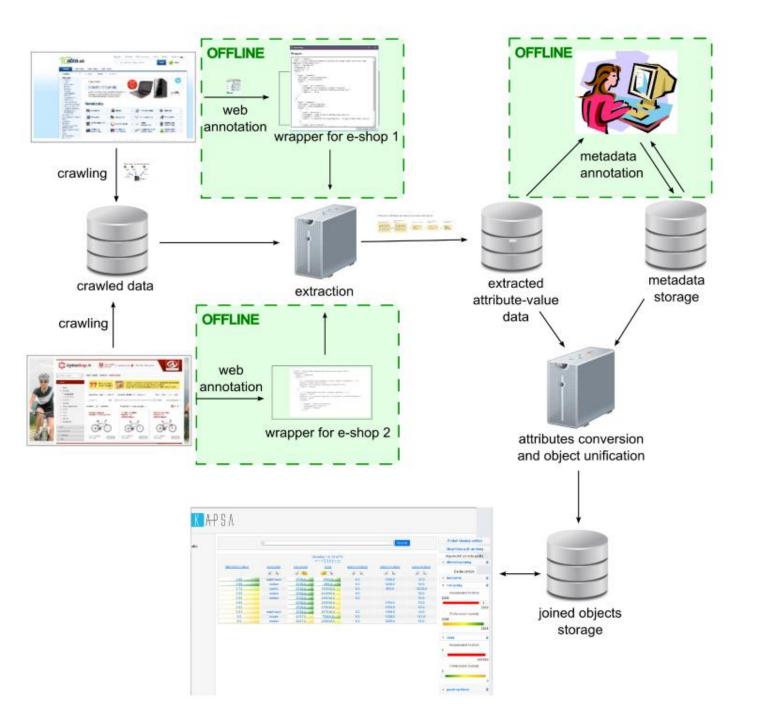
# Project Kapsa

### Peter Gurský Pavol Jozef Šafárik University in Košice Faculty of science, Institute of computer science

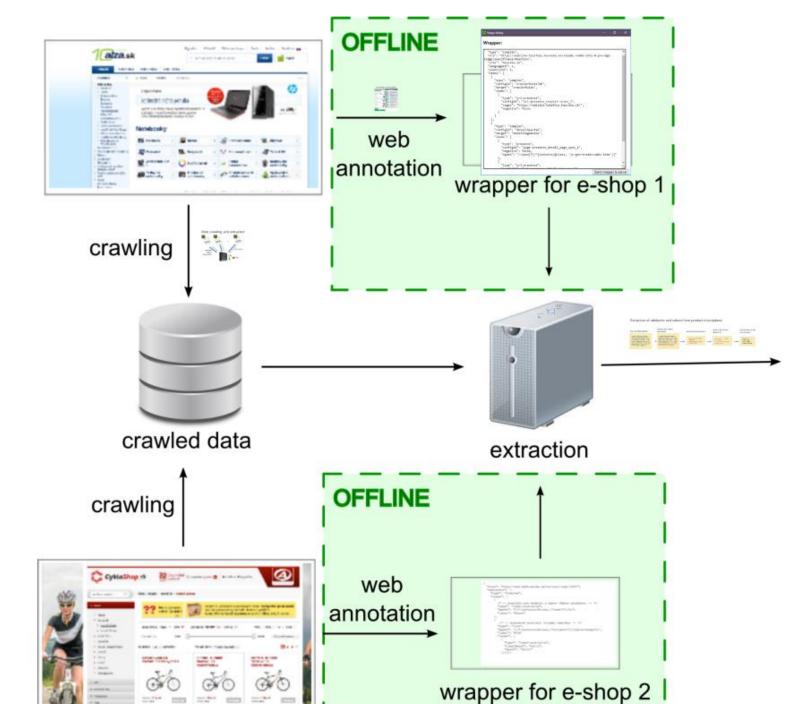
#### Cocreators:

Michal Vaško, Milan Vereščák, Martin Šumák, Július Malčovský, Róbert Novotný, Veronika Vaneková, Vladimír Chabaľ, Libor Mališ, Jozef Kuper, Peter Šinaľ, Peter Kál, František Trusa, Michaela Linková, Patrik Sedlák, Dávid Varga, Rudolf Pavel, Matej Perejda, Samuel Koprda, Olha Charna, Yevhenii Kozhevin

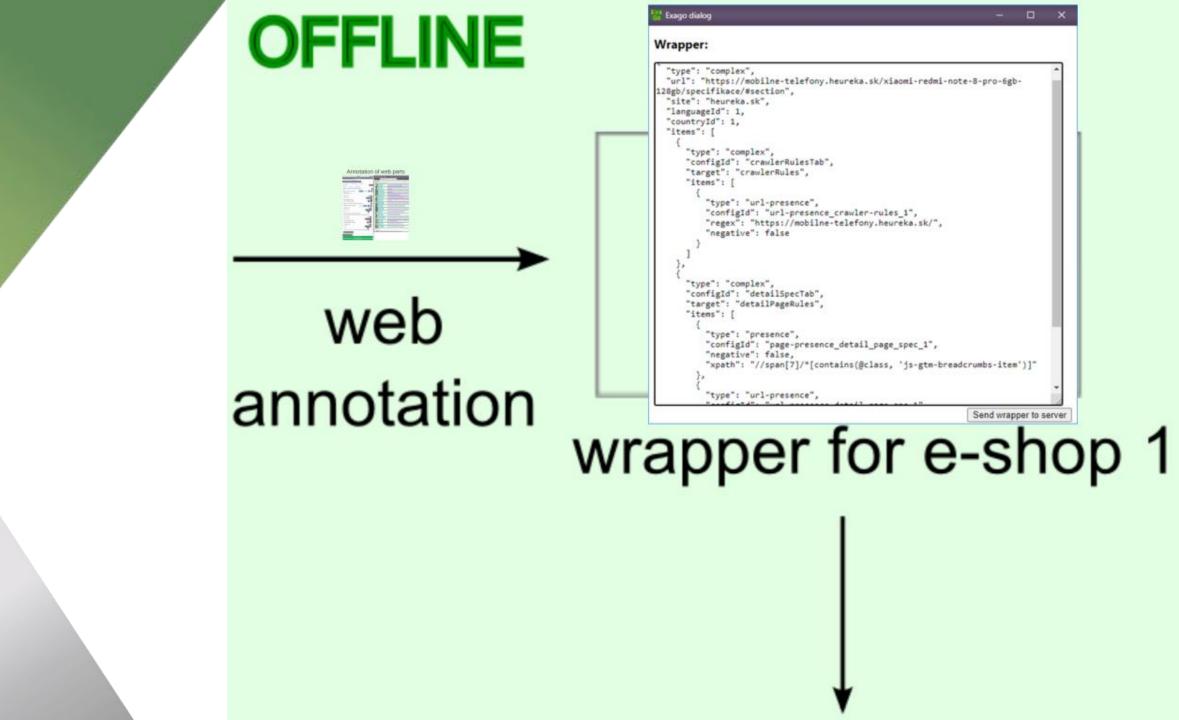














## Annotation of web parts

Start new annotation rt of URL indicating e-shop: imdb.com	Eloud Atlas (2012) - IMDb × + imdb.com/title/t1371111/7re_ EB @ ☆ ♦ ∧ ● 9 > •
Show and send wrapper Open crawler manager	Cast
Crawler rules Detail page spec. Annotation	Cast overview, first billed only:
Start page   Add click	Tom Hanks Dr. Henry Goosel Hotel Manager Isaac Sachs
Attributes Comments General data	Cavenoish Look-a-Like Actor / Zachry
bject: can contain many attribute-value pairs	Halle Berry Native Woman Docasta Ayrs / Luisa Rey / Indian
Attribute name and value pair Manual XPath -	Dim Broadbert Proscient 2 Timothy Cavera
Attribute name: Herci	Hugo Weaving Haskell Moore (Tadeusz Kesselring / Bill Smoke)
List of items:	Adam Ewing/Poor Hotel Guest/Megan's Dad/P Chang Adam (Zechry's Brother-in-Law)
Choose parent or child 1 1 Choose between children 0	Doona Bae Tilda/Megan's Mom/Mexican Woman/(Sonmi-4
Object: can contain many attribute-value pairs	Ren Whishaw Cabin Boy / Robert Frobisher / Store Clerk/ Geor
Attribute name and value pair Manual XPath	Keth David Kupaka/ Joe Napier/An-kor Apis/ Prescient
Attribute name	Dames D'Arcy Young Rufus Sixsmith /Old Rufus Sixsmith /Nurs
xPath: td[2]/a	
RegEx:	Kun Zhou Talbot (Hotel Manager)/ Yoona-939/ Rose
More than element found by the XPath! Result: ["Tom Hanks", "Halle Berry", "Jim Broadbent", "Hugo Weaving", "Jim Stur	David Gyzsi Auton/Lester Rey/Duophysite
List of items:	Susan Swandon Madame Horros Older Ursula Yusouf Suleiman
xPath: td[4]/a	
Choose parent or child 1 1 Choose between children 0	Hugh Grant Rev. Giles Horroy / Horel Heavy / Lloyd Hooks / D
Simple value	Robert Fyle Old Salty Dog/Mr. Meeks/Prescient 1
xPath:	Martin Wuttke Mr. Boerhaave/Guard/Leary the Healer
RegEx: 0	
Result: ["Dr. Henry Goose", "Hotel Manager", Isaac Sachs", "Dermot Hoggins", "	See full cast >
attribute name-value pair	
attribute name-value pair	
Extraction Preview	



### Features

- annotation of various objects on webpages
- support for
  - one object on one page
  - many objects on one page
  - one object on many pages
  - pagination
  - hierarchical structures
- semi-automatic annotation with mouse clicks, with possibility to precise the rules manually
  - generation of XPath a regular expressions
  - correction of rules by possitive and negative examples
  - nesting and emergence to/from elements
- extraction from URL link and HTML structures



## Semiautomatic annotation

- identification of repeating structures
- modified MDR algorithm
  - breadth-first search
  - using class and id attributes of <div> and <span> elements
  - Levenshtein distance of DOM tree levels, not whole subtrees





## Future annotation features

- infinite scrolling
- forms filling
- CSS selectors



# wrapper for e-shop '

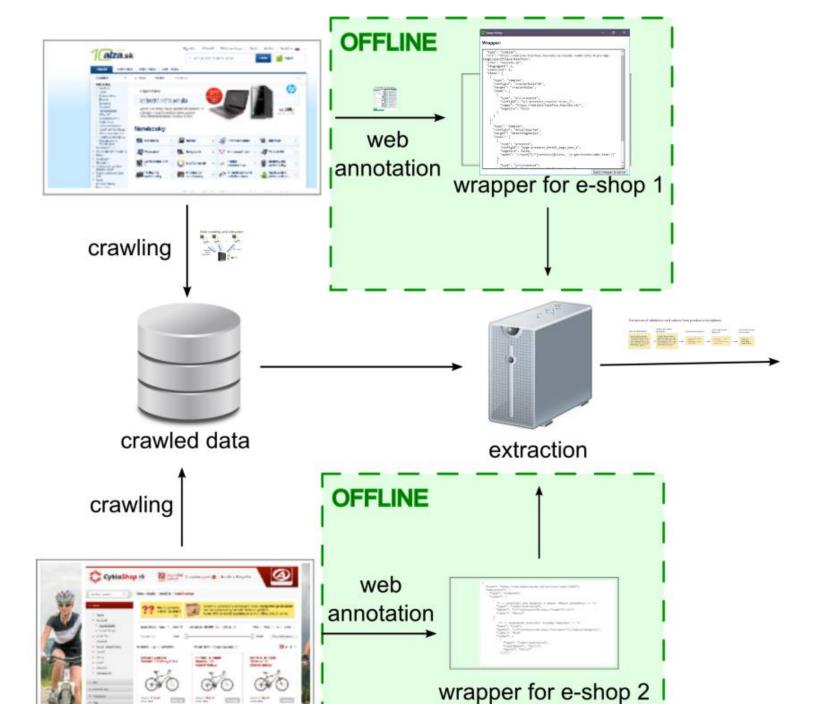
```
"type": "complex",
  "url": "https://mobilne-telefony.heureka.sk/xiaomi-redmi-note-8-pro-6gb-
128gb/specifikace/#section",
  "site": "heureka.sk",
  "languageId": 1,
  "countryId": 1,
  "items": [
      "type": "complex",
      "configId": "crawlerRulesTab",
      "target": "crawlerRules",
      "items": [
          "type": "url-presence",
          "configId": "url-presence crawler-rules 1",
          "regex": "https://mobilne-telefony.heureka.sk/",
          "negative": false
      "type": "complex",
      "configId": "detailSpecTab",
      "target": "detailPageRules",
      "items": [
          "type": "presence",
          "configId": "page-presence_detail_page_spec_1",
          "negative": false,
          "xpath": "//span[7]/*[contains(@class, 'js-gtm-breadcrumbs-item')]"
          "type": "url-presence",
```

Exago dialog

Wrapper:

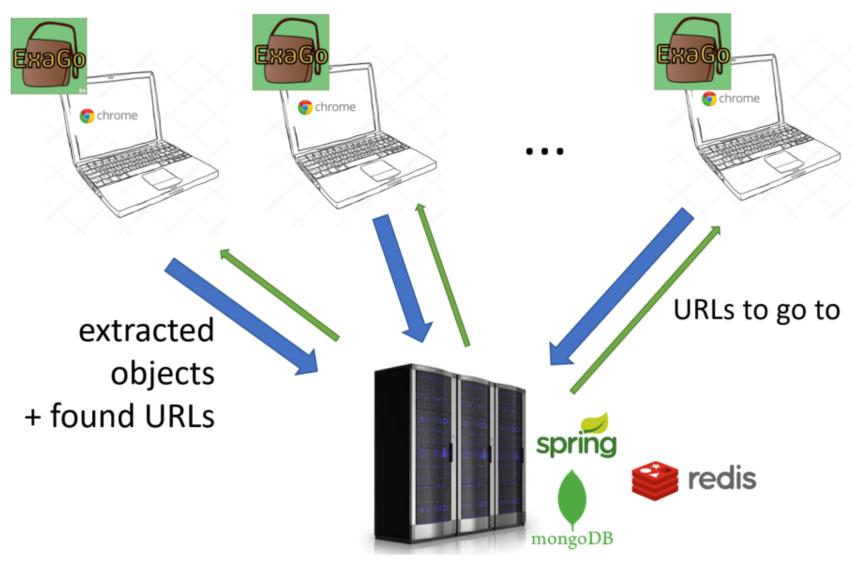
Send wrapper to server

×

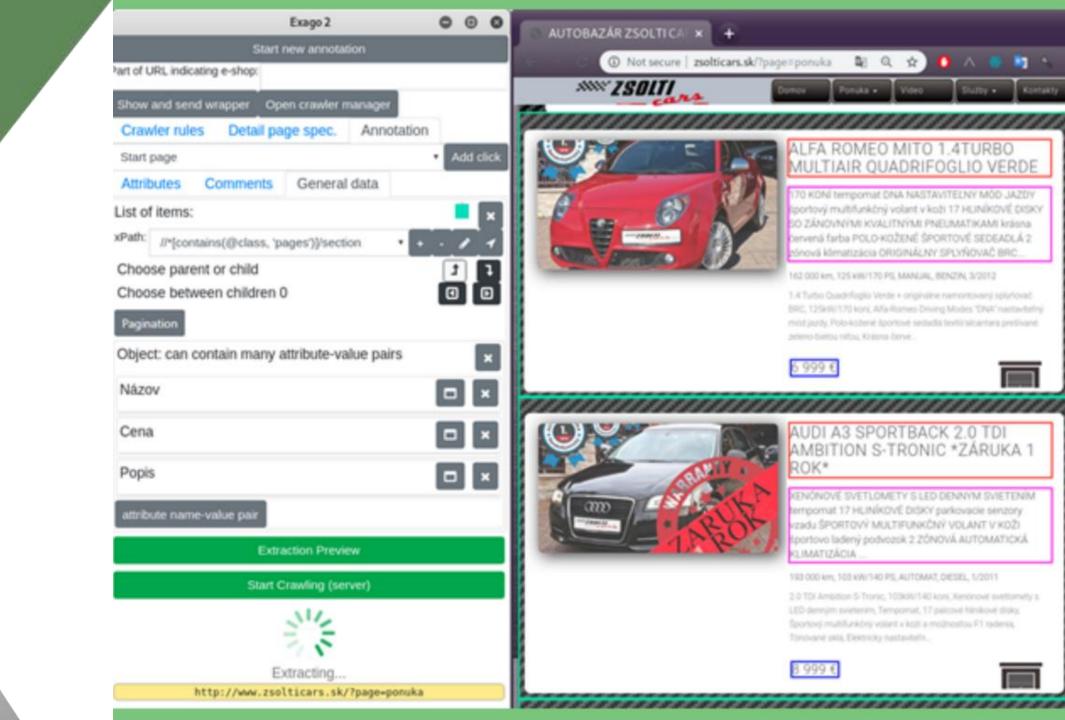




## Web crawling and extraction









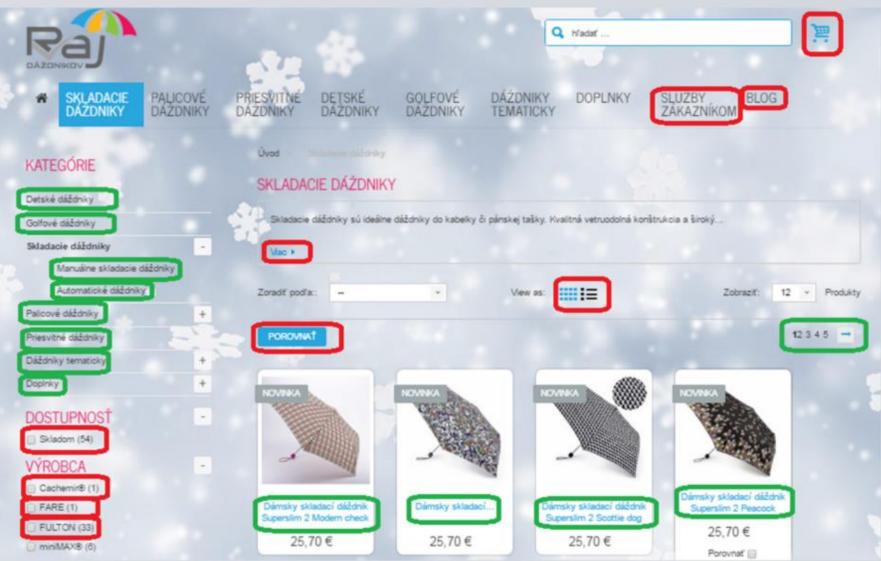
### Repeating webportal data extraction

- crawling with focus to relevant pages
  - other pages can provide links to relevant pages (or not)
    - relevant pages: lists of products, detail pages
    - irrelevant pages: shopping card, contact information, terms of use, ...
  - optimization of crawling process shorten processing time significantly





## **Relevant links**



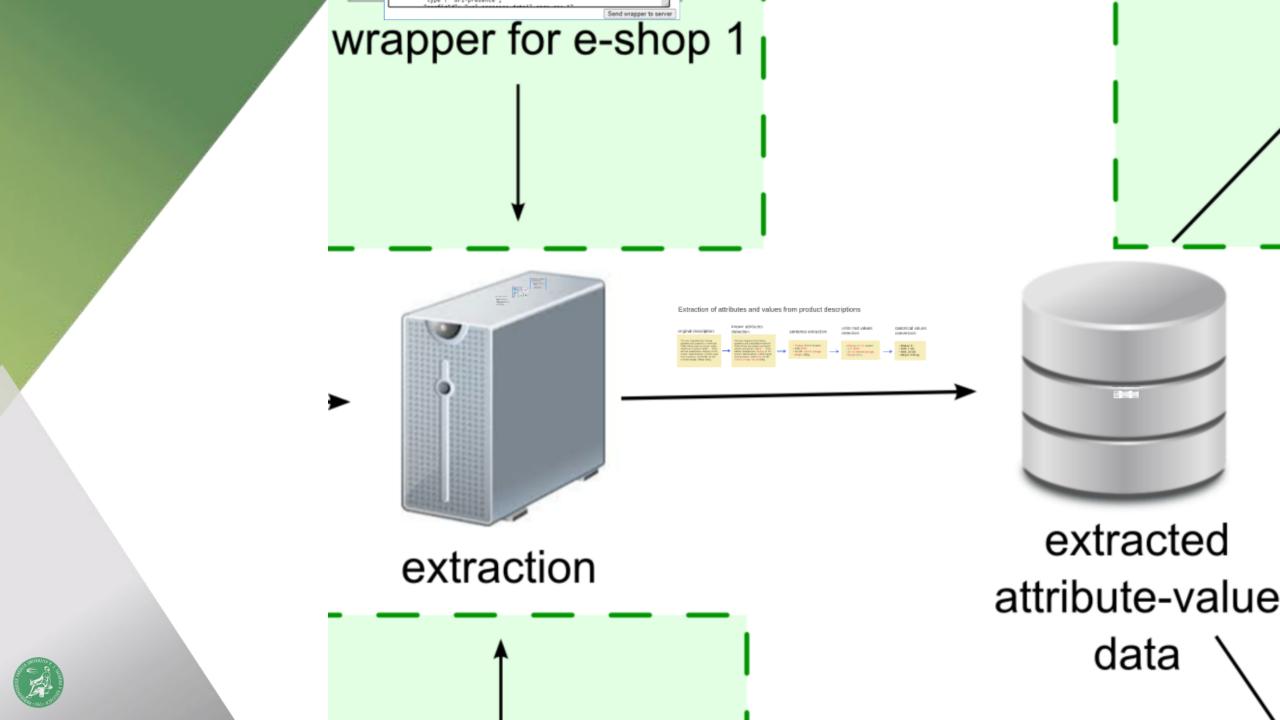


## Web scrapping: Future work

- limiting data transfer for browser background scrapping
  - amount of processed data in my extraction tasks should correlate with scrapping for other people
- scheduling
- postprocessing of extracted data
  - mapping from HTML snippets to specific structures
  - aggregation of data
- dashboard with reports of current tasks
- API for accessing the results
- publication to Chrome web store







#### Extraction of attributes and values from product descriptions







## Extraction of attributes and values

### original description:

The two integrated front-facing speakers and powerfully immersive Dolby Atmos sound give you louder, clearer and dynamic audio . . . Even without headphones. Display: 8 inch screen. Specifications: 1.3GHz quad core processor. 1GB RAM. 16 GB internal storage. Weight 550g.

## known attributes detection:

The two integrated front-facing speakers and powerfully immersive Dolby Atmos sound give you louder, clearer and dynamic audio . . . Even without headphones. Display: 8 inch screen. Specifications: 1.3GHz quad core processor. 1GB RAM. 16 GB internal storage. Weight 550g.



## ributes and values from product de

## known attributes detection:

 $\rightarrow$ 

The two integrated front-facing speakers and powerfully immersive Dolby Atmos sound give you louder, clearer and dynamic audio . . . Even without headphones. Display: 8 inch screen. Specifications: 1.3GHz quad core processor. 1GB RAM. 16 GB internal storage. Weight 550g.

### sentence extraction:

- Display: 8 inch screen.
- 1GB **RAM**.
- 16 GB internal storage.
- Weight 550g.



## rom product descriptions

### sentence extraction:

units nad values detection:

- Display: 8 inch screen.
- 1GB **RAM**.

- 16 GB internal storage.
  - Weight 550g.

• Display: 8 inch screen.

- **1**GB RAM.
- 16 GB internal storage.
- Weight 550g.



scriptions

units nad values detection:

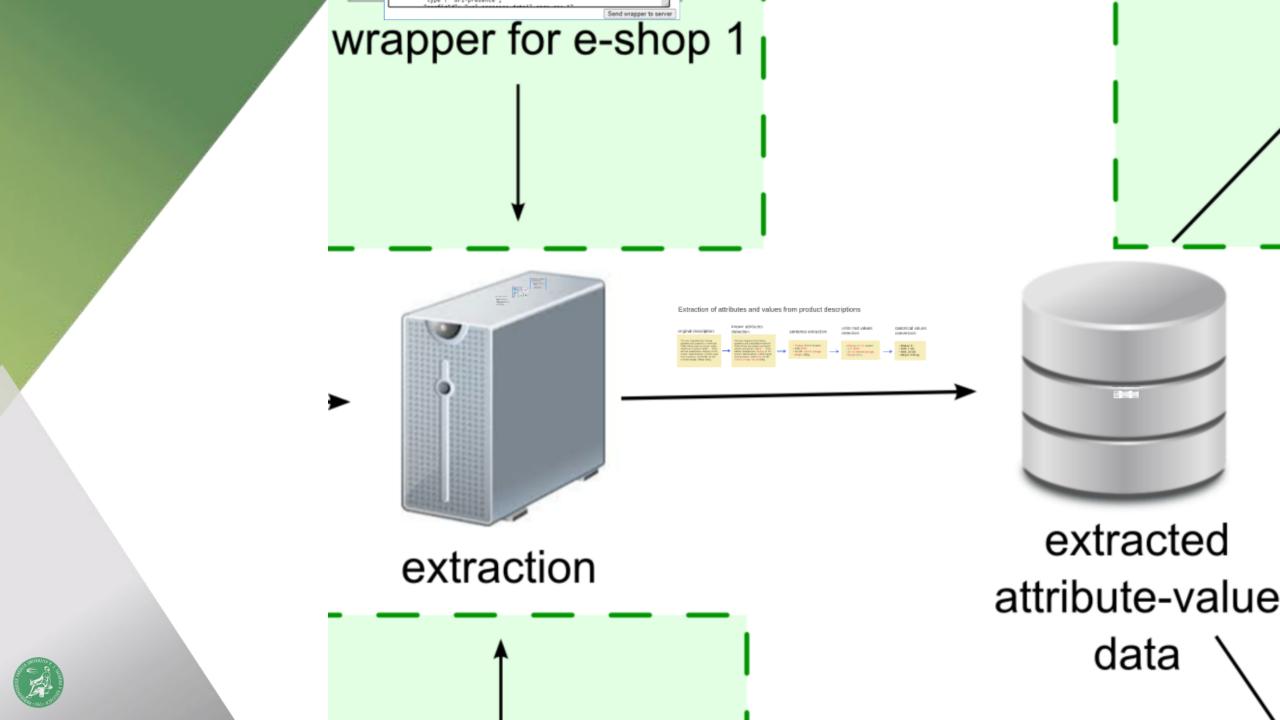
## canonical values conversion:

- Display: 8 inch screen.
- 1GB RAM.
- 16 GB internal storage.
- Weight 550g.

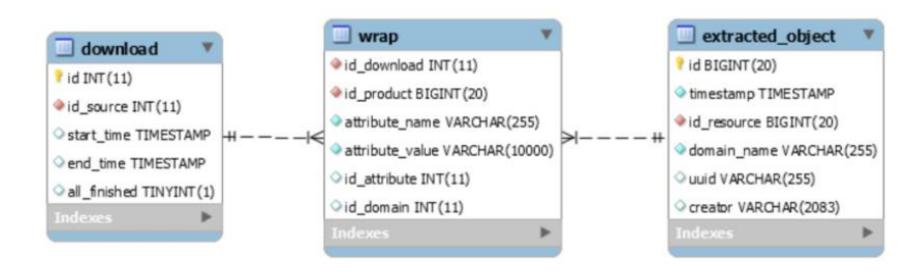


- Display: 8 "
- RAM: 1 GB
- HDD: 16 GB
- Weight: 0.55 kg.



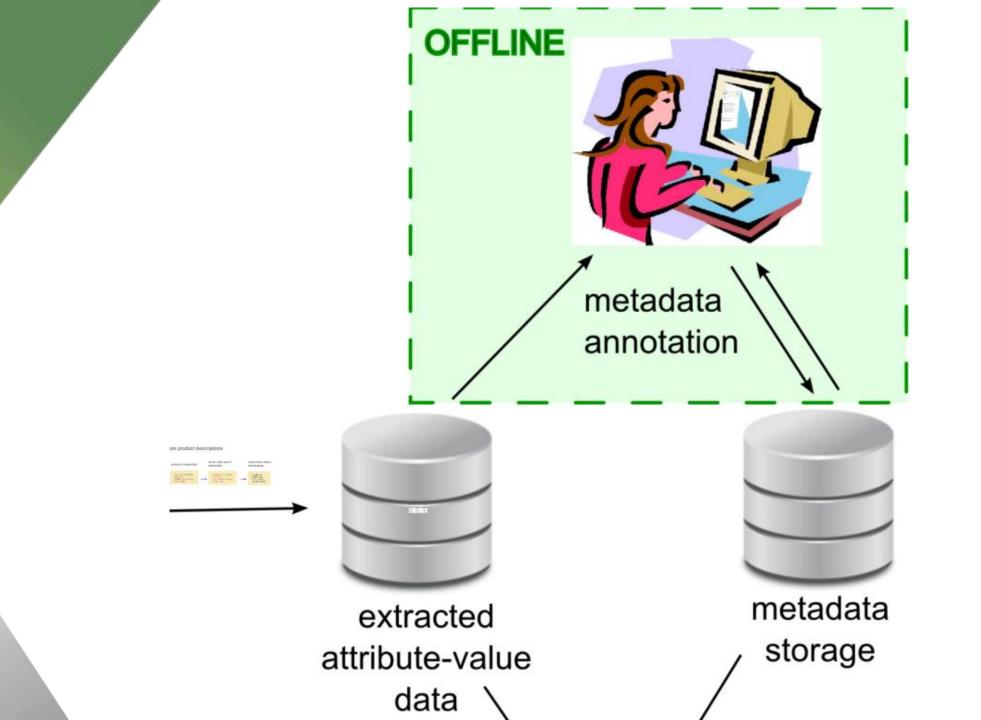


## Canonical structure for product attributes









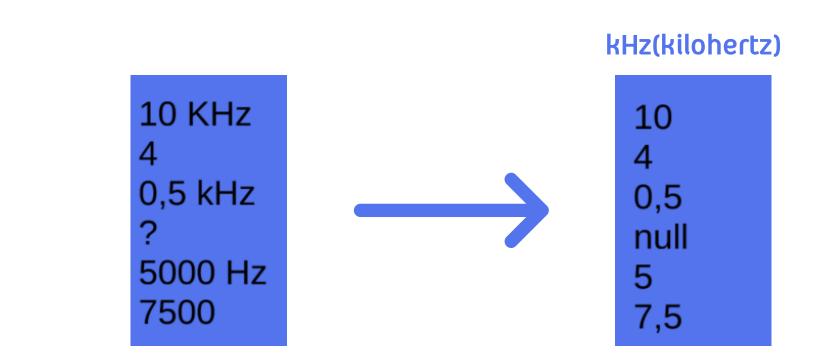


### Metadata annotation: tasks

- Metadata to "always successful" automatic conversion to corresponding domain, attribute, type and form
- Attribute name and domain name mapping
  - synonyms, different languages, different precision of domain specification
- Identifying attribute types
  - continuous, discrete ordinal, nominal, hierarchical, text, datetime, string,...
- Value formats
  - localized formatting, units
- Enumerated attribute types
  - extraction of possible values







- regular expressions for decomposition to value and unit
- default unit

Units

- unit abbreviation variants
- unit conversions
- hidden units
- buggy values notification





## Admin page

Doména Tablety	Derivované atribúty Atribút				
ibút 🗸 Der 🖌 Obyë	Nový derivovaný atribút výšika širka hĺbka	Objem			
	Meno derivovaného atribút	u:	Výraz pre hodnotu:		Výsledné hodnoty atribútu
e Podpora NFC *	Objem		{{\$v * \$s * \$h}} mm3		rysterile notificty at route
Podpora pamä?ovej	Premenné:				427413.6 mm3 (6)
	Meno	Extrahované z	Regulárny výraz	Skupina regexu	206624.7360000003 mm3 (5)
WiFi Prodná kamera					122616.0 mm3 (2)
Procesor	\$name	mena atribútu 💌	(Rozmery.*VxSxH.*)	з Х	206779.32 mm3 (2)
Rozlíšenie displeja					213300.0 mm3 (2)
Rozmery (VxŠxH)	\$v	hodnoty atribú 💌	([o-9,]+)\W*x\W*[o-9,]+\W*x\W*[o-9,]+	1 ×	226590.0000000003 mm3 (2)
Senzory -					237184.2 mm3 (2)
	\$5	hodnoty atribú 💌	[0-9,]+\W*x\W*([0-9,]+)\W*x\W*[0-9,]+	3 ×	259804.1600000006 mm3 (2)
					264000.0 mm3 (2) 264132.39999999997 mm3 (2)
dnoty atribútu	sh	hodnoty atribú 💌	[0-9,]+\W*x\W*[0-9,]+\W*x\W*([0-9,]+)	1 X	299757.6000000003 mm3 (2)
-	911	inventory at the	10 301 (n x(n 10 301 (n x(n 110 301)		329171.98600000003 mm3 (2)
x 178 x 9,2 mm (6) 3 x 113,7 x 9,6 mm (5) =					10203.52 mm3 (1)
3 x 113,7 x 9,6 mm (5)	Pridať premennú				4



# Semiautomated domain and attribute identification using full-text search (J.Kuper)

- Preventing same attribute or domain to be twice in the system
- "Schema unification" problem
- Our solution
  - values of attributes are more immune to synonyms and language differences
  - using fulltext index to find most similar products to currently annotated products with the same domain value
  - extract "promissing" domains and attributes from returned products
  - decomposition of non-atomic attributes

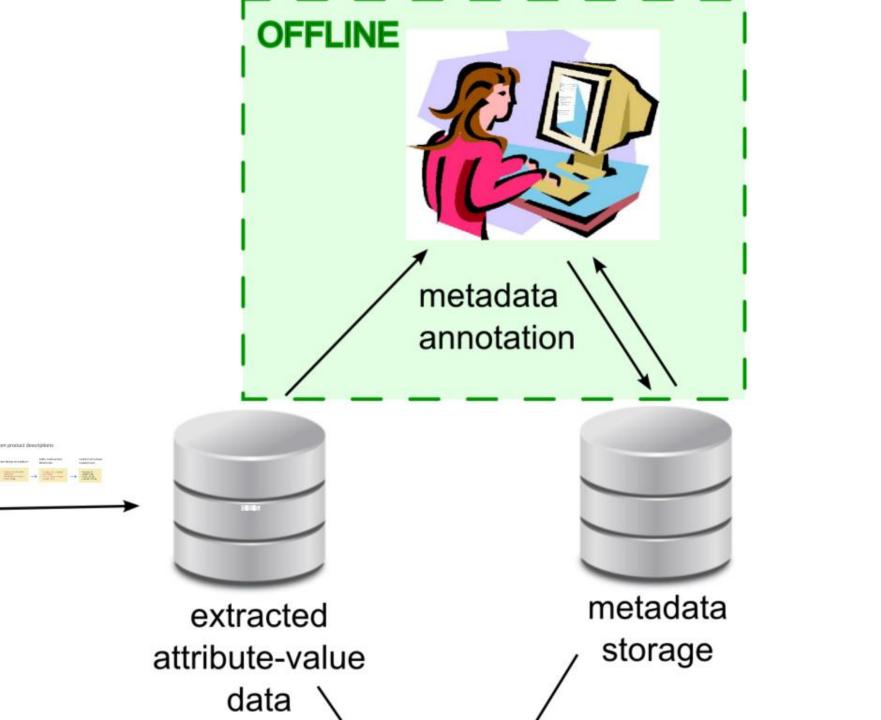




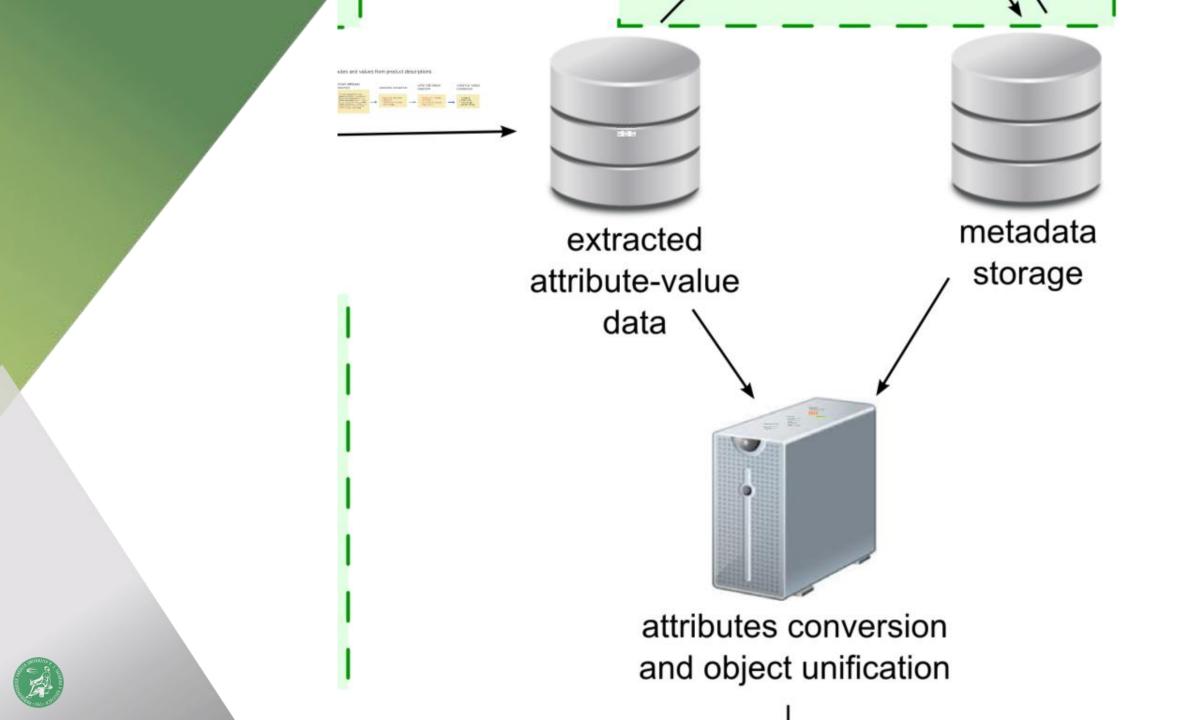
### Metadata annotation : open problems

- ordinal string type (XS, X,M,L,XL,..; C,B,A,A+,A++) and their conversions
- handling unsuccessful conversions









## Product unification

• First step : conversion from pile of "object – attribute – string value" rows to table for each domain having row for each extracted product with attribute values of appropriate type

wrap

id_download	id_product	attribute_name	attribute_value
10	235	domain	notebook
10	235	HDD	500 GB
10	235	size	13"
10	236	domain	notebook
10	236	HDD	1 TB
10	236	size	11

wrap\_joined\_notebook

id_source	id_product_wrap	id_product	HDD	size
5	235	null	500	13
5	236	null	1000	11





### Product unification

- second step: find same products and assign them common id
  - new products have new id
  - similar products have new id and a hint for administrator
- 3 types of attributes
  - determining (ISBN, EAN, local web id)
  - relevant (name, weight, size,...)
  - other (description, price, on stock,...)
- source dependance
  - local web id, price,.. are source dependent
  - name, weight, size,... are source independent

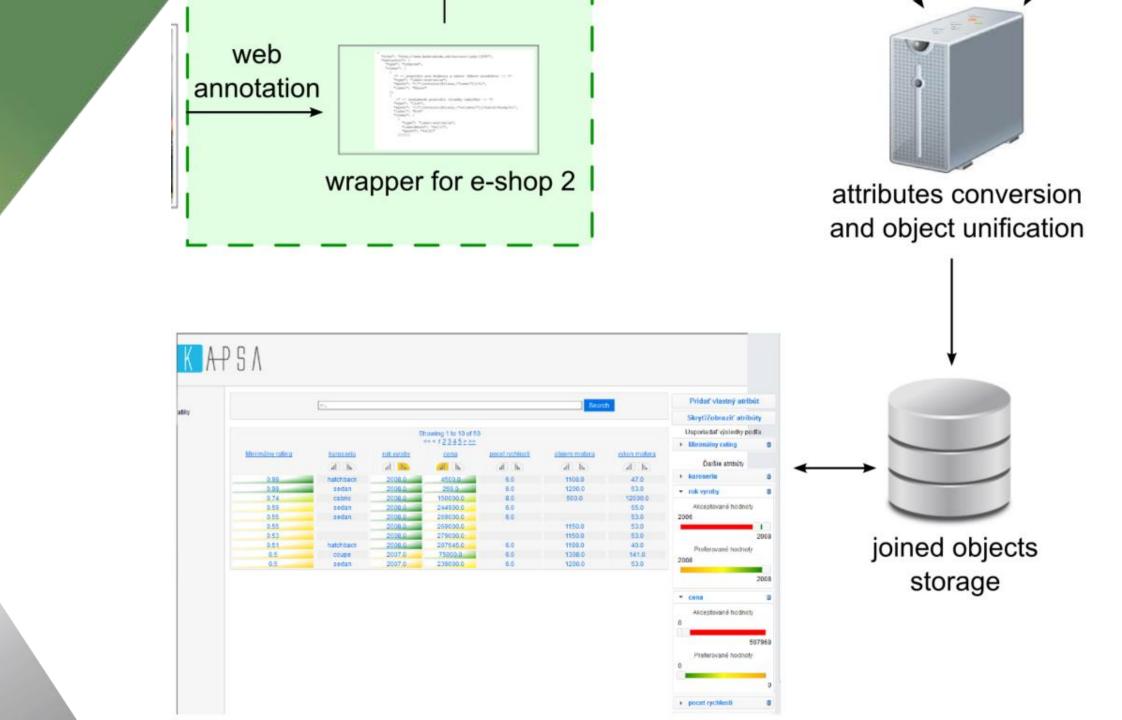


## Product unification: open problems

- administrator UI to handle similar objects
- hierarchy of domains
- many missing values -> false equality
  - using product images









		(Fe				Searc	5	Pridať vlastný atribú	t
atiky								Skryt/Zobrazit' atribút	y.
			Sh	owing 1 to 10 of 50				Usporiadať výsledky podľ	a
			-	<<12345222				Minimålny rating	1
	Minimálny rating	karoseria	rok wroby	cena	pocet rychlosti	objem motora	wkon motora		
		al la	al In	al 16	al la	all Ite	al h	Ďalšie atribúty	
	0.99	hatchback	2008.0	4500.0	6.0	1108.0	47.0	karoseria	1
	0.99	sedan	2008.0	269.0	6.0	1200.0	53.0	<ul> <li>rok vyroby</li> </ul>	
	0.74	cabrio	2008.0	150000,0	8.0	500.0	12000.0	- Tok syrooy	- 10
	0.59	sedan	2008.0	244900.0	6.0		55.0	Akceptované hodnoty	
	0.55	sedan	2008.0	259000.0	6.0		53.0	2006	
	0.55		2008.0	259000.0		1150.0	53.0		I.
	0.53		2008.0	279000.0		1150.0	53.0		200
	0.51	hatchback	2008.0	287645.0	6.0	1198.0	40.0	Preferované hodnoty	
	0.5	coupe	2007.0	75000.0	6.0	1308.0	141.0		
	0.5	sedan	2007.0	239000.0	6.0	1200.0	53.0	2008	





### Full-text with structural result

#### Matching domains Bezdrôtové audio (3) Hmotnosť (1) Názov (2) Konektory (3) Sluchadla (2) cena (2) Operacny system (2) Opercny system (2) Názov (2) cena (1) Výrobca (12) Uhlopriečka displeja (5) Lokálny identifikátor (1) TV a video (10) Názov (10) ✓ → Mobilny telefon (36) Názov (36) ✓ → Hifi veza (5) Názov (5)

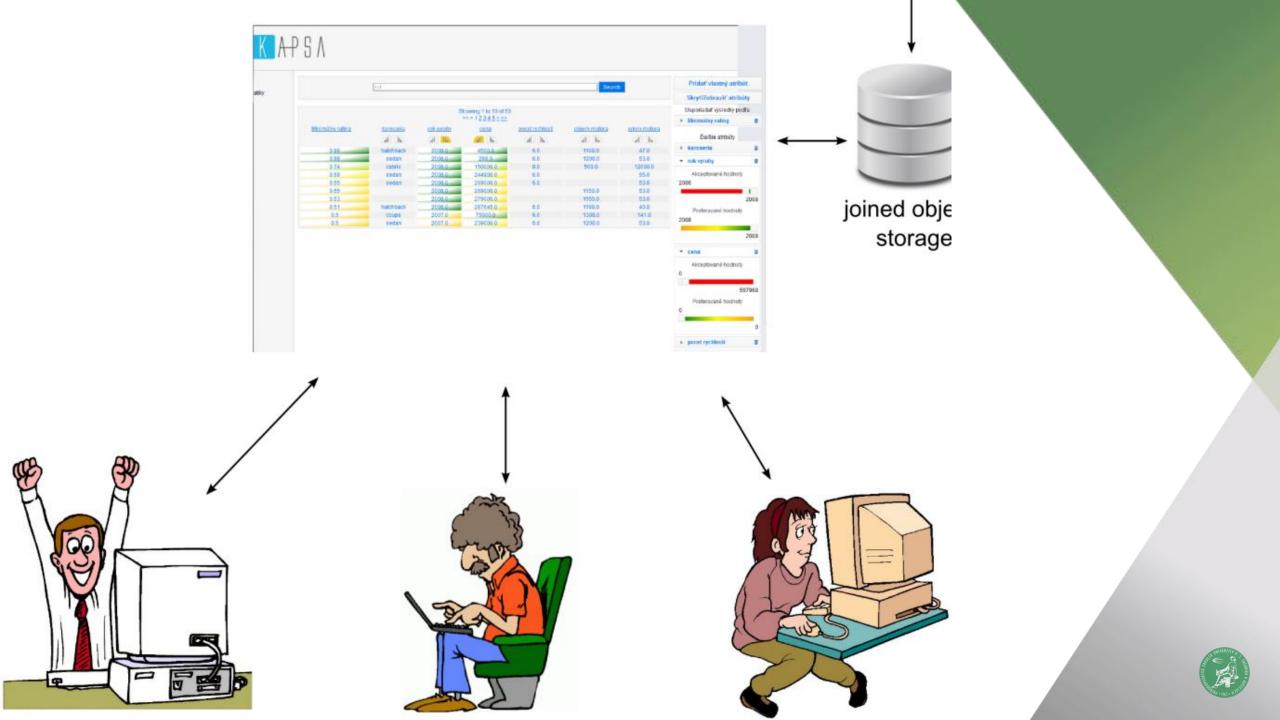
ľ	-Search	for
	-	

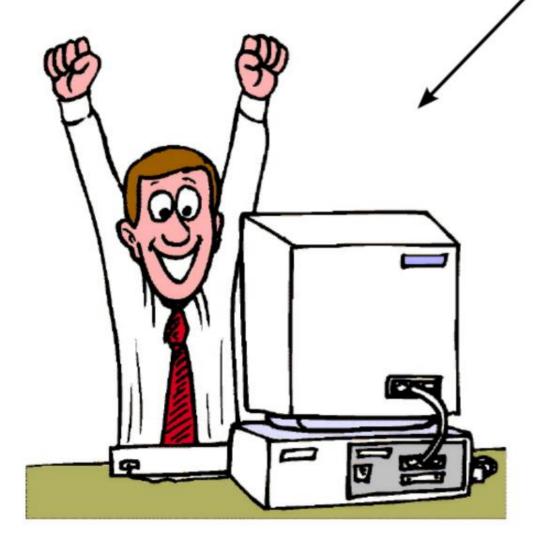
Samsung 7"

Search

#### Product name Domain Bezdrôtové Id produktu = 69914 domain = [Bezdrôtové audio, null] Bluetoott Samsung WAM551 audio Id produktu = 68642 domain = [Tablety, Tablets] Bluetooth - ano Samsung Tablety WAM551 Samsung WAM551 Id produktu = 68646 domain = [Tablety, Tablets] Bluetooth - áno Samsung Tablety WAM551 Samsung WAM551 Id produktu = 68652 domain = [Tablety, Tablets] Bluetooth - áno Samsung Tablety WAM551 Samsung WAM551 Id produktu = 68628 domain = [Tablety, Tablets] Bluetooth - áno Samsung Tablety WAM551 Samsung WAM551 Id produktu = 70017 domain = [Hifi veza, null] Bluetooth - áno W Samsung MX-Hifi veza H630 Samsung MX-H630 Samsung MX-Id produktu = 68616 domain = [Tablety, Tablets] Bluetooth - áno Tablety H630 - Samsung MX-H630 Id produktu = 68619 domain = [Tablety, Tablets] Bluetooth - áno Samsung MX-Tablety - Samsung MX-H630 H630 Id produktu = 68622 domain = [Tablety, Tablets] Bluetooth - áno Samsung MX-Tablety - Samsung MX-H630 H630 Id produktu = 70197 domain = [Sluchadla, null] Bluetooth - áno ) Philips SHE2001 Sluchadla identifikátor - 78618 Názov - Philips SHE2001







## Thank you ! Any questions?



## Thank you for tour attention



Jesenná 5. Košice ics.science.upjs.sk ui@upjs.sk

