PUBLIO SEARCH

Search Specifications

Customer: Publications Office
Contract reference: FC 10666
Subject: PUBLIO Search Specifications
Version: 1.01
Document reference: PUBLIO Search Specifications
## DOCUMENT REVISION HISTORY

### DOCUMENT HISTORY

<table>
<thead>
<tr>
<th>Version</th>
<th>Release Date</th>
<th>Author</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>06/09/2022</td>
<td></td>
<td>Initial version</td>
</tr>
<tr>
<td>1.01</td>
<td>16/09/2022</td>
<td></td>
<td>Updated after feedback from OP and Deloitte</td>
</tr>
</tbody>
</table>

### REVIEWS

<table>
<thead>
<tr>
<th>Version</th>
<th>Release Date</th>
<th>Reviewer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>06/09/2022</td>
<td></td>
<td>Delivery Manager</td>
</tr>
<tr>
<td>1.01</td>
<td>16/09/2022</td>
<td></td>
<td>Delivery Manager</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>WP1 – TE9 – Export periodically all authors</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>WP1 – TE9 – Update of the Autocomplete authors</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Appendix</td>
<td>9</td>
</tr>
<tr>
<td>4.1</td>
<td>Test scenarios</td>
<td>9</td>
</tr>
<tr>
<td>4.1.1</td>
<td>WP1 – TE9 – Export periodically all authors</td>
<td>9</td>
</tr>
<tr>
<td>4.1.2</td>
<td>WP1 – TE9 – Update of the Autocomplete authors</td>
<td>9</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: Glossary of terms............................................................................................................................................ 5
Table 2: TE9 Test Export Periodically all Authors ........................................................................................................ 9
Table 3: TE9 Test Update of the Autocomplete Authors.............................................................................................. 9
# 1 INTRODUCTION

Work Package 1 (WP1) and Work Package 3 (WP3) are two of the three work packages planned for the augmentation of the Publico assistant in 2022.

WP1 and WP3 include developments on the Search layer that will be discussed in this document. For each development we will cover the product specifications and technical details.

## Glossary of terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author</strong></td>
<td>Refers to the authors of the documents available in the Publications Office collections. It includes both institutional and private authors. The Author Entity is the component LUIS uses to extract an author value from a user utterance.</td>
</tr>
<tr>
<td><strong>Entity</strong></td>
<td>Entities allow to extract key data from user utterances in Language Understanding (LUIS) apps. Extracted entity data is used by the client application to fulfill customer requests. The entity represents a word or phrase inside the utterance that we want extracted. Entities describe information relevant to the intent, and sometimes they are essential for applications to perform their task.</td>
</tr>
<tr>
<td><strong>LUIS</strong></td>
<td>Language Understanding (LUIS) is a cloud-based conversational AI service that applies custom machine-learning intelligence to a user’s conversational, natural language text to predict overall meaning, and pull out relevant, detailed information. LUIS provides access through its custom portal, APIs and SDK client libraries</td>
</tr>
<tr>
<td><strong>MUL</strong></td>
<td>Language ISO code for multiple languages</td>
</tr>
<tr>
<td><strong>Transliterations</strong></td>
<td>Transliteration is a type of conversion of a text from one script to another that involves swapping letters (thus trans- + liter-) in predictable ways. In search it has for purpose to normalize letters into a unified and limited set of letters. Removal of diacritics is for example part of the transliteration mechanisms (removal of glyph added to a letter)</td>
</tr>
</tbody>
</table>

*Table 1: Glossary of terms*
2 WP1 – TE9 – EXPORT PERIODICALLY ALL AUTHORS

The Publio assistant requires the publications authors to train them as a machine learned entity in LUIS.

The corporate authors and the personal authors will be exported from the publications index as a JSON file and stored on an Azure Blob storage accessible by the Publio assistant.

The JSON will contain:

- Corporate authors and personal authors part of the publications index.
- The labels, short labels and acronyms.
- The information if the corporate author is deprecated.

The export will be configured to automatically run every week on Saturday at 06:00 UTC.

Example of authors:

```json
{
    "resource": "corporate-body.PUBL",
    "labels": {
        "HUN": "Az Európai Unió Kiadóhivatala",
        "LAV": "Eiropas Savienības Publikāciju birojs",
        "SWE": "Europeiska unionens publikationsbyrå",
        "GLE": "Oifig Foilseacháin an Aontais Eorpaigh",
        "ELL": "Υπηρεσία Εκδόσεως της Ευρωπαϊκής Ένωσης",
        "SPA": "Oficina de Publicaciones de la Unión Europea",
        "EST": "Euroopa Liidu Väljaannete Talitus",
        "NLD": "Bureau voor publicaties van de Europese Unie",
        "FIN": "Euroopan unionin julkaisutoimisto",
        "POL": "Urząd Publikacji Unii Europejskiej",
        "HRV": "Ured za publikacije Europske unije",
        "CES": "Úřad pro publikace Evropské unie",
        "DAN": "Den Europæiske Unions Publikationskontor",
        "RON": "Oficiul pentru Publicații al Uniunii Europene",
        "POR": "Serviço das Publicações da União Europeia",
        "SLK": "Úrad pre vydávanie publikácií Európskej únie",
        "FRA": "Office des publications de l’Union européenne",
        "MLT": "Uffiċċju tal-Pubblikazzjonijiet tal-Unjoni Ewropea",
        "DEU": "Amt für Veröffentlichungen der Europäischen Union",
        "LIT": "Europos Sąjungos leidinių biuras",
        "ITA": "Ufficio delle pubblicazioni dell’Unione europea",
        "BUL": "Служба за публикации на Европейския съюз",
        "SLV": "Urad za publikacije Evropske unije",
        "ENG": "Publications Office of the European Union"
    }
}
```
3 WP1 – TE9 – UPDATE OF THE AUTOCOMPLETE AUTHORS

The autocomplete author endpoint is currently used in the OP Portal advanced search to help the user finding an author. It is built by extracting all the current authors (personal and corporate) associated to the publications index.

The autocompletion has the following capabilities:

- Support the completion on acronyms and labels that are concatenated “OP – Publications Office of the European Union”
- Support of the transliterations.

The Publio assistant will have to check if an author is valid by providing the label, the short label, or the acronyms. The autocomplete authors will be used for that check. To fulfil the assistant requirement, the support of the short label will be added to the autocomplete authors.

Short label will be proposed as an autocomplete suggestion following the same principle than the labels: the acronym will be prepended to the short label. For example, for the “Directorate-General for Budget” having for acronym “BUDG”:

- Label: “BUDG - Directorate-General for Budget”.
- Short label: “BUDG – DG Budget”.

"SLK" : "OP",
"FRA" : "OP",
"MLT" : "OP",
"DEU" : "OP",
"LIT" : "OP",
"ITA" : "OP",
"BUL" : "OP",
"SLV" : "OP",
"ENG" : "OP"
},
"deprecated" : false
}
4 APPENDIX

4.1 TEST SCENARIOS

4.1.1 WP1 – TE9 – EXPORT PERIODICALLY ALL AUTHORS

<table>
<thead>
<tr>
<th>Step</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Verify the file is exported weekly and it contains the corporate and personal authors</td>
</tr>
<tr>
<td>2</td>
<td>Verify that the cardinality of corporate authors and personal authors are equals to the publications index. This verification will happen right after the weekly export to ensure a consistent comparison.</td>
</tr>
<tr>
<td>3</td>
<td>Verify by sampling (10 checks on corporate and personal) that authors contain the information defined in the publications index</td>
</tr>
</tbody>
</table>

_Table 2: TE9 Test Export Periodically all Authors_

4.1.2 WP1 – TE9 – UPDATE OF THE AUTOCOMPLETE AUTHORS

<table>
<thead>
<tr>
<th>Step</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Verify that an author can be search by its short label with the acronym prepended.</td>
</tr>
</tbody>
</table>

_Table 3: TE9 Test Update of the Autocomplete Authors_