IDOL Government Eduction Package

Software Version 12.10

Technical Note

Document Release Date: October 2021
Software Release Date: October 2021

Legal notices

Copyright 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 Government Eduction Package contains tools that allow you to find governmental document markings and other information in your data, to help you comply with data management restrictions.

The IDOL Government Eduction Package uses IDOL Eduction Grammars, on page 7 (.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 Government Eduction Package describe different kinds of government document markings, so that you can find these in your data.

Data Sources

The IDOL Government Eduction Package contains a variety of different kinds of entities to describe governmental document markings. 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.

Australian Government Email Markings

Australian government email markings information comes from the Australian government Protective Security Policy Framework, documented on https://www.protectivesecurity.gov.au.

Global Legal Entity Identifiers

Legal Entity Identifier data is collected from the Global Legal Entity Identifiers Foundation (https://www.gleif.org). Landmark data has been drawn from public sources, such as Wikipedia.

Export Numbers

Various export identifiers and codes are collected from US Government sources, the Bureau of Industry and Security (https://www.bis.doc.gov/index.php/regulations/) and the International Trade Administration (https://2016.export.gov/faq/eg_main_017509.asp). Some knowledge is also drawn from Wikipedia.

US Government CUI Markings

CUI fields are collected from US government archives, https://www.archives.gov/cui/registry/). The grammar patterns are drawn from Information Security Oversight Office documents.

US Department of Defense Markings

The Department of Defense (DOD) markings are collected from the US Department of Defense Information Security Program: Marking of Information Manual, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/520001m_vol2.pdf?ver=2020-08-04-112507-683.

Country and Language Support

The IDOL Government Eduction Package contains grammars that apply to data from different countries.

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
au	Australia
us	United States of America

IDOL Eduction Grammars

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

The IDOL Government Eduction Package includes a default configuration file, which includes the basic required settings that you need to use the GOV grammars.

> **NOTE:** If you create your own configuration file, you must include some of the settings in the default configuration file, such as post-processing (see Configure Post Processing, below).

Configure Post Processing

When you use the IDOL Government Eduction Package Eduction grammars it is essential to configure a Lua post-processing task to run the script `gov_postprocessing.lua`. This script contains post-processing to normalize scoring.

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

The default configuration file provided in the IDOL Government Eduction Package includes a suitable post-processing task. If you use a different configuration, you must add the post-processing task to your Eduction configuration. For example:

```
[Eduction]
PostProcessingTask0=MyPostProcessingSection

[MyPostProcessingSection]
Type=Lua
Script=scripts/gov_postprocessing.lua
Entities=gov/*
```

For more information about configuring post-processing tasks, 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.

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"). Each match is given a score value so that you can filter the results.

As described in Entity Context, on the previous page, 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 a 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 GOV post-processing script (see Configure Post Processing, on the previous page) includes a step to validate matches (for example, it can validate some ID numbers 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 Government Eduction 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 and Programming 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 Government Eduction Package Eduction Grammars require tangible characters to be set in order to perform correctly (see the descriptions of the entities in Eduction Grammar Reference, below).

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*.

Eduction Grammar Reference

The following tables describe the grammar files that are available in the IDOL Government Eduction package, and the entities that each provides.

au_email_markings.ecr

Entity	Description
gov/document_markings/au_email/caveat/context/au	A security handling caveat, warning of additional protection to the classification level, with context. For example "CAVEAT=SH:NATIONAL-CABINET". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/caveat/nocontext/au	A security handling caveat, warning of additional protection to the classification level, without context. For example "SH:NATIONAL-CABINET".
gov/document_markings/au_email/caveat/landmark/au	A security handling caveat landmark. For example "CAVEAT".

Entity	Description
gov/document_markings/au_email/classification_level/context/au	A security classification level, with context. For example "SEC=PROTECTED". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/classification_level/nocontext/au	A security classification level, without context. For example "PROTECTED".
gov/document_markings/au_email/classification_level/landmark/au	A security classification level landmark. For example "SEC".
gov/document_markings/au_email/deprecated_classification_level/nocontext/au	A classification level marker that was part of a standard which was replaced in Oct 2018, without context. For example "HIGHLY PROTECTED". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/dissemination_limiting_marker/nocontext/au	A security dissemination marker part of a standard which was replaced in Oct 2018, without context. For example " For Official Use Only". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/downto_level/context/au	The classification level after expiration, with context. For example "DOWNTO=OFFICIAL:Sensitive" This entity returns components. See Components, on page 15.
gov/document_markings/au_email/downto_level/landmark/au	A landmark for the classification level after expiration. For example "DOWNTO".
gov/document_markings/au_email/expires/context/au	A classification level expiration date or event, with context. For example "EXPIRES=2019-01-07". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/expires/nocontext/au	A classification level expiration date or event, without context. For example "2019-01-07".
gov/document_markings/au_email/expires/landmark/au	A classification level expiration date or event landmark. For example "EXPIRES".
gov/document_markings/au_email/header_marking/context/au	A complete internet message header extension security marking conforming to the specifications. For example " [SEC=OFFICIAL, CAVEAT=SH:DELICATE SOURCE,

Entity	Description
	ACCESS=Personal-Privacy, ACCESS=Legal-Privilege]". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/information_management_marker/context/au	An information management marker used for non-security related restrictions, with context. For example "ACCESS=Legal-Privilege". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/information_management_marker/nocontext/au	An information management marker used for non-security related restrictions, without context. For example "Legal-Privilege".
gov/document_markings/au_email/information_management_marker/landmark/au	An information management marker landmark. For example "ACCESS".
gov/document_markings/au_email/namespace/context/au	Namespace of terms used in protective marking, with context. For example "NS=gov.au". This entity returns components. See Components, on page 15.
gov/gov/document_markings/au_email/namespace/nocontext/au	Namespace of terms used in protective marking, without context. For example "gov.au".
gov/document_markings/au_email/namespace/landmark/au	A namespace landmark. For example "NS".
gov/document_markings/au_email/note/context/au	A supplementary note or comment, with context. For example "NOTE=a useful comment". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/note/landmark/au	A supplementary note or comment landmark. For example "NOTE".
gov/document_markings/au_email/origin/context/au	Email address of the person who originally classified the email, with context. For example "ORIGIN=jane.doe@example.gov.au". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/origin/nocontext/au	Email address of the person who originally classified the email, without context. For example "jane.doe@example.gov.au".
gov/document_markings/au_	An origin email address landmark. For example "ORIGIN".

Entity	Description
email/origin/landmark/au	
gov/document_markings/au_email/portion_marking/nocontext/au	A marker indicating the security classification level of a portion of a document, without context. For example " (PROTECTED)".
gov/document_markings/au_email/subject_field_marking/context/au	A complete email subject field securiy marking conforming to the specifications. For example "VER=2018.4, NS=gov.au, SEC=TOP SECRET, CAVEAT=RI:REL AUS/USA/FRA, ORIGIN=jane.doe@example.gov.au" This entity returns components. See Components, on page 15.
gov/document_markings/au_email/version/context/au	Version of the protective marking specification used, with context. For example "VERSION=2018.4". This entity returns components. See Components, on page 15.
gov/document_markings/au_email/version/nocontext/au	Version of the protective marking specification used, without context. For example "2018.4".
gov/document_markings/au_email/version/landmark/au	A protective marking specification version landmark. For example "VERSION=".

entity_identifiers.ecr

Entity	Description
gov/entity_identifiers/lei/context	A Legal Entity Identifier, with context. For example "LEI: 4469000001AVO26P9X86" or "RECHTSTRÄGER-KENNUNG: EVK05KS7XY1DEII3R011".
gov/entity_identifiers/lei/nocontext	A Legal Entity Identifier, without context. For example "558600FNC30A8J9EGQ54" or "378900F4A0A690EA6735".
gov/entity_identifiers/lei/landmark	A Legal Entity Identifier landmark, in many languages. For example "Legal Entity Identifier" or "とりひきしゅたいしきべつし".

us_dod_markings.ecr

Entity	Description
gov/document_markings/us_dod/banner_line/nocontext/us	Banner lines in US DoD (Department of Defense) documents. These banner lines appear in the

Entity	Description
	header and footer of every page in the document. For example "(S//REL TO USA, GBR/DISPLAY ONLY AFG)".
gov/document_markings/us_dod/portion_marking/nocontext/us	Portion markings in US DoD documents. Portion markings are used to mark individual sentence, sections and paragraphs. For example "TOP SECRET//SI/TK//RELIDO".
gov/document_markings/us_dod/classification_authority_block/declassify/nocontext/us	The 'Declassify on' directive in the classification authority block in US DoD documents. For example "Declassify On: 25X 20320109"
gov/document_markings/us_dod/classification_authority_block/downgrade/nocontext/us	The 'Downgrade to' directive in the classification authority block in US DoD documents. For example "Downgrade To: CONFIDENTIAL on 20121231".
gov/document_markings/us_dod/classification_authority_block/reason/nocontext/us	The 'Reason' directive in the classification authority block in US DoD documents. For example "Reason: 1.4(a)(b)(c)".
gov/document_markings/us_dod/classification_authority_block/classified_by_header/nocontext/us	The 'Classified by' header in the classification authority block in US DoD documents. For example "Classified By:".
gov/document_markings/us_dod/classification_authority_block/derived_from_header/nocontext/us	The 'Derived from' header in the classification authority block in US DoD documents. For example "Derived From:".

us_cui_markings.ecr

Entity	Description
gov/document_markings/us_cui/basic/nocontext/us	The basic CUI (Controlled Unclassified Information) marking. For example "CONTROLLED//TSCA//NOFORN". This entity returns components. See Components, on page 15.
gov/document_markings/us_cui/specified/nocontext/us	The specified CUI marking. For example "CUI//SP-INTEL/SP-CTI/IFNC/AG//FEDCON/REL TO USA, GBR". This entity returns components. See Components, on page 15.
gov/document_markings/us_cui/dissemination_list/nocontext/us	The list of dissemination targets, marked by a landmark such as "Dissemintation List". For example "Dissemination List: Office of Intelligence and Analysis, Department of Homeland Security

Entity	Description
	Counterterrorism Division, Federal Bureau of Investigation." This value is normalized to "Office of Intelligence and Analysis, Department of Homeland Security Counterterrorism Division, Federal Bureau of Investigation."
gov/document_markings/us_cui/controlled_by/nocontext/us	The name of the department that a document is controlled by, marked by a landmark such as "Controlled by". For example "Controlled by: Department of Good Works, Security and Inspection Division, 2025554567." This value is normalized to "Department of Good Works, Security and Inspection Division, 2025554567."

number_export_us.ecr

Entity	Description
gov/number/export/eccn/context/us	The Export Control Classification Number (ECCN), with context. For example "ECCN: 5A002.a.1".
gov/number/export/eccn/nocontext/us	The ECCN, without context. For example "5A002.a.1".
gov/number/export/eccn/landmark/us	An ECCN landmark. For example "ECCN"
gov/number/export/ccl/context/us	The Commerce Control List (CCL) designation, with context. (The CCL is a superset of the ECCN.) For example "CCL: Annex to Cat 1 List of Explosives 45".
gov/number/export/ccl/nocontext/us	The CCL designantion, without context. For example "Annex to Cat 1 List of Explosives 45"
gov/number/export/ccl/landmark/us	A CCL landmark.
gov/number/export/ear_exception_enc/context/us	An Export Administration Regulations (EAR) exceptions for encryption (ENC) number, with context. For example "License Exception: 740.17(b)(2)".
gov/number/export/ear_exception_enc/nocontext/us	An EAR exceptions for ENC number, without context. For example "740.17(b)(2)".
gov/number/export/ear_exception_enc/landmark/us	An EAR exceptions for ENC landmark. For example "License Exception".
gov/number/export/ccats/context/us	A Commodity Classification Automated Tracking System (CCATS) code, with context. For example "CCAT: G144401".
gov/number/export/ccats/nocontext/us	A CCATS code, without context. For example "G144401".

Entity	Description
gov/number/export/ccats/landmark/us	A CCATS code landmark. For example "CCAT".
gov/number/export/hts/context/us	A Harmonized Tariff Schedule (HTS) code, with context. For example "Harmonized Tariff (US): 8542.31.0000".
gov/number/export/hts/nocontext/us	An HTS code, without context. For example "8542.31.0000".
gov/number/export/hts/landmark/us	An HTS code landmark. For example "Harmonized Tariff (US)".
gov/number/export/schedule_b/context/us	A Schedule B code, with context. For example "U.S. Schedule B: 8542.31.0000".
gov/number/export/schedule_b/nocontext/us	A Schedule B code, without context. For example "8542.31.0000".
gov/number/export/schedule_b/landmark/us	A Schedule B code landmark. For example "U.S. Schedule B"

Components

Some of the GOV entities extract *components* as well as whole matches. Components are parts of a match that can provide useful information.

The following sections list the components available for particular entities.

Australian Email Markings Components

Component Name	Notes
CLASSIFICATION_LEVEL	
DOWNTO_CLASSIFICATION_LEVEL	
DEPRECATED_CLASSIFICATION_LEVEL	
DISSEMINATION_LIMITING_MARKER	
CAVEAT	
NFORMATION_MANAGEMENT_MARKER	
EXPIRY_DATE	
EXPIRY_EVENT	

Component Name	Notes
VERSION	
NAMESPACE	
NOTE	
ORIGIN	

The following examples demonstrate the use of these components.

- VER=2018.4, NS=gov.au, SEC=TOP SECRET, CAVEAT=RI:REL AUS/USA/FRA, ORIGIN=jane.doe@example.gov.au

  ```
  VERSION: 2018.4
  NAMESPACE: gov.au
  CLASSIFICATION_LEVEL: TOP SECRET
  CAVEAT: RI:REL
  ORIGIN: jane.doe@example.gov.au
  ```

- [SEC=OFFICIAL, CAVEAT=SH:DELICATE SOURCE, ACCESS=Personal-Privacy, ACCESS=Legal-Privilege]

  ```
  CLASSIFICATION_LEVEL: OFFICIAL
  CAVEAT: SH:DELICATE SOURCE
  INFORMATION_MANAGEMENT_MARKER: Personal-Privacy
  INFORMATION_MANAGEMENT_MARKER: Legal-Privilege
  ```

- [SEC=PROTECTED, EXPIRES=2019-07-01, DOWNTO=OFFICIAL]

  ```
  CLASSIFICATION_LEVEL: PROTECTED
  EXPIRY_DATE: 2019-07-01
  DOWNTO_CLASSIFICATION_LEVEL: OFFICIAL
  ```

- [SEC=PROTECTED, EXPIRES=Legislation Published, DOWNTO=OFFICIAL, NOTE=Some comment]

  ```
  CLASSIFICATION_LEVEL: PROTECTED
  EXPIRY_EVENT: Legislation Published
  DOWNTO_CLASSIFICATION_LEVEL: OFFICIAL
  NOTE: Some comment
  ```

- `DEPRECATED_CLASSIFICATION_LEVEL: HIGHLY PROTECTED`
- `DISSEMINATION_LIMITING_MARKER: For Official Use Only`

US CUI Markings Components

Component Name	Notes
CUI_CATEGORY	A basic or specified CUI category.
CUI_DISSEMINATION	A dissemination target in a dissemination list.

The following examples demonstrate the use of these components.

- CONTROLLED//TSCA//NOFORN

  ```
  CUI_CATEGORY: TSCA
  CUI_DISSEMINATION: NOFORN
  ```

- CUI//SP-INTEL/SP-CTI/IFNC/AG//FEDCON/REL TO USA, GBR

  ```
  CUI_CATEGORY: SP-INTEL
  CUI_CATEGORY: SP-CTI
  CUI_CATEGORY: IFNC
  CUI_CATEGORY: AG
  CUI_DISSEMINATION: FEDCON
  CUI_DISSEMINATION: REL TO USA, GBR
  ```

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 Micro Focus IDOL Government Eduction Package 12.10 Technical Note

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!