



Ontology-based Web Information Extraction in Practice

eRecruitment – eTourism - eProcurement



FWF

Japan-Austria Joint Workshop on “ICT”

Tokyo, October 18-19, 2010



Institute for
Application Oriented Knowledge Processing



a.Univ.-Prof. Dr. DI Birgit Pröll
bproell@faw.jku.at



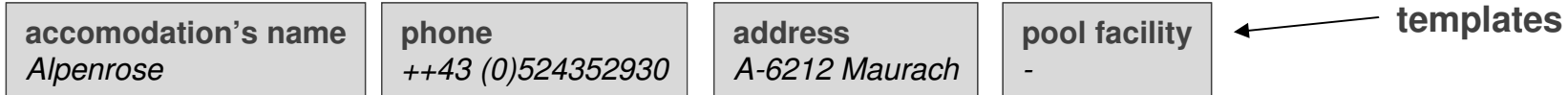
Contents

- **Motivation**
- Web Information Extraction (WebIE) by Examples
 - General Architecture
 - Web Crawler
 - Ontology Aware WebIE
 - Structure Analysis: Page Segmentation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work



Web Information Extraction (WebIE)

...extracting structured data from Web pages



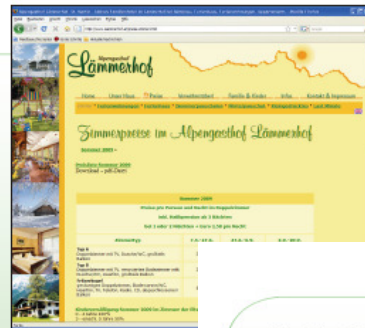
The screenshot shows the website for Alpenrose, a wellness resort. Key elements and annotations include:

- Header:** Alpenrose logo with 'superior' tagline.
- Navigation:** Location & Arrivals, Home, Wellness, etc.
- Main Content:**
 - Hotel Name:** Alpine Spa Hotel **** Haus Hirt, Bad Gastein (circled in red).
 - Address:** DER STERNSTEINHOF, A-6212 Maurach (circled in red).
 - Phone:** Telefon ++3(0)7213/6365 (circled in red).
 - Facilities:** Swimming Pool, Indoor Swimming Pool, Dry Heat Sauna (circled in red).
 - Additional Facilities:** Car Park, Pets welcome, Restaurant, Guest Lounge, Hotel Bar, Beauty Farm, Floor Service, Garden/Private Grounds.
- Search Room Form:** Check In: 26.09.2009, Check Out: 03.10.2009, Room(s): 1, Room type: (All), Adults: 2, Children: 0.
- Footer:** 2008 Wellnessresidenz Alpenrose, Wolfgang Kostenzer GmbH, A-6212 Maurach/Achensee (circled in red).



WebIE Projects in cooperation with Austrian Industry

TourIE
Tiscover AG



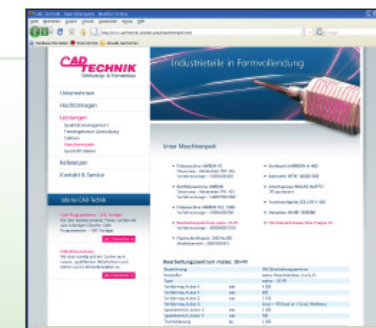
Application area
eTourism

JobOlize
JoinVision E-Services
GmbH
FFG (grant 813202)



Application area
eRecruitment

Marlies
Tech2select GmbH
FFG (grant 817789)



Application area
eManufacturing, supply-chain-management

Projects' Requirements and Approach Taken

Some WeBLE peculiarities in the given projects

- Heterogeneously designed Web pages
- Mixture of (semi-)structured data and full text
- Significant structural aspects, e.g.,
 - location of information on Web page
 - information „hidden“ in Web tables
- Information scattered over several Web pages
- Web site evolution



WeBLE Approaches

- Screen scraping approaches (wrapper generation)
- Automatically trainable systems (machine learning)
- Knowledge-engineering approach
+ Web crawler + structural analysis + ...

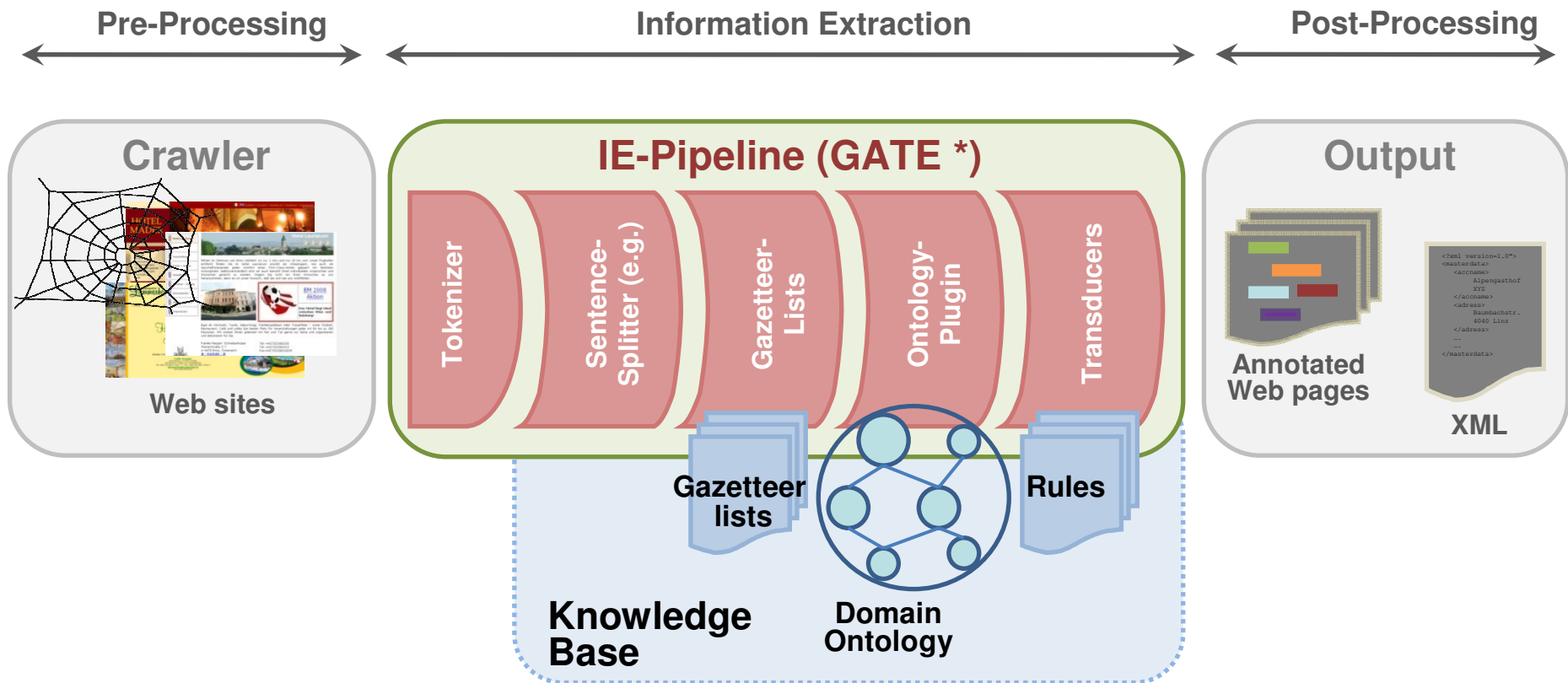
[Appelt et al., 1999]



Contents

- Motivation
- Web Information Extraction (WebIE) by Examples
 - General Architecture
 - Web Crawler
 - Ontology Aware WebIE
 - Structure Analysis: Page Segmentation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work

Overall Architecture



*) [Cunningham et al, 2006]

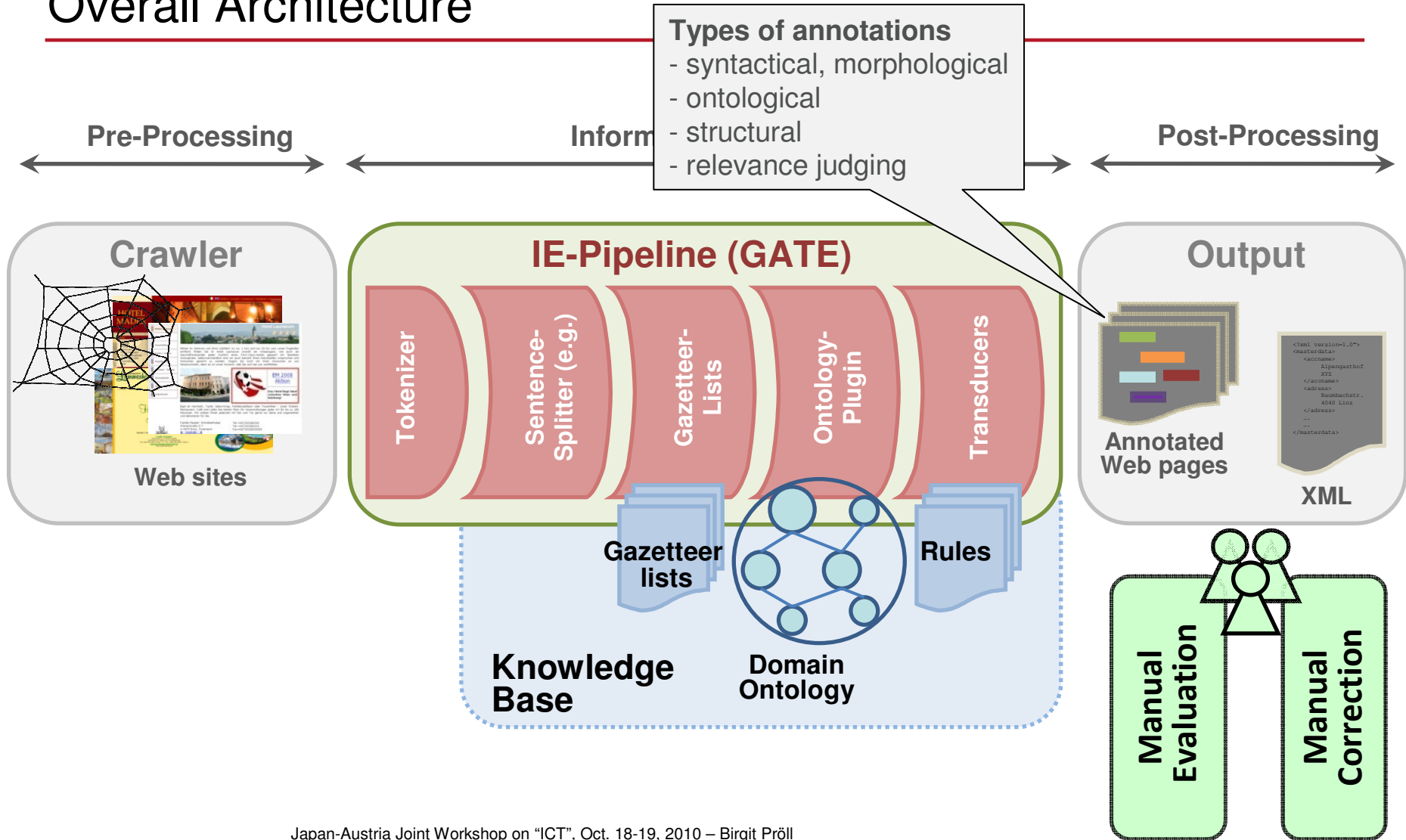
Web Crawler



- Collects relevant Web pages
- Classifies Web pages
 - Home page, price pages, location pages, etc.
 - Based on Support Vector Machine
- Recognises language
 - Using meta-tags and an n-gram based algorithm



Overall Architecture





Regular Expressions & Gazetteer Lookup

Phone: +43 (0)5243 52930
Fax: +43 (0)5243 5466
info@alpenrose.at

Telefon +43(0)7213/6365
Fax +43 (0)7213/6365-8
info@sternsteinhof.at

Tel: 0043/6432/84 75
Fax: 0043/6432/84 75 70
e-mail: info@alpina-hotel.com
internet: www.alpina-hotel.com

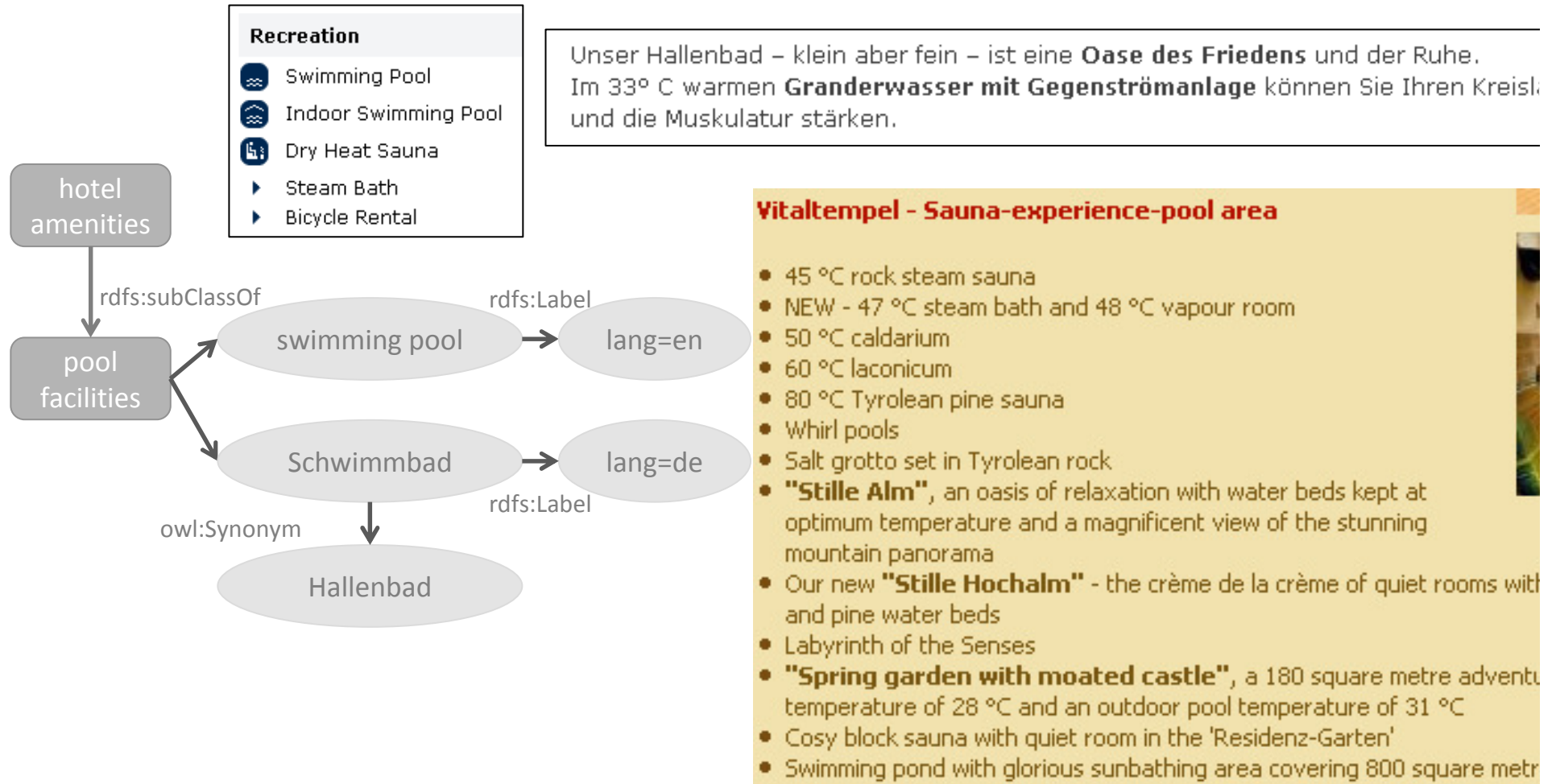
Rule: Phone1

```
(  
  {Token.string=="+"}  
  {Token.kind==number}  
  ({SpaceToken.kind==space})*  
  {Token.string=="("}  
  {Token.kind==number}  
  {Token.string==")"}  
  (({SpaceToken.kind==space})*  
  {Token.kind==number})*  
):phone  
→  
:phone.MyPhone={}
```

Gazetteer list 'phone keywords'

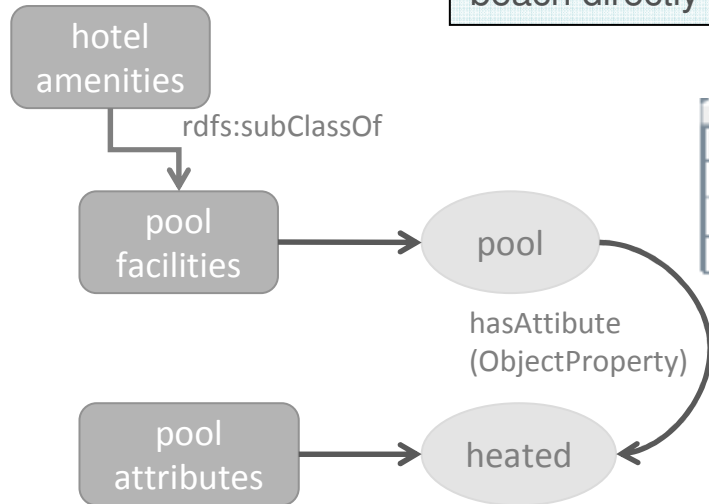
Phone
Telephone
Tel.
Tel:
Tel.:
Telefon

Ontology-Aware Entity Recognition (1/2)



Ontology-Aware Entity Recognition (2/2)

We offer a wonderful 2500m2 wellness area, lead by a trained wellness team. Indoor swimming pools, new heated natural outdoor pool with sandy beach, open air whirlpool with a wonderful view of lake Caldaro, large sauna world, and our private beach directly at lake Caldaro, full fill all wishes!



Type	Set	Start	End	Id	Features
MyPool		1404	1409	19916	{attrib0=natural, attrib1= , attrib2=heated, attrib
MyPool		1437	1446	19917	{attrib0=open air, attrib1= }
MyPool		2769	2773	19918	{attrib0=indoor, attrib1= }

```

Rule: PoolsWithAttributes
(
  {{PoolAttribute}}
  +:attributes
  {{PoolFacility}}:facilities
)
→
  
```

```

RHS (↓List attributes, ↓Annotation facilities){
  create new FeatureMap features;
  for(Annotation attribute : attributes){
    add attribute to features;
  }
  create new Annotation MyPool;
  add features to MyPool;
  return MyPool;
}
  
```



Structure Analysis: Web Page Segmentation

Top part

MyMATRIX™
Your career. We take it personally.

HOME ▾

For more than two decades, MATRIX Resources has connected great people and achieved great solutions. Serving the IT marketplace exclusively, MATRIX is a premier IT staffing and solutions provider to the business community, as well as the preferred partner for career assistance by job seeking IT professionals.

In a Hurry? Send us your resume and we'll find the right job for you!

Content part

Position Information

Senior JAVA Developer Job Number: PATL672960

Date Posted: 07/01/2008

Location: 30339, GA 30339

Status: Permanent

Compensation: \$85,000 to \$95,000

Career Level: Not Specified

Education Level: Not Specified

Reference Number: PATL672960

Senior JAVA Developer Needed to Make an Impact in a Small Growing Organization!

Join a dynamic team as a Senior JAVA developer for a leading-edge integrated advertising platform in a Linux/MySQL environment. To be considered, you must have solid Object Oriented development experience. Must have experience taking ownership of development initiatives for entire project lifecycle, the ability to work independently and with a team and the ability to conceptualize, design and implement business requirements. PERMANENT Position. Apply today!

Required Experience

- JAVA
- Object Oriented Development
- Design, implement and take ownership of business requirements
- Team Player - yet, ability to work independently, think strategically, with strong design skills

Required Skills

- Java
- MySQL
- Preferred skills include: LAMP, Ruby on Rails, Linux

Benefits/Perks

- Flexible hours, work from home 1-2 days a week
- Very independent organization

Application Requirements

Client will only consider local candidates. Client requires Green Card or U.S. Citizenship for this position.

Please Note: Your resume will never be submitted to a client company without your prior knowledge and consent to

powered by Typo3

Apply Now Email this job to a friend

Bottom part

Contact Technical Support. Copyright © 2003-2008 MATRIX Resources, Inc.

templates

job title
Senior Java Developer

IT skills + level
*JAVA + perfect
MySQL + basic*

operation area
SW programming, testing

language skills
English fluently

contact
-

[Debnath et al., 2005]
[Chakrabarti et al., 2007]



Structure Analysis: Block Identification

Senior JAVA Developer

Job Num

Content part

Senior JAVA Developer Needed to Make an Impact in a Small Growing Organization!

Join a dynamic team as a Senior JAVA developer for a leading-edge integrated advertising platform in a Linux/MySQL environment. To be considered, you must have solid Object Oriented development experience. Must have experience taking ownership of development initiatives for entire project lifecycle, the ability to work independently and with a team and the ability to conceptualize, design and implement business requirements. PERMANENT Position! Apply today!

Block

Responsibilities

- JAVA Perform detailed software design, documentation, development and testing for Java
- Analyze customer needs and develop overall concept and design objectives

Responsibilities

Block

Required Skills

- Extensive knowledge of Java
- MySQL
- Preferred skills include: LAMP, Ruby on Rails, Linux

Requirements

Block

Benefits/Perks

- Flexible hours, work from home 1-2 days a week
- Very independent organization

Offer

Block

Application Requirements

Client will only consider local candidates. Client requires Green Card or U.S. Citizenship for this position.

Block

Structure Analysis: Table Data Extraction in Marlies

machine type	description	∅ (mm)	to module	weight max.	toothing quality
toothing machines:					
Pfauter P 400	hobbing machine	15 - 400	6		8 - 9
Pfauter P 900	hobbing machine	30 - 900	10		8 - 9
Pfauter P 900	hobbing machine	30 - 900	10		8 - 9
Pfauter PE 1200	hobbing machine	30 - 1200	20		6 - 7
Pfauter P 1250	hobbing machine	50 - 1250	16		8 - 9
Lorenz SN 8	shaping machine	30 - 750	10		8 - 9

Result	
Hobbing machine - Pfauter P 400	Diameter (min): 15 mm Diameter (max): 400 mm

- ... measure
- ... unit
- ... machine

Structure Analysis: Table Data Extraction in TourIE

Summer 2009

Prices per person, per night in the double room

including half-board - min. stay 3 nights

for 1 or 2 nights surcharge of Euro 3,50 per person per night *

Room type	7.5. - 27.6.	27.6. - 6.9.	6.9. - 20.9.
Type A double room with TV, shower/WC, partly with a balcony	34,00	36,00	33,00
Type B double room with TV, bath room with shower/WC, hairdryer, partly with a balcony	37,00	38,00	35,00
Type Fritzerkogel bigger double room with 1 double bed and 1 extra bed, bath tube/WC, hairdryer, TV, telephone, radio, CD, balcony	42,00	43,00	40,00

Childrens reduction summer 2009 in the room of the parents:
 0 - 2 years 100% *
 3 - 5 years 50%
 6-12 years 30%
 13-18 years 20%

Annotations:

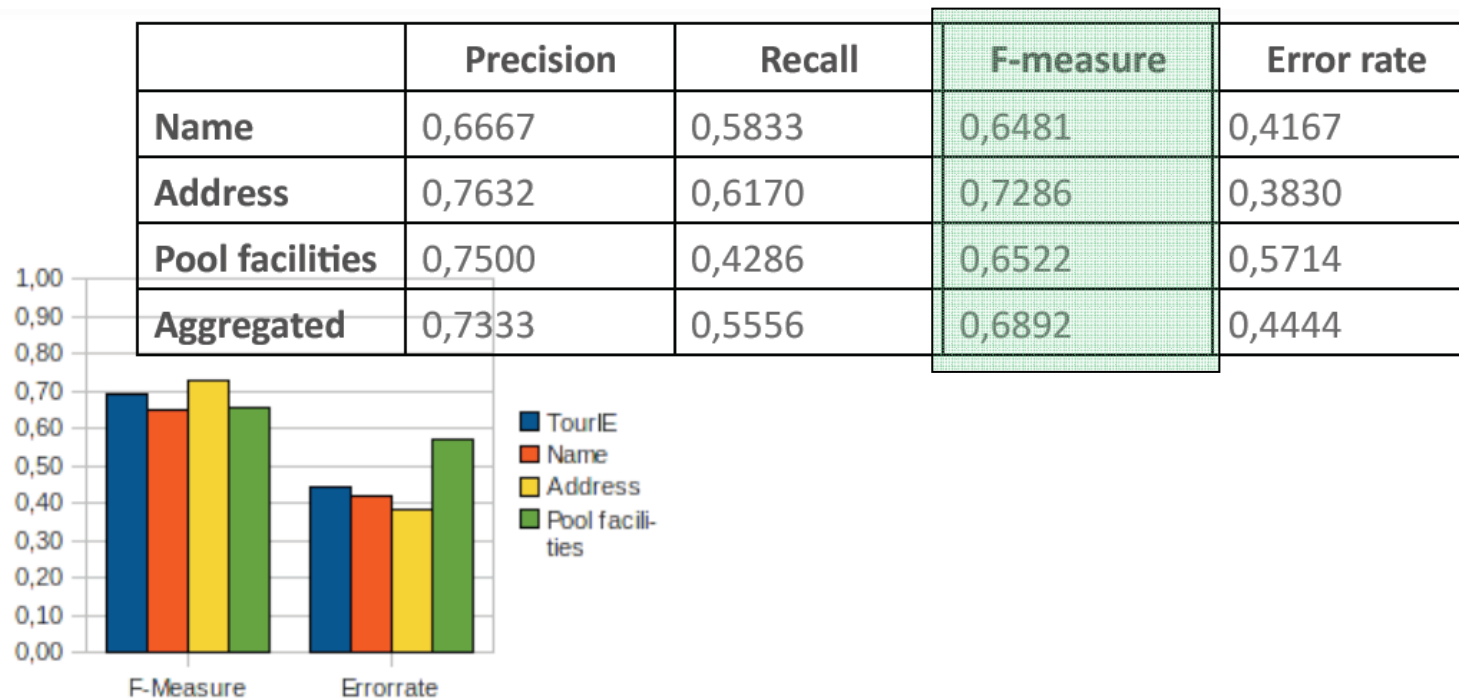
- price domain
- board
- time period
- room name + room type
- price value without currency
- room name + room type
- detailed description of room and its amenities



Contents

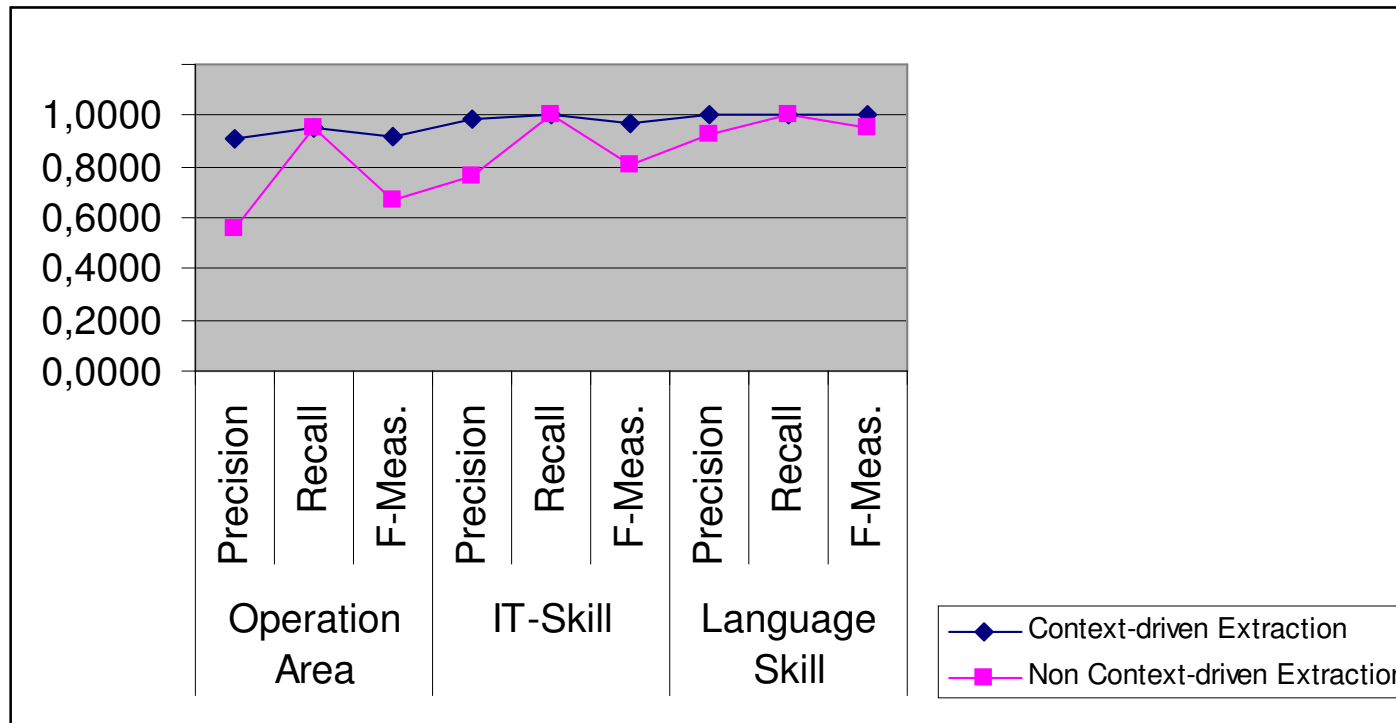
- Motivation
- Web Information Extraction (WebIE) by Examples
 - General Architecture
 - Web Crawler
 - Ontology Aware WebIE
 - Structure Analysis: Page Segmentation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work

Evaluation: TourIE



- Evaluation results were satisfactory with respect to the preliminary study.
- Pool facility extraction quality was poor because of incomplete ontology.

Evaluation: JobOlife



→ Page segmentation & block identification considerably rises precision.



Evaluation: Marlies

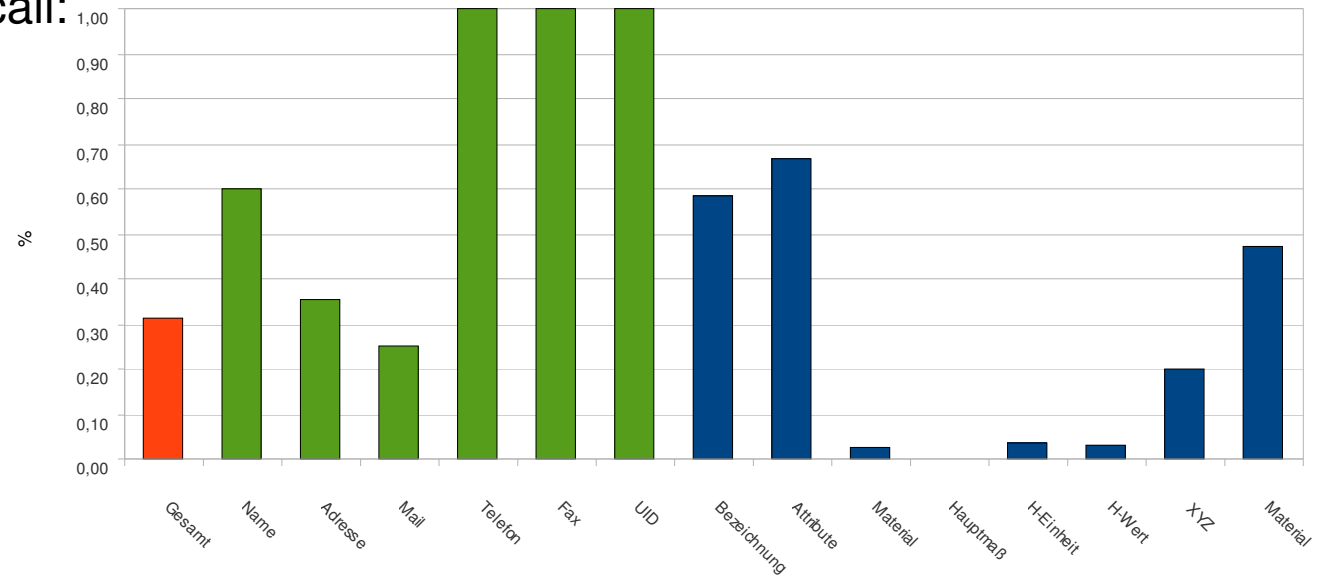
Marlies Ontology

Classes: 2313

Instances: 2661

Assignments of object properties to instances: 42791

Preliminary results for recall:



→ Work in progress (e.g., table extraction).



Manual Correction via Rich Client GUI

The screenshot shows a Mozilla Firefox browser window displaying a job advertisement for 'Junior Software Developer (m/f)' on the 'jobOlive' website. The browser's address bar shows 'IT jobs and projects for freelancers and it experts - Mozilla Firefox'. The page content includes a 'Jobs' section with a company description, a job title, 'Main Responsibilities', 'Requirements', and 'We offer' sections. The 'Requirements' section lists skills like C#, ASP.NET, Java, Web services, XML, SVG, and Javascript, along with database knowledge (MySQL, Oracle) and a university degree in computer science. The 'We offer' section lists benefits like cutting-edge technologies, a challenging environment, and attractive benefits. A rich client GUI is overlaid on the left side of the browser window, featuring a 'Send to Annotation Server' button, an 'Annotation List' table, and 'Annotation Details' for a selected 'Skill' annotation. The 'Annotation List' table has columns for 'Active', 'Color', 'Element', 'Hits', and 'Description'. The 'Annotation Details' section shows a 'Tag' of 'Skill' and 'Attributes' with a rule: 'relevance->80 id->mysql level->50 rule->Skill...'. Below this, there is a 'Language' dropdown menu with options: English (eng), German (ger), Slovenian (slv), French (fra), English (eng) (selected), and Spanish (spa). A green checkmark button is visible next to the selected language. The browser's menu bar includes 'Datei', 'Bearbeiten', 'Ansicht', 'Chronik', 'Lesezeichen', 'Extras', and 'Hilfe'. The browser's status bar at the bottom indicates 'powered by Typo3'.

Active	Color	Element	Hits	Description
<input checked="" type="checkbox"/>		Attributes		
<input checked="" type="checkbox"/>		Titel	1	Titel des Jobangebots
<input checked="" type="checkbox"/>		Skill	11	Job-Anforderungen
<input checked="" type="checkbox"/>		Language	1	Sprachanforderungen
<input checked="" type="checkbox"/>		Structure		
<input checked="" type="checkbox"/>		Block	0	Semantic blocks

Tag	Attributes
Skill	relevance->80 id->mysql level->50 rule->Skill...
Skill	relevance->80 id->ora level->50 rule->Skillswi...
Language	id->eng level->g rule->AdjLang

Language: English (eng)
Level: German (ger)
Slovenian (slv)
French (fra)
English (eng)
Spanish (spa)



Contents

- Motivation
- Web Information Extraction (WebIE) by Examples
 - General Architecture
 - Web Crawler
 - Ontology Aware WebIE
 - Structure Analysis: Page Segmentation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work

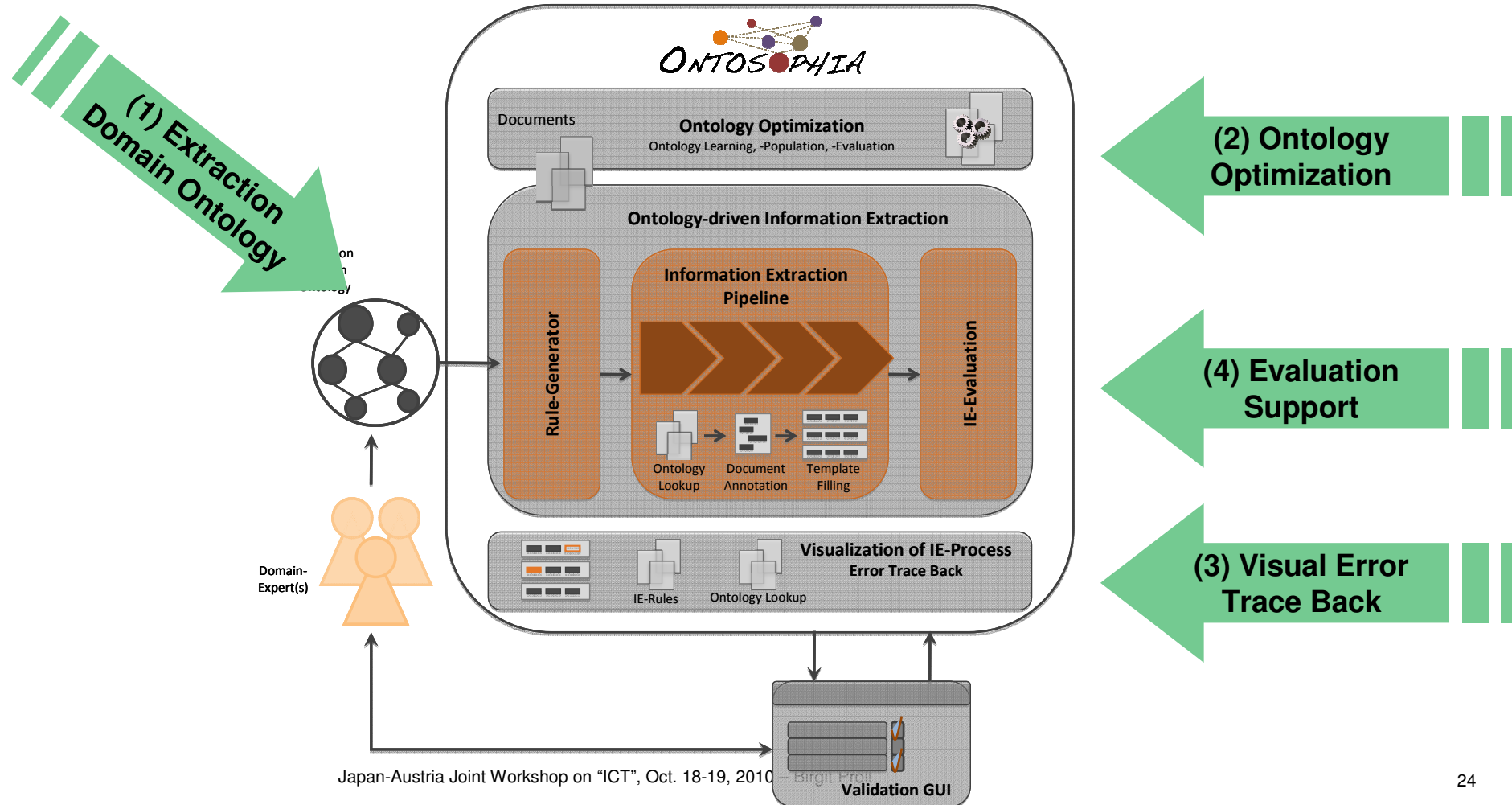


Lessons Learned

- Today's Web pages do not adhere to standards or semantic Web proposals.
 - Only a few RDF resources available; proposed microformats rarely used
 - Poor HTML, e.g, tables used for layout purposes
 - Web 2.0 coded Web pages in progress; content-based image retrieval & OCR
- Development & maintenance of knowledge-based WebIE systems is expensive.
 - Domain experts & knowledge engineers are needed.
 - Rule-coding is tedious and errorprone.
 - Evaluation of numerous methods & algorithms; multiplied due to multilinguality
 - Manual evaluation is time consuming.
- WebIE performance considerably depends on quality of domain ontology.
- We have to observe (evolving) legal issues
 - Robots exclusion standard, Sitemap etc.
 - Further processing of extracted data

Future Work: Ontosophia

Ontology-driven IE Supported by (Semi-) Automatic Corrective Feedback





Thank you for your Attention!

Acknowledgements to:



Christina
Feilmayr



Stefan
Parzer



Christina
Buttinger



Michael
Guttenbrunner