From Partially Structured Documents to Relations

Utilizing and Extending Best Practices in Table Recognition for Automatic Information Extraction in Spreadsheets

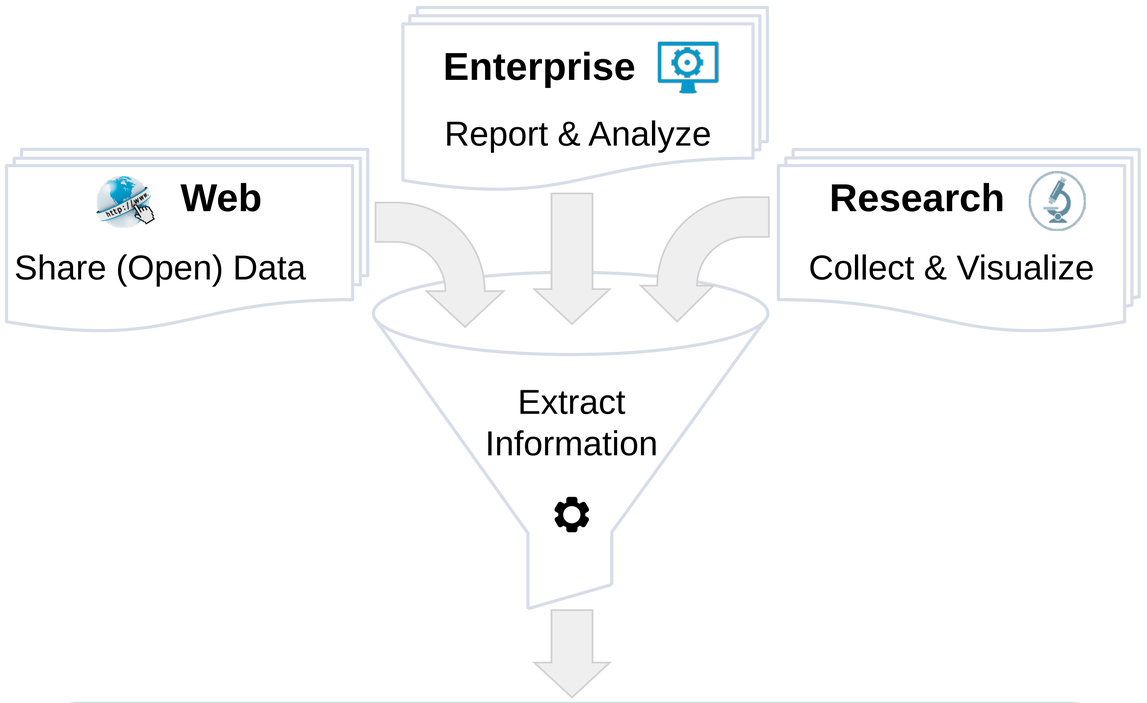
Elvis Koci
elvis.koci@tu-dresden.de

TB2 — Roles in Software Development

2nd COHORT — Prof. Lehner, Prof. Schill

Funded by ITABI DOC

Motivation



Benefits:

- Increased reusability
- Organizational learning
- Better governance
- Machine readable content
- Legacy systems integration
- Minimal user interaction

Proposed Solution

Number of items (in units) sold per region

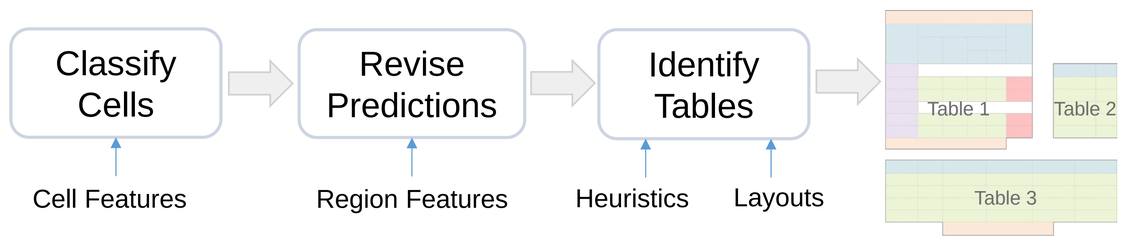
Location	Item				Total
	Mouse	Monitor	Adapter		
			VGA	HDMI	
Europe					
Spain	500	200	85	61	846
France	465	169	80	80	794
Asia					
China	422	163	90	44	719
Vietnam	473	182	74	55	784

Item “Keyboard” is omitted. Check next sheet.

Items	Code
Mouse	abs12f
Monitor	tbyx1f
Adaptors	3rbsfz
VGA	3rbsfz1
HDMI	3rbsfz2

Customer	Country	Industry	Contact Person	Phone	Email
Bravo	Spain	Retail	Juan García	+34 934124602	j.g@bravo.es
Sonra	France	Non Profit	Clément Dubois	+33 509758351	c.d@sonra.fr
Ambra	Vietnam	IT	Pham Van Duc	+84 129678432	pvd@ambra.vn
Cheetah	China	Retail	Li Wei	+86 254372941	lw@cheetah.cn

Short list of customers



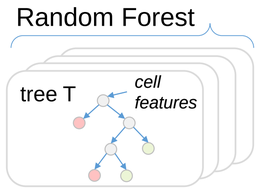
Cell Classification

Worksheet

Features

- *Type String*
- *Is Bold Font*
- *No Underscore*
- *Center Aligned*
- *Default Fill Color*
- *Thick Left Border*
- *Not Aggregation*
- *Not Referenced*
- *Right is Empty*
- *Top Same Style*

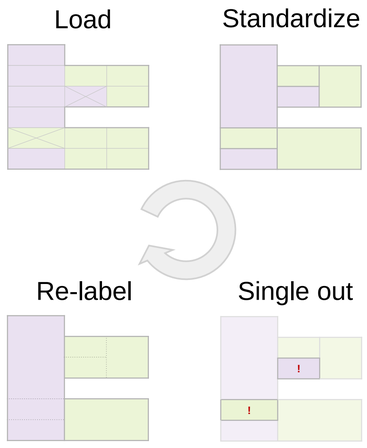
Random Forest



Layout Classes

Data, Derived, Header, Attributes, Metadata

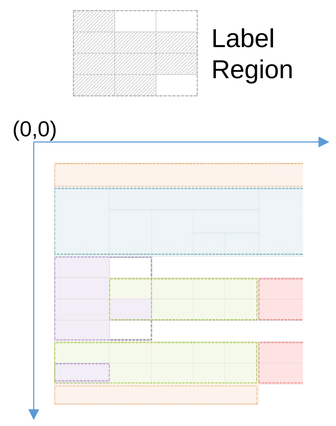
Misclassification Handling



Approach:

- Examine irregularities
- From neighbor regions:
  - Pinpoint misclassified
  - Predict true class

Table Identification



TIRS: framework of 8 heuristics steps:

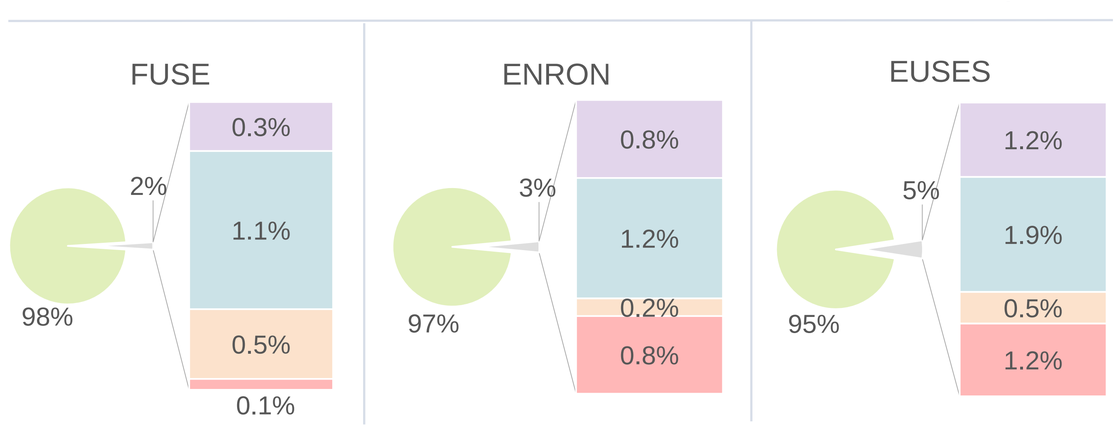
- Multiple tables
- Various types
- Irregular layouts

Regions make tables:

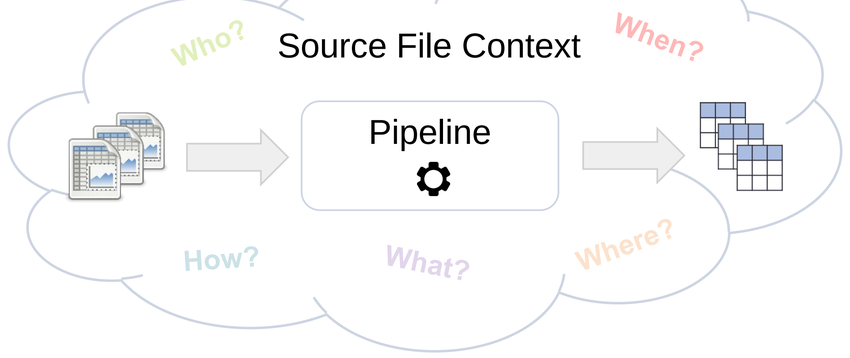
- Intersection
- Distance
- Location

Annotated Dataset

Contains **465** sheets, **898** tables, **>800,000** cells, from three corpora:



Adaptive Process



To maximize accuracy tailor:

- Features
- Heuristics
- Layouts

Dynamically Reconfigurable

[1] Elvis Koci, Maik Thiele, Oscar Romero, and Wolfgang Lehner. A machine learning approach for layout inference in spreadsheets. IC3K’16, p. 77–88, SciTePress, 2016.
[2] Elvis Koci, Maik Thiele, Oscar Romero, and Wolfgang Lehner. Table identification and reconstruction in spreadsheets. CAiSE’17. (Accepted Paper)
[3] Elvis Koci, Maik Thiele, Oscar Romero, and Wolfgang Lehner. Cell Classification for Layout Recognition in Spreadsheets. Chapter in CCIS book series. (In Press)

DFG
RoSI - Research Training Group
Funding: DFG (GRK 1907)
Spokesman: Prof. Dr.-Ing. Wolfgang Lehner
Project Website: rosi-project.org

DRESDEN concept
Exzellenz aus Wissenschaft und Kultur