IDOL GDPR Package

Software Version 12.1

Technical Note

Document Release Date: October 2018
Software Release Date: October 2018

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.

You can check for more recent versions of a document through the MySupport portal. Many areas of the portal, including the one for documentation, require you to sign in with a Software Passport. If you need a Passport, you can create one when prompted to sign in.

Additionally, if you subscribe to the appropriate product support service, you will receive new or updated editions of documentation. Contact your Micro Focus sales representative for details.

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 with a Software Passport. If you need a Passport, 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 GDPR Package contains tools that allow you to find personal identifiable information in your data, to help you comply with the General Data Protection Regulation (GDPR).

The IDOL GDPR 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 GDPR grammars contain details of different kinds of personally identifiable information, to allow you to find these values in text.

- 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.

The following section describes the data sources that have been used to compile the GDPR grammars and IDX files.

Data Sources

The IDOL GDPR Package contains a variety of different kinds of entities to describe personally identifiable information that is protected by 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 across the EEA 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 country subject to GDPR. 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 of the EEA, 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 across the EEA 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 countries of the EEA 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, on the previous page).

Telephone Number

The general schemes for the creation of telephone 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 number is used to strengthen the likelihood of the match.

National Identification Number

Each country of the EEA 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 16 of the countries of the EEA.

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 languages subject to GDPR. 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 GDPR Package in version 12.1.

- To simplify the results returned from Eduction for national ID numbers, the `national_id.ecr` grammar file has been updated so that all entities match the naming pattern `gdpr/id/`*`CC`* or `gdpr/id/nocontext/`*`CC`*, where *`CC`* is a country code. Some of the entities now have components which provide more information about what was matched. Component information is always available through the `edkmatch` object passed to post-processing Lua scripts, but if you want to see component information in the output from an Eduction Server or the edktool command-line utility, set the configuration parameter `EnableComponents=TRUE`.
- The IDOL GDPR Package includes a script to validate health ID numbers, national ID numbers, and tax 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. For information about using the validation script, see Validate ID Numbers, on page 15.
- The Eduction grammar for telephone numbers has been improved, to increase accuracy and precision.
- The Eduction grammar for names has been improved to increase precision. In addition, the single 'first name' and 'surname' entities are no longer exposed to avoid single-name false positives.
- The relevance values of various entity types have been improved to allow more accurate thresholding.

Country and Language Support

The IDOL GDPR Package contains grammars and IDX files that apply to data from several 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
be	Belgium
bg	Bulgaria
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
hu	Hungary
ie	Ireland
is	Iceland
it	Italy

Country Code	Country
li	Liechtenstein
lt	Lithuania
lu	Luxembourg
lv	Latvia
mt	Malta
nl	Netherlands
no	Norway
pl	Poland
pt	Portugal
ro	Romania
se	Sweden
si	Slovenia
sk	Slovakia

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
cat	Catalan
cze	Czech
dan	Danish
dut	Dutch
eng	English
est	Estonian
fin	Finnish

Language Code	Language
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
rum	Romanian
slo	Slovak
slv	Slovenian
spa	Spanish
swe	Swedish

IDOL Eduction Grammars

The following section describes the GDPR Eduction grammars available in the IDOL GDPR 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 GDPR grammars with Eduction, you must have a license that enables them. To obtain a license, contact Micro Focus Support.

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 negative (that is, it misses fewer matches).

If you configure Eduction to use both versions of the entity, it gives the context-based entities a higher score in the results.

GDPR Grammar Reference

The following table describes the grammar files that are available in the GDPR 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 9.
- the abbreviation LLL refers to a three-letter language code. For a list of available languages, see Languages, on page 10.

File	Entity	Description
address.ecr	gdpr/address/CC	A postal address.
date.ecr	gdpr/date/dob/LLL	A date of birth, written numerically or using words.
	gdpr/date/nocontext/LLL	A calendar date, written numerically or using words, without context.
driving.ecr	gdpr/driving/CC	A driving license number with context.
	gdpr/driving/nocontext/CC	A driving license number, without context.
health.ecr	gdpr/health/ehic/CC	An EHIC personal identification number with context.
	gdpr/health/ehic/nocontext/CC	An EHIC personal identification number without context.
	gdpr/health/tarjeta_sanitaria/es	A Spanish health insurance card number with context.
	gdpr/health/nhs/gb	A British NHS number with context.
	gdpr/health/carte_vitale/fr	A French Carte Vitale number with context.
mrtd.ecr	gdpr/mrtd/mrp	A machine readable passport.
	gdpr/mrtd/mrotd/td1	A machine readable TD1-size travel document.

File	Entity	Description
name.ecr	gdpr/name/CC	A full personal name.
national_id.ecr	gdpr/id/CC	A national identity number with context.
	gdpr/id/nocontext/CC	A national identity number without context.
passport.ecr	gdpr/passport/CC	A passport number with context.
	gdpr/passport/nocontext/CC	A passport number without context.
postcode.ecr	gdpr/postcode/CC	A postal code with context.
	gdpr/postcode/nocontext/CC	A postal code without context.
telephone.ecr	gdpr/telephone/CC	A telephone number with context. **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 the next page.
	gdpr/telephone/nocontext/CC	A telephone number without context. **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 the next page.
tin.ecr	gdpr/tin/CC	A tax identification number with context.
	gdpr/tin/nocontext/CC	A tax identification number without context.

Configure Tangible Characters

`TangibleCharacters` is a configuration parameter than 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 GDPR Package Eduction Grammars require tangible characters to be set in order to perform correctly (see the descriptions of the entities in GDPR Grammar Reference, on page 13).

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*.

Validate ID Numbers

The IDOL GDPR Package includes a script 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.

To run the validation script, add a post-processing task to your Eduction configuration. For example:

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

[ValidateUsingChecksum]
Type=Lua
```

```
Script=scripts/gdpr_postprocessing.lua
Entities=gdpr/*
```

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

The following tables list the entities that are validated by the script.

Health ID numbers (`health.ecr`)
gdpr/health/nhs/gb

National ID numbers (`national_id.ecr`)		
gdpr/id/at	gdpr/id/nocontext/at	Only the SSN component is validated.
gdpr/id/be	gdpr/id/nocontext/be	
gdpr/id/bg	gdpr/id/nocontext/bg	
gdpr/id/cz	gdpr/id/nocontext/cz	
gdpr/id/ee	gdpr/id/nocontext/ee	
gdpr/id/es	gdpr/id/nocontext/es	
gdpr/id/fi	gdpr/id/nocontext/fi	
gdpr/id/fr	gdpr/id/nocontext/fr	
gdpr/id/gr	gdpr/id/nocontext/gr	Only the AMKA component is validated.
gdpr/id/hr	gdpr/id/nocontext/hr	
gdpr/id/hu	gdpr/id/nocontext/hu	Only the PIN component is validated.
gdpr/id/ie	gdpr/id/nocontext/ie	
gdpr/id/is	gdpr/id/nocontext/is	
gdpr/id/it	gdpr/id/nocontext/it	
gdpr/id/lt	gdpr/id/nocontext/lt	
gdpr/id/lu	gdpr/id/nocontext/lu	
gdpr/id/nl	gdpr/id/nocontext/nl	
gdpr/id/no	gdpr/id/nocontext/no	
gdpr/id/pl	gdpr/id/nocontext/pl	
gdpr/id/pt	gdpr/id/nocontext/pt	
gdpr/id/ro	gdpr/id/nocontext/ro	

gdpr/id/si	gdpr/id/nocontext/si	
gdpr/id/se	gdpr/id/nocontext/se	
gdpr/id/sk	gdpr/id/nocontext/sk	

Tax ID numbers (`tin.ecr`)	
gdpr/tin/at	gdpr/tin/nocontext/at
gdpr/tin/be	gdpr/tin/nocontext/be
gdpr/tin/bg	gdpr/tin/nocontext/bg
gdpr/tin/cy	gdpr/tin/nocontext/cy
gdpr/tin/de	gdpr/tin/nocontext/de
gdpr/tin/dk	gdpr/tin/nocontext/dk
gdpr/tin/ee	gdpr/tin/nocontext/ee
gdpr/tin/es	gdpr/tin/nocontext/es
gdpr/tin/fi	gdpr/tin/nocontext/fi
gdpr/tin/fr	gdpr/tin/nocontext/fr
gdpr/tin/hr	gdpr/tin/nocontext/hr
gdpr/tin/hu	gdpr/tin/nocontext/hu
gdpr/tin/ie	gdpr/tin/nocontext/ie
gdpr/tin/it	gdpr/tin/nocontext/it
gdpr/tin/lt	gdpr/tin/nocontext/lt
gdpr/tin/lu	gdpr/tin/nocontext/lu
gdpr/tin/mt	gdpr/tin/nocontext/mt
gdpr/tin/nl	gdpr/tin/nocontext/nl
gdpr/tin/pl	gdpr/tin/nocontext/pl
gdpr/tin/pt	gdpr/tin/nocontext/pt
gdpr/tin/se	gdpr/tin/nocontext/se
gdpr/tin/si	gdpr/tin/nocontext/si
gdpr/tin/sk	gdpr/tin/nocontext/sk

Choose a Threshold to 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"). To do this, matches are given a 'Score' value.

Low-likelihood matches can be removed using the `MinScore` configuration parameter, as entities matching with a Score below this value are not returned. This parameter has a default value of `0.5`, which has been chosen to return only entities with a relatively high likelihood of being a useful match.

For example, an otherwise context-free date ("January 18, 1998") may be returned by the Date Of Birth entity type with a Score of 0.4. But if there is context to suggest that it is indeed a date of birth (for example, "DOB: January 18, 1998") then the Score should be above 0.5.

You should set the `MinScore` parameter in order to achieve the desired balance between precision and recall when matching entities.

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

There is an IDX file for each of the supported languages (see Languages, on page 10).

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 GDPR Package 12.1)

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!