

Prepájanie multimediálnych metadát

DP

Bc. Peter Gašpar

Vedúci: Ing. Jakub Šimko, PhD.

DOMÉNA MULTIMÉDIÍ

PROBLÉMOVÁ DOMÉNA

TV program (a jeho entity)

SÚČASNÝ STAV

dominujú **statické informácie**

ZDROJ DYNAMIKY

používatelia a ich aktivita na **sociálnych sieťach**

TV Stanica

9:00 Relácia A

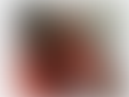
10:00 Relácia B

10:30 Relácia C

11:30 Relácia D

 ešte lepšie to dal Mišo Hudák s tým pánom zvonárom v jednom z minulých dielov:) toho uhádol myslím bez predchádzajúcich otázok a rovno sa to spýtal 😊

Páči sa mi to · Odpovedať · 👍 3 · 23 hod.

 Presne tak, ved zábava chyba v TV, myslím na úrovni zábava. Nie farma atď

Páči sa mi to · Odpovedať · 👍 3 · Včera o 15:41

- rôzne druhy príspevkov
- správanie televízií a divákov
- **problémy:**
mnohojazyčnosť, preklepy, nespisovnosť,
nestabilné správanie

Ako nájsť prepojenie medzi TV
programom a sociálnou sieťou?

CIELE PRÁCE

1

Vytvoriť metódu **mapovania** medzi televíznym programom a príspevkom sociálnej siete.

TV Stanica	
9:00	Relácia A
10:00	Relácia B
10:30	Relácia C
11:30	Relácia D

?

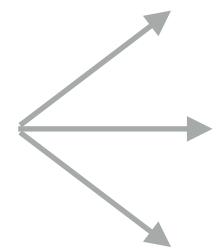


2

Využiť vytvorené mapovanie na extrakciu ďalších metadát.

EXISTUJÚCE PRÍSTUPY

- ➔ **POUŽITÉ ČRTY**
text (obsah) príspevku, hashtagy,
obsah zo znalostných báz
#hashtag
→ problém novej položky, slovenské dáta
- ➔ **PRÁCA S PRIRODZENÝM JAZYKOM**
n-gramy, TF-IDF
ng, gr, ra, am, my
→ slabá odolnosť voči preklepom
- ➔ **METÓDY MAPOVANIA**
prístupy založené na pravidlách,
strojové učenie
→ potreba anotovaných dát

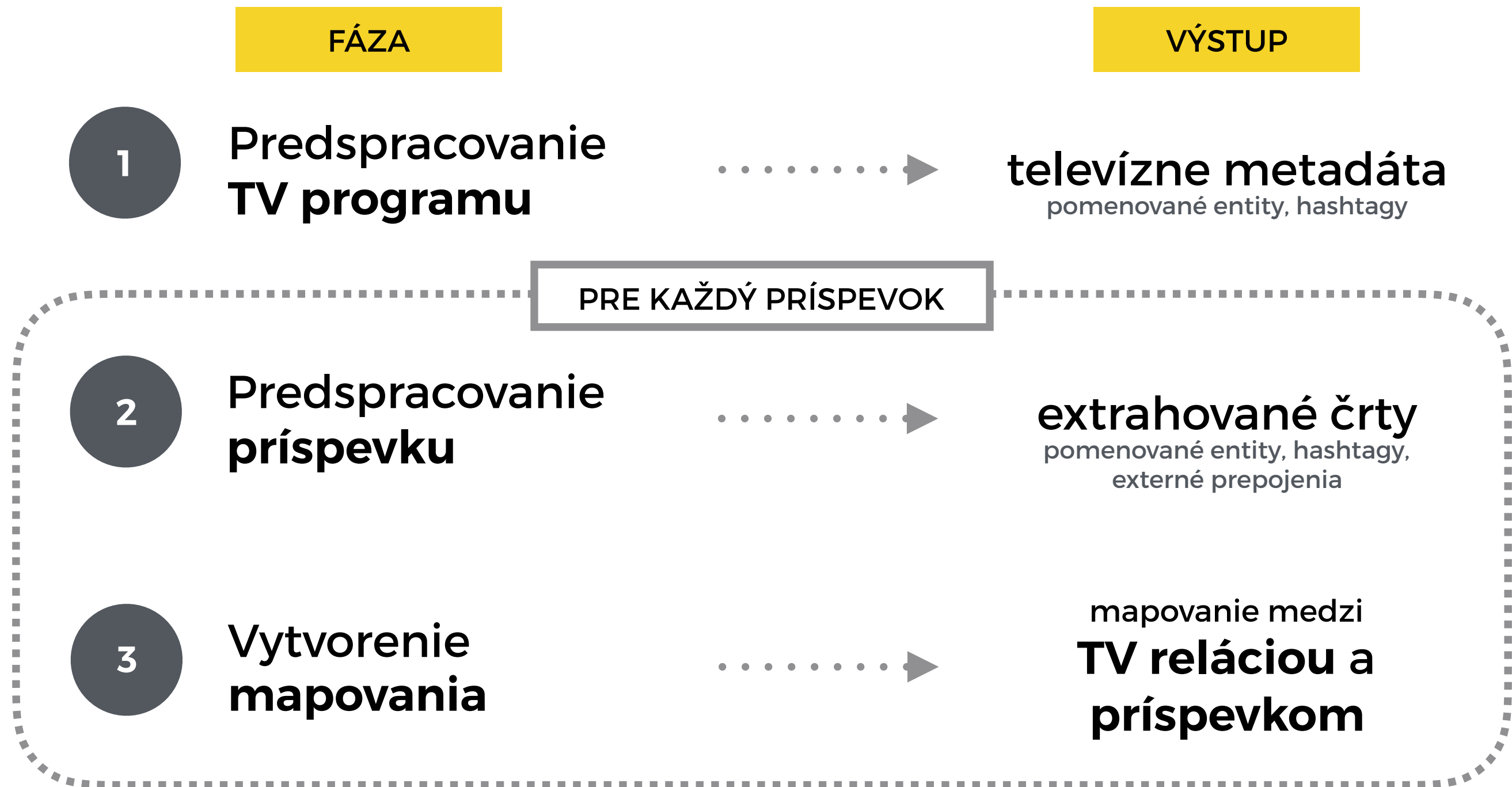


PRÍSPEVOK ZO SOCIÁLNEJ SIETE

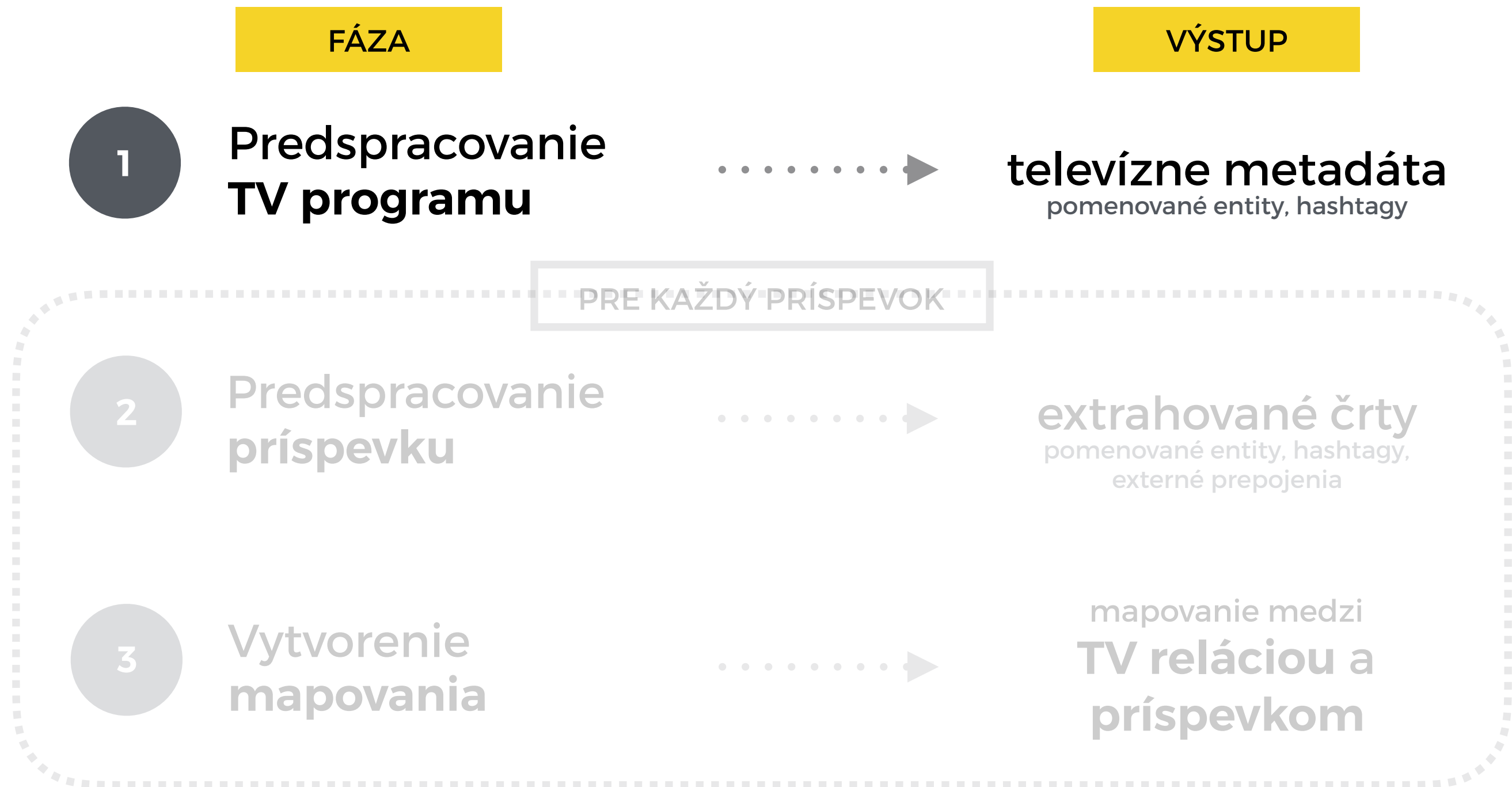


HLAVNÁ MYŠLIENKA:
Podobné črty vieme získať z TV programu.

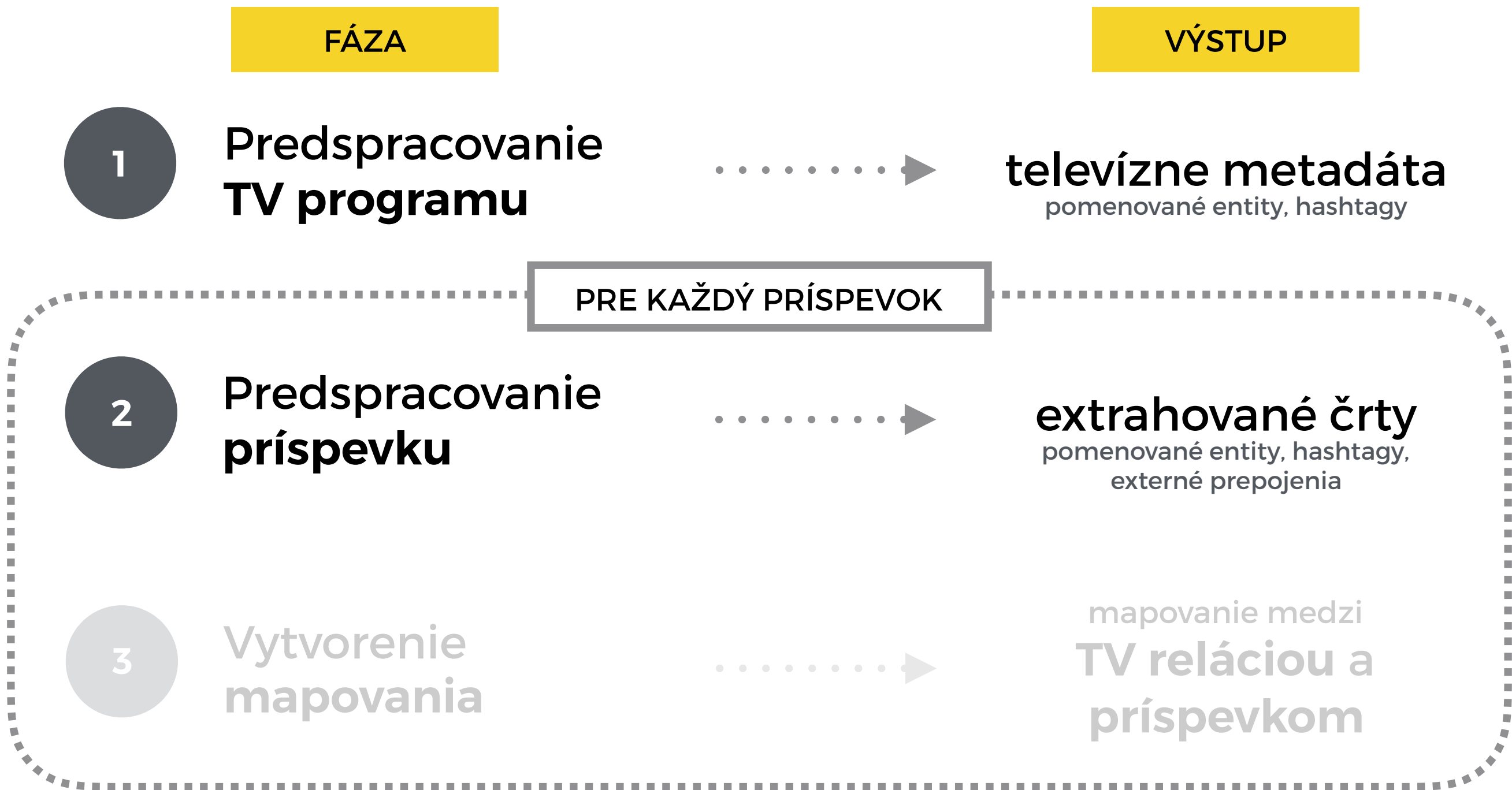
NÁVRH PROCESU MAPOVANIA



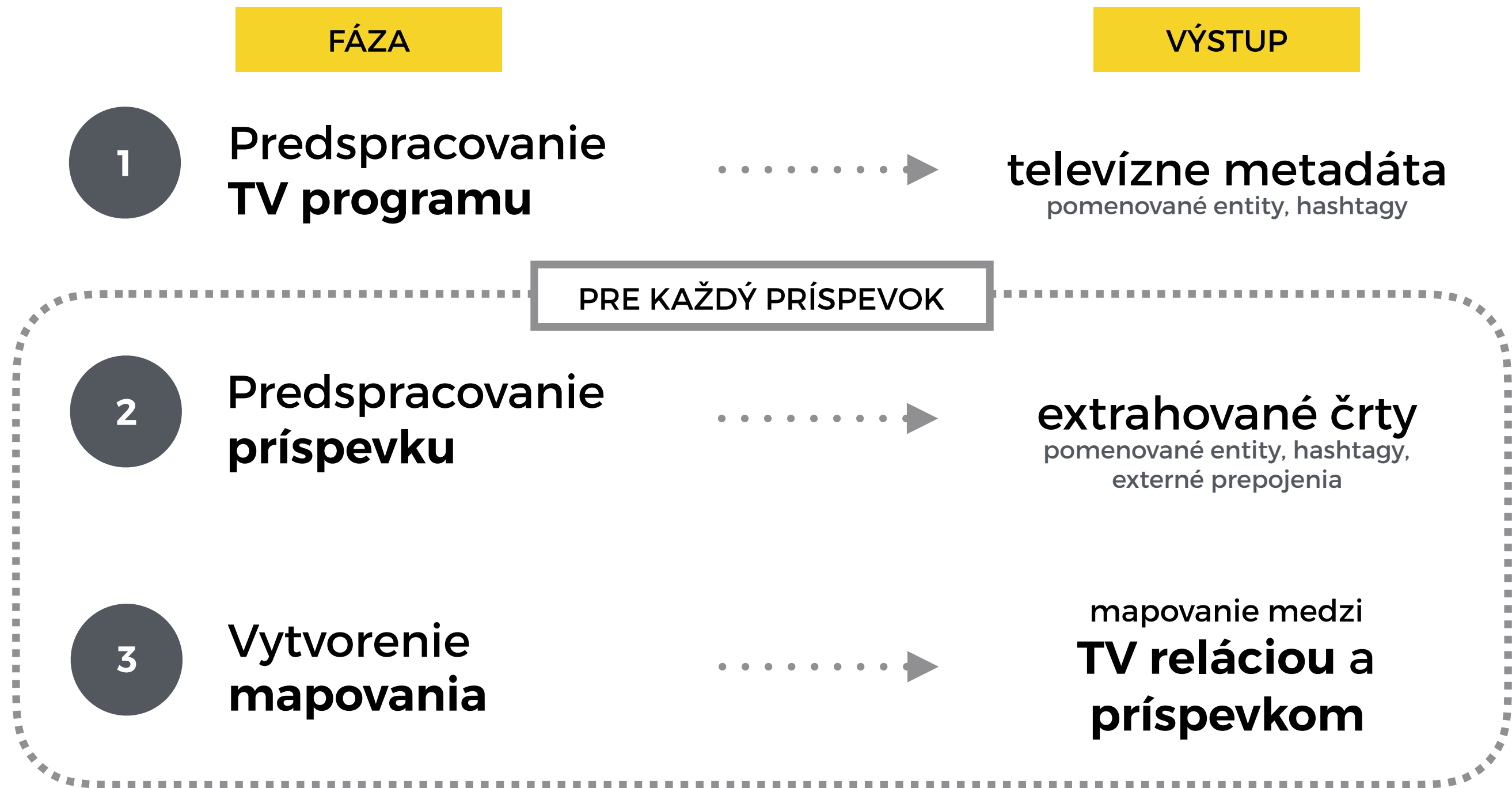
NÁVRH PROCESU MAPOVANIA



NÁVRH PROCESU MAPOVANIA



NÁVRH PROCESU MAPOVANIA



PREDSPRACOVANIE TV PROGRAMU

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

GENEROVANIE **TELEVÍZNYCH METADÁT:**

TV Stanica

9:00 Relácia A

10:00 Relácia B

10:30 Relácia C

11:30 Relácia D

→ Pomenované entity

Teória veľkého tresku► teoria velkeho tresku

→ Hashtagy

Teória veľkého tresku► teoriavelkehotresku,
tv

→ Prefixové slová

Teória veľkého tresku► teor
Čo ja viem► coja

PREDSPRACOVANIE PRÍSPEVKU

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

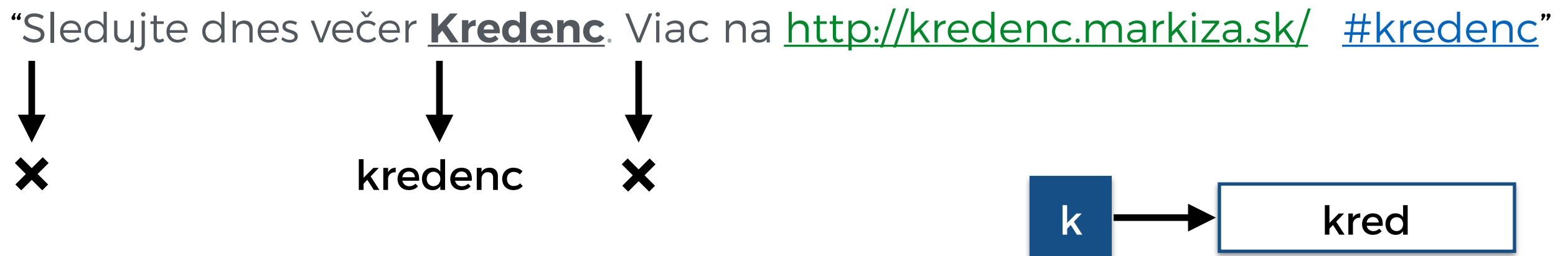
3 Vytvorenie
mapovania

EXTRAKCIA ČŔT Z PRÍSPEVKU:

POMENOVANÁ ENTITA

Generujeme **slovné n-gramy** zo slov s veľkým začiatočným písmenom.

Podmienka: prítomné v prefixovom indexe



PREDSPRACOVANIE PRÍSPEVKU

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

EXTRAKCIA ČŔT Z PRÍSPEVKU:

HASHTAG

Slová začínajúce na #.

“Sledujte dnes večer **Kredenc**. Viac na <http://kredenc.markiza.sk/> [#kredenc](https://www.facebook.com/kredenc)”



kredenc

PREDSPRACOVANIE PRÍSPEVKU

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

EXTRAKCIA ČŔT Z PRÍSPEVKU:

EXTERNÉ PREPOJENIE

Rozdelenie URL na doménu, subdoménu a podčasti.

Súčasťou procesu je aj expanzia skrátenej URL adresy
(napr. <http://bit.ly/1WDF9VE> -> <http://www.fiit.stuba.sk>).

<http://kredenc.markiza.sk/> → kredenc, markiza

<http://relacie.markiza.sk/cl/kredenc/> → relacie, markiza,
cl, kredenc

PREDSPRACOVANIE PRÍSPEVKU

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

EXTRAKCIA ČŔT Z PRÍSPEVKU:

POMENOVANÁ ENTITA

generujeme **slovné n-gramy** zo slov s veľkým
začiatočným písmenom

HASHTAG

slová začínajúce na #

EXTERNÉ PREPOJENIE

rozdelenie URL na doménu,
subdoménu a podčasti

“Sledujte dnes večer **Kredenc**. Viac na <http://kredenc.markiza.sk/> [#kredenc](#)”

VYTVORENIE MAPOVANIA

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

TELEVÍZNE METADÁTA

PRÍSPEVOK SOCIÁLNEJ SIETE

POMENOVANÉ ENTITY

POMENOVANÉ ENTITY

HASHTAGY

HASHTAGY

ČASTI EXTERNÉHO PREPOJENIA

VYTVORENIE MAPOVANIA

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

TELEVÍZNE METADÁTA

PRÍSPEVOK SOCIÁLNEJ SIETE

POMENOVANÉ ENTITY

POMENOVANÉ ENTITY

HASHTAGY

HASHTAGY

ČASTI EXTERNÉHO PREPOJENIA

metrika podobnosti kandidátov:
n-gramová podobnosť
(Kondrak, 2005)

VYTVORENIE MAPOVANIA

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

TELEVÍZNE METADÁTA

PRÍSPEVOK SOCIÁLNEJ SIETE

POMENOVANÉ ENTITY

POMENOVANÉ ENTITY

HASHTAGY

HASHTAGY

ČASTI EXTERNÉHO PREPOJENIA

metrika podobnosti kandidátov:
binárna podobnosť

= 1, ak sa dva hashtagy líšia v najviac 1 znaku
= 0, inak

VYTVORENIE MAPOVANIA

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

TELEVÍZNE METADÁTA

PRÍSPEVOK SOCIÁLNEJ SIETE

POMENOVANÉ ENTITY

POMENOVANÉ ENTITY

HASHTAGY

HASHTAGY

ČASTI EXTERNÉHO PREPOJENIA

metrika podobnosti kandidátov:
n-gramová podobnosť s prefixom

/superstar2016/ ←.....→ superstar
/CBStopnews/ ←.....→ topnews
/supeeerstar/ ←.....→ superstar

VYTVORENIE MAPOVANIA

1 Predspracovanie
TV programu

2 Predspracovanie
príspevku

3 Vytvorenie
mapovania

PRÍSPEVOK SOCIÁLNEJ SIETE

extrahovaná
pomenovaná entita

TELEVÍZNE METADÁTA

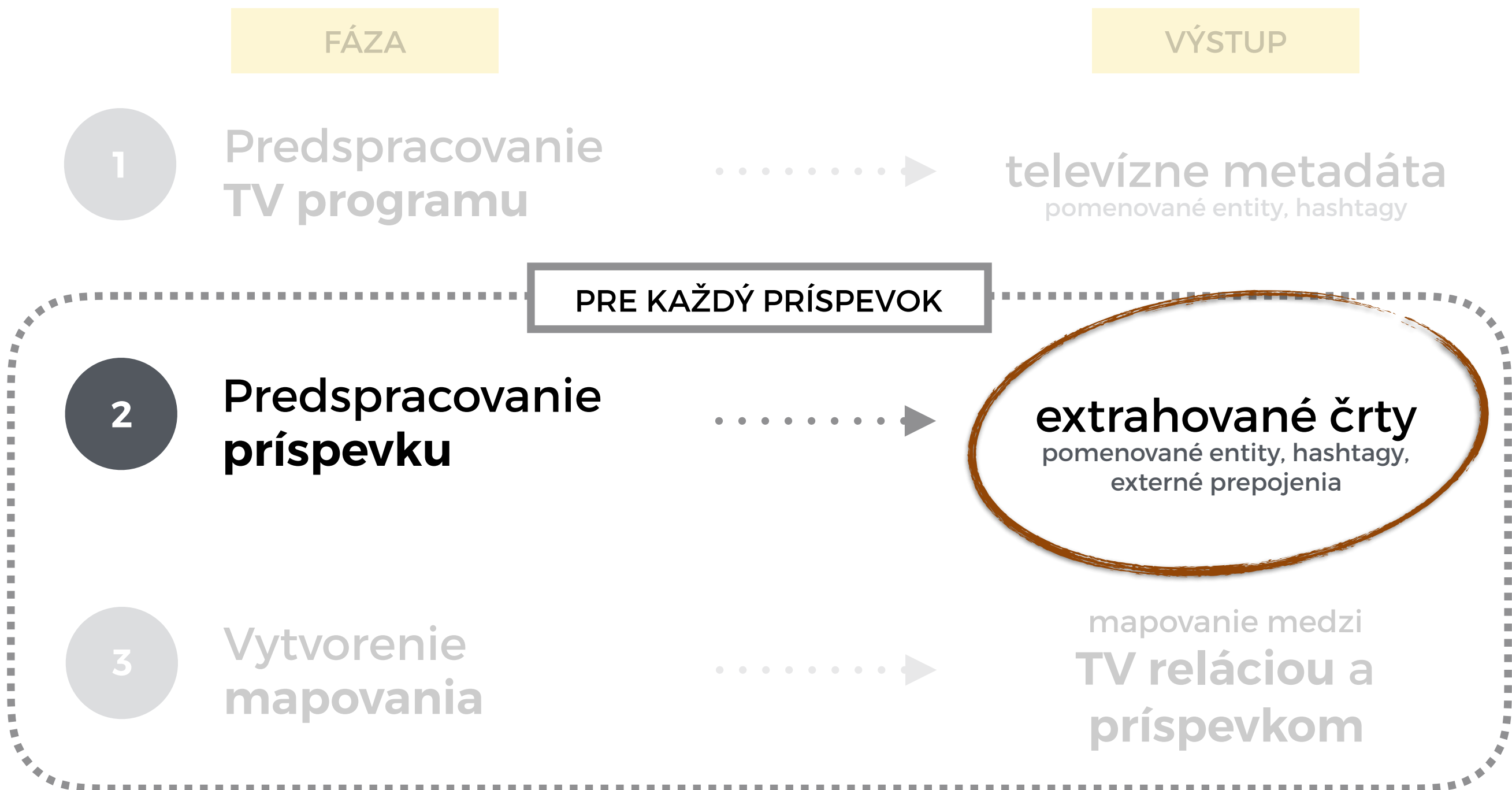


invertovaný index
pomenovaných entít

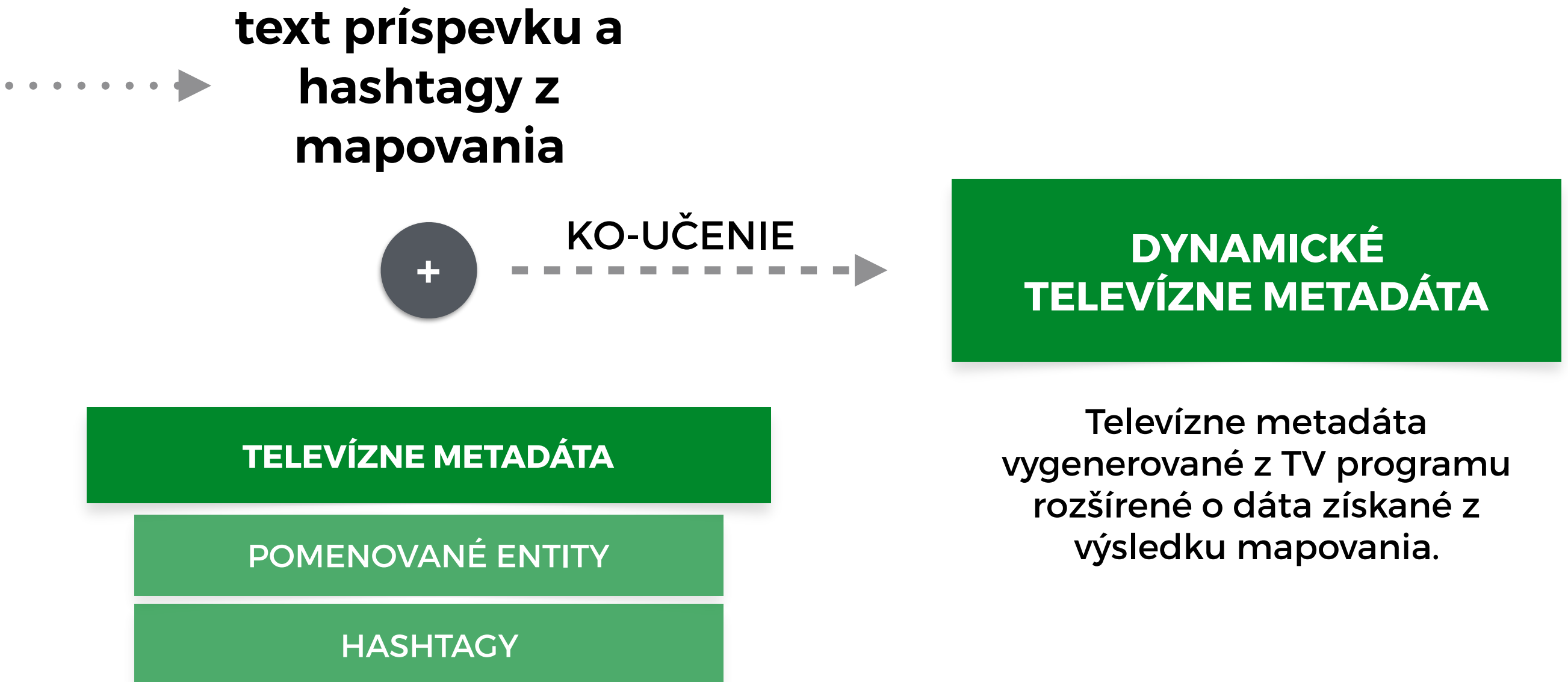
- 1) Vypočítame mieru súvislosti.
- 2) Získame televízne relácie.
- 3) Vyberieme vhodné relácie.



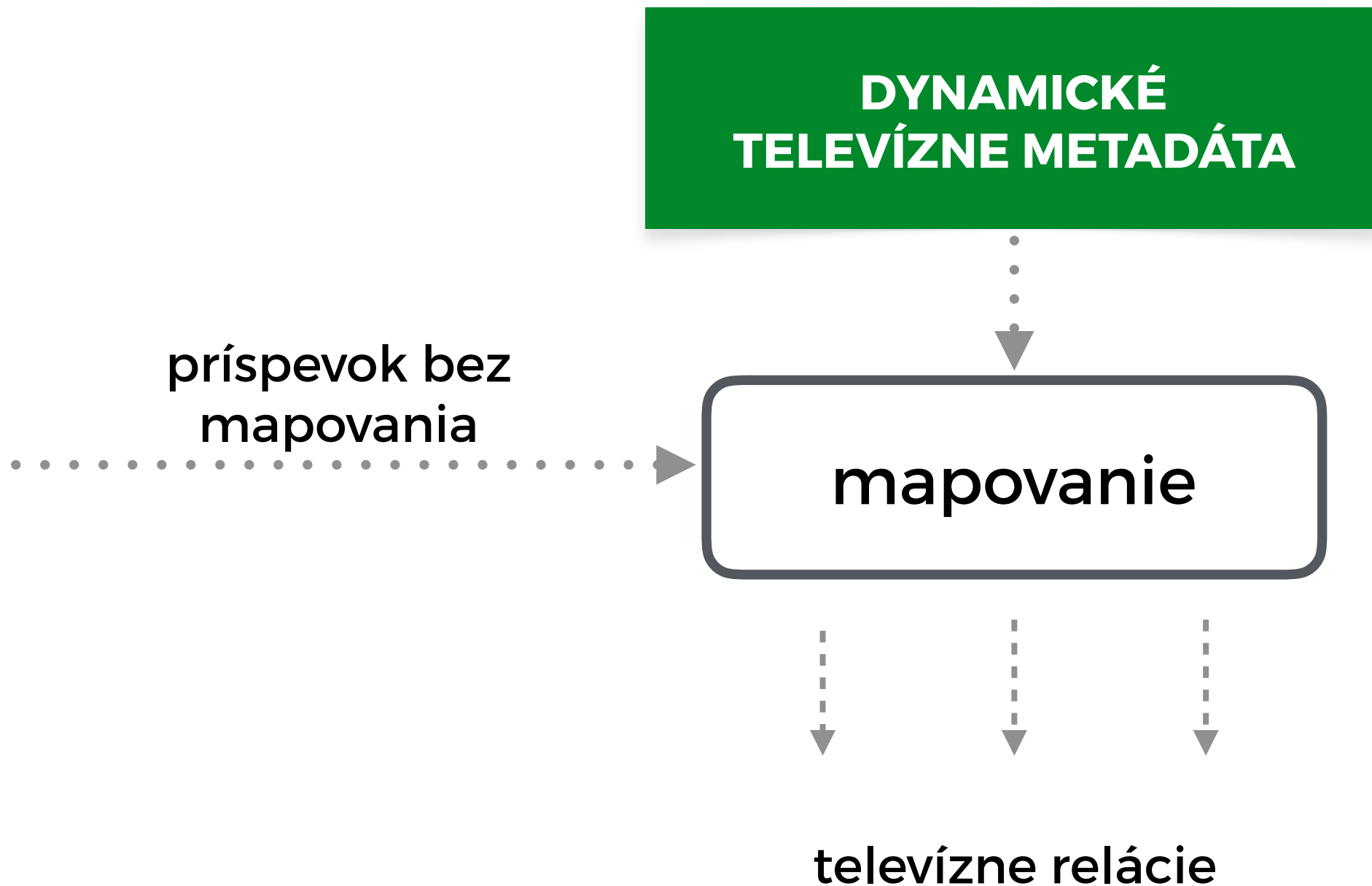
NÁVRH PROCESU MAPOVANIA



Problém: príspevok nevieme namapovať.



KO-UČENIE (2)



NAŠA HYPOTÉZA

Navrhnuté vylepšenia:

- prefixový index,
- pridanie externých prepojení,
- ko-učenie.

prinesú zlepšenie skúmanej metriky.

PRÍPRAVA DÁTOVEJ MNOŽINY

- získanie **TV programu** (10 staníc)
slovenské: Markiza, JOJ
zahraničné (anglické): CBS, NBC, Fox, History, CW, USA, Showtime, HBO
- stiahnutie a manuálna anotácia **príspevkov zo sociálnej siete Facebook**
- spolu **1300 príspevkov**
500 slovenské, 800 anglické

NÁVRH EXPERIMENTU (2)

METRIKA OVERENIA

- zamerali sme sa na **presnosť**
- usporiadanie príspevkov do **tried** (= TV relácia)

$$\text{presnosť} = \frac{\text{počet správne namapovaných príspevkov}}{\text{počet vrátených príspevkov}}$$

- výsledná presnosť — vážený priemer (podľa veľkosti triedy)

STANOVENIE PARAMETROV (2)

Pri výsledkom mapovaní sme zavrhovali mapovanie s mierou súvislosti pod **hranicou** (hraničná súvislosť).

Hranicu sme stanovili experimentálne na **THRES = 0.70**.

	Bez hranice	THRES = 0.70
Bez prefixového indexu	0.29	0.92
S prefixovým indexom	0.58	0.93

Pri niektorých televíziách priniesol prefixový index zlepšenie o 8%.

POROVNaNIE NASTAVENÍ

→ Samostatné črty

Hashtagy 0.30

Externé prepojenia 0.51

Pomenované entity 0.85

→ Kombinované črty

Hashtagy 0.67
Externé prepojenia

Pomenované entity 0.86
Hashtagy

Pomenované entity 0.91
Externé prepojenia

→ Úplné nastavenie

Pomenované entity 0.93
Hashtagy
Externé prepojenia

ĎALŠIE EXPERIMENTY

- **Príspevky bez mapovania (bez anotácie)**
87 / 89 úspešných (97.75% úspešnosť)
- **Prínos ko-učenia**
mapovanie nevrátilo výsledok pre 51 príspevkov
ko-učenie namapovalo 18 príspevkov (35,29% úsp.)
- **Jazyková závislosť a presnosť**

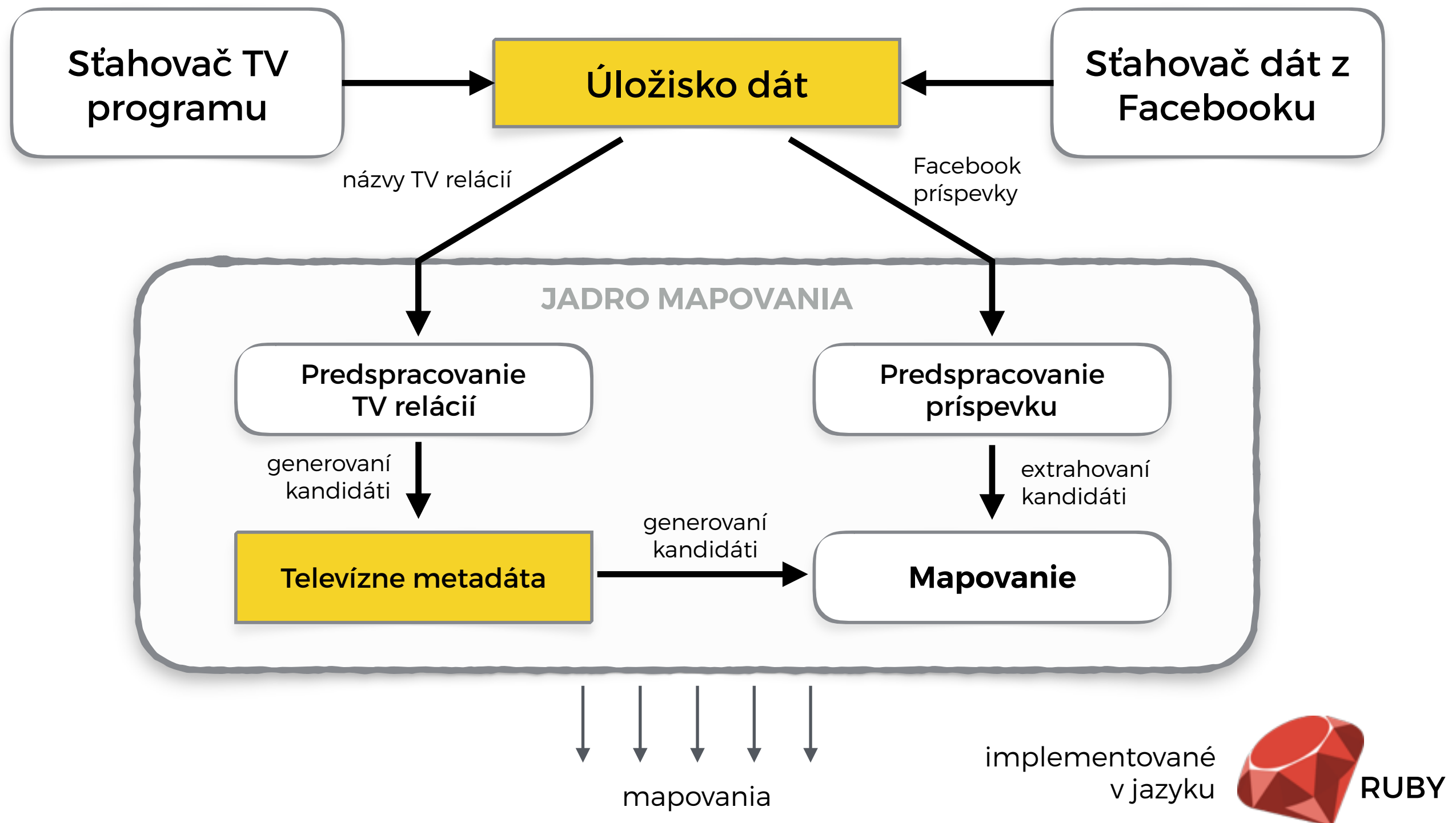
slovenské príspevky	0.90 (500)
anglické príspevky	0.93 (800)

ZHODNOTENIE VÝSLEDKOV

- ➔ **Nová črta príspevku – externé prepojenie**
zlepšuje presnosť v porovnaní s klasickými prístupmi
- ➔ **Minimalizácia vstupných dát**
absencia strojového učenia a problému novej položky
vstup: TV program a priradenie TV stanice profilu na soc. sieti
jazykovo nezávislé
- ➔ **Ko-učenie ako perspektíva**

- Existujúce prístupy využívajú strojové učenie na vytvorenie mapovania medzi TV programom a sociálnou sieťou.
- Vykonali sme **analýzu** príspevkov zo sociálnej siete.
- Identifikovali sme zaujímavú črtu príspevku – **externé prepojenie**.
- Navrhli a overili sme metódu mapovania využívajúcu **pomenované entity, hashtagy a externé prepojenia**.
- Metóda je jazykovo-nezávislá, nevyžaduje si strojové učenie a dokáže sa vysporiadať s novou položkou.

PREHL'AD RIEŠENIA



ZLATÝ ŠTANDARD

Každému príspevku sme prideliť množinu súvisiacich TV relácií na základe:

- **textu príspevku** (pomenované entity, hashtagy, postavy),
- **komentárov**,
- **interných (sociálna sieť) a externých odkazov.**

Obmedzenie:

- **TV relácie iba analyzovanej TV stanice.**

ZLATÝ ŠTANDARD (2)

@NBC

TAKES
ON INS



NBC

28 marec · 🌐

Follow along as **Nick Gehlfuss** takes over our Instagram today!
<http://instagram.com/nbctv/>



#CHICAGOMED

NBC

please ... A. D. The Bible Continues is the perfect series continues! no one had ever focused on what happened after Jesus Christ ... even for non-Catholics is series is very good, why has to do with the Roman yoke of time ... is a historical marvel! we know that was canceled, but only asks them to reconsider, and think well, and so, hopefully, can resume.
[#PleaseResumeADTheBibleContinues](#)
👍 2 · 28 marec, 10:45

It is time for every struggling American to get behind Bernie Sanders. He is fighting the elite, for rigging the economic and judicial system against us. Let's end the greedy billionaire suppression of all Americans. Vote prosperity for all, vote Bernie Sanders.
👍 4 · 28 marec, 9:46

He looks like Bob Saget
😊 not a bad thing, just an observation.
18 april, 9:54

Fail now
👍 1 · 28 marec, 18:53

😊
👍 1 · 28 marec, 13:04

ZLATÝ ŠTANDARD (3)

NBC
28 marec · 🌐

Follow along as **Nick Gehlfuss** takes over our Instagram today!
<http://instagram.com/nbctv/>

Nick Gehlfuss
10 máj · 🌐

Only 2 more episodes in Season 1! Check. It. Out. **#ChicagoMed**

please ... A. D. The Bible Continues is the perfect series continues! no one had ever focused on what happened after Jesus Christ ... even for non-Catholics is series is very good, why has to do with the Roman yoke of time ... is a historical marvel! we know that was canceled, but only asks them to reconsider, and think well

He looks like Bob Saget
😊 not a bad thing, just an observation.
18 april, 9:54

Fail now
👍 1 · 28 marec, 18:53

👍 1 · 28 marec, 13:04

#CHICAGOMED

NBC