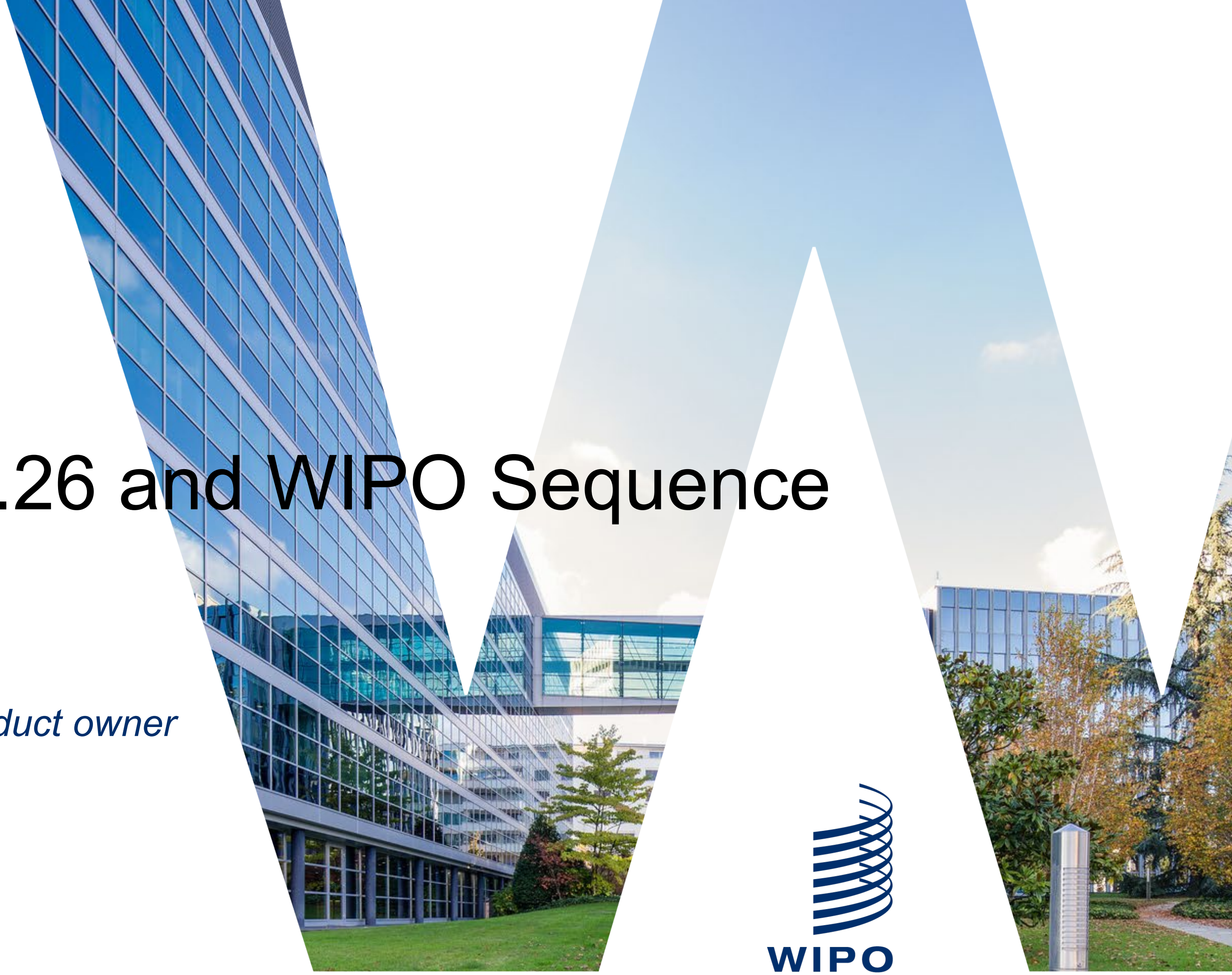


WIPO Standard ST.26 and WIPO Sequence

Basic Introduction

Emma Francis

IP Data Expert & WIPO Sequence product owner

WIPO Standards Team

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

A sequence listing…

- Contains nucleotide and/or amino acid sequences disclosed in a patent application and forms part of the description
- Includes descriptive information about each sequence known as annotations
- Allows for the sequence data of an invention to be searchable:
  - Inside an IP Office
  - In publicly available databases (INSDC databases)
- Conforms to the requirements of the relevant WIPO Standard (WIPO ST.25 or ST.26)

So why a new WIPO Standard?

Sequence listings were filed compliant with WIPO ST.25

However:

- WIPO ST.25 format is not compliant with INSDC requirements, so data is lost when entered into public databases
- WIPO ST.25 rules are not clear enough, and IP Offices worldwide interpret and enforce the rules differently
- Sequence types that are common today are not covered by WIPO ST.25 rules (nucleotide analogs, D-amino acids, branched sequences) and therefore are not present in searchable databases
- Data is unstructured – WIPO ST.25 format is difficult to use for automated validation and data exchange

WIPO Standard ST.26: Benefits

Acceptance of a single sequence listing worldwide

Standard ensures agreement amongst IP Offices

Enhanced submission quality

Data compatibility with INSDC database

More sequence data is searchable e.g., D-amino acids

Clarifies what sequence disclosures are required

Clarifies what sequence disclosures are permitted

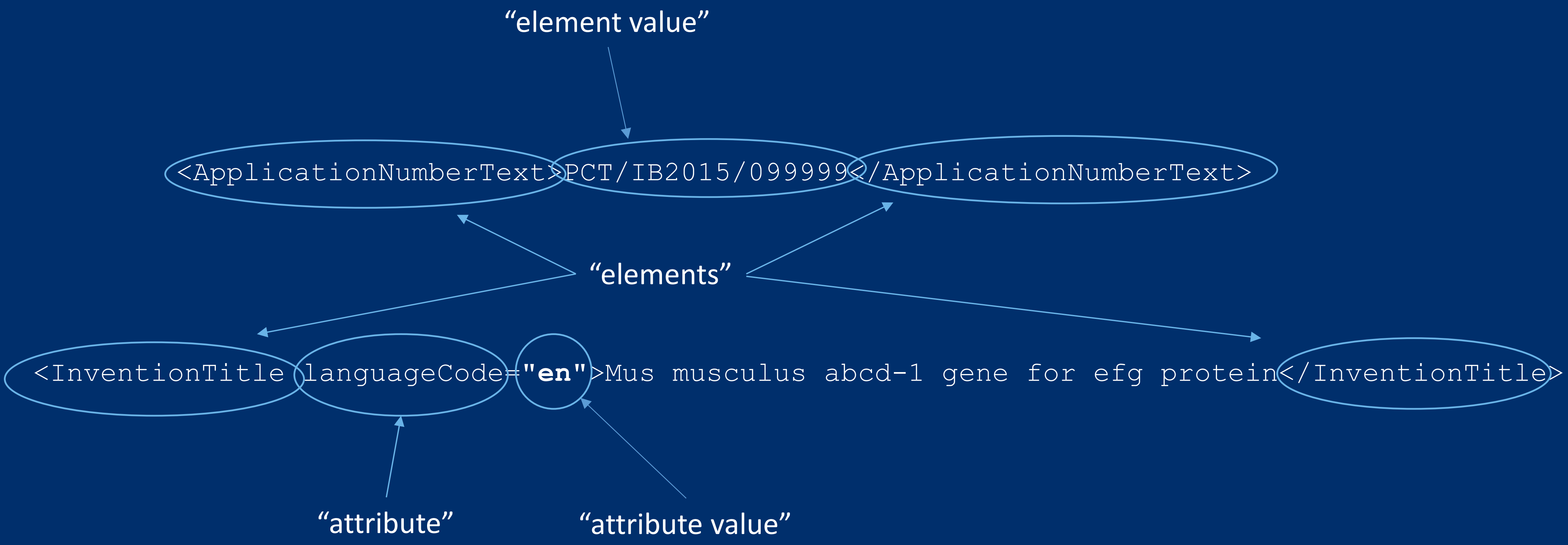
Standardization of feature locations, annotations and qualifiers

WIPO Standard ST.26: Contents

- **Main Body** – Requirements for inclusion/representation
- **Annex I** – Controlled vocabulary based on INSDC
- **Annex II** – ST.26 Document Type DTD
- **Annex III** – Example ST.26 Sequence Listing XML file
- **Annex IV** – Character Subset of Basic Latin Code for ST.26 XML Instance
- **Annex V** – INDS Data Exchange Requirements (IPOs only)
- **Annex VI** – Guidance Document with Examples
- **Appendix to Annex VI** –XML file including all of the sequence disclosures exemplified in Annex VI
- **Annex VII** – Recommendation for the Transformation of a Sequence Listing from ST.25 to ST.26

XML Basics

- XML = eXtensible Markup Language
- Information is ‘tagged’ using descriptive elements and attributes
- Standardized means of data exchange that is both human and machine-readable
- DTD = Document Type Definition - defines the structure and the legal elements and attributes of an XML document



“element value”

```
<ApplicationNumberText>PCT/IB2015/099999</ApplicationNumberText>
```

“elements”

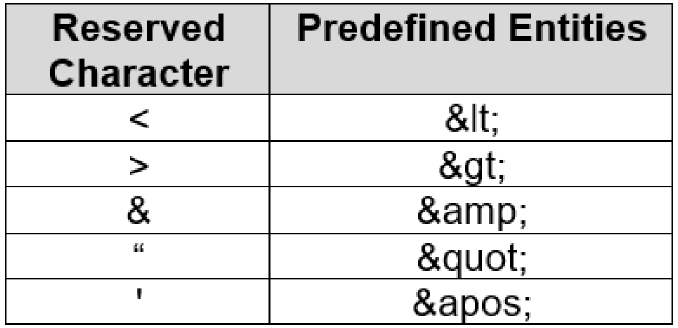
```
<InventionTitle languageCode="en">Mus musculus abcd-1 gene for efg protein</InventionTitle>
```

“attribute”

“attribute value”

XML basics: reserved characters

…must be replaced by their predefined entities in an element value.



Reserved Character	Predefined Entities
<	<
>	>
&	&
“	"
'	'

Example: feature location is “<50..62”:

× <INSDFeature_location><50..62</INSDFeature_location>

✓ <INSDFeature_location><50..62</INSDFeature_location>

WIPO ST.26: XML document

- Must be provided as one XML 1.0 format file
- Must validate against the WIPO ST.26 DTD (Annex II) and business rules derived from the content of the Standard
- Must be encoded using Unicode UTF-8
- Structure of the ST.26 sequence listing:
  - XML declaration:

```
<?xml version="1.0" encoding="UTF-8"?>
```

  - Document type (DOCTYPE) declaration:

```
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD
Sequence Listing 1.3//EN" "ST26SequenceListing_V1_3.dtd">
```

  - Root element
    - General information part
    - Sequence data part

XML declaration

```
<?xml version="1.0" encoding="UTF-8"?>
```

DOCTYPE declaration

```
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing 1.3//EN"
"ST26SequenceListing V1 3.dtd">
```

Root element

```
<ST26SequenceListing dtdVersion="V1_3" fileName="st26-annex-iii-sequence-listing-specimen.xml"
softwareName="WIPO Sequence" softwareVersion="1.2.0" productionDate="2022-01-07"
originalFreeTextLanguageCode="ja" nonEnglishFreeTextLanguageCode="de">
```

General information

```
        <ApplicationIdentification>
                <IPOfficeCode>IB</IPOfficeCode>
                <ApplicationNumberText>PCT/IB2015/099999</ApplicationNumberText>
                <FilingDate>2015-01-31</FilingDate>
        </ApplicationIdentification>
        <ApplicantFileReference>AB123</ApplicantFileReference>
        <EarliestPriorityApplicationIdentification>
                <IPOfficeCode>IB</IPOfficeCode>
                <ApplicationNumberText>PCT/IB2014/111111</ApplicationNumberText>
                <FilingDate>2014-01-30</FilingDate>
        </EarliestPriorityApplicationIdentification>
        <ApplicantName languageCode="ja">出願製薬株式会社</ApplicantName>
        <ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>
        <InventorName languageCode="ja">特許 太郎</InventorName>
        <InventorNameLatin>Taro Tokkyo</InventorNameLatin>
        <InventionTitle languageCode="ja">efgタンパク質をコードするマウスabcd-1遺伝子</InventionTitle>
        <InventionTitle languageCode="en">Mus musculus abcd-1 gene for efg protein</InventionTitle>
        <SequenceTotalQuantity>11</SequenceTotalQuantity>
```

Sequence data

```
        <SequenceData sequenceIDNumber="1">
                <INSDSeq>
                        <INSDSeq_length>133</INSDSeq_length>
                        <INSDSeq_moltype>DNA</INSDSeq_moltype>
                        <INSDSeq_division>PAT</INSDSeq_division>
                                                …
```

WIPO ST.26 General Information (1)

- The “**Application Identification**” section:

- application number, filing date, and IP office code are mandatory if known

- otherwise, just the applicant file reference is sufficient

- The “**Priority Application**” section

- only one priority application can be included in the sequence listing, and it must be the <u>earliest</u> priority application;

- mandatory where priority is claimed

- The “**Applicant and Inventor Name**” section
  - only one applicant name and one inventor name may be included in the sequence listing, and they must be the “primary” applicant and inventor;

- however multiple applicants/inventors can be saved in a project;
- applicant name is mandatory; inventor name is optional;
- a language code for applicant and inventor names is mandatory;
- if the applicant and/or inventor name contains non-Unicode Basic Latin characters, then a transliteration or translation into Basic Latin characters must be included

WIPO ST.26 General Information (2)

- The “**Invention Title**” section
  - at least one invention title in the language of filing is mandatory;
  - additional titles in other languages may be included;
  - a language code is mandatory for each title;
- The “**Sequence Total Quantity**” element
  - mandatory;
  - the total must include skipped sequences

WIPO ST.26 sequences… what must be included?

For nucleotide sequences:

- 10 or more “specifically defined” and “enumerated” residues
- Include sequences with nucleotide analogs such as peptide nuclei acids (PNAs and glycol nucleic acids (GNAs)

For amino acid sequences:

- 4 or more “specifically defined” and “enumerated” residues
- Include sequences with D-amino acids
- Linear regions of branches sequences are required to be included in a sequence listing

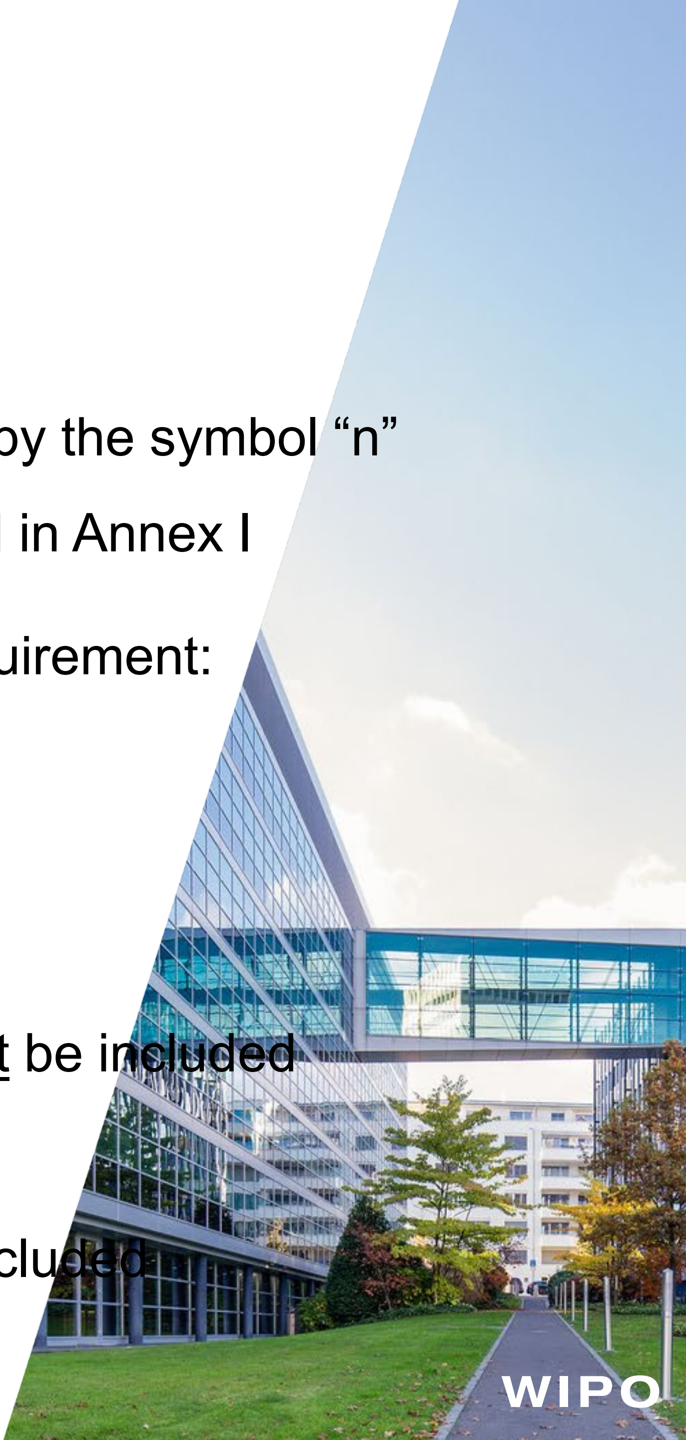
What is specifically defined?

What is a “specifically defined” nucleotide or amino acid?

- “specifically defined” means any nucleotide other than those represented by the symbol “n” and any amino acid other than those represented by the symbol “X”, listed in Annex I
- only “specifically defined” residues count towards the minimum length requirement:
  - 10 or more specifically defined nucleotides; or,
  - 4 or more specifically defined amino acids

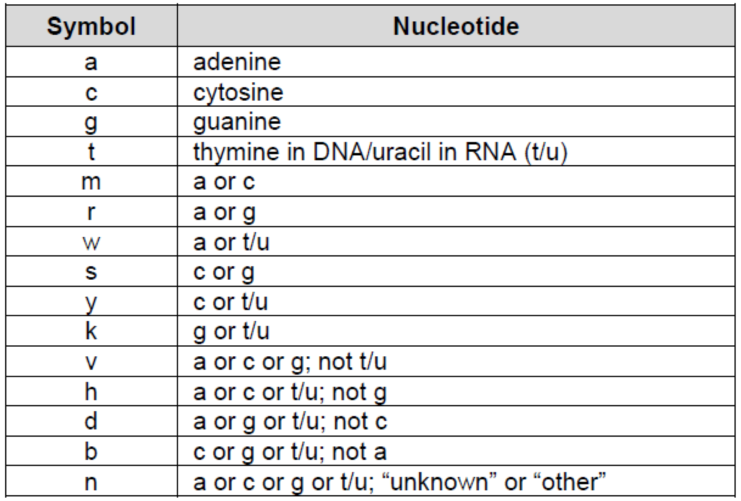
× 5’- a**n**ctggcaa**n** – 3’ only 8 specifically defined nucleotides; must not be included

✓ 5’- agctggcaat – 3’ ten specifically defined nucleotides; must be included



WIPO ST.26 – nucleotides & amino acids

- Nucleotide sequences:
  - all lower-case symbols;
  - no spaces, no numbering;
  - no “u” symbols; “t” represents uracil in RNA
  - “n” has a default value of “any one of ‘a’, ‘c’, ‘g’, or ‘t/u’”
- Amino acid sequences:
  - all single letter, upper case symbols;
  - no spaces, no numbering;
  - “X” has a default value of “any one of ‘A’, ‘R’, ‘N’, ‘D’, ‘C’, ‘Q’, ‘E’, ‘G’, ‘H’, ‘I’, ‘L’, ‘K’, ‘M’, ‘F’, ‘P’, ‘O’, ‘S’, ‘U’, ‘T’, ‘W’, ‘Y’, or ‘V’”



Symbol	Nucleotide
a	adenine
c	cytosine
g	guanine
t	thymine in DNA/uracil in RNA (t/u)
m	a or c
r	a or g
w	a or t/u
s	c or g
y	c or t/u
k	g or t/u
v	a or c or g; not t/u
h	a or c or t/u; not g
d	a or g or t/u; not c
b	c or g or t/u; not a
n	a or c or g or t/u; “unknown” or “other”

WIPO ST.26 - Annotation

Feature Keys and Qualifiers

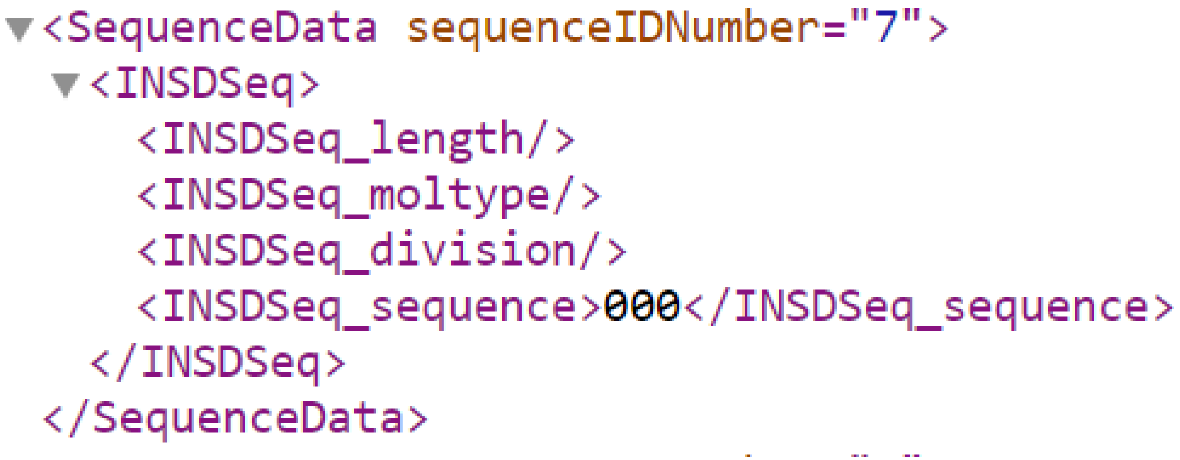
- Annotation in sequence listings comprises of a **feature key** at a particular **feature location** within the sequence length where additional details are provided using **qualifiers**
- In addition to the mandatory “source” feature, applicants can add multiple optional features to further describe the sequence:
  - different feature keys for nucleotide sequences and amino acid sequences;
  - each feature may have one or more optional qualifiers, and may have a mandatory qualifier
  - The full list is defined in Annex I

WIPO ST.26 Sequences – mandatory elements

- Every sequence must have:
  - a “**source”** feature with both a mandatory “**organism**” qualifier and **“mol_type”** qualifier, spanning the entire sequence
  - a unique SEQ ID NO. or Sequence ID Number
  - The molecule type (DNA, RNA or AA)
  - a series of residues or ‘000’ for a skipped sequence
- For an “organism” qualifier, the qualifier value can be:
  - Latin genus and species name e.g., “Mus musculus”
  - Genus name followed by “sp.” e.g., “Mus sp.”
  - Virus name e.g., “Torque teno virus 1”
  - “unidentified”
  - “synthetic construct”
  - Note: Common names, such as “mouse,” must **not** be used as the organism name. If desired, common names can be included in a note qualifier.

WIPO ST.26 – skipped sequences

- <u>Skipped sequences</u>: allow an applicant to delete sequence data from a sequence listing without the need to renumber subsequence sequences.
  - INSDSeq_length, INSDSeq_moltype, INSDSeq_division present, but with no value;
  - No feature table and no source feature;
  - Sequence element must have the value “000”



```xml
<SequenceData sequenceIDNumber="7">
  <INSDSeq>
    <INSDSeq_length/>
    <INSDSeq_moltype/>
    <INSDSeq_division/>
    <INSDSeq_sequence>000</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

WIPO Sequence: Introduction

- Desktop tool developed by WIPO to support authoring, validation, and generation of ST.26 compliant sequence listings
- Member states requested WIPO develop this common tool for all Offices and applicants at international, national and regional level
- Use of WIPO Sequence simplifies ST.26 XML creation with a user-friendly interface: no need to ever directly edit an XML file
- Beta versions released in 2020 and 2021 and production release 2.3.0

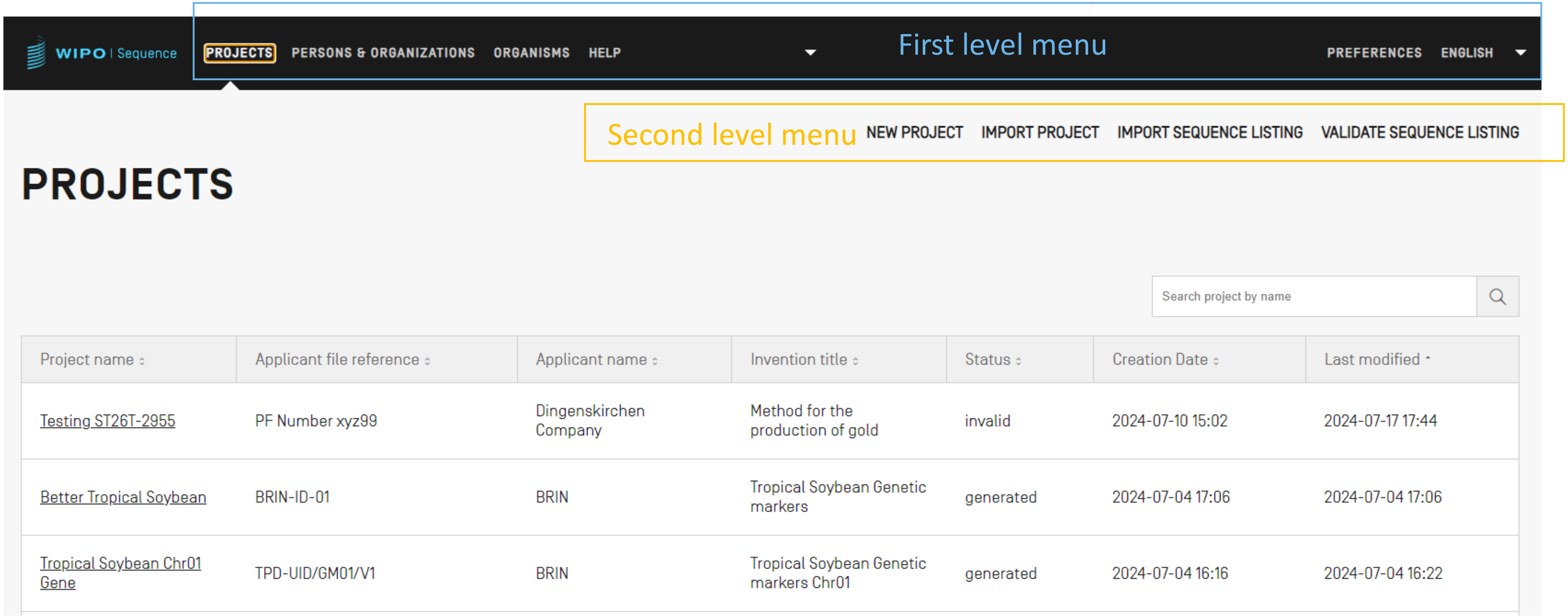
WIPO Sequence: Installation

- All installation packages and user support documentation can be download from the WIPO Sequence homepage at:
  - https://www.wipo.int/standards/en/sequence
- WIPO Sequence desktop tool is available for MS Windows, Mac OSX and Linux operation systems
- An auto-update functionality will prompt the user when a new version is available

!! WIPO Sequence: Important !!

- All generated sequence listings and project data entered into WIPO sequence are stored **locally** (on the user’s computer)
- Projects can be created and sequence listings generated while offline but the auto-update feature requires a connection to the internet
- WIPO Sequence is distributed for the free use of all applicants without the provision of the source code. A Terms of Use has recently been published on our website.

Project home



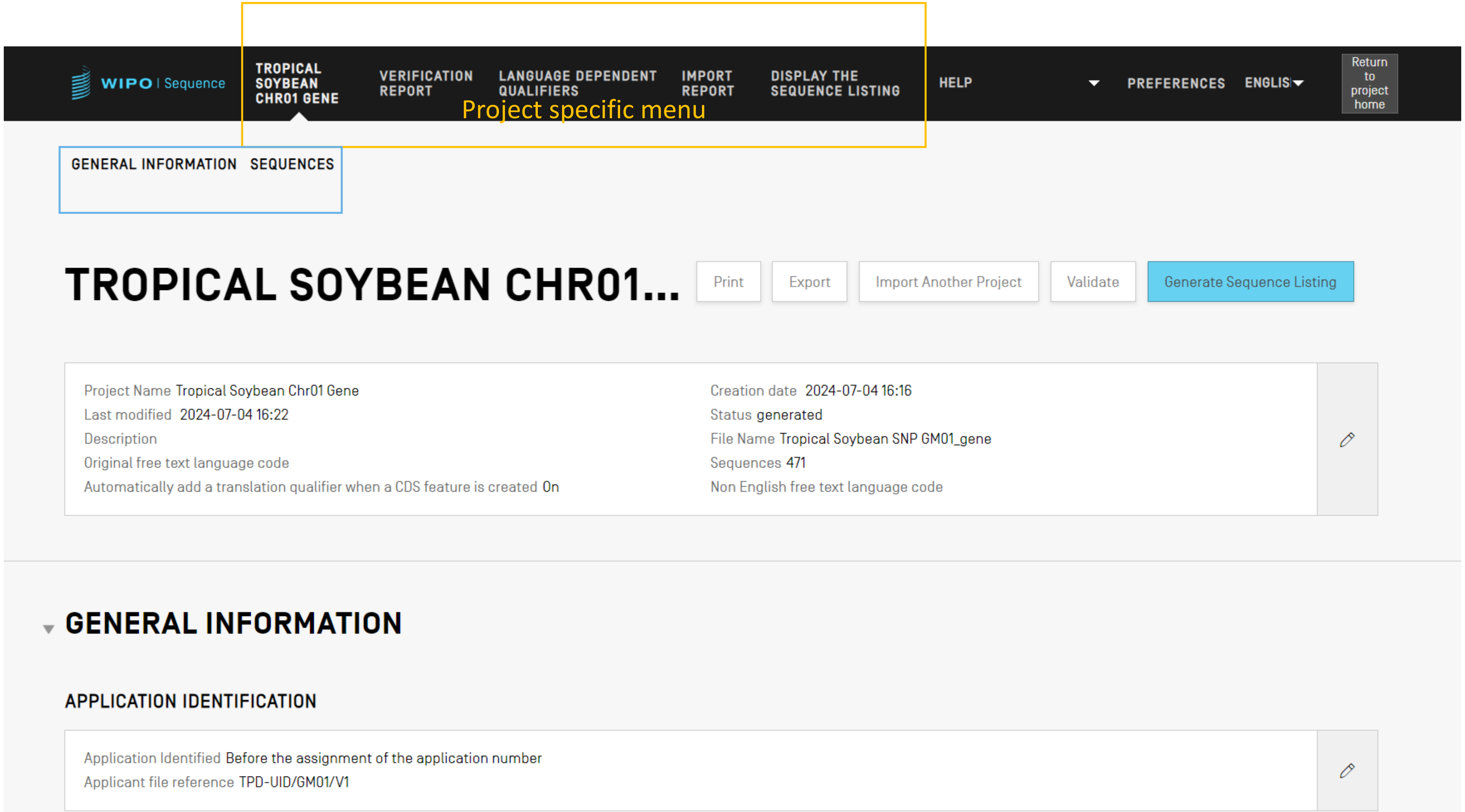
First level menu

Second level menu

PROJECTS

Project name	Applicant file reference	Applicant name	Invention title	Status	Creation Date	Last modified
Testing ST26T-2955	PF Number xyz99	Dingenskirchen Company	Method for the production of gold	invalid	2024-07-10 15:02	2024-07-17 17:44
Better Tropical Soybean	BRIN-ID-01	BRIN	Tropical Soybean Genetic markers	generated	2024-07-04 17:06	2024-07-04 17:06
Tropical Soybean Chr01 Gene	TPD-UID/GM01/V1	BRIN	Tropical Soybean Genetic markers Chr01	generated	2024-07-04 16:16	2024-07-04 16:22

Project detail



TROPICAL SOYBEAN CHR01...

Project Name Tropical Soybean Chr01 Gene
Last modified 2024-07-04 16:22
Description
Original free text language code
Automatically add a translation qualifier when a CDS feature is created On

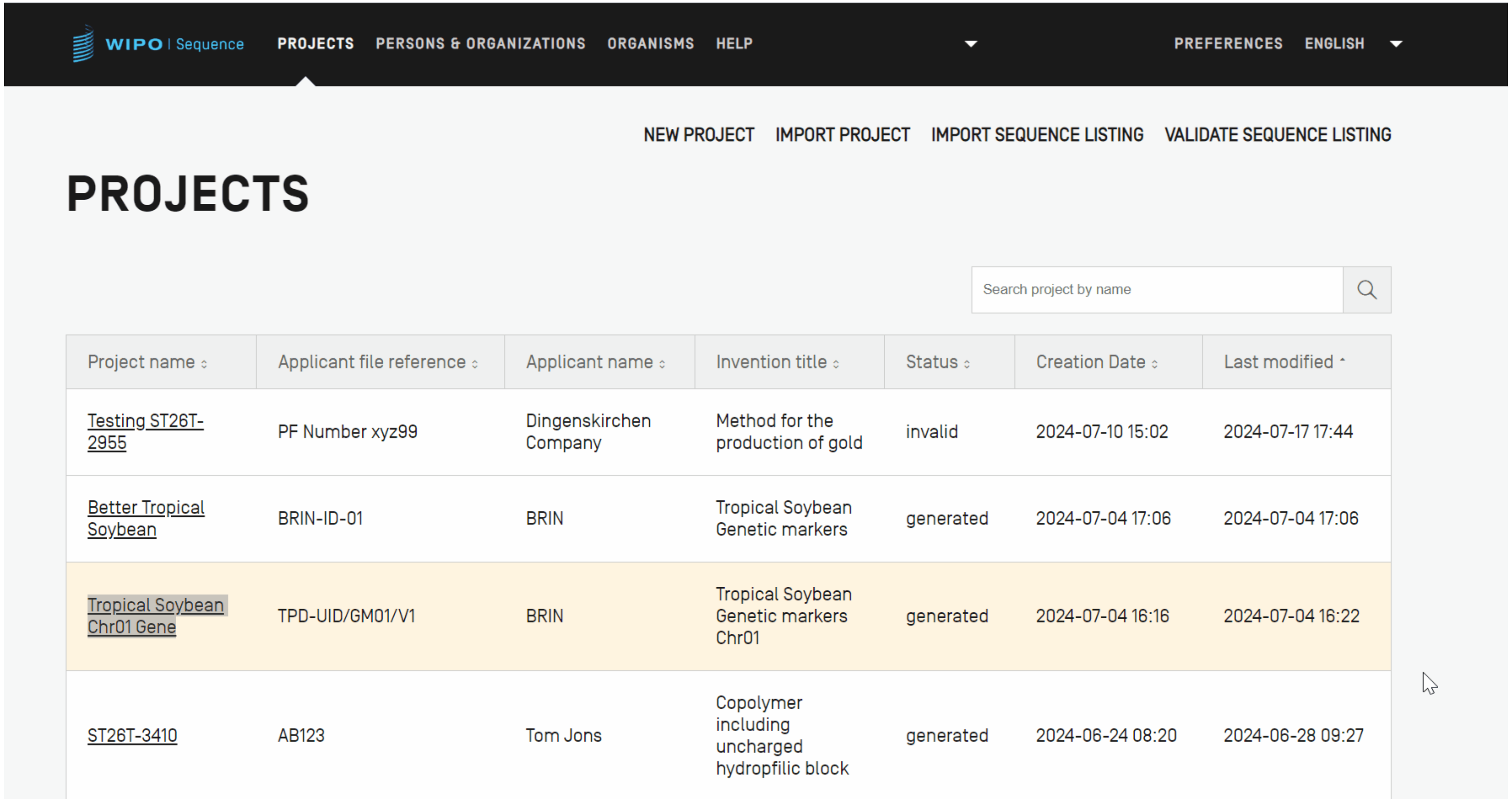
Creation date 2024-07-04 16:16
Status generated
File Name Tropical Soybean SNP GM01_gene
Sequences 471
Non English free text language code

GENERAL INFORMATION

APPLICATION IDENTIFICATION

Application Identified Before the assignment of the application number
Applicant file reference TPD-UID/GM01/V1

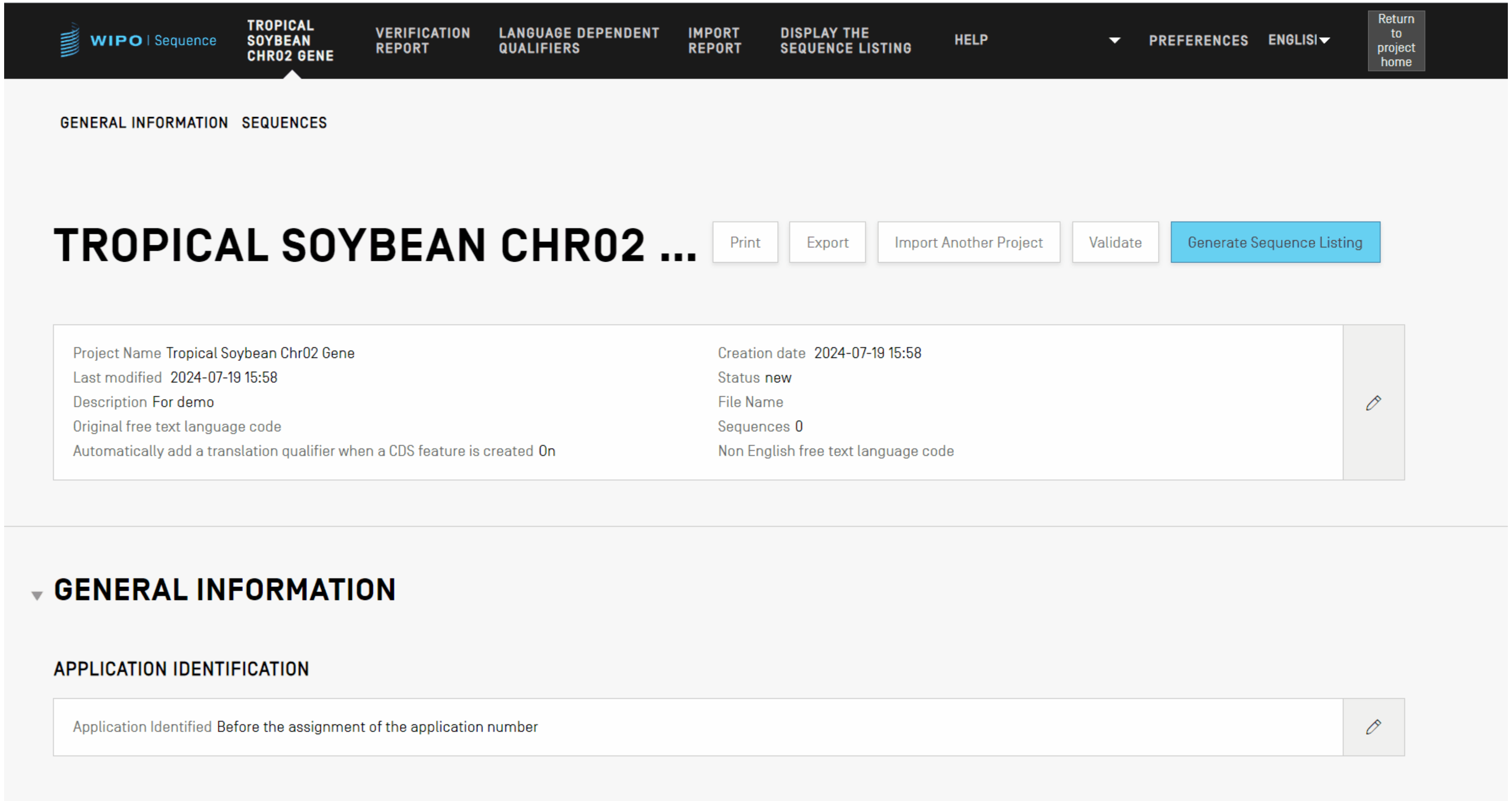
Getting started: Create a project



PROJECTS

Project name	Applicant file reference	Applicant name	Invention title	Status	Creation Date	Last modified
Testing ST26T-2955	PF Number xyz99	Dingenskirchen Company	Method for the production of gold	invalid	2024-07-10 15:02	2024-07-17 17:44
Better Tropical Soybean	BRIN-ID-01	BRIN	Tropical Soybean Genetic markers	generated	2024-07-04 17:06	2024-07-04 17:06
Tropical Soybean Chr01 Gene	TPD-UID/GM01/V1	BRIN	Tropical Soybean Genetic markers Chr01	generated	2024-07-04 16:16	2024-07-04 16:22
ST26T-3410	AB123	Tom Jons	Copolymer including uncharged hydropfilic block	generated	2024-06-24 08:20	2024-06-28 09:27

Provide project details: general information



TROPICAL SOYBEAN CHR02 ...

Project Name Tropical Soybean Chr02 Gene
Last modified 2024-07-19 15:58
Description For demo
Original free text language code
Automatically add a translation qualifier when a CDS feature is created On

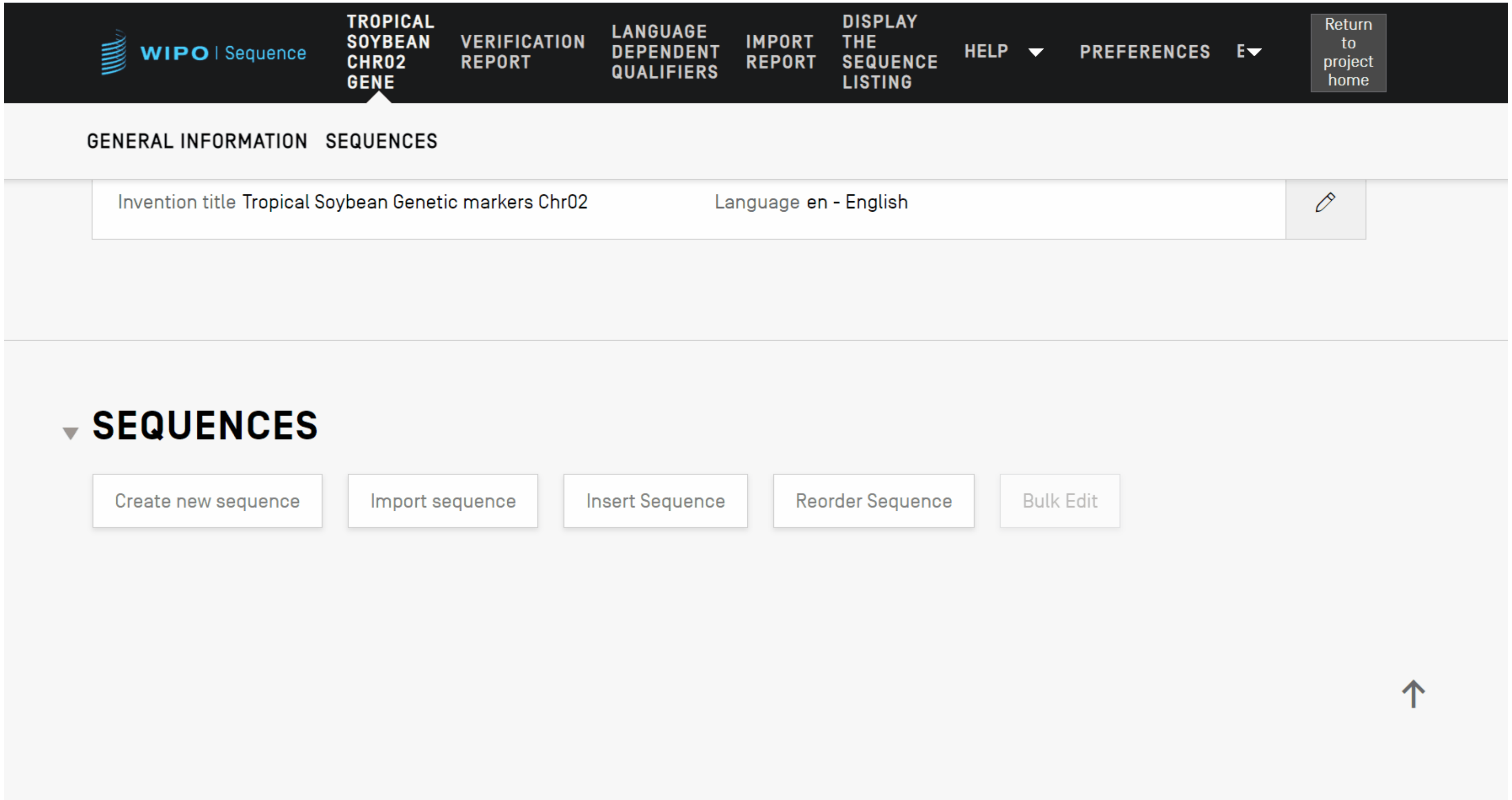
Creation date 2024-07-19 15:58
Status new
File Name
Sequences 0
Non English free text language code

GENERAL INFORMATION

APPLICATION IDENTIFICATION

Application Identified Before the assignment of the application number

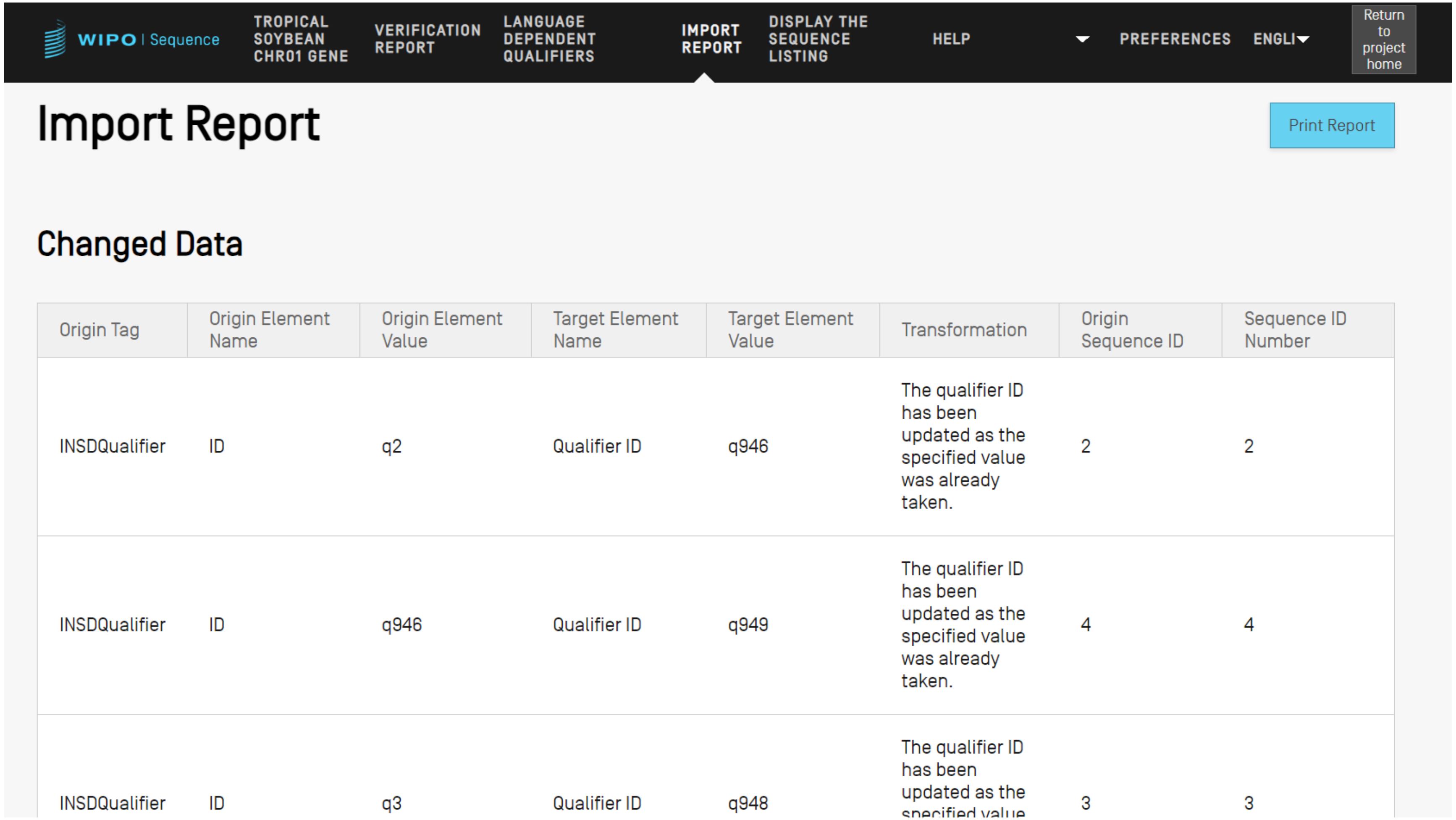
Import a sequence data into a project



Invention title Tropical Soybean Genetic markers Chr02 Language en - English

SEQUENCES

Import report/change data report

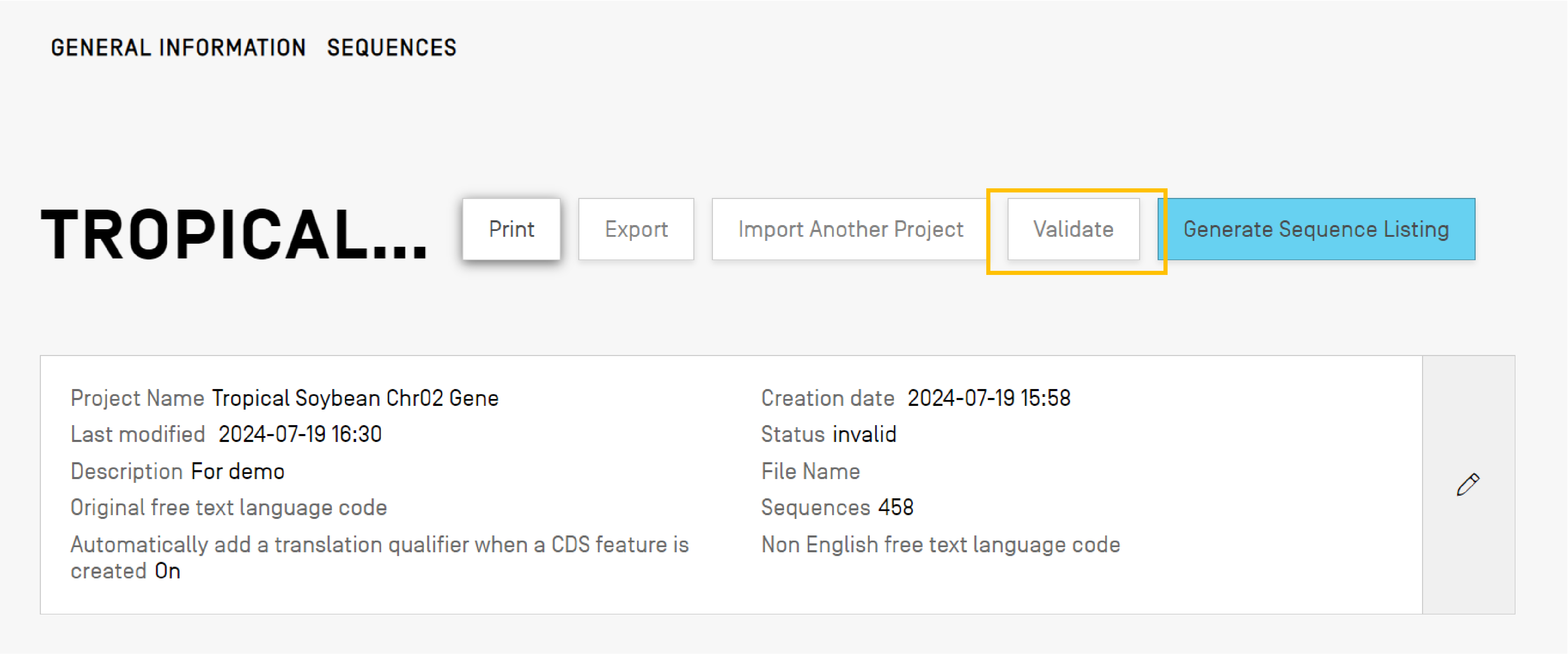


Import Report

Changed Data

Origin Tag	Origin Element Name	Origin Element Value	Target Element Name	Target Element Value	Transformation	Origin Sequence ID	Sequence ID Number
INSDQualifier	ID	q2	Qualifier ID	q946	The qualifier ID has been updated as the specified value was already taken.	2	2
INSDQualifier	ID	q946	Qualifier ID	q949	The qualifier ID has been updated as the specified value was already taken.	4	4
INSDQualifier	ID	q3	Qualifier ID	q948	The qualifier ID has been updated as the specified value	3	3

Validate sequence listing



Project Name Tropical Soybean Chr02 Gene

Last modified 2024-07-19 16:30

Description For demo

Original free text language code

Automatically add a translation qualifier when a CDS feature is created On

Creation date 2024-07-19 15:58

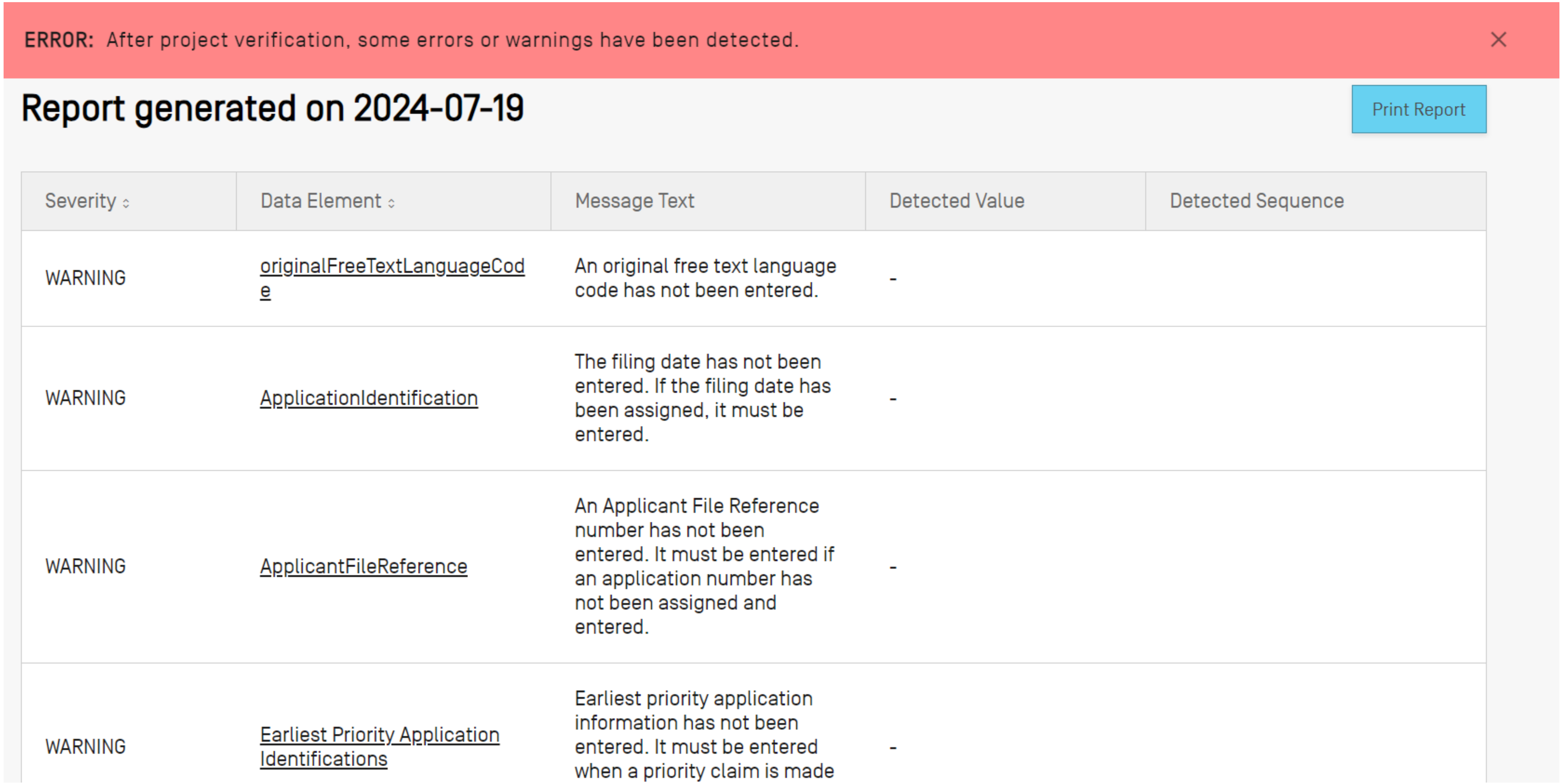
Status invalid

File Name

Sequences 458

Non English free text language code

Verification report: errors and warnings



ERROR: After project verification, some errors or warnings have been detected.

Report generated on 2024-07-19

Severity ⇕	Data Element ⇕	Message Text	Detected Value	Detected Sequence
WARNING	originalFreeTextLanguageCode	An original free text language code has not been entered.	-	
WARNING	ApplicationIdentification	The filing date has not been entered. If the filing date has been assigned, it must be entered.	-	
WARNING	ApplicantFileReference	An Applicant File Reference number has not been entered. It must be entered if an application number has not been assigned and entered.	-	
WARNING	Earliest Priority Application Identifications	Earliest priority application information has not been entered. It must be entered when a priority claim is made	-	

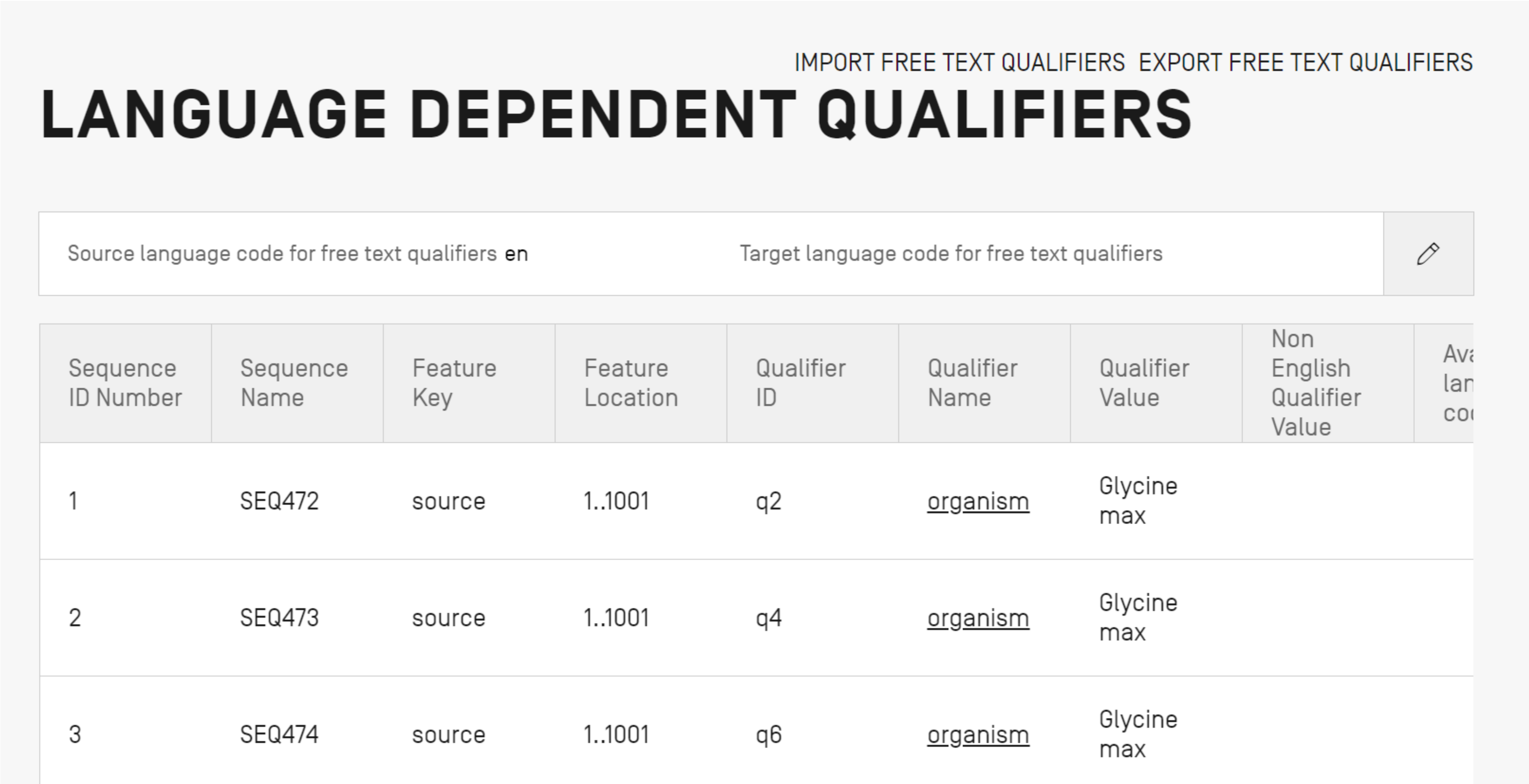
Errors and warnings

- WARNING: can be ignored but <u>should</u> be addressed by manual review
- ERROR: <u>must</u> be addressed by user
- Link in verification report to specific component within the sequence which was the cause of the error/warning
- Project must be validated again for this error to be removed from the report
- The verification report can be printed for later reference

Language dependent free text qualifiers

- Free text: "*is a type of value format for certain qualifiers, presented in the form of a descriptive text phrase or other specified format*"- ref: WIPO ST.26
- One of two types:
    - Language dependent (may require a translation) e.g., note
    - Language independent e.g., allele
- Must not exceed 1000 characters
- See WIPO ST.26 Annex I Sections 6 and 8: complete list of language dependent free text qualifiers

Import/export XLIFF



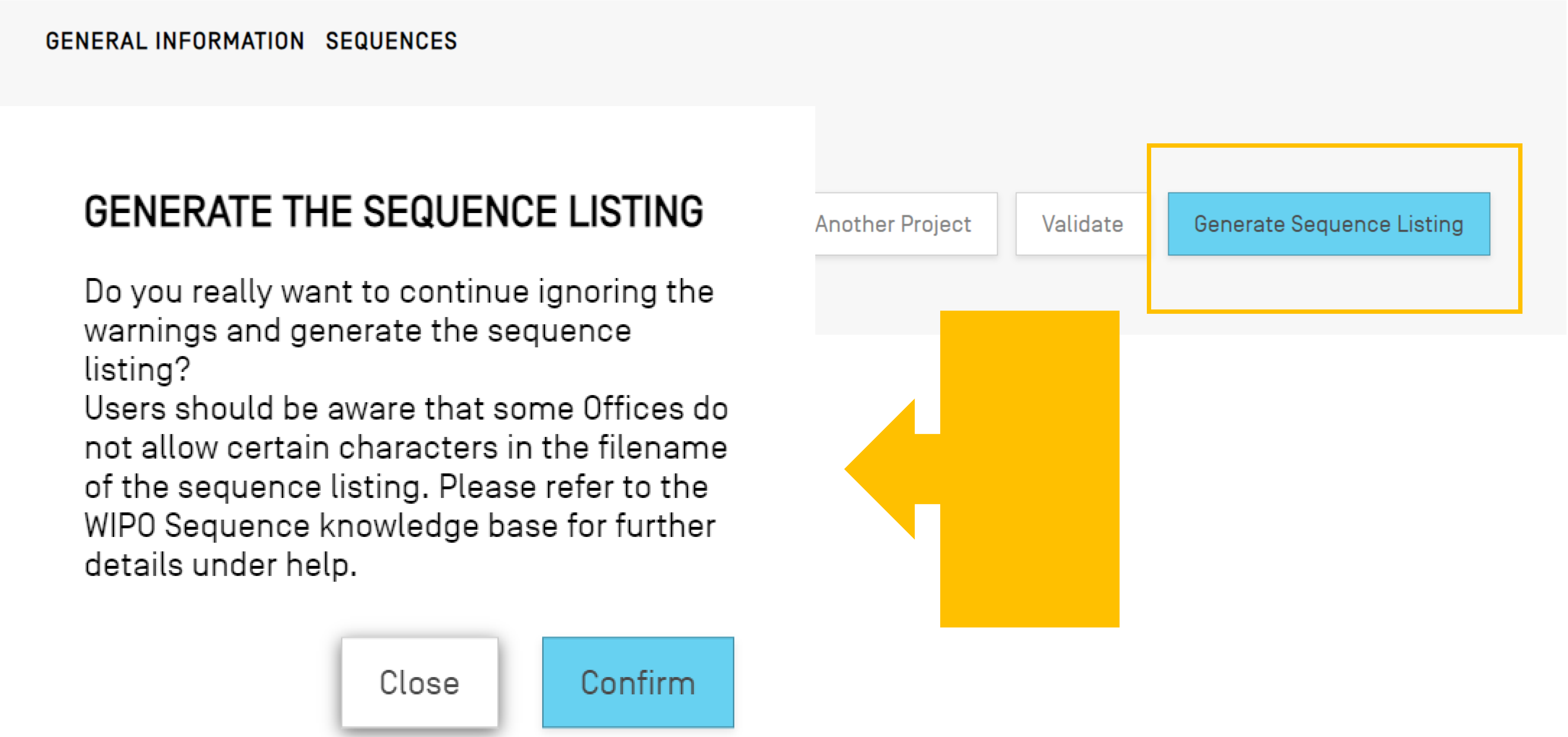
LANGUAGE DEPENDENT QUALIFIERS

Source language code for free text qualifiers en

Target language code for free text qualifiers

Sequence ID Number	Sequence Name	Feature Key	Feature Location	Qualifier ID	Qualifier Name	Qualifier Value	Non English Qualifier Value	Ava lan co
1	SEQ472	source	1..1001	q2	organism	Glycine max		
2	SEQ473	source	1..1001	q4	organism	Glycine max		
3	SEQ474	source	1..1001	q6	organism	Glycine max		

Generate sequence listing



GENERATE THE SEQUENCE LISTING

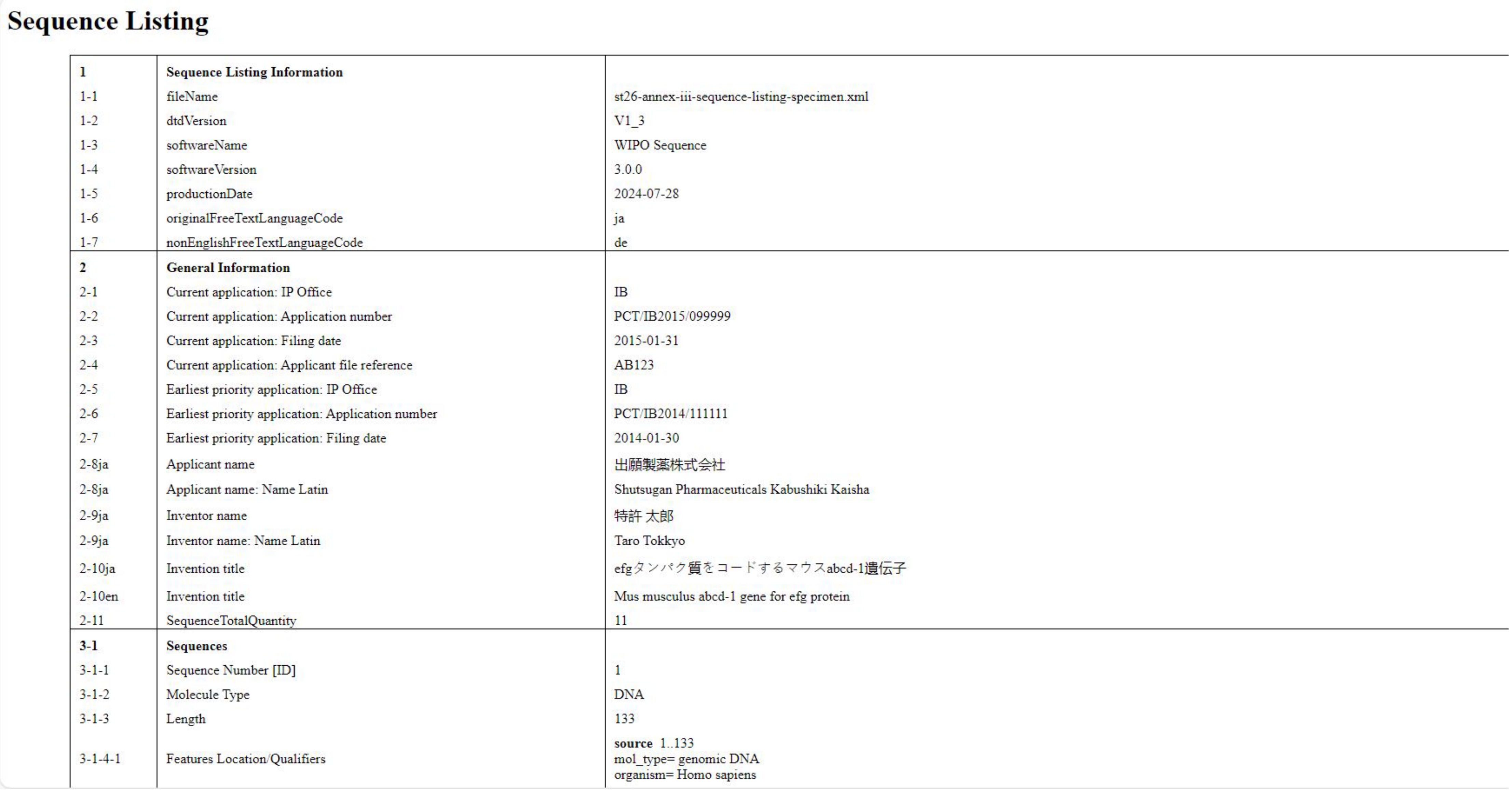
Do you really want to continue ignoring the warnings and generate the sequence listing?
Users should be aware that some Offices do not allow certain characters in the filename of the sequence listing. Please refer to the WIPO Sequence knowledge base for further details under help.

```xml
<ST26SequenceListing dtdVersion="V1_3" fileName="st26-annex-iii-sequence-listing-specimen.xml" softwareName="WIPO Sequence" softwareVersion="1.2.0" productionDate="2022-01-07" originalFreeTextLanguageCode="ja" nonEnglishFreeTextLanguageCode="de">
<ApplicationIdentification>
    <IPOfficeCode>IB</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2015/099999</ApplicationNumberText>
    <FilingDate>2015-01-31</FilingDate>
</ApplicationIdentification>
<ApplicantFileReference>AB123</ApplicantFileReference>
<EarliestPriorityApplicationIdentification>
<IPOfficeCode>IB</IPOfficeCode>
<ApplicationNumberText>PCT/IB2014/111111</ApplicationNumberText>
<FilingDate>2014-01-30</FilingDate>
</EarliestPriorityApplicationIdentification>
<ApplicantName languageCode="ja">出願製薬株式会社</ApplicantName>
<ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>
<InventorName languageCode="ja">特許 太郎</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle languageCode="ja">efgタンパク質をコードするマウスabcd-1遺伝子</InventionTitle>
<InventionTitle languageCode="en">Mus musculus abcd-1 gene for efg protein </InventionTitle>
<SequenceTotalQuantity>11</SequenceTotalQuantity>
<SequenceData sequenceIDNumber="1">
    <INSDSeq>
        <INSDSeq_length>133</INSDSeq_length>
        <INSDSeq_moltype>DNA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>source</INSDFeature_key>
                <INSDFeature_location>1..133</INSDFeature_location>
                    <INSDFeature_quals>
                        <INSDQualifier>
                            <INSDQualifier_name>organism</INSDQualifier_name>
                            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
                        </INSDQualifier>
                        <INSDQualifier>
                            <INSDQualifier_name>mol_type</INSDQualifier_name>
                            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
                        </INSDQualifier>
                    </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
    <INSDSeq_sequence>atgaaattaaaacataaaarggatgataaaatgagatttgatataaaaaaggttttagagttagcagagaaggattttgagacggcatggagagagacaagggcattaataaaggataaacatattgacaata
    </INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
```

Display sequence listing in txt format

```
Sequence Listing Information:

        dtdVersion: V1_3

        fileName: st26-annex-iii-sequence-listing-specimen.xml

        softwareName: WIPO Sequence

        softwareVersion: 3.0.0

        productionDate: 2024-07-28

General Information:

        Current application / IP Office: IB

        Current application / Application number: PCT/IB2015/099999

        Current application / Filing date: 2015-01-31

        Current application / Applicant file reference: AB123

        Earliest priority application / IP Office: IB

        Earliest priority application / Application number: PCT/IB2014/111111

        Earliest priority application / Filing date: 2014-01-30

        Applicant name: 出願製薬株式会社

        Applicant name / Language: ja

        Applicant name / Name Latin: Shutsugan Pharmaceuticals Kabushiki Kaisha

        Inventor name: 特許 太郎

        Inventor name / Language: ja

        Inventor name / Name Latin: Taro Tokkyo

        Invention title: efgタンパク質をコードするマウスabcd-1遺伝子 ( ja )

        Invention title: Mus musculus abcd-1 gene for efg protein ( en )

        SequenceTotalQuantity: 11

Sequences:

        Sequence Number (ID): 1

        Length: 133

        Molecule Type: DNA

        Features Location/Qualifiers:

                - source, 1..133

                        > mol_type, genomic DNA

                        > organism, Homo sapiens

        Residues:

        atgaaattaa  aacataaaar  ggatgataaa  atgagatttg  atataaaaaa  ggttttagag  60

        ttagcagaga  aggattttga  gacggcatgg  agagagacaa  gggcattaat  aaaggataaa  120

        catattgaca  ata                                                         133
```

Display sequence listing in HTML format



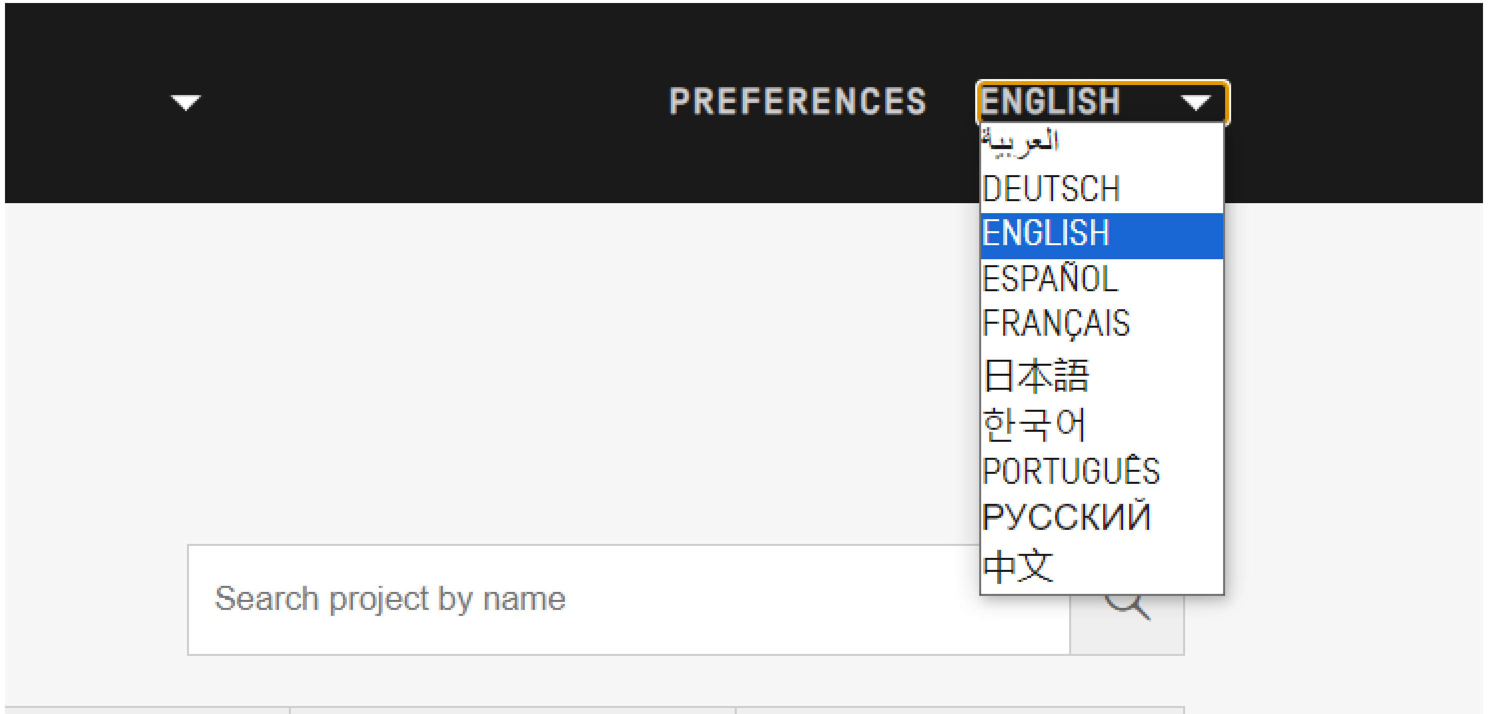
Sequence Listing

1	**Sequence Listing Information**	
1-1	fileName	st26-annex-iii-sequence-listing-specimen.xml
1-2	dtdVersion	V1_3
1-3	softwareName	WIPO Sequence
1-4	softwareVersion	3.0.0
1-5	productionDate	2024-07-28
1-6	originalFreeTextLanguageCode	ja
1-7	nonEnglishFreeTextLanguageCode	de
2	**General Information**	
2-1	Current application: IP Office	IB
2-2	Current application: Application number	PCT/IB2015/099999
2-3	Current application: Filing date	2015-01-31
2-4	Current application: Applicant file reference	AB123
2-5	Earliest priority application: IP Office	IB
2-6	Earliest priority application: Application number	PCT/IB2014/111111
2-7	Earliest priority application: Filing date	2014-01-30
2-8ja	Applicant name	出願製薬株式会社
2-8ja	Applicant name: Name Latin	Shutsugan Pharmaceuticals Kabushiki Kaisha
2-9ja	Inventor name	特許 太郎
2-9ja	Inventor name: Name Latin	Taro Tokkyo
2-10ja	Invention title	efgタンパク質をコードするマウスabcd-1遺伝子
2-10en	Invention title	Mus musculus abcd-1 gene for efg protein
2-11	SequenceTotalQuantity	11
3-1	**Sequences**	
3-1-1	Sequence Number [ID]	1
3-1-2	Molecule Type	DNA
3-1-3	Length	133
3-1-4-1	Features Location/Qualifiers	**source** 1..133 mol_type= genomic DNA organism= Homo sapiens

Languages available

- The WIPO Sequence desktop tool provides all labels and notification messages in each of the 10 PCT languages:
  - English, French, Spanish, Arabic, Russian, Chinese, German, Portuguese, Korean and Japanese
- The WIPO Sequence home page/desktop user manual is also provided in these 10 languages
- The language for the interface can be set up in the top right-hand corner of the screen (shown next page)
- The default GUI language can now be set under ‘Preferences’
- The verification report will also be generated in this language

Setting interface language



Further resources:

- WIPO Sequence homepage
- WIPO Standard ST.26
- WIPO Standards webinar recordings
- Contact standards@wipo.int

Photo credits: