IDOL PII Package

Software Version 12.4

Technical Note

Document Release Date: October 2019
Software Release Date: October 2019

Legal notices

Copyright notice

The only warranties for products and services of Micro Focus and its affiliates and licensors (“Micro Focus”) are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Micro Focus shall not be liable for technical or editorial errors or omissions contained herein. The information contained herein is subject to change without notice.

Documentation updates

The title page of this document contains the following identifying information:

- Software Version number, which indicates the software version.
- Document Release Date, which changes each time the document is updated.
- Software Release Date, which indicates the release date of this version of the software.

To check for updated documentation, visit https://www.microfocus.com/support-and-services/documentation/.

Support

Visit the MySupport portal to access contact information and details about the products, services, and support that Micro Focus offers.

This portal also provides customer self-solve capabilities. It gives you a fast and efficient way to access interactive technical support tools needed to manage your business. As a valued support customer, you can benefit by using the MySupport portal to:

- Search for knowledge documents of interest
- Access product documentation
- View software vulnerability alerts
- Enter into discussions with other software customers
- Download software patches
- Manage software licenses, downloads, and support contracts
- Submit and track service requests
- Contact customer support
- View information about all services that Support offers

Many areas of the portal require you to sign in. If you need an account, you can create one when prompted to sign in. To learn about the different access levels the portal uses, see the Access Levels descriptions.

Contents

Introduction

The IDOL PII Package contains tools that allow you to find personal identifiable information (PII) in your data, to help you comply with regulations such as the General Data Protection Regulation (GDPR).

The IDOL PII Package has two types of tools:

- IDOL Eduction Grammars (.ecr files). IDOL Eduction is a tool for finding entities (small pieces of information such as names and phone numbers) in text. Eduction grammars contain descriptions of the entities. In some cases, this might be a list of fixed values (such as names), and in others it might be pattern matching tools that find data of a particular type (such as a set of digits that make up a phone number). The Eduction grammars included in the IDOL PII Package describe different kinds of personally identifiable information, so that you can find these in your data.
- IDOL AgentBoolean IDX. IDOL AgentBoolean is a method of storing entities and querying for them that uses the IDOL Agentstore component (a specially configured IDOL Content component), rather than Eduction. The IDX files are index files that contain the details of the entities, which you can index into the IDOL Agentstore component.

Data Sources

The IDOL PII Package contains a variety of different kinds of entities to describe personally identifiable information that is protected by regulations such as GDPR. The following sections provide some information about how this information is compiled.

For all of these types of information, as much test data is acquired as possible to test the recall metric of the algorithms. Many millions of examples are run through the grammars to ensure that all patterns in usage are covered.

Names

An international database containing over 100 million individuals is analyzed to identify the structure and characteristics of names in each country. In doing so, extensive lists of the frequencies of occurrence of given names and family names are used to generate strong identification grammars for names.

In addition, rules are included to handle linguistic information, such as transliteration (for example, from the Cyrillic or Greek alphabets), or the use or removal of diacritic marks.

Date of Birth

A large corpus of documents from public sources is processed to analyze the occurrence and format of dates for each supported country. In this way, coverage of all common and less-common formats is

built up, while enabling a *likelihood* measure to indicate the confidence that the characters identified are a date of birth, rather than an unrelated date or other alphanumeric string.

Postal Codes

For each country, the publications of the national Postal Services are used as the authoritative source on the postal code.

In addition, testing against widely-gathered examples allows the identification and inclusion of non-standard formats and common errors (such as mixing the letter O with the digit 0), with an appropriately adjusted likelihood measure.

Addresses

The identification of addresses consists of a number of steps, each of which is used as additional evidence that a piece of text represents a postal address. These are:

1. The format of the text.
2. The house number / street-name portion.
3. The village / town / county / region portion.
4. The postal code.

These components are not necessarily always present for a particular address, but each is taken as evidence that the text does indeed contain an address, combining to form an overall likelihood.

- Few countries have prescribed formats for addresses, while most have conventions defined by the national Postal Service that is generally adhered to, but also frequently ignored.

  The IDOL Web Connector is used to gather many millions of web documents to identify candidate addresses in each applicable country. From there, the variety of formats that are used in practice are identified. In addition, any recommendations published by the national Postal Services are also used.
- For the street-address portion, the extensive OpenStreetMap project is used, and a database of every named street in each of the supported countries is obtained and analyzed. From this database, rules are derived to allow the identification of the vast majority of street-address strings.
- The de facto authority for geographical place names is the GeoNames database, with 11 million locations identified by data including country, population and type. In particular the *type* field is used to generate complete lists of populated settlements and administrative regions (such as county / department / region) for the countries that frequently use those in addresses. In addition, the names are available in different character sets and transliteration schemes to ensure internationalization.
- The patterns derived for matching Postal Codes are also used here (see Postal Codes, above).

Telephone Number

The general schemes for the creation of telephone numbers and fax numbers are readily available from the appropriate government department of each country. However, the formats of such numbers when written down varies considerably within a country, and even more so when numbers are referred to in a foreign document.

The strategy for creating comprehensive phone number matching grammars is centered on several key methods:

- Knowledge of the national scheme for assigning numbers.
- Databases of international and area codes in each country, obtained from authoritative sources.
- Analysis of many millions of examples of the usage of telephone numbers, obtained from a wide variety of public sources.

This final point is the most important. Only through examination of real-world usage of such numbers is the full range of formats obtained for each country.

The proximity of keywords indicating that the digits represent a telephone or fax number is used to strengthen the likelihood of the match.

National Identification Number

Each country has a different scheme for the use of National Identification. For countries with National ID cards, the format of the number is derived from governmental sources. In other countries, the formats of National Health, National Social Security, or National Insurance numbers are obtained from governmental sites, with the exception of a few cases in which other sources are used.

Tax Identification Number (TIN)

Each country in the European Union uses a Tax Identification Number. Grammars are used to identify these using rules laid down by the European TIN Portal, published by the European Commission.

The *strength* of the format (that is, the likelihood of false positives) and the proximity of each format to key TIN-related terms allows the calculation of a likelihood measure, where high likelihood items are stronger indicators that a TIN is present, as opposed to an unrelated number that happens to be in the same format.

Passport Number

The format of the national passport numbers is not as widely available as other such numbers. However, authoritative government documents are acquired for the formats of passport numbers in the majority of supported countries.

In other cases, non-governmental sources and the examination of examples have allowed grammars to be created for each country. In all cases, the presence of keywords and phrases in appropriate

languages in proximity to the number are used to increase the likelihood of the match and to reduce the number of false positives.

In addition, grammars to identify Machine-Readable travel documents such as the MROTD and MRP have been added.

Driving License

As with passport numbers, not all governments have published the scheme used in the numbering of Driving Licenses. The format of the number is obtained for the majority of relevant countries, with the remainder derived from secondary sources and from analysis of example numbers.

Medical

Documents that contain mention of medical procedures or conditions are identified with the Medical categories, available in each of the supported languages. The categories are generated from the Medical Subject Headings (MeSH) taxonomy published by the United States National Library of Medicine using the C hierarchy (diseases and conditions).

New in this Release

This section describes the enhancements to the IDOL PII Package in version 12.4.

- The IDOL PII Package now includes resources for Brazil. A complete set of entities are available to extract information including addresses and postcodes, driving license numbers, names, national ID numbers, health numbers (CNS), passport numbers, telephone numbers, and tax identification numbers.
- The IDOL PII Package now includes resources for Switzerland. A complete set of entities are available to extract information including addresses and postcodes, dates, driving license numbers, names, national ID numbers (AHV), health insurance card numbers (Swiss and EHIC), passport numbers, telephone numbers, and tax identification numbers.
- The IDOL PII Package now supports Eduction prefiltering to improve performance for certain grammars. MicroFocus recommends that you use prefilter tasks for the address.ecr and combined_address.ecr grammars. See Configure Pre-Filtering, on page 14.
- The IDOL PII Package now includes new ejr versions of some grammars, which are performance-optimized for cases when the expected match density is low (that is, it is expected that less than 10% of the input to Eduction is a valid match). The following new grammars are available: date.ejr, driving.ejr, health.jr, mrtd.ejr, national_id.ejr, nationality.ejr, passport.ejr, postcode.ejr, tin.ejr, and travel.ejr. These grammars contain the same entities as the ecr equivalent. See ECR and EJR Grammars, on page 15.
- The sample code included with the PII package has been updated to demonstrate how to programmatically detect all forms of PII. For details, see the included `README.txt` file in the `edk_samples` directory of your IDOL PII Package installation.
- Scoring for the address grammar has been improved. More addresses now score less than one. In general, a score of one is given to an address that includes a numbered, common format street address (for example 23 North Road), a known city (for example London), and a postal code in a viable format for the country (for example SW1A 2AA). Deviations from this form lead to score penalties. The ordering of these elements varies by country.
- Checksum validation is now performed on matches of the `pii/mrtd/mrp` and `pii/mrtd/mrotd/td1` entities.
- Slovenian national ID entities now match EMŠO numbers with embedded region codes other than 50 (Slovenia). This change finds legacy ID numbers issued under other registers, which might still be in use.
- United States SSN values now match the SSN landmark when it is adjacent to the number without a space, for example SSN12456789 (or ssn123456789).
- The IDOL PII Package now includes a new grammar, `travel.ecr` to match US passport card numbers.
- The IDOL PII Package now includes a new grammar, `nationality.ecr`, to match nationalities and countries. For all supported countries, this grammar finds nationalities in the native language. It also finds nationalities for all countries in English.

Resolved Issues

This section lists the resolved issues in the IDOL PII Package version 12.4.

- The pii/passport/context entities could return matches for overlong numbers where the final portion matched the pattern. For example:

  ```
  Passport: 1234567890
  ```

  matched the GB passport entity, despite having ten digits rather than nine (the returned NUMBER component was 234567890).

Country and Language Support

The IDOL PII Package contains grammars and IDX files that apply to data from many countries and languages.

Country Codes

For data that corresponds to a particular country, the Eduction grammars identify each country by using the ISO 3166-1 alpha-2 country codes. The following countries are supported:

Country Code	Country
at	Austria
au	Australia
be	Belgium
bg	Bulgaria
br	Brazil
ca	Canada
ch	Switzerland
cy	Cyprus
cz	Czech Republic
de	Germany
dk	Denmark
ee	Estonia
es	Spain
fi	Finland
fr	France
gb	United Kingdom (England, Wales, Scotland, and Northern Ireland)
gr	Greece
hr	Croatia

Country Code	Country
hu	Hungary
ie	Ireland
is	Iceland
it	Italy
li	Liechtenstein
lt	Lithuania
lu	Luxembourg
lv	Latvia
mt	Malta
nl	Netherlands
no	Norway
nz	New Zealand
pl	Poland
pt	Portugal
ro	Romania
se	Sweden
si	Slovenia
sk	Slovakia
tr	Turkey
us	United States of America

Languages

For data that corresponds to a particular language, the Eduction grammars and AgentBoolean IDX files identify each language by using the ISO 639-2/B language codes. The following languages are supported:

Language Code	Language
bul	Bulgarian

Language Code	Language
cat	Catalan
cze	Czech
dan	Danish
dut	Dutch
eng	English
est	Estonian
fin	Finnish
fre	French
ger	German
gle	Irish
gre	Greek
hrv	Croatian
hun	Hungarian
ice	Icelandic
ita	Italian
lav	Latvian
lit	Lithuanian
mlt	Maltese
nor	Norwegian
pol	Polish
por	Portuguese
roh	Romansh
rum	Romanian
slo	Slovak
slv	Slovenian
spa	Spanish
swe	Swedish
tur	Turkish

IDOL Eduction Grammars

The following section describes the Eduction grammars available in the IDOL PII Package.

You can use these grammars with IDOL Eduction, by using Eduction Server, the edktool command-line utility, or the Eduction SDK. For more information, refer to the *IDOL Eduction User Guide* and the *Eduction SDK Programming Guide*.

> **IMPORTANT:** To use the Eduction grammars in the IDOL PII Package, you must have a license that enables them. To obtain a license, contact Micro Focus Support.

Configure Post Processing

When you use the IDOL PII Package Eduction grammars it is essential to configure a Lua post-processing task to run the script `pii_postprocessing.lua`. This script contains post-processing to improve results for various entities, such as stop list filtering, entity name mapping for combined grammars (see Combined Entities, on page 25), ambiguous landmark detection (see Ambiguous Entities, on page 32) and checksum validation (see Validated ID Numbers, on page 30).

> **IMPORTANT:** If you do not run this script, you might encounter unexpected behavior.

Add a post-processing task to your Eduction configuration. For example:

```
[PostProcessingTasks]
NumTasks=1
Task0=MyPostProcessingSection

[MyPostProcessingSection]
Type=Lua
Script=scripts/pii_postprocessing.lua
Entities=pii/*,gdpr/*
```

For more information about configuring post-processing tasks, refer to the *Eduction User and Programming Guide*.

Configure Pre-Filtering

Pre-filtering allows the IDOL PII Package to run a quick initial check to find potential matches in your input text. It then selects match windows around these potential matches, reducing the amount of text that it must match against your grammars. This process can improve the performance in certain cases.

Micro Focus recommends that you use the following pre-filtering configuration with the `address.ecr` and `combined_address.ecr` grammars.

```
[Eduction]
PrefilterTask0=AddressPrefilter

[AddressPrefilter]
Regex=\d{1,7}
WindowCharsBeforeMatch=100
WindowCharsAfterMatch=100
```

> **NOTE:** Pre-filter tasks run for all configured entities, so you must configure it only for the appropriate entities to ensure that it does not affect the results for other entities.

For more information about pre-filtering, refer to the *Eduction User and Programming Guide*.

Entity Context

Some of the entities are available in two versions, with and without context. The context-based entities match the entity when it occurs in an easily identifiable location in text. For example, it might match a telephone number that occurs next to the prefix **Phone:**.

The entities that do not have context attempt to match the entity wherever it occurs. This version might over-match significantly (that is, it is likely to return values that are similar to the entity patterns, such a number that is not a telephone number). However, it also reduces the number of false negatives (that is, it misses fewer matches).

You can configure Eduction to use both versions of an entity; matches located with context are given a higher score in the results.

ECR and EJR Grammars

Some grammars are available in two formats, ECR and EJR. In these cases, both formats contain the same entities for extraction, and the format that you use depends on your input data.

EJR files are performance-optimized for cases where the expected match density in your input text is low. Micro Focus recommends that you use EJR files when you expect less than 10% of the input text to be valid matches. In all other cases, use the ECR files.

When you use EJR grammars, you must run them in a separate matching engine to any ECR grammars, although you can run multiple EJR grammars in the same engine.

For example, the following configuration is allowed:

```
ResourceFiles=passport.ejr,date.ejr
```

You cannot set `ResourceFiles=passport.ejr,date.ecr`.

Balance Precision and Recall

In many cases, Eduction is able to locate entities that are ambiguous, such as a postal code which is simply a five-digit number. In some situations it is desirable to match as many entities as possible ("high recall") and in others only entities with a high likelihood of being a useful match ("high precision"). So that you can filter the results produced by Eduction, each match is given a 'score' value.

As described in Entity Context, matches located by an entity that requires context are assigned higher scores than matches located by the corresponding entity without context. Most matches extracted without context have a score of `0.4`. For example, a context-free date ("January 18, 1998") might be returned by the Date Of Birth entity with a score of `0.4`. But with context to suggest that it is indeed a date of birth ("DOB: January 18, 1998"), the score should be above `0.5`.

The PII post-processing script (see Configure Post Processing, on page 14) includes a step to validate matches (some ID numbers can be validated by calculating a checksum). The script increases the score of matches that have valid checksums, because this is an indication that the match is more likely to be genuine. Any match that has an invalid checksum is immediately discarded because it cannot be genuine.

When you configure Eduction, use the parameters `MinScore` and `PostProcessThreshold` to achieve the desired balance between precision and recall. Eduction discards any match with a score lower than `MinScore`. Matches with scores that meet or exceed `MinScore` are then processed by post-processing tasks. After post-processing has finished, Eduction discards any match with a score lower than `PostProcessThreshold`.

In the example configuration that is included with the IDOL PII Package, `MinScore` is set to `0.4` and `PostProcessThreshold` is set to `0.5`. These values have been chosen to return results only if they have a relatively high likelihood of being a useful match. Any match that is located without context can proceed to post-processing, but, unless its score is increased through successful validation, it is then discarded. If you prefer to maximize recall rather than precision, you can reduce or remove these thresholds.

For more information about Eduction configuration parameters, refer to the *Eduction User Guide*.

Configure Tangible Characters

`TangibleCharacters` is a configuration parameter that you can set when using the Eduction SDK, the Eduction Server, or the Eduction command-line utility (`edktool`). It specifies a list of characters to treat as part of a word, rather than as word boundaries.

Some of the entities in the IDOL PII Package Eduction Grammars require tangible characters to be set in order to perform correctly (see the descriptions of the entities in Eduction Grammar Reference, on page 18).

When you use Eduction to search for matches, `TangibleCharacters` applies across all of your chosen entities. If you use multiple entities that have different recommended tangible character sets, you might need to take some extra steps. For example:

- If you are using the Eduction SDK, create a separate EDK engine for each distinct set of tangible characters, and configure the tangible characters for the engine using the appropriate API call:

C	`EdkSetTangibleCharacters`
Java	`EDKEngine.setTangibleCharacters`

  After configuring an engine with the correct tangible characters, you can add the relevant entities. You will need to create a session from each engine to process your input text.
- If you are using an Eduction Server, send a separate action (`EduceFromText` or `EduceFromFile`) for each distinct set of tangible characters. In each action, set the `TangibleCharacters` and `Entities` action parameters to specify which set of tangible characters and which entities to use.
- If you are using the command line `edktool`, create a separate configuration file for each distinct set of tangible characters and associated entities, and process your input text once with each configuration file.

For more information about the `TangibleCharacters` configuration parameter, refer to the *Eduction User Guide*.

Customize Stop Lists

The IDOL PII Package post-processing script (see Configure Post Processing, on page 14) uses stop lists to discard matches that are likely to be false positives. You can add entries to the stop lists, or remove entries, by modifying the following files.

- `scripts/address_stoplist.lua` contains a list of common words that are likely to indicate a false positive when returned as the `STREET` or `CITY` component of an address match.
- `scripts/names_stoplist.lua` contains two stop lists to discard names. In the first stop list, each component is plausible but the entire match is likely to be a false positive, for example "Christian Church" or "Norman Conquest". The second stop list contains common words that are likely to indicate a false positive when returned as either the `FORENAME` or `SURNAME` component of a name match. The stop lists in this file can be customized such that a name can be considered a false positive in one country but not another.

Eduction Grammar Reference

The following table describes the grammar files that are available in the IDOL PII Package, and the entities that each provides.

In the entity names:

- the abbreviation CC refers to a two-letter country code. For a list of available country codes, see Country Codes, on page 11.
- the abbreviation LLL refers to a three-letter language code. For a list of available languages, see Languages, on page 12.

TIP: You can use the Eduction parameter `EntityN` to specify which entities you want to extract. This parameter accepts wildcards, so you can extract entities of a specific type for all supported countries or languages. For example, to match postal addresses for all countries specify a value of `pii/address/??`. To match dates of birth in all languages, specify `pii/date/dob/context/???`.

Some grammars are available in two formats, ECR and EJR. For more information about which to use, see ECR and EJR Grammars, on page 15.

NOTE: The IDOL PII Package is backwards-compatible with the IDOL GDPR package. You can continue to use existing configurations that use entity names such as `gdpr/address/CC` or `gdpr/telephone/CC`. These entities are similar to the corresponding `pii/*` entity, but are limited to countries in the GDPR region. However, Micro Focus recommends that you use the `pii/*` entities instead, so that Eduction extracts matches for all supported countries.

File	Entity	Description
address.ecr	pii/address/CC	A postal address. In general, a score of one is given to an address that includes a numbered, common format street address (for example "23 North Road"), a known city (for example "London"), and a postal code in a viable format for the country (for example "SW1A 2AA"). Deviations from this form lead to score penalties. The ordering of these elements varies by country. Micro Focus recommends that you use pre-filtering to improve the performance for this grammar. See Configure Pre-Filtering, on page 14.

File	Entity	Description
		Example matches: "Schlosshoferstrasse 20, 1210 Vienna", "Avenida Juan Xxiii 20, 41006, Sevilla", "162-168 Regent Street, London, W1B 5TG".
date.ecr date.ejr	pii/date/dob/context/LLL	A date of birth, written numerically or using words. For example "date of birth 1/1/2018", "GEBOORTEDATUM: 01/01/2018"
	pii/date/nocontext/LLL	A calendar date, written numerically or using words, without context. For example "01.03.1918", "2018_01_01", "вторник, 30 октомври 2018".
	pii/date/dob/landmark/LLL	A date of birth landmark, such as "DOB" or "Fecha de nacimiento".
driving.ecr driving.ejr	pii/driving/context/CC	A driving license number with context. For example: "australian automobile association: 103 805 501", or "driver's license: A234567890". This entity matches both the driving license number, and the personal number or driver number, if present. On the standard European driving license, these are fields 5 and 4d.
	pii/driving/nocontext/CC	A driving license number, without context.
	pii/driving/landmark/CC	A driving license landmark, such as "Driver's license" or "Driving Licence".
health.ecr health.ejr	pii/health/ehic/context/CC	An EHIC personal identification number with context. For example "EHIC: UK 1234 5678 " or "TSE: 123456789012".
	pii/health/ehic/nocontext/CC	An EHIC personal identification number without context. For example "123456-789A".
	pii/health/ehic/landmark/CC	An EHIC landmark, such as "EHIC" or "EHIC PIN".
	pii/health/ehic/context/all	An EHIC personal identification number with context, for EU countries and Switzerland.
	pii/health/ehic/nocontext/all	An EHIC personal identification number without context, for EU countries and Switzerland.

File	Entity	Description
	pii/health/ehic/landmark/all	An EHIC landmark, such as "EHIC" or "EHIC PIN", for EU countries and Switzerland.
	pii/health/id/context/au	An Australian Medicard (card) number or Individual Healthcare Identifier (IHI) with context. For example "Medicare Card Number: 3501 80315 1-6".
	pii/health/id/context/br	A Brazilian Cartão Nacional de Saúde (CNS, also known as SUS) number with context, for example "CNS: 190129759240018".
	pii/health/id/context/ca	A Canadian health insurance (card) number with context. For example "health insurance: 12345-6789", or "assurance-maladie: 12345-6789".
	pii/health/id/context/ch	A Swiss health insurance card number with context. For example "Schweizerische Krankenversicherungskarte: 12345678901234567890".
	pii/health/id/context/es	A Spanish health insurance card number with context. For example "CatSalut: ABCD 1 123456 12 1".
	pii/health/id/context/fr	A French Carte Vitale number with context. For example "INSEE: 187090100100141".
	pii/health/id/context/gb	A British NHS number with context. For example "NHS Number: 943 476 5919".
	pii/health/id/context/nz	A New Zealand National Health Index (NHI) number with context. For example "NHI Number: CGC2720".
	pii/health/id/context/us	A US health insurance number with context. For example "Medicare ID: 1EG4-TE5-MK72".
	pii/heath/id/nocontext/CC	A health number, such as a British NHS number or French Carte Vitale number, without context.
	pii/health/id/landmark/CC	A health number landmark, such as "NHS number" or "Medicare ID".
mrtd.ecr	pii/mrtd/mrp	A machine readable passport. For example

File	Entity	Description
mrtd.ejr		"P<GBRUK<SPECIMEN<<ANGELA<ZOE<<<<<<<<<<<<<<< 5333244280GBR8812049F2509286<<<<<<<<<<<<<<00"
	pii/mrtd/mrotd/td1	A machine readable TD1-size travel document. For example "IDD<<T220001293<<<<<<<<<<<<<< 6408125<2010315D<<<<<<<<<<<<4 MUSTERMANN<<ERIKA<<<<<<<<<<<<"
name.ecr	pii/name/CC	A full personal name, in title case or upper case.
national_id.ecr national_id.ejr	pii/id/context/CC	A national identity number with context. For information about the supported ID numbers, see Supported National ID Numbers, on page 28.
	pii/id/nocontext/CC	A national identity number without context. For information about the supported ID numbers, see Supported National ID Numbers, on page 28.
	pii/id/landmark/CC	A national identity number landmark, such as "National insurance number" or "Social security number".

File	Entity	Description
nationality.ecr[1] nationality.ejr	pii/nationality/adj/context/CC	A nationality adjective with context. For example, "Nationality: British".
	pii/nationality/adj/nocontext/CC	A nationality adjective without context. For example, "British".
	pii/nationality/adj/landmark/CC	A nationality adjective landmark. For example, "Nationality".
	pii/nationality/noun/context/CC	A nationality noun with context. For example, "Country: Britain".
	pii/nationality/noun/nocontext/CC	A nationality noun without context. For example: "Britain".
	pii/nationality/noun/landmark/CC	A nationality noun landmark. For example, "Country".
	pii/nationality/any/context/CC	Any combination of nationality adjective and noun landmark and value. For example, "Country: British", or "Nationality: British".
	pii/nationality/any/nocontext/CC	Any nationality adjective or noun. For example, "Britain" or "British".
	pii/nationality/any/landmark/CC	Any nationality adjective or landmark. For example, "Nationality" or "Country".
passport.ecr passport.ejr	pii/passport/context/CC	A passport number with context. For example "Passport number: 533324428", "Passport Number: P4366918", or "italian passaporti AA5275702".
	pii/passport/nocontext/CC	A passport number without context. For example "533324428", "C015918", or "14CV28142".
	pii/passport/landmark/CC	A passport landmark, such as "Passport" or "Pasaporte". For information about cases where the landmark and passport number do not match or have an ambiguous match, see Ambiguous Entities, on page 32.
postcode.ecr postcode.ejr	pii/postcode/context/CC	A postal code with context. For example "PLZ: 1210", "Poštanski broj: 10000", or "Cod poştal: 235200".

[1]This grammar matches nationalities for all countries in English, and for each supported country in their native language.

File	Entity	Description
	pii/postcode/nocontext/CC	A postal code without context. For example "2700-439 AMADORA", "75018", or "W1B 5TG".
	pii/postcode/landmark/CC	A postal code landmark, such as "Postcode" or "Postleitzahl".
telephone.ecr	pii/telephone/context/CC	A telephone number with context. For example "Tel: +44 1234 224050", "Telephone: (204)-243-9955", or "numéro de téléphone: +1-902-861-7000". **NOTE:** To ensure that this entity performs correctly, set your `TangibleCharacters` configuration to include the following characters: `()+-`. For more information, see Configure Tangible Characters, on page 16.
	pii/telephone/nocontext/CC	A telephone number without context. For example: "+39 055 326 43 11", or "44 20 7499 9000". **NOTE:** To ensure that this entity performs correctly, set your `TangibleCharacters` configuration to include the following characters: `()+-`. For more information, see Configure Tangible Characters, on page 16.
	pii/telephone/landmark/CC	A telephone number landmark, such as "Tel:" or "Telefon No".
tin.ecr tin.ejr	pii/tin/context/CC	A tax identification number with context. For example "ITIN: 911-92-3333", or "TIN-numre: 101111113".
	pii/tin/nocontext/CC	A tax identification number without context. For example "756.3047.5009.62", or "Z1234567R".
	pii/tin/landmark/CC	A tax identification number landmark, such as "ITIN" or "TIN-numre".
travel.ecr travel.ejr	pii/travel/context/us	A US passport card number with context. For example "Passport card number: C12345678".

File	Entity	Description
	pii/travel/nocontext/us	A US passport card number without context. For example "C12345678".
	pii/travel/landmark/us	A US passport card number landmark. For example "Passport card number".

Combined Entities

In addition to the entities described in the Eduction Grammar Reference, on page 18, the IDOL PII Package includes grammar files that contain "combined" entities. These files are named `combined_*.ecr` and the entities match addresses, dates, driving license numbers, and so on, from multiple countries.

- The entities that end in `/all` match data for any supported country or language.
- The entities that end in `/gdpr` match data for any supported country or language subject to GDPR.

For example:

- Using `pii/address/all` from `combined_address.ecr` matches a postal address from any country. This is similar to using the `address.ecr` grammar file and extracting `pii/address/??`.
- Using `pii/address/gdpr` from `combined_address.ecr` matches a postal address from any country subject to GDPR. This is similar to using the `address.ecr` grammar file and extracting `gdpr/address/??`.
- Using `pii/date/dob/context/all` from `combined_date.ecr` matches a date of birth written numerically or using words in any language. This is similar to using the `date.ecr` grammar file and extracting `pii/date/dob/context/???`.

The combined (`/all` and `/gdpr`) entities provide a significant improvement in processing speed when you extract matches for all countries or languages.

You must run the script `pii_postprocessing.lua` as a post-processing task (see Configure Post Processing, on page 14). Running the script ensures that the entity names returned by Eduction contain the relevant country code or language code. For example, if a UK postal address is found, the entity name in the returned match is still `pii/address/gb`, and not `pii/address/all`.

The combined grammar files might produce fewer matches, because only a single match is returned in cases where the same characters in the input text would match multiple countries or languages.

File	Entity
combined_address.ecr	pii/address/all
	pii/address/gdpr
combined_date.ecr	pii/date/dob/context/all
	pii/date/dob/landmark/all
	pii/date/dob/context/gdpr
	pii/date/dob/landmark/gdpr
	pii/date/nocontext/all
	pii/date/nocontext/gdpr

File	Entity
combined_driving.ecr	pii/driving/context/all
	pii/driving/nocontext/all
	pii/driving/landmark/all
	pii/driving/context/gdpr
	pii/driving/nocontext/gdpr
	pii/driving/landmark/gdpr
combined_health.ecr	pii/health/ehic/context/gdpr
	pii/health/ehic/nocontext/gdpr
	pii/health/ehic/landmark/gdpr
	pii/health/id/context/all
	pii/health/id/nocontext/all
	pii/health/id/landmark/all
	pii/health/id/context/gdpr
	pii/health/id/nocontext/gdpr
	pii/health/id/landmark/gdpr
combined_name.ecr	pii/name/all
	pii/name/gdpr
combined_national_id.ecr	pii/id/context/all
	pii/id/nocontext/all
	pii/id/landmark/all
	pii/id/context/gdpr
	pii/id/nocontext/gdpr
	pii/id/landmark/gdpr

File	Entity
combined_passport.ecr	pii/passport/context/all
	pii/passport/nocontext/all
	pii/passport/landmark/all
	pii/passport/context/gdpr
	pii/passport/nocontext/gdpr
	pii/passport/landmark/gdpr
combined_postcode.ecr	pii/postcode/context/all
	pii/postcode/nocontext/all
	pii/postcode/landmark/all
	pii/postcode/context/gdpr
	pii/postcode/nocontext/gdpr
	pii/postcode/landmark/gdpr
combined_telephone.ecr	pii/telephone/context/all
	pii/telephone/nocontext/all
	pii/telephone/landmark/all
	pii/telephone/context/gdpr
	pii/telephone/nocontext/gdpr
	pii/telephone/landmark/gdpr
combined_tin.ecr	pii/tin/context/all
	pii/tin/nocontext/all
	pii/tin/landmark/all
	pii/tin/context/gdpr
	pii/tin/nocontext/gdpr
	pii/tin/landmark/gdpr

Supported National ID Numbers

The following table lists the national ID numbers that are supported by the `pii/id/context/CC` and `pii/id/nocontext/CC` Eduction entities.

Country	Supported national identity numbers	Example context	Example Match
Australia	ImmiCard	ImmiCard	AMS123456
Austria	SSN (social security number) CRR	ASVG	1788011550
Belgium	NRN (numéro de registre national)	numéro national	85 07 30 033 28
Bulgaria	EGN (Uniform Civil Number)	EGN	8032056031
Brazil	Registro de Identidade Civil (RIC) Registro Geral (RG)	RIC Registro Geral	12345678901 56.843.539-4
Canada	Canadian national ID (Social Insurance Number)	numéro d'assurance sociale	159749357
Croatia	OIB (Osobni identifikacijski broj)	OIB	79423753532
Cyprus	Identity card number	Αριθμός ταυτότητας	3861811-2
Czech republic	Rodné číslo	rodné číslo	7360285163
Denmark	CPR	legitimation	011118-0001
Estonia	IK (isikukood)	Isikukood	37605030299
Finland	Henkilötunnus (Personal identity code)	Henkilötunnus	311280-888Y
France	INSEE code	Code INSEE	187090100100141
Germany	National ID serial number	Personalausweis	T22000129
Greece	National ID card AMKA (social security number)	AMKA	13121199999
Hungary	Personal Identification Number ID card number	Nemzeti személyazonosító jel	58709189997
Iceland	Kennitala	Kennitala	1809872079

Ireland	PPSN (personal public service number)	PPSN	1234567TW
Italy	Codice Fiscale	codice fiscale	RSS MRA 74D22 A001Q
Latvia	Personas kods	personas kods	121282-11212
Liechtenstein	Identitätskarte	Personalausweis	ID98754015
Lithuania	Asmens Kodas	asmens kodas	38409152012
Luxembourg	National ID card number Identity card number	Steuernummer	1893120105732
Malta	ID card number	ID card	9999999M
Netherlands	BSN	legitimatiebewijs	269740533
Norway	Fødselsnummer D-nummer H-nummer FH-nummer	Fødselsnummer	18098749914
Poland	PESEL	PESEL	44051401359
Portugal	Número de identificação civil Cartão de cidadão number Número de Identificação de Segurança Social	NIC	118666070
Romania	Cod Numeric Personal	CNP	1800101221144
Slovakia	Rodné číslo ID card number	rodné číslo	7360285163
Slovenia	Enotna matična številka občana	EMŠO	1809987504991
Spain	DNI NIE	DNI	00000000T
Sweden	Personnummer Samordningsnummer	personnummer	870918-9990
Turkey	Turkish Identification Number	türkiye cumhuriyeti kimlik numarası	98768109974
UK	National Insurance Number	National Insurance Number	AB 12 34 56 A

US	US Social Security Number	SSN	111-22-3333

Validated ID Numbers

The script `pii_postprocessing.lua` (see Configure Post Processing, on page 14) includes steps to validate ID numbers that are found by Eduction. This improves accuracy by discarding results that match the pattern for a valid ID number, but cannot be genuine because they have an invalid checksum. The script increases the score for matches that have a valid checksum, because this is an indication that the match is more likely to be genuine.

The following tables list the entities that are validated.

Health ID numbers (`health.ecr`)	
pii/health/id/context/au	
pii/health/id/context/br	
pii/health/id/context/gb	
pii/health/id/context/fr	Validated using the INSEE checksum
pii/health/id/context/nz	

National ID numbers (`national_id.ecr`)		
pii/id/context/au	pii/id/nocontext/au	
pii/id/context/at	pii/id/nocontext/at	Only the SSN component is validated.
pii/id/context/be	pii/id/nocontext/be	
pii/id/context/bg	pii/id/nocontext/bg	
pii/id/context/ca	pii/id/nocontext/ca	
pii/id/context/cz	pii/id/nocontext/cz	
pii/id/context/ee	pii/id/nocontext/ee	
pii/id/context/es	pii/id/nocontext/es	
pii/id/context/fi	pii/id/nocontext/fi	
pii/id/context/fr	pii/id/nocontext/fr	
pii/id/context/gr	pii/id/nocontext/gr	Only the AMKA component is validated.
pii/id/context/hr	pii/id/nocontext/hr	
pii/id/context/hu	pii/id/nocontext/hu	Only the PIN component is validated.

pii/id/context/ie	pii/id/nocontext/ie	
pii/id/context/is	pii/id/nocontext/is	
pii/id/context/it	pii/id/nocontext/it	
pii/id/context/lt	pii/id/nocontext/lt	
pii/id/context/lu	pii/id/nocontext/lu	
pii/id/context/nl	pii/id/nocontext/nl	
pii/id/context/no	pii/id/nocontext/no	
pii/id/context/pl	pii/id/nocontext/pl	
pii/id/context/pt	pii/id/nocontext/pt	
pii/id/context/ro	pii/id/nocontext/ro	
pii/id/context/si	pii/id/nocontext/si	
pii/id/context/se	pii/id/nocontext/se	
pii/id/context/sk	pii/id/nocontext/sk	Only the Rodné číslo component is validated.
pii/id/context/tr	pii/id/nocontext/tr	

Tax ID numbers (`tin.ecr`)		
pii/tin/context/at	pii/tin/nocontext/at	
pii/tin/context/au	pii/tin/nocontext/au	
pii/tin/context/be	pii/tin/nocontext/be	
pii/tin/context/bg	pii/tin/nocontext/bg	
pii/tin/context/br	pii/tin/nocontext/br	Cadastro de Pessoas Físicas (CPF) Cadastro Nacional de Pessoa Jurídica (CNPJ)
pii/tin/context/ca	pii/tin/nocontext/ca	
pii/tin/context/cy	pii/tin/nocontext/cy	
pii/tin/context/de	pii/tin/nocontext/de	
pii/tin/context/dk	pii/tin/nocontext/dk	
pii/tin/context/ee	pii/tin/nocontext/ee	

pii/tin/context/es	pii/tin/nocontext/es	
pii/tin/context/fi	pii/tin/nocontext/fi	
pii/tin/context/fr	pii/tin/nocontext/fr	
pii/tin/context/hr	pii/tin/nocontext/hr	
pii/tin/context/hu	pii/tin/nocontext/hu	
pii/tin/context/ie	pii/tin/nocontext/ie	
pii/tin/context/it	pii/tin/nocontext/it	
pii/tin/context/lt	pii/tin/nocontext/lt	
pii/tin/context/lu	pii/tin/nocontext/lu	
pii/tin/context/mt	pii/tin/nocontext/mt	
pii/tin/context/nl	pii/tin/nocontext/nl	
pii/tin/context/nz	pii/tin/nocontext/nz	Inland Revenue Department Number
pii/tin/context/pl	pii/tin/nocontext/pl	
pii/tin/context/pt	pii/tin/nocontext/pt	
pii/tin/context/se	pii/tin/nocontext/se	
pii/tin/context/si	pii/tin/nocontext/si	
pii/tin/context/sk	pii/tin/nocontext/sk	

Machine readable passport numbers (`mrtd.ecr`)
pii/mrtd/mrp
pii/mrtd/mrotd/td1

Ambiguous Entities

For some entities, IDOL PII Package cannot always unambiguously determine the country of origin for a value. For some of these cases, it can return an ambiguous result.

Cross-Language Passport Landmarks

The IDOL PII Package allows cross-language passport landmarks, so that it detects passport numbers provided in languages that do not belong to the associated passport country.

For example, the text "Oma passi on P 4366918" contains *passi*, which is Finnish for passport, and the number P 4366918, which is an Austrian passport number. The PII grammar identifies this as an Austrian passport number and returns the entity `pii/passport/at`.

In some cases, the country of origin is ambiguous. In this case, the grammar attempts to identify all possible countries and returns an entity with the label `ambiguous`.

Example 1

"Mon passeport est LA080402"

In this example, both the landmark text *passeport*, and the passport number could be from either Belgium, Canada, or Luxembourg. This example returns the entity `pii/passport/ambiguous/be_ca_lu` to represent all three possibilities.

Example 2

"Vegabref mitt er AA5275702"

In this example, *Vegabref* is Icelandic, but AA5275702 could be a passport number for several countries, not including Iceland. This example returns the entity `pii/passport/ambiguous/au_fi_ie_it_lv_pl_sk_si_gr_hu_ee_nl_de_us02` to represent all the possibilities.

NOTE: The us02 option in this response means that this pattern scores 0.2 as a US passport pattern.

TIP: In this example, if the text belonged to a language from one of the possible countries, the passport number would not be considered ambiguous. For example, "Oma passi on AA5275702" (where *passi* is Finnish for passport), returns the entity `pii/passport/fi`, because passi applies only to Finnish, and not to any of the other countries where the passport number is valid.

Ambiguous Driving License Matches

There are some countries that have some overlap in driving license number, and where the languages are the same it is not possible to identify which country a particular number comes from. In this case, the grammar attempts to identify all possible countries and returns an entity with the label `ambiguous`. For example:

```
pii/driving/ambiguous/au_ie_us
pii/driving/ambiguous/au_nz_us
pii/driving/ambiguous/au_us
pii/driving/ambiguous/mt_us
```

IDOL AgentBoolean IDX

IDOL AgentBoolean provides another way of finding pieces of information in text. In this case, you index the entities that you want to find into an IDOL Agentstore component.

The IDOL Agentstore component is a specially configured IDOL Content component. It uses IDOL AgentBoolean queries for entity matching.

When you use AgentBoolean for entity matching, each entity becomes a document in Agentstore. You then send a piece of text as a query to Agentstore, and it returns the entity documents that match the text.

The IDOL PII Package contains several IDX documents that describe entities for medical data, which you can use as another tool to find data that is protected by regulations such as GDPR. There is an IDX file for each of the supported languages (see Languages, on page 12).

The package also contains example Agentstore configuration files to allow you to set up your Agentstore component more easily.

After you configure and set up your Agentstore, you can index the IDX documents and use Agentstore for entity matching.

For more information about how to set up and use IDOL querying, refer to the *IDOL Server Administration Guide* and the *IDOL Content Component Reference*.

Send documentation feedback

If you have comments about this document, you can contact the documentation team by email. If an email client is configured on this system, click the link above and an email window opens with the following information in the subject line:

Feedback on Technical Note (Micro Focus IDOL PII Package 12.4)

Add your feedback to the email and click **Send**.

If no email client is available, copy the information above to a new message in a web mail client, and send your feedback to swpdl.idoldocsfeedback@microfocus.com.

We appreciate your feedback!