Unstructured → Structured Data Conversion

...

Story structure

1. Introduction and description of the problem
2. How are the others approaching the problem?
3. Our approach
    a. The focus area in the targeted document
    b. Identifying the keys: Frequency Analysis
    c. Identifying the values: Exploiting the document structure → html element paths
    d. Improving the precision: Data Rescue

The ideal world: case 1

Title of the contract	Building a new bridge in Cambridge
CPV	90291010 - Construction work
Estimated price	£ 1 000 000 excluding taxes
Estimated date	10/03/2017

The ideal world: case 2

Section 1: Title and description of the contract

1.1 Title of the contract

Building a new bridge in Cambridge

1.2 ...

Section 2: Estimated price and value

...

The ideal world: case 3

Title of the contract: Building a new bridge in Cambridge

Value of the contract = £1 000 000

...

But the world is rarely ideal …

PESQUISA > CONTRATO

Detalhe do Contrato

Imprimir

Data de publicação no BASE	11-02-2016
Tipo(s) de contrato	Aquisição de bens móveis
Tipo de procedimento	Ajuste directo
Descrição	Aquisição de 1 ecógrafo para o Serviço de Cirurgia Vascular do Hospital Divino Espírito Santo de Ponta Delgada, EPER
Fundamentação	Artigo 20.º, n.º 1, alínea a) do Código dos Contratos Públicos
Fundamentação da necessidade de recurso ao ajuste direto (se aplicável)	ausência de recursos próprios
Entidade adjudicante - Nome, NIF	Hospital do Divino Espírito Santo de Ponta Delgada, E. P. E. R. (512103097)
Entidade adjudicatária - Nome, NIF	General Electric Portuguesa, SA (500357129)
Objeto do Contrato	Aquisição de 1 ecógrafo para o Serviço de Cirurgia Vascular do Hospital Divino Espírito Santo de Ponta Delgada, EPER
Procedimento Centralizado	-
CPV	33112000-8, Equipamento de imagiologia por ecos, ultra-sons e doppler
Data de celebração do contrato	28-01-2016
Preço contratual	19.750,00 €
Prazo de execução	40 dias (1 mês e 9 dias)
Local de execução - País, Distrito, Concelho	Portugal, Região Autónoma dos Açores, Ponta Delgada
Concorrentes	-
Anúncio	-
Incrementos superiores a 15%	-
Documentos	Decreto_Legislativo_RegionalN.15-2009-A.pdf
Observações	-

Execução do Contrato

Data de fecho do contrato	-
Preço total efetivo	-
Causas das alterações ao prazo	-
Causas das alterações ao preço	-

Support

Do you have any questions or do you need support?
Frequently asked questions
Contact Simap-Helpdesk

Modify the search

You can modify your search criterias or refine the search

Keyword:

Place of contract performance:

Periode of time:

- Today
- Current week
- Current year
- from
- to
- Total (last 3 years)

Search

Your results

You searched for: Date of publication : 08.02.2016 - 11.02.2016

Registration: Registration

New search « previous notice | next notice » | back to list

10.02.2016 | Project ID 136329 | Notice no. 901753 | Invitation to tender

Appel d'offres

Date de publication Simap : 10.02.2016

1. Pouvoir adjudicateur

1.1 Nom officiel et adresse du pouvoir adjudicateur
Service demandeur/Entité adjudicatrice : Office fédéral des routes - Infrastructure routière Ouest Filiale Estavayer-le-Lac
Service organisateur/Entité organisatrice : Office fédéral des routes - Infrastructure routière Ouest
Filiale Estavayer-le-Lac, à l'attention de Gestion des projets, Place de la Gare 7, 1470 Estavayer-le-Lac, Suisse, Téléphone: +41 58 461 87 11, Fax: +41 58 461 87 90, E-mail: marchespublics.estavayer@astra.admin.ch

1.2 Les offres sont à envoyer à l'adresse suivante
Office fédéral des routes - Infrastructure routière Ouest
Filiale Estavayer-le-Lac, à l'attention de N09 Ref.PS Beausite, PS Colondalles et Baye de Clarens - Mandataire APR, Place de la Gare 7, 1470 Estavayer-le-Lac, Suisse, Téléphone: +41 58 461 87 11, Fax: +41 58 461 87 90, E-mail: marchespublics.estavayer@astra.admin.ch

1.3 Délai souhaité pour poser des questions par écrit
28.02.2016
Remarques : Les questions doivent être formulées de manière anonyme sur le forum de Simap (www.simap.ch) de la soumission concernée. Les réponses seront données exclusivement par l'intermédiaire de cette même plate-forme jusqu'au 04.03.2016. Le téléchargement des réponses relève de la responsabilité exclusive des soumissionnaires. Il ne sera envoyé aucun avertissement. Les questions reçues hors délai ne seront pas traitées.

1.4 Délai de clôture pour le dépôt des offres
Date: 21.03.2016 **Heure:** 23:59, **Délais spécifiques et exigences formelles:** Dossier complet sur support papier (2 exemplaires) et numérique sur 2 clés USB dans une enveloppe cachetée portant le numéro / l'intitulé officiel du projet avec la mention «Ne pas ouvrir, documents d'appel d'offres».

En cas d'envoi postal (au moins en courrier A), le timbre postal ou le code barres de l'office de la Poste suisse ou du bureau de poste étranger officiellement reconnu déterminent si le délai de remise est respecté (l'affranchissement effectué par les entreprises n'est pas considéré comme un timbre postal).

En cas de remise en mains propres, l'offre doit être déposée à la loge de l'OFROU, filiale d'Estavayer-le-Lac, au plus tard dans le délai indiqué ci-dessus, pendant les heures d'ouverture (horaires : 8h00 - 12h00 et 13h30 - 17h00), contre remise d'un accusé de réception (adresse au point 1.2).

Les offres reçues par fax ou par courriel ne seront pas prises en compte.

Dans tous les cas, il incombe au soumissionnaire d'apporter la preuve qu'il a remis son offre dans les délais. Les offres déposées hors délai ne sauraient être prises en considération et sont renvoyées à leur expéditeur sans avoir été ouvertes.

1.6 Genre de pouvoir adjudicateur
Confédération (Administration fédérale centrale)

1.7 Mode de procédure choisi
Procédure ouverte

1.8 Genre de marché

ODDÍL 1: IDENTIFIKACE ZADAVATELE

Předvyplnit údaje o subjektu podle názvu

Název*	VOP CZ, s.p.				
Poštovní adresa*	Dukelská 102				
Obec*	Šenov u Nového	**PSČ**	742 42	**Stát***	CZ
Kontaktní místa	Dukelská 102			**Telefon**	+420 55678322
K rukám	Jana Gablerová				
E-mail (je-li k dispozici)	verejne.zakazky@vop.cz			**Fax**	+420 55670173

ODDÍL 2: IDENTIFIKACE VEŘEJNÉ ZAKÁZKY NA ZÁKLADĚ KTERÉ BYLA RÁMCOVÁ SMLOUVA UZAVŘENA

Název veřejné zakázky přidělený zadavatelem*	Dodávky plechů 019/3/2014		
Evidenční číslo veřejné zakázky ve VVZ*	493831		
Zakázka dělená na části (§ 98 ZVZ)	■ Ano	**Počet částí zakázky**	

ODDÍL 3: IDENTIFIKACE RÁMCOVÉ SMLOUVY

Název rámcové smlouvy přidělený zadavatelem*	Dodávky plechů - VZ dělená na 9 částí, rámcové smlouvy uzavřeny s více účastníky: RS 199 - 207/2014 /V/3/3/ŘÚF-150 Části VZ jsou rozepsány v přílohách - Oddíl IV. Tyto přílohy jsou vyplněny pouze u částí ze kterých bylo v daném kvartálu plněno.
Evidenční číslo rámcové smlouvy přidělené zadavatelem	199/2014/V/3/3/ŘÚF-150
Pořadové číslo rámcové smlouvy (v případě, že byla uzavřena na základě VZ dělené na části)	1
Počet účastníků rámcové smlouvy	10
Doba trvání rámcové smlouvy **od** 21/04/2015 (dd/mm/rrrr) **do** 21/04/2016 (dd/mm/rrrr)	

IV IEDAĻA: PAPILDU INFORMĀCIJA

IV.1. Paziņojuma saturs

☐ Nepilnīga procedūra ■ Labojums ☐ Papildinājums

IV.2. Informācija par nepilnīgu līguma piešķiršanas procedūru (*attiecīgā gadījumā atzīmējiet tik lodziņus, cik nepieciešams)*

Līguma piešķiršanas procedūra tika pārtraukta. ☐

Līguma piešķiršanas procedūra tika izsludināta par neveiksmīgu. ☐

Līgums netika piešķirts. ☐

Līgums varētu tikt publicēts atkārtoti. ☐

IV.3. Informācija par to, kādēļ un kur veicami labojumi vai papildinājumi *(attiecīgā gadījumā, lai atzīmētu vietu tekstā vai datumus, kas jāizlabo vai jāpapildina, lūdzu, vienmēr norādiet paziņojuma par līgumu vai paziņojuma par metu konkursu attiecīgo iedaļas punkta un/vai apakšpunkta numuru)*

IV.3.1) Grozījums pasūtītāja iesniegtajā sākotnējā informācijā

■ **Paziņojuma publikācija neatbilst pasūtītāja sniegtajai sākotnējai informācijai**

☐ **Abi iemesli**

IV.3.2) Paziņojumā par līgumu vai paziņojumā par metu konkursu

☐ **Iepirkuma procedūras dokumentos** *(sīkākai informācijai skatīt saistītos attiecīgos iepirkuma procedūras dokumentus)*

☐ **Abos** *(sīkākai informācijai skatīt saistītos attiecīgos iepirkuma procedūras dokumentus)*

Iepirkuma dokumentus var saņemt

(✓) Iepriekšaminētajā adresē ( ) Citādi (lūdzu aizpildiet šīs veidlapas A pielikumu)

Piedāvājumi jāiesniedz

(✓) Iepriekšaminētajā adresē ( ) Citādi (lūdzu aizpildiet šīs veidlapas A pielikumu)

I.2) Pasūtītāja veids un galvenā (ās) darbības joma (as)

( ) Ministrija vai jebkura cita valsts vai federāla iestāde, ieskaitot to reģionālās vai vietējās apakšnodaļas

( ) Valsts vai federālā aģentūra / birojs

( ) Reģionāla vai vietēja iestāde

(✓) Reģionāla vai vietēja aģentūra/birojs

( ) Publisko tiesību subjekts

( ) Eiropas institūcija/aģentūra vai starptautiska organizācija

( ) Cits:

( ) Vispārēji sabiedriskie pakalpojumi

( ) Aizsardzība

( ) Sabiedriskā kārtība un drošība

( ) Vide

( ) Ekonomika un finanses

( ) Veselība

( ) Dzīvokļu un komunālā saimniecība

( ) Sociālā aizsardzība

(✓) Atpūta, kultūra un reliģija

( ) Izglītība

( ) Cita:

Pasūtītājs veic iepirkumu citu pasūtītāju vajadzībām

(Ja "Jā" sīkāku informāciju par minētajiem pasūtītājiem var sniegt pielikumā A)

( ) Jā (✓) Nē

But the world is rarely ideal … (conclusion)

information outside the “main” area

values without a key, usually outside the focus area

variations on document structure → inverted sections (keys), missing sections

variations over key names

- contract type
- contract types
- the contract type

spelling mistakes in section names (French case)

How are the others approaching the problem?

Short state of art (1) - manual parsing

using XQuery* (or similar querying languages to parse the document) + any programing language (e.g. Python + Beautifulsoup)

Problem 1: Not very robust on structural changes

Problem 2: Difficult to find low frequency fields

Problem 3: Parsers not generic enough to work cross-country

Short state of art (2) - Transformers

XSLT* (Extensible Stylesheet Transformations), supported by major programming languages

Problem 1: Slightly more robust on structural changes than XQuery (depends on how the rules are written)

Problem 2 + **Problem 3** : not resolved

Problem 4: Extremely difficult to debug

*Also Turing complete

Short state of art (3) - HTML simplification (HTML tidy)

In theory, HTML is XML compatible (except the embedded scripts and stylesheets)

HTML simplification eliminates all the incompatible HTML tags + simplifies the formatting (removes the <b>, <i>, css formatting, etc)

Assumes that the HTML code is well structured for simplification

 field :
 value
 field:
 ...

Works well with HTML tables

Short state of art (4) - supervised learning

keys + values are annotated in a number of documents

feature processing

- feature dictionary → feature encoding (LSA, TF-IDF, binary encoding)
- stemmers or lemmatisers for dictionary reduction
- bag of words vs multi-keyword approach

hierarchical classifiers (multi-svm, multi-NN, or a mix of multiple classifiers)

- determine the key
- separator

Short state of art (5) - Entity-Relation mining

DeepDive-like approaches

use an ontology-like format for entities that can identify the potential actions to be executed against other entities: Wisci(-pedia), DBPedia

Obama → Person, President

France → Country, Place

action-visit **in** {“visit”, ...}

Person + action-visit + Place → relation

“President Obama went to visit a small city in France, known for its architecture”

Could be used for our problem, but we need to build the key-value dataset in an

Short state of art (6) - Structure mining

"Extracting structured data from Web Pages", Arvind Arasu, Hector Garcia-Molina, Stanford University

Exploiting repetitive document structure → creating a template

Frequent fields → promoted as keys (automatically)

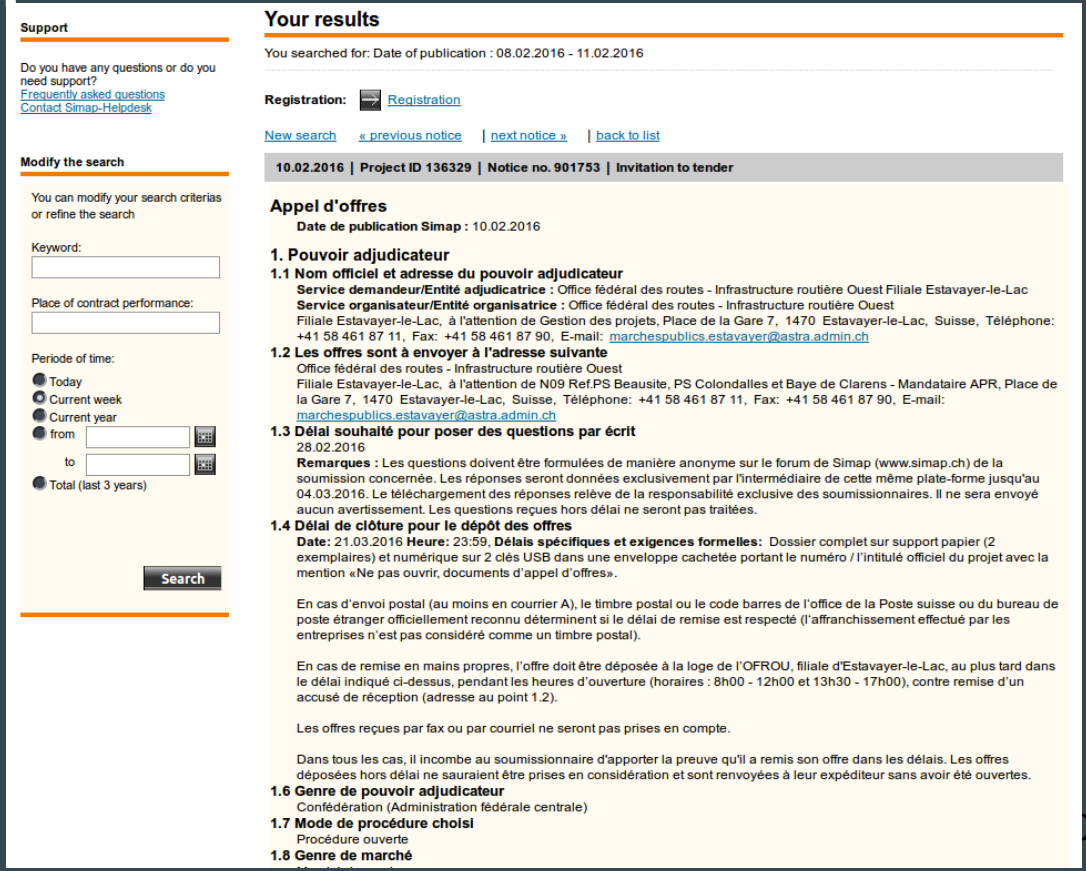
Works well with machine generated web pages, using the same template (Amazon products)

Problem? Does not work well with variations over the template

What did we do?

Finding the page viewpoint

→ (re-) centering the page on useful information
→ could be done automatically



Support

Do you have any questions or do you need support?
Frequently asked questions
Contact Simap-Helpdesk

Modify the search

You can modify your search criterias or refine the search

Keyword:

Place of contract performance:

Periode of time:

- Today
- Current week
- Current year
- from
- to
- Total (last 3 years)

Search

Your results

You searched for: Date of publication : 08.02.2016 - 11.02.2016

Registration: Registration

New search « previous notice | next notice » | back to list

10.02.2016 | Project ID 136329 | Notice no. 901753 | Invitation to tender

Appel d'offres

Date de publication Simap : 10.02.2016

1. Pouvoir adjudicateur

1.1 Nom officiel et adresse du pouvoir adjudicateur
Service demandeur/Entité adjudicatrice : Office fédéral des routes - Infrastructure routière Ouest Filiale Estavayer-le-Lac
Service organisateur/Entité organisatrice : Office fédéral des routes - Infrastructure routière Ouest
Filiale Estavayer-le-Lac, à l'attention de Gestion des projets, Place de la Gare 7, 1470 Estavayer-le-Lac, Suisse, Téléphone: +41 58 461 87 11, Fax: +41 58 461 87 90, E-mail: marchespublics.estavayer@astra.admin.ch

1.2 Les offres sont à envoyer à l'adresse suivante
Office fédéral des routes - Infrastructure routière Ouest
Filiale Estavayer-le-Lac, à l'attention de N09 Ref.PS Beausite, PS Colondalles et Baye de Clarens - Mandataire APR, Place de la Gare 7, 1470 Estavayer-le-Lac, Suisse, Téléphone: +41 58 461 87 11, Fax: +41 58 461 87 90, E-mail: marchespublics.estavayer@astra.admin.ch

1.3 Délai souhaité pour poser des questions par écrit
28.02.2016
Remarques : Les questions doivent être formulées de manière anonyme sur le forum de Simap (www.simap.ch) de la soumission concernée. Les réponses seront données exclusivement par l'intermédiaire de cette même plate-forme jusqu'au 04.03.2016. Le téléchargement des réponses relève de la responsabilité exclusive des soumissionnaires. Il ne sera envoyé aucun avertissement. Les questions reçues hors délai ne seront pas traitées.

1.4 Délai de clôture pour le dépôt des offres
Date: 21.03.2016 **Heure:** 23:59, **Délais spécifiques et exigences formelles:** Dossier complet sur support papier (2 exemplaires) et numérique sur 2 clés USB dans une enveloppe cachetée portant le numéro / l'intitulé officiel du projet avec la mention «Ne pas ouvrir, documents d'appel d'offres».

En cas d'envoi postal (au moins en courrier A), le timbre postal ou le code barres de l'office de la Poste suisse ou du bureau de poste étranger officiellement reconnu déterminent si le délai de remise est respecté (l'affranchissement effectué par les entreprises n'est pas considéré comme un timbre postal).

En cas de remise en mains propres, l'offre doit être déposée à la loge de l'OFROU, filiale d'Estavayer-le-Lac, au plus tard dans le délai indiqué ci-dessus, pendant les heures d'ouverture (horaires : 8h00 - 12h00 et 13h30 - 17h00), contre remise d'un accusé de réception (adresse au point 1.2).

Les offres reçues par fax ou par courriel ne seront pas prises en compte.

Dans tous les cas, il incombe au soumissionnaire d'apporter la preuve qu'il a remis son offre dans les délais. Les offres déposées hors délai ne sauraient être prises en considération et sont renvoyées à leur expéditeur sans avoir été ouvertes.

1.6 Genre de pouvoir adjudicateur
Confédération (Administration fédérale centrale)

1.7 Mode de procédure choisi
Procédure ouverte

1.8 Genre de marché

Algorithm sketch

a synonym dictionary approach with “fuzzy” key comparison

for a given document ***d***
for every path ***p*** *in* ***d***
determine if ***p*** *is a key* ← *fuzzy comparison*
if ***p*** *is not key*
// p may be a value
if ***p*** *is similar to* ***previous key***
p is a value related to previous key
create ***(previous key, p.text)*** *pair*

Determining the key

A frequency analysis is performed on a few sampled documents → candidate keys

- similar keys (contract type example) are automatically grouped into a synonym set
- potentially, logical synonyms can be also grouped into the same synonym set
-  not all the high frequency synsets are keys: “download document”, “print document”, contract types, procedure types, etc. → requires manual filtering after automatic collection

The comparison with the keys during runtime is done with the same similarity algorithm

A successfully identified key at runtime will have:

- a html element path: *document|body|div|div|table|tr|td*
- a text value: *some text*

Determining the value

A value is:

- not a key
- not an ignored element
- the html path of a value is related to the html path of a key → similarity measure

all the similarities are computed using inverted Levenshtein distances (alternatively we can use set-based distances, bit similarities, etc)

Improving the accuracy - Data Rescuers

A set of preprocessing and postprocessing rules (in various stages of the algorithm) to determine if a html element can be reused

Each country has its own set of rules:

- CZ - rules for extracting the value out of the html elements (list, checkboxes, input fields)
- EE - the contract type is a custom-made title without a key
- CH - in the document header the document type, announcement id and other information is concatenated
- BG - contract number and contract type does not have a key
- HR - some documents (contract awards) have an empty contract type

Conclusion

This approach is more robust to structure/document variations
Easy to debug
Easy to tune in case something goes wrong → dictionary tuning

Issues:

- Difficult to assess the accuracy → no ground truth
- Flat vs hierarchical structure → CZ has a hierarchical structure of the keys → document model has to be (automatically) built
- For documents with a very high variability (FR, HU, etc), building the dictionary becomes more manual → the keys cannot be distinguished from the values