

Extracting structure from non-standard texts

Gábor Recski
TU Wien

`gabor.recski@tuwien.ac.at`

ReMeP Workhsop on Inclusion of Non-Standard Sources in
Legal Searches
06/23/2020

Non-standard sources

18. Verordnung der Kommunikationsbehörde Austria (KommAustria) zur näheren Festlegung der Auswahlgrundsätze für die Erteilung terrestrischer Multiplex-Zulassungen für digitalen Hörfunk im Standard DAB+ 2017 (MUX-Auswahlgrundsätzeverordnung DAB+ 2017 – MUX-AG-V DAB+ 2017)

Aufgrund des § 15a Abs. 2 und 3 in Verbindung mit § 31 Abs. 2 Privatradiogesetz (PrR-G), BGBl. I Nr. 20/2001 idF BGBl. I Nr. 86/2015, wird verordnet:

Gegenstand

§ 1. Diese Verordnung legt die Auswahlgrundsätze gemäß § 15a Abs. 1 Privatradiogesetz (PrR-G), BGBl. I Nr. 20/2001 idF BGBl. I Nr. 86/2015, und die erforderlichen Unterlagen zur Glaubhaftmachung der Erfüllung der finanziellen Voraussetzungen nach § 15 Abs. 2 PrR-G für die Erteilung von Zulassungen zu Errichtung und Betrieb von terrestrischen Multiplex-Plattformen zur Verbreitung von digitalem terrestrischem Hörfunk im Standard DAB+ näher fest.

Begriffsbestimmungen

§ 2. Im Sinne dieser Verordnung bezeichnet der Ausdruck

1. Ballungsraum: ein städtisches Siedlungsgebiet mit einer Bevölkerung von mehr als 30.000 Einwohnern;
2. DAB+: einen Übertragungsstandard für digitalen Hörfunk entsprechend der aktuellen Spezifikationen des European Telecommunications Standards Institute;
3. Simulcast: die zeitgleiche Ausstrahlung eines Hörfunkprogramms im selben Gebiet auf digital terrestrischem und analog terrestrischem Weg;
4. Zusatzdienst: einen Dienst, der zusätzlich zu den angebotenen Hörfunkprogrammen über die Multiplex-Plattform verbreitet wird;
5. Kapazitätseinheit: eine rechnerische Größe zur Bestimmung der Anzahl der verbreitbaren Programme und Zusatzdienste über eine Multiplex-Plattform.

Jahrgang 2020

Kundgemacht am 29. Mai 2020

www.stadt-salzburg.at

49. Kundmachung

Wirtschaftsförderungen - Richtlinie für die Nahversorgerförderung

GZ: MD/04/47600/2018/005

Wirtschaftsförderungen Richtlinie für die Nahversorgerförderung

Kundmachung

Der Gemeinderat der Landeshauptstadt Salzburg hat am 13. Mai 2020 die Richtlinie zur Gewährung der Nahversorgerförderung beschlossen.

Richtlinien

Der Landeshauptstadt Salzburg für die Nahversorgerförderung

1. Förderungsziel

Verordnung über die versicherungsmathematische Bestätigung, den Erläuterungsbericht und den Angemessenheitsbericht des Verantwortlichen Aktuars (Aktuarverordnung - AktuarV)

[Nichtamtliches Inhaltsverzeichnis](#)

AktuarV

Ausfertigungsdatum: 18.04.2016

Vollzeit:

Aktuarverordnung vom 18. April 2016 (BGBl. I S. 776), die durch Artikel 6 Absatz 8 des Gesetzes vom 19. Dezember 2018 (BGBl. I S. 2672) geändert worden ist

Stand: Geändert durch Art. 6 Abs. 8 G v. 19.12.2018 | 2672

Näheres zur Standangabe finden Sie im Menü unter [Hinweise](#)

Fußnote

(+++ Textnachweis ab: 22.4.2016 +++)
(+++ Zur Anwendung vgl. §§ 1, 7 +++)

[Nichtamtliches Inhaltsverzeichnis](#)

Eingangsformel









Auf Grund des § 145 Absatz 4 in Verbindung mit Absatz 5 Satz 2, auch in Verbindung mit § 161 Absatz 1, § 162, § 212 Absatz 1, § 219 Absatz 1 und § 234 Absatz 1 des Versicherungsaufsichtsgesetzes vom 1. April 2015 (BGBl. I S. 434), von denen § 161 Absatz 1 durch Artikel 14 Nummer 2 des Gesetzes vom 2. November 2015 (BGBl. I S. 1864) geändert worden ist, verordnet das Bundesministerium der Finanzen:

[Nichtamtliches Inhaltsverzeichnis](#)

§ 1 Anwendungsbereich



But we want structure!

Suchworte	<input type="text"/>
Titel, Abkürzung	<input type="text"/>
Paragraf von	<input type="text"/> bis <input type="text"/>
Artikel von	<input type="text"/> bis <input type="text"/>
Anlage von	<input type="text"/> bis <input type="text"/>
Kundmachungsorgan	<input type="text"/>  Nr. <input type="text"/>
Typ	<input type="text"/>
Index	<input type="text"/>
Unterzeichnungsdatum	<input type="text" value="TT.MM.JJJJ"/> 
	<input checked="" type="radio"/> Suche nach Fassung
Fassung vom	<input type="text" value="22.06.2020"/> 
	<input type="radio"/> Suche nach zeitlichem Geltungsbereich
Inkrafttretensdatum von	<input type="text"/>  bis <input type="text"/> 
Außerkrafttretensdatum von	<input type="text"/>  bis <input type="text"/> 
Neu/geändert im RIS seit	<input type="text"/> 

And all we have is this:

```
42 5 0 obj^M
43 <</Type/Font/Subtype/TrueType/Name/F1/BaseFont/Times#20New#20Roman/Encoding/WinAnsiEncoding
44 endobj^M      5 <title>AktuarV - Verordnung über die versicherungsmathematische Bestätigung, den Erläuter
45 6 0 obj^M      .css" rel="stylesheet" type="text/css" media="screen">
46 <</Type/FontDes 6 <link href="./bafin_files/print_aiz_einzel.css" rel="stylesheet" type="text/css" media="p
568 -216 2000 € 7 <link href="./bafin_files/jnhtml.css" rel="stylesheet" type="text/css">
47 endobj^M      8 </head>
48 7 0 obj^M      9 <body>
49 <</Type/Font/Su10 <a name="Seitenanfang"></a>
50 endobj^M      11 <div id="fTop">
51 8 0 obj^M      12 <div id="fTop1"></div>
52 <</Type/FontDes13 <div id="fTop2"></div>
ox[ -558 -216 214 </div>
53 endobj^M      15 <div id="fTopLogos"><area shape="rect" alt="Logo Bun
55 <</Type/Font/Su erbraucherschutz, die Seite wird in einem eigenen Fenster angezeigt." coords="1,0,214,136
56 endobj^M      s zum Bundesamt für Justiz, die Seite wird in einem eigenen Fenster angezeigt." coords="2
57 10 0 obj^M      17 </div>
58 [ 11 0 R ] ^M 18 <div id="nav_12793">
59 endobj^M      19 <ul>
60 11 0 obj^M      20 <li id="grDarst5"><a href="https://gesetze-im-internet.questionpro.eu/" title="Die Ku
61 <</BaseFont/Tir t von Gesetzen</a></li>
62 endobj^M      21 </ul>
63 12 0 obj^M      22 </div>
64 <</Ordering(Ide23 <div id="level2">
65 endobj^M      24 <div id="content_12793">
66 13 0 obj^M      25 <div id="container">
67 <</Type/FontDes26 <div id="paddingLR12">
ox[ -558 -216 227
68 endobj^M      28 <div class="jnnorm" id="BJNR077600016" title="Rahmen"><div class="jnheader"> <h1><span cl
69 14 0 obj^M      emessenheitsbericht des Verantwortlichen Aktuars</span> <span class="jnkurzueamtabk"> (Ak
70 <</Type/Font/Su tamtliches Inhaltsverzeichnis</a><p>AktuarV</p><p>Ausfertigungsdatum: 18.04.2016</p> </di
71 endobj^M      tikel 6 Absatz 8 des Gesetzes vom 19. Dezember 2018 (BGBl. I S. 2672) geändert worden ist
72 15 0 obj^M      "90%"></colgroup><tbody><tr><td class="StandangabeStand">Stand:</td><td>Geändert durch Ar
er <a href="https://www.gesetze-im-internet.de/hinweise.html#stand">Hinweise</a><br></div
29 <div class="infussnote" id="BJNR077600016 FN"><h4>Fußnote</h4><div><div class="inhtml"><d
```

Structure extraction steps

Structure extraction steps

Which parts do we keep?

Structure extraction steps

Which parts do we keep?

layout detection, OCR, boilerplate removal, deduplication

Structure extraction steps

Which parts do we keep?

layout detection, OCR, boilerplate removal, deduplication

Where are sections, paragraphs, sentences, words?

Structure extraction steps

Which parts do we keep?

layout detection, OCR, boilerplate removal, deduplication

Where are sections, paragraphs, sentences, words?

document segmentation, sentence splitting, tokenization

Structure extraction steps

Which parts do we keep?

layout detection, OCR, boilerplate removal, deduplication

Where are sections, paragraphs, sentences, words?

document segmentation, sentence splitting, tokenization

What's it about?

Structure extraction steps

Which parts do we keep?

layout detection, OCR, boilerplate removal, deduplication

Where are sections, paragraphs, sentences, words?

document segmentation, sentence splitting, tokenization

What's it about?

keyword extraction, named entity recognition, event detection

Structure extraction steps

Which parts do we keep?

layout detection, OCR, boilerplate removal, deduplication

Where are sections, paragraphs, sentences, words?

document segmentation, sentence splitting, tokenization

What's it about?

keyword extraction, named entity recognition, event detection

Any of these can be trivial but also very hard!

General advice

Data/NLP specialists should rank tasks by **doability/risk**

Data/NLP specialists should rank tasks by **doability/risk**

Domain experts should rank tasks by **value/potential**

Data/NLP specialists should rank tasks by **doability/risk**

Domain experts should rank tasks by **value/potential**

Everyone should spend some time **looking at the data**

Data/NLP specialists should rank tasks by **doability/risk**

Domain experts should rank tasks by **value/potential**

Everyone should spend some time **looking at the data**

Assumptions should be **tested early, tested often**

Thank you for your attention!

E-mail: `gabor.recski@tuwien.ac.at`