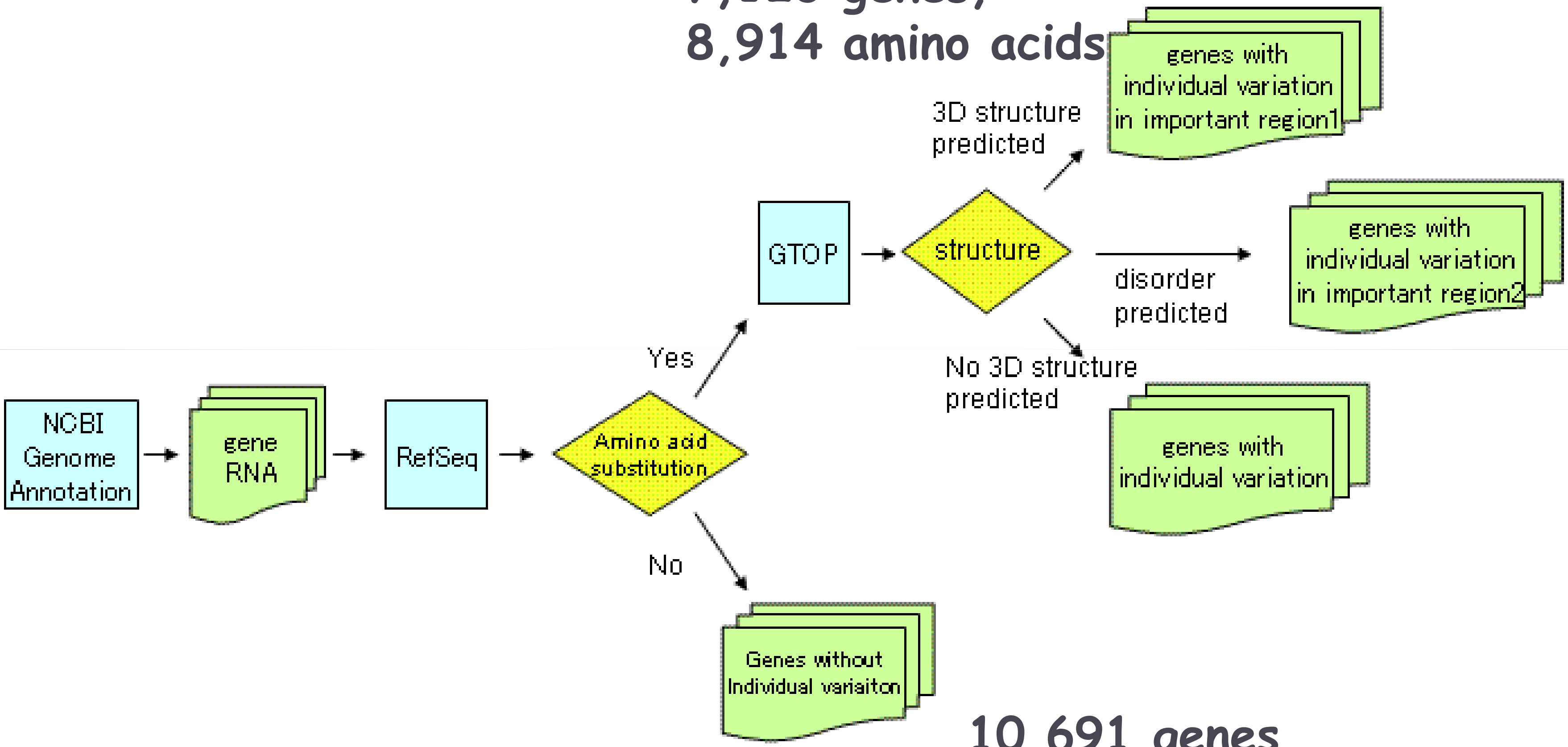
Genes **without** the amino acid substitution are important for drug development.

Genes **with** amino acid substitution and 3D structure modification are also important because it causes the disease.

SNPs workflow chart and result

There is a possibility that
it causes the disease.

7,123 genes,
8,914 amino acids



NCBI Genome Annotation → gene RNA → RefSeq → Amino acid substitution

Yes → GTOP → structure

3D structure predicted → genes with individual variation in important region1

disorder predicted → genes with individual variation in important region2

No 3D structure predicted → genes with individual variation

No → Genes without Individual variaiton

Important for the drug development

10,691 genes,
14,571 amino acids

Frequently posted questions and comments:

- Is it useful?
- Who will use?
- Biologists will not use.
- Neither will bioinformaticians do.
- It is only for OTAKU.

OUTREACH

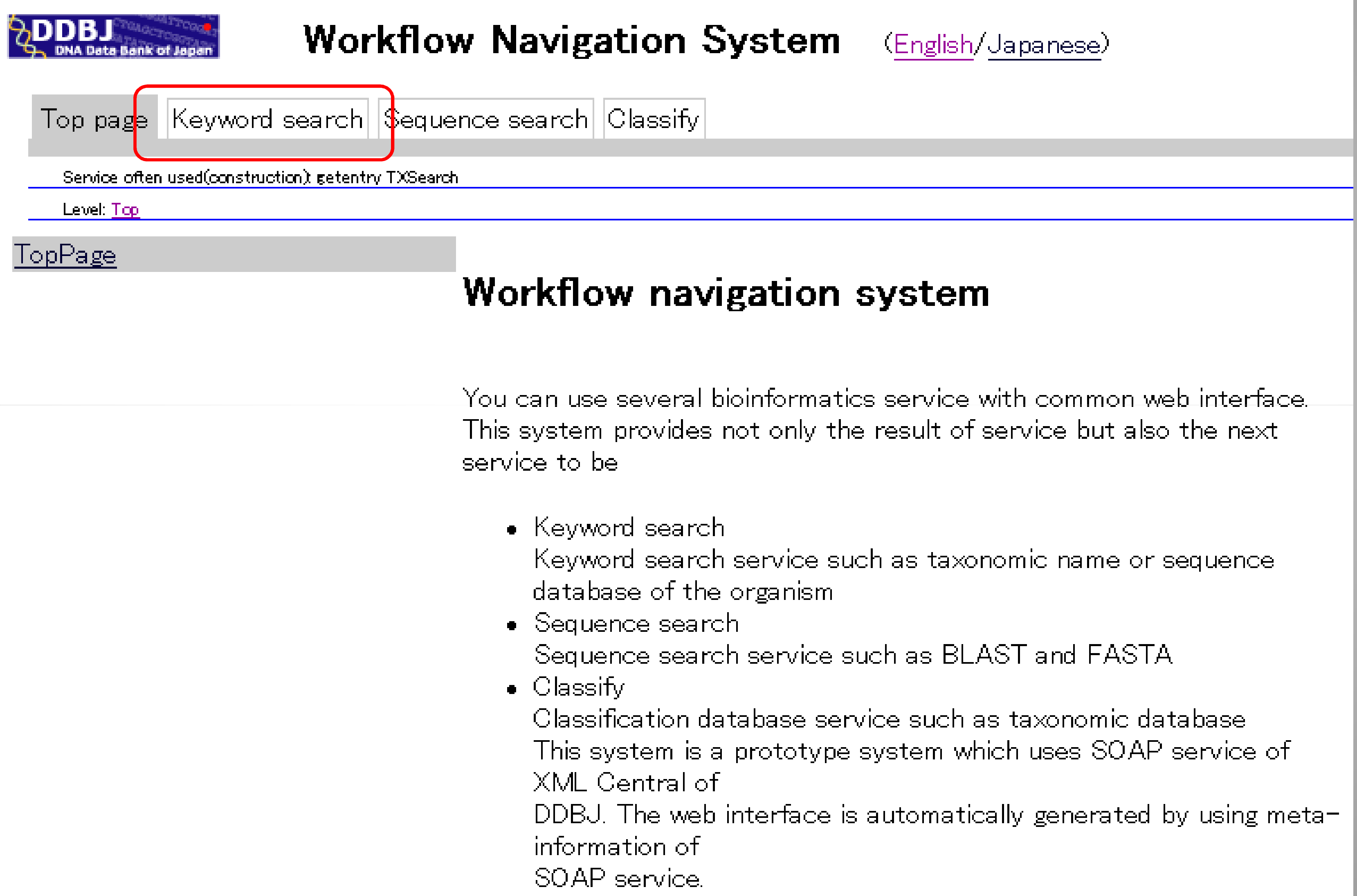
FYI: OTAKU

- The term was popularized in the English speaking world in William Gibson's 1996 novel *Idoru*, which has several references to otaku. In particular, the term was defined as **'pathological-techno-fetishist-with-social-deficit'**.
- In an April 2001 edition of *The Observer*, William Gibson explained his view of the term:

  The otaku, **the passionate obsessive, the information age's embodiment of the connoisseur, more concerned with the accumulation of data than of objects**- - - Understanding otaku -hood, I think, is one of the keys to understanding the culture of the web.

WORKFLOW NAVIGATION SYSTEM FOR OUTREACH

http://cyclamen.ddbj.nig.ac.jp/



Workflow Navigation System

Service often used(construction): getentry TXSearch

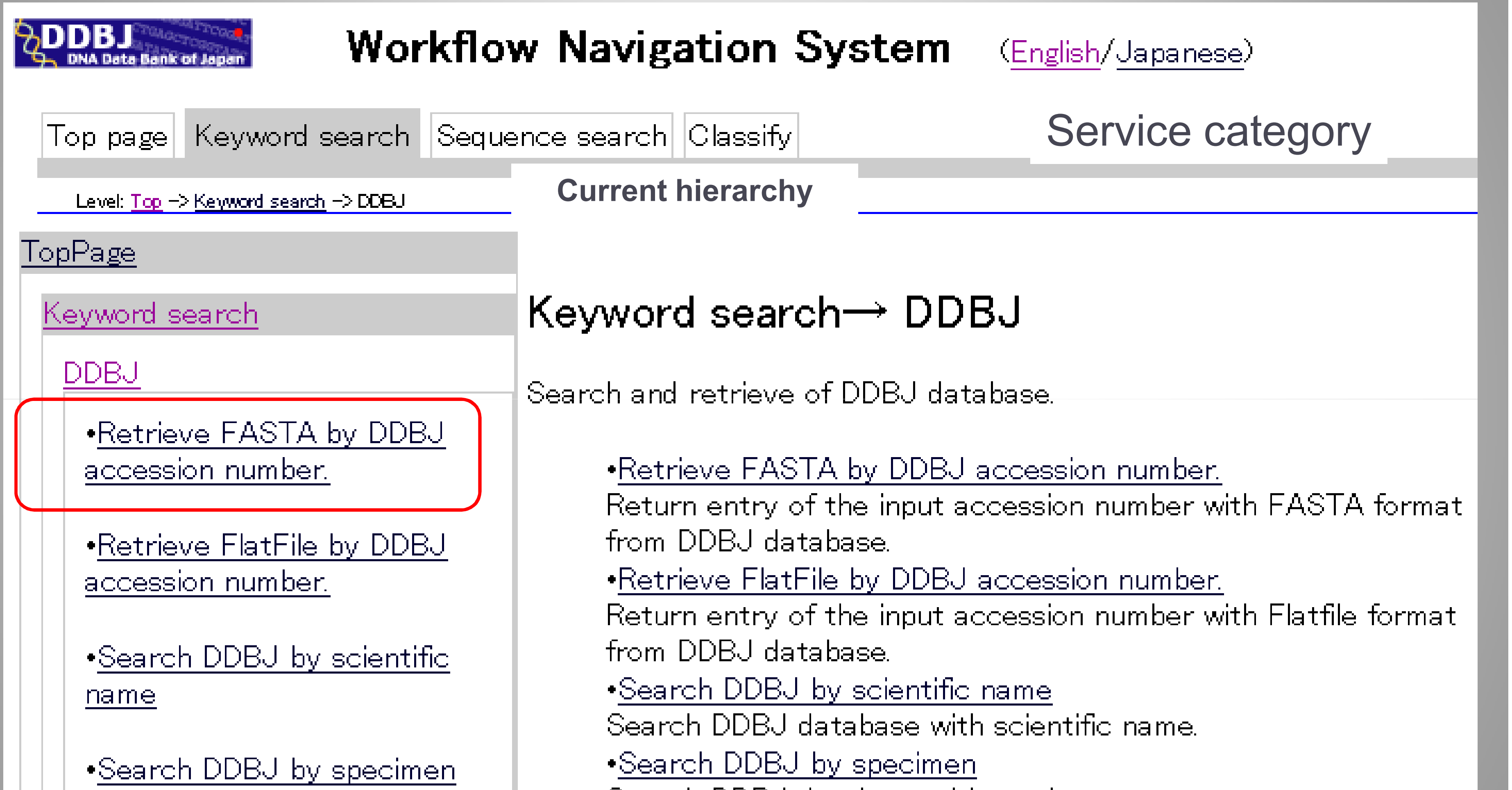
Level: Top

Workflow navigation system

You can use several bioinformatics service with common web interface.
This system provides not only the result of service but also the next service to be

- Keyword search
  Keyword search service such as taxonomic name or sequence database of the organism
- Sequence search
  Sequence search service such as BLAST and FASTA
- Classify
  Classification database service such as taxonomic database
  This system is a prototype system which uses SOAP service of XML Central of
  DDBJ. The web interface is automatically generated by using meta-information of
  SOAP service.

Categorization of services



Workflow Navigation System (English/Japanese)

Service category

Level: Top -> Keyword search -> DDBJ

Current hierarchy

- Retrieve FASTA by DDBJ accession number.
- Retrieve FlatFile by DDBJ accession number.
- Search DDBJ by scientific name
- Search DDBJ by specimen

Keyword search→ DDBJ

Search and retrieve of DDBJ database.

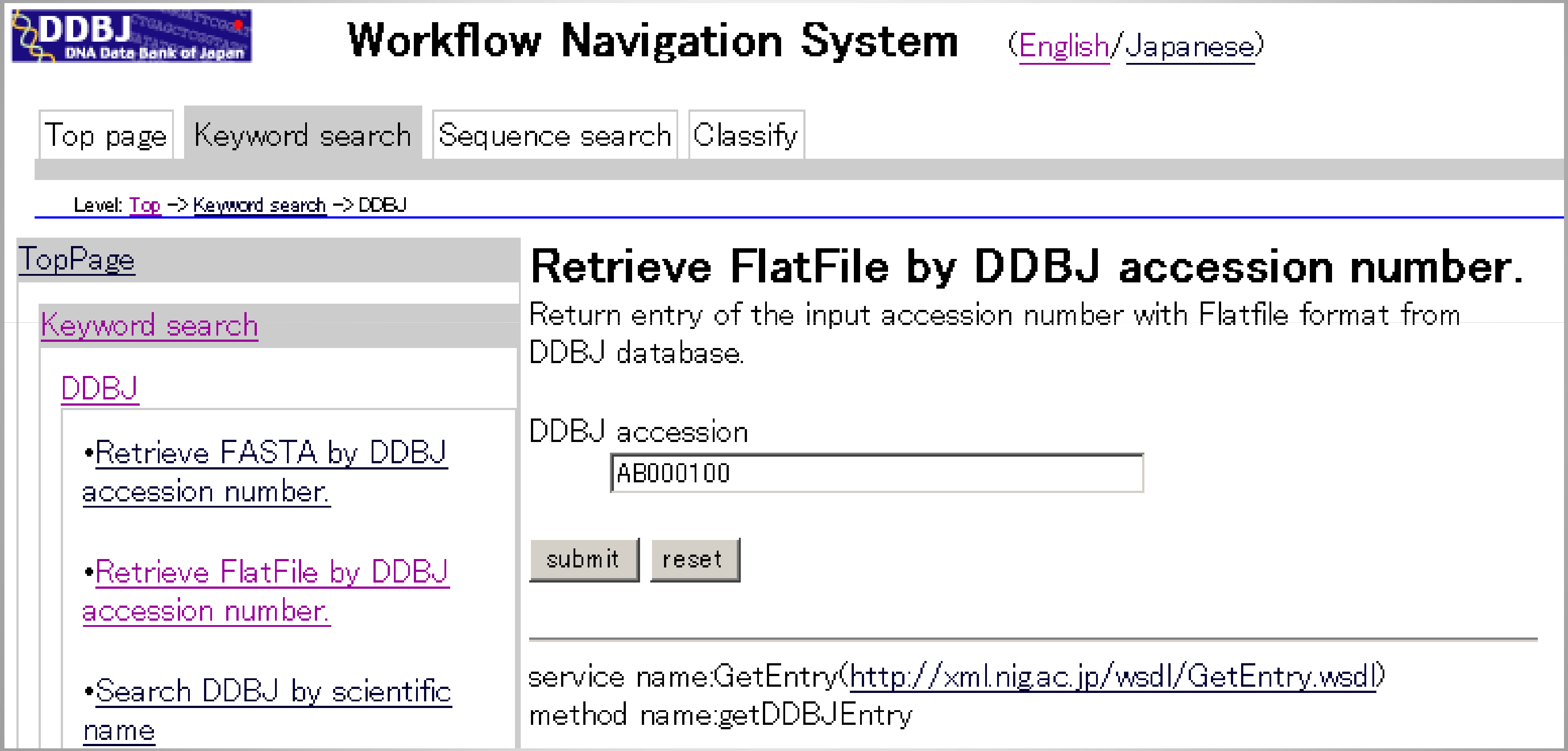
- Retrieve FASTA by DDBJ accession number.
  Return entry of the input accession number with FASTA format from DDBJ database.
- Retrieve FlatFile by DDBJ accession number.
  Return entry of the input accession number with Flatfile format from DDBJ database.
- Search DDBJ by scientific name
  Search DDBJ database with scientific name.
- Search DDBJ by specimen

Sub-categories

List of methods

Web interface created by WABI

The Web interface is mechanically generated by WABI



Workflow Navigation System (English/Japanese)

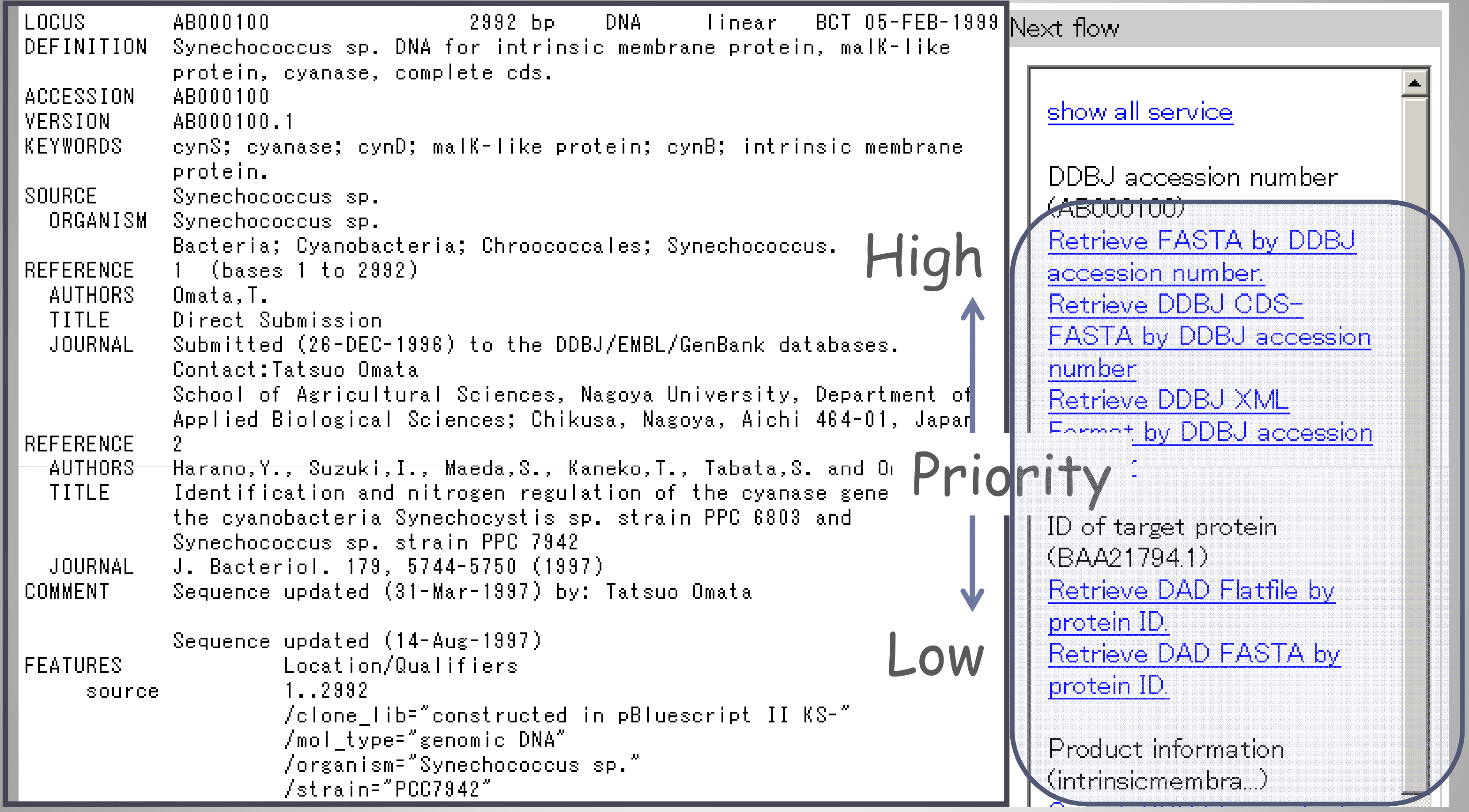
Retrieve FlatFile by DDBJ accession number.

Return entry of the input accession number with Flatfile format from DDBJ database.

DDBJ accession: AB000100

service name:GetEntry(http://xml.nig.ac.jp/wsdl/GetEntry.wsdl)
method name:getDDBJEntry

Interconnection of WABIs



```
LOCUS       AB000100               2992 bp    DNA     linear   BCT 05-FEB-1999
DEFINITION  Synechococcus sp. DNA for intrinsic membrane protein, malK-like
            protein, cyanase, complete cds.
ACCESSION   AB000100
VERSION     AB000100.1
KEYWORDS    cynS; cyanase; cynD; malK-like protein; cynB; intrinsic membrane
            protein.
SOURCE      Synechococcus sp.
  ORGANISM  Synechococcus sp.
            Bacteria; Cyanobacteria; Chroococcales; Synechococcus.
REFERENCE   1  (bases 1 to 2992)
  AUTHORS   Omata,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-DEC-1996) to the DDBJ/EMBL/GenBank databases.
            Contact:Tatsuo Omata
            School of Agricultural Sciences, Nagoya University, Department of
            Applied Biological Sciences; Chikusa, Nagoya, Aichi 464-01, Japan
REFERENCE   2
  AUTHORS   Harano,Y., Suzuki,I., Maeda,S., Kaneko,T., Tabata,S. and Oi
  TITLE     Identification and nitrogen regulation of the cyanase gene
            the cyanobacteria Synechocystis sp. strain PPC 6803 and
            Synechococcus sp. strain PPC 7942
  JOURNAL   J. Bacteriol. 179, 5744-5750 (1997)
COMMENT     Sequence updated (31-Mar-1997) by: Tatsuo Omata

            Sequence updated (14-Aug-1997)
FEATURES             Location/Qualifiers
     source          1..2992
                     /clone_lib="constructed in pBluescript II KS-"
                     /mol_type="genomic DNA"
                     /organism="Synechococcus sp."
                     /strain="PCC7942"
```

Next flow

show all service

DDBJ accession number
(AB000100)
Retrieve FASTA by DDBJ accession number.
Retrieve DDBJ CDS-FASTA by DDBJ accession number
Retrieve DDBJ XML Format by DDBJ accession :

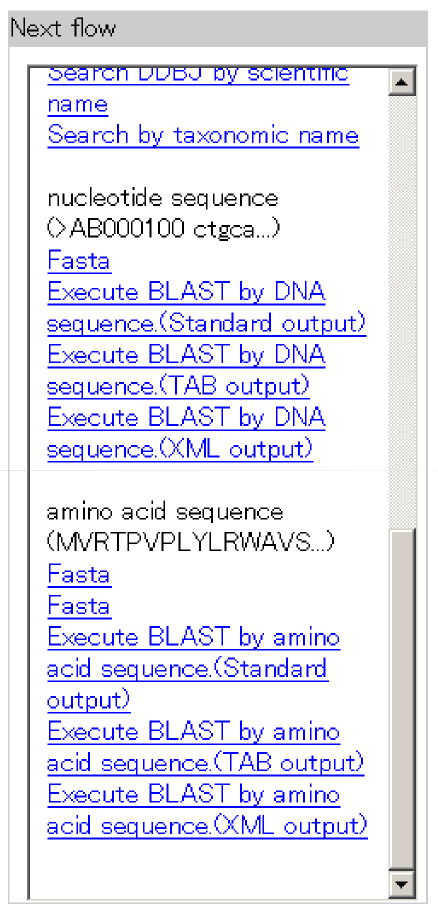
ID of target protein
(BAA21794.1)
Retrieve DAD Flatfile by protein ID.
Retrieve DAD FASTA by protein ID.

Product information
(intrinsicmembra...)

The result of the previous service

Services applicable to the result in the left

The result is mechanically given to the next service selected

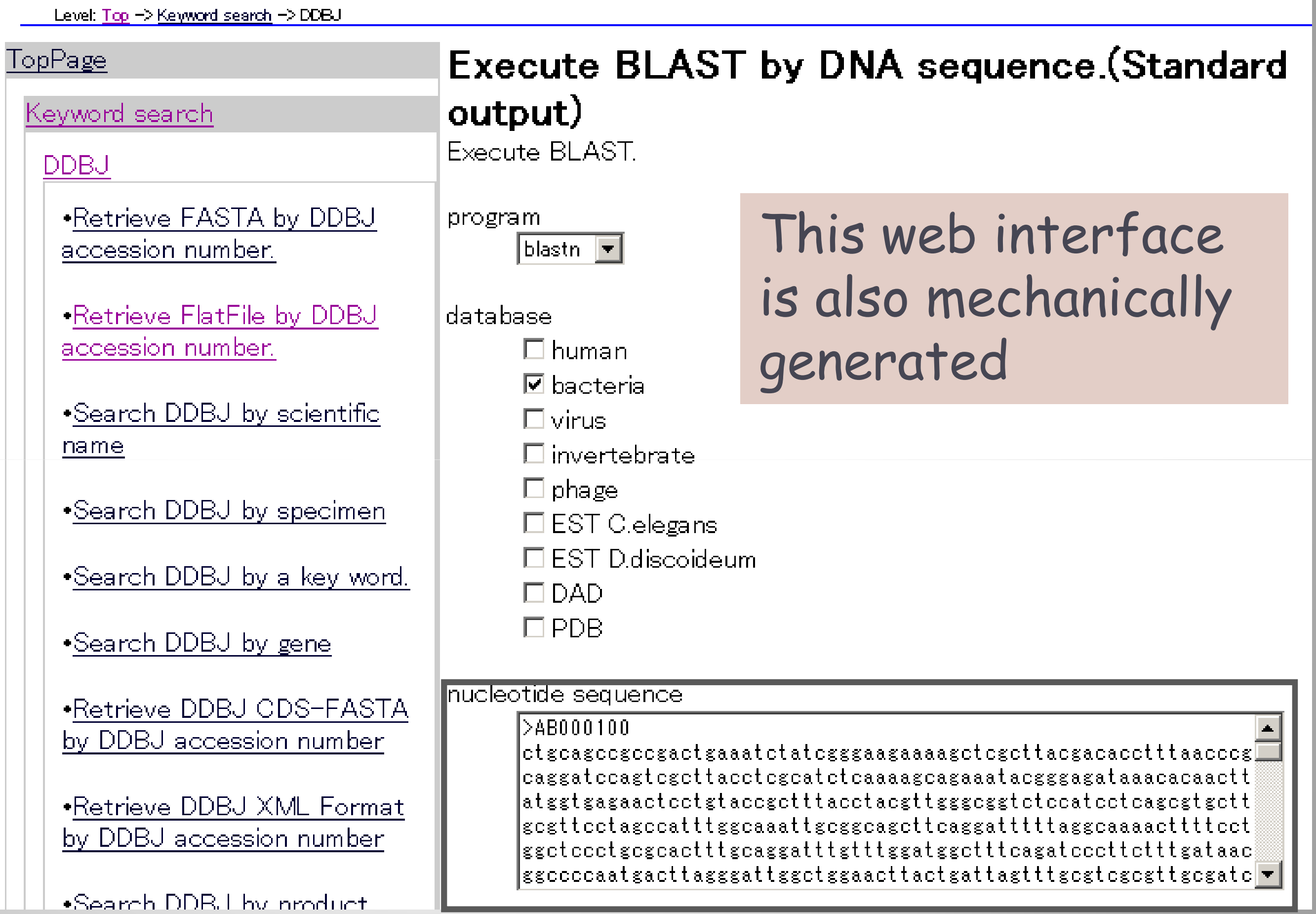


Next flow

Search DDBJ by scientific name
Search by taxonomic name

nucleotide sequence
(>AB000100 ctgca...)
Fasta
Execute BLAST by DNA sequence.(Standard output)
Execute BLAST by DNA sequence.(TAB output)
Execute BLAST by DNA sequence.(XML output)

amino acid sequence
(MVRTPVPLYLRWAVS...)
Fasta
Fasta
Execute BLAST by amino acid sequence.(Standard output)
Execute BLAST by amino acid sequence.(TAB output)
Execute BLAST by amino acid sequence.(XML output)



Level: Top -> Keyword search -> DDBJ

TopPage

Keyword search

DDBJ

- Retrieve FASTA by DDBJ accession number.
- Retrieve FlatFile by DDBJ accession number.
- Search DDBJ by scientific name
- Search DDBJ by specimen
- Search DDBJ by a key word.
- Search DDBJ by gene
- Retrieve DDBJ CDS-FASTA by DDBJ accession number
- Retrieve DDBJ XML Format by DDBJ accession number
- Search DDBJ by product

Execute BLAST by DNA sequence.(Standard output)

Execute BLAST.

program
blastn ▼

database
☐ human
☑ bacteria
☐ virus
☐ invertebrate
☐ phage
☐ EST C.elegans
☐ EST D.discoideum
☐ DAD
☐ PDB

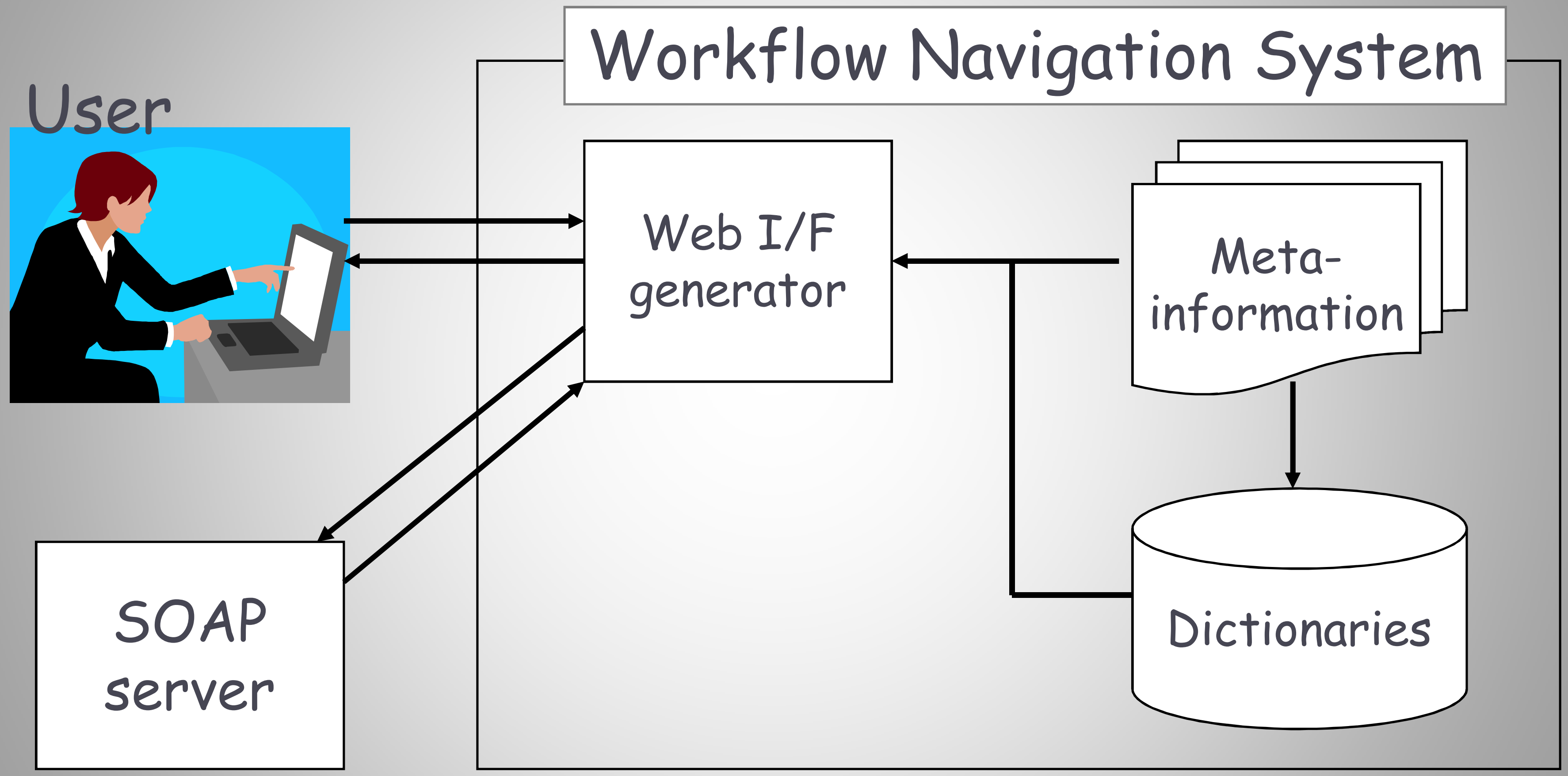
nucleotide sequence

```
>AB000100
ctgcagccgccgactgaaatctatcggaagaaaagctcgcttacgacacctttaacccg
caggatccagtcgcttacctcgcatctcaaaagcagaaatacgggagataaacacaactt
atggtgagaactcctgtaccgctttacctacgttgggcggtctccatcctcagcgtgctt
gcgttcctagccatttggcaaattgcggcagcttcaggattttaggcaaaacttttcct
ggctccctgcgcactttgcaggatttgtttggatggctttcagatcccttctttgataac
ggccccaatgacttaggattggctggaacttactgattagtttgcgtcgcgttgcgatc
```

This web interface is also mechanically generated

The result of the previous service

System architecture and components



User

Workflow Navigation System

Web I/F generator

Meta-information

Dictionaries

SOAP server

20

Three dictionaries (ontology)

- Service category
  - All the services are categorized based on the service category dictionary.
- Input/Output
  - Each service has several input and output types.
  - The input/output dictionary is used to connect services.
- HTML
  - HTML tags in the web interface are defined
  - The HTML dictionary is referred from the input/output dictionary.

Meta-information of a service

- Define (refer to) the category and input/output of service
- Define the web service information such as WSDL, URL and method name to call the SOAP with input parameters.
- Define description, initial value and example values which are displayed in the web interface.

Example of meta-information and dictionary

In the case of getFASTA_DDBJEntry(accession) method:

- Category: Keyword search -> DDBJ
  (refer to category dictionary)
- Input type: DDBJ accession (refer to in/out dictionary)
- Output type: FASTA format (refer to in/out dictionary)
- WSDL: http://xml.nig.ac.jp/wsdl/GetEntry.wsdl
- Other information such as description, initial value, …

Web interface is mechanically generated

23

“getentry” and “BLAST” is connected by “sequence”

getentry:getDDBJEntry

Input:
DDBJ Accession

Output:
DDBJ Flatfile

BLAST:searchSimple

Input:
sequence

Output:
BLAST Result

DDBJ accession
organism
sequence

DDBJ Flatfile has sequence defined in the input/output dictionary.

24

Service category dictionary

```
<?xml version="1.0" ?>
<dictionary>
   <category
   name_en="Keyword search“
   description_en="Keyword search service …
   />
   <category
   name_en="DDBJ"
   description_en="Search and retrieve of DDBJ database."
   >
      <parent name_jp=“Keyword search“ />
   </category>
```

Input/Output dictionary

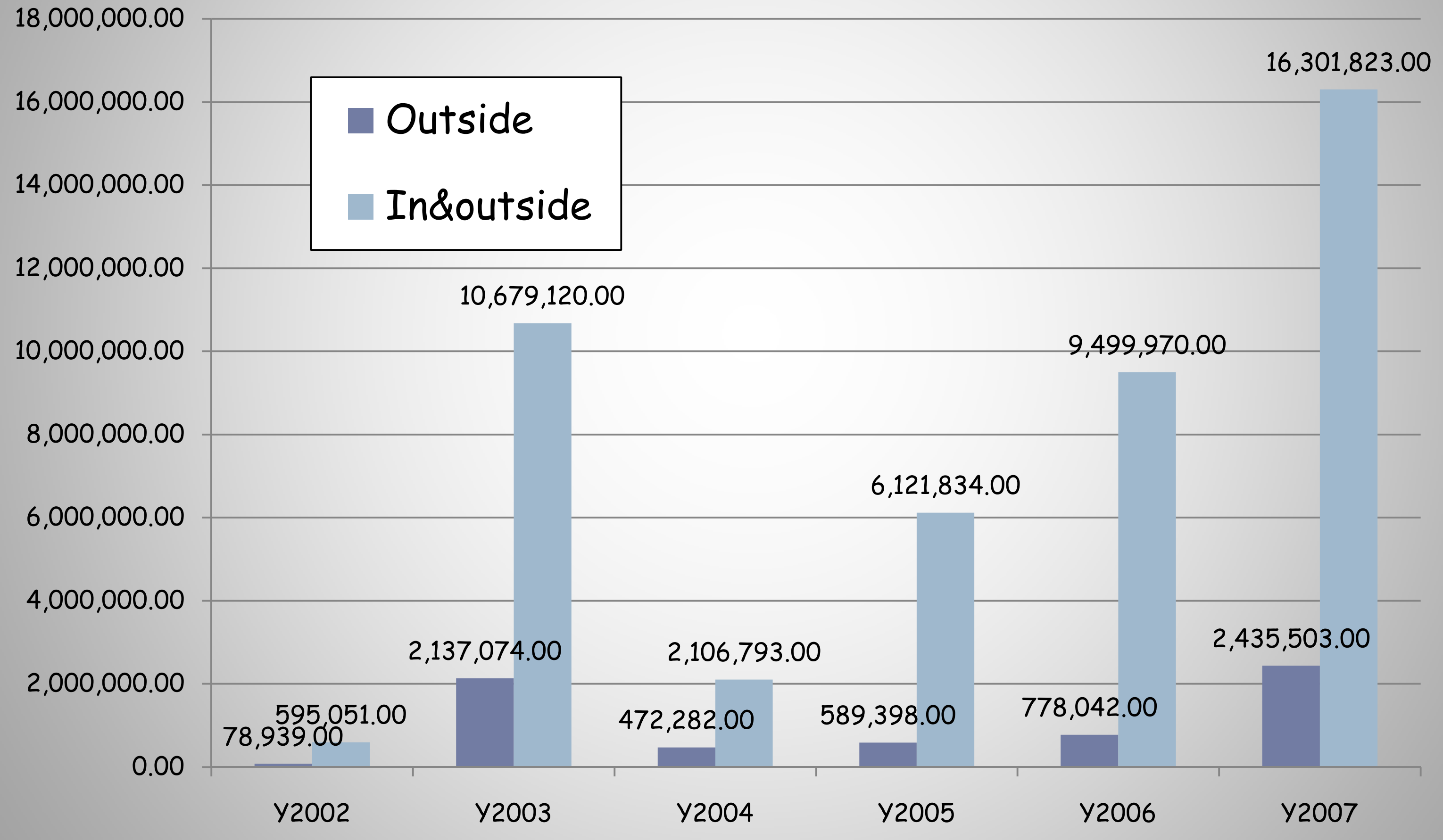
```
<?xml version="1.0" ?>
<dictionary>
    <content
    name=“DDBJ FlatFile“
    description_en=“DDBJ Flat File”>
        <parent name=“FlatFile” />
        <parser type=“LOCAL” class_name=“DDBJFFParser” />
        <data_list>
            <data name=“DDBJ Accession”
                    method=“getAccession” />
            <data name=“organism”
                    method=“getOrganismFromFF” />
            <data name=“nucleotide sequence” method=“…” />
            <data name=“gene” method=“getGeneFromFF” />
        </data_list>
    </content>
    …
```

HTML dictionary

```
<?xml version="1.0" ?>
<dictionary>
    <content
    name="text"
    description_en="The display type of text is
          displayed with HTML tag input type and text">
       <tag_list>
          <tag name="text" />
       </tag_list>
    </content>
    …
```

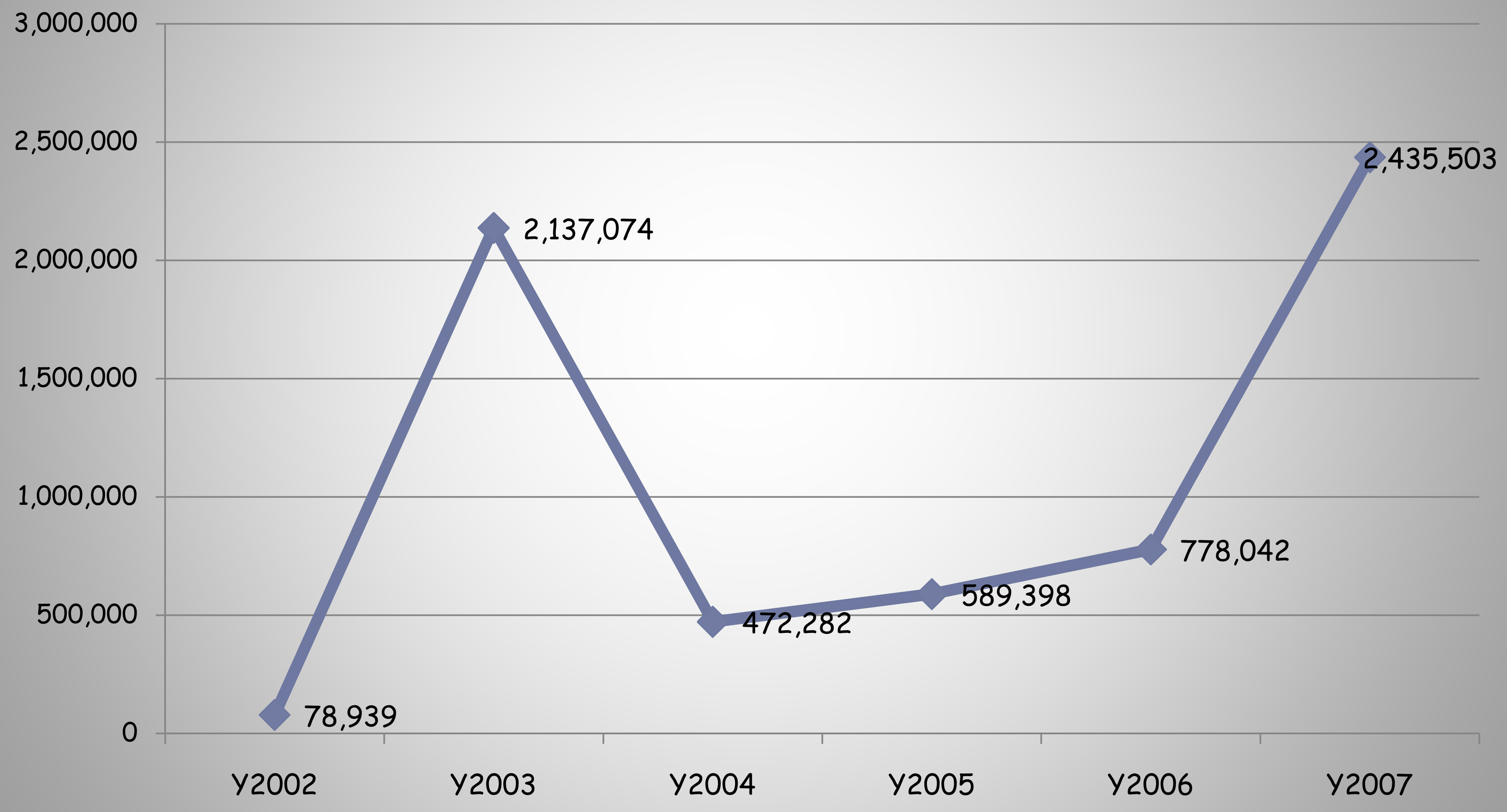
Beyond OTAKU

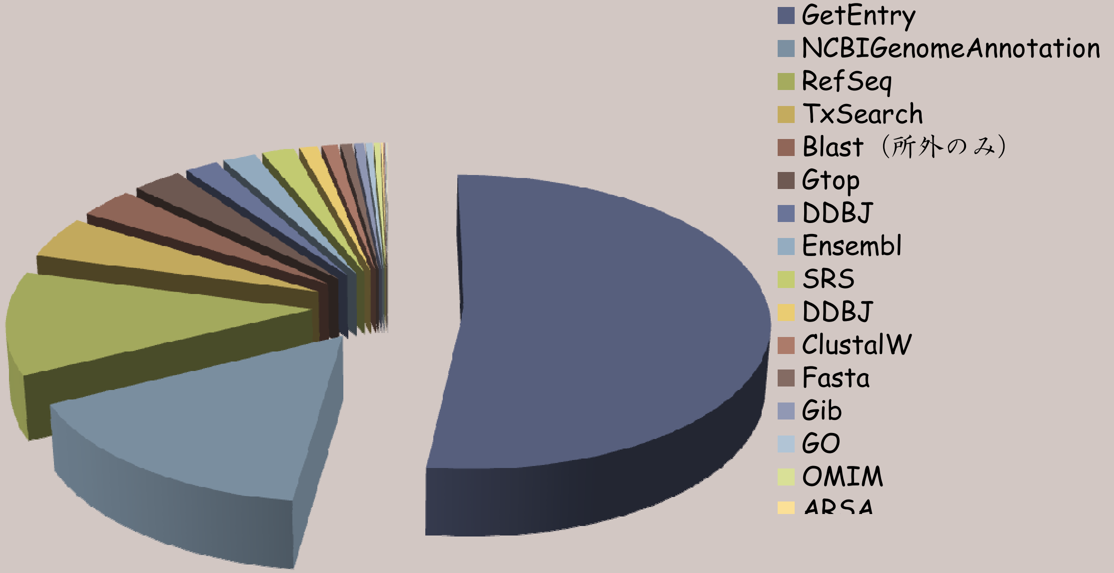
Usage in and outside NIG



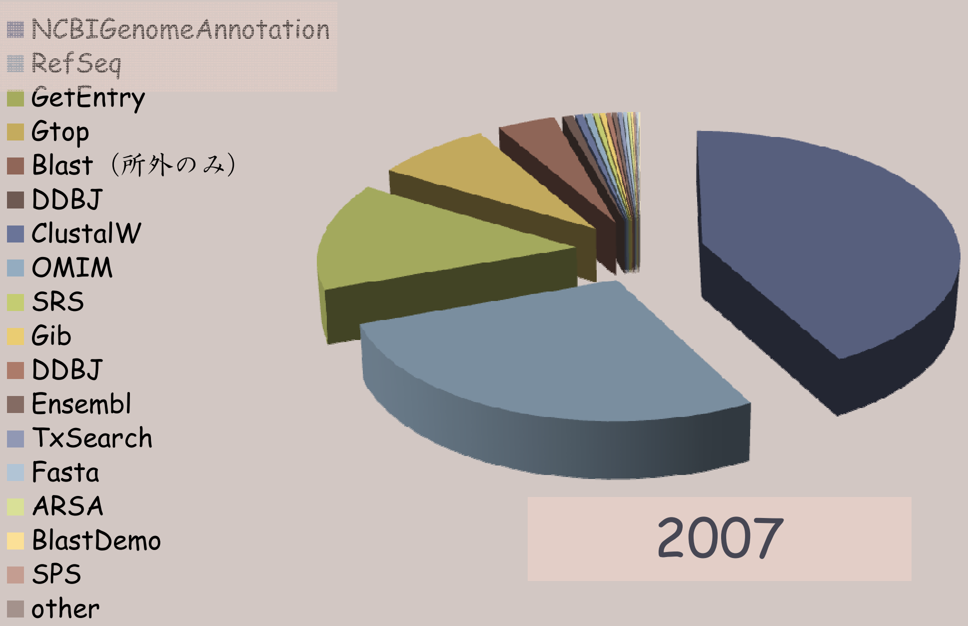
Beyond OTAKU

Usage outside NIG





2002~2007



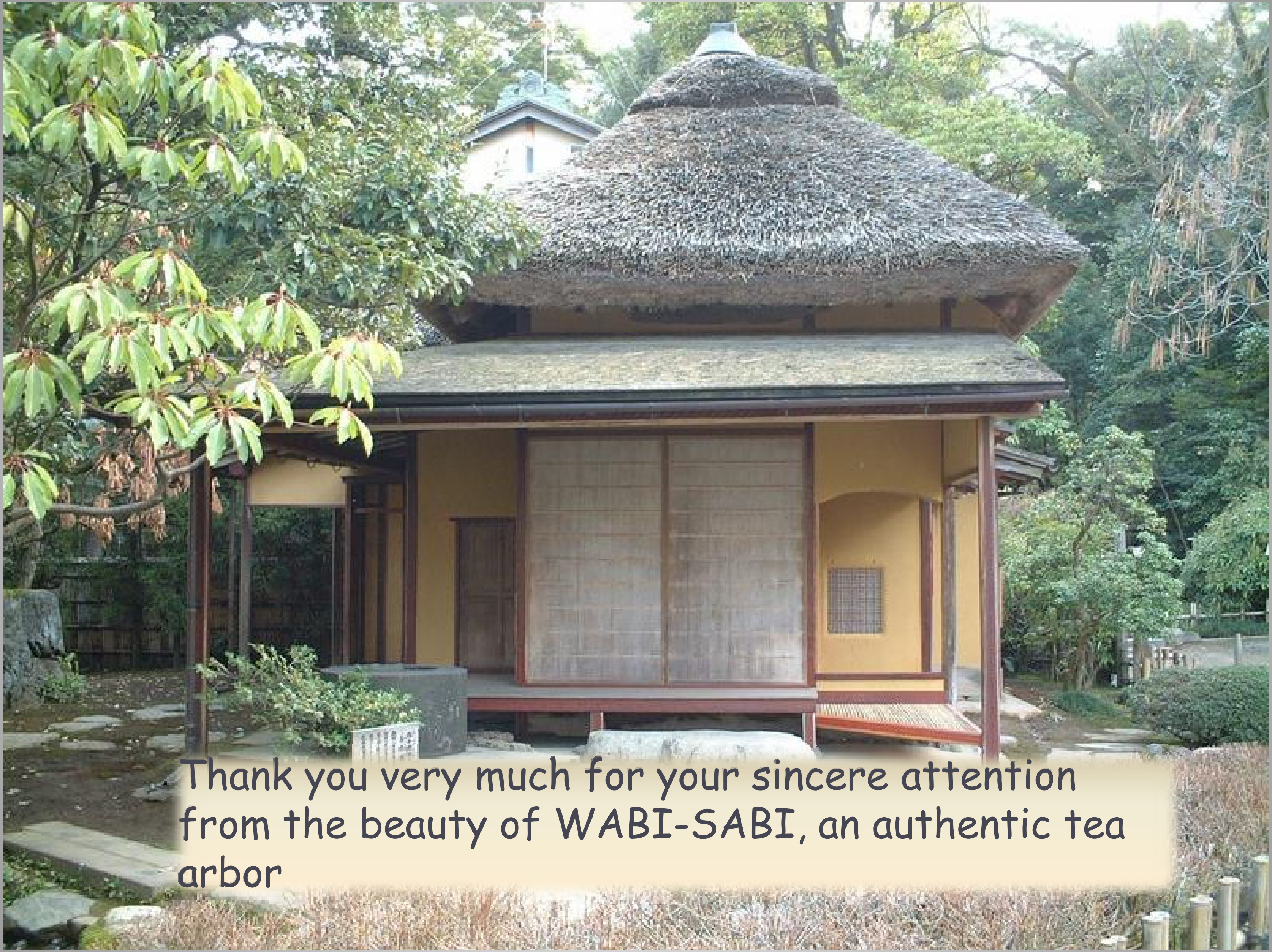
2007

From WABI to SABI

Plainly stated, the basic tenets of wabi-sabi are **simple is best, less is more**. Leonard Koren in his book "Wabi-Sabi: for Artists, Designers, Poets and Philosophers (Stonebridge Press, 1995) describes it as **a beauty----**

The common dictionaries and meta-information will be discussed in the BioHackathon. OTAKUs will follow up the discussion to make programs possible to find and interconnect Web APIs in 3 years, 5years or 10years.

SABI WILL STAND FOR SEMANTIC WEB API FOR BIOLOGY SOMETIME.



Thank you very much for your sincere attention from the beauty of WABI-SABI, an authentic tea arbor