

Lehounký úvod: Data o využívanosti EIZ

zjednodušeně a nesprávně též „uživatelské
statistiky“

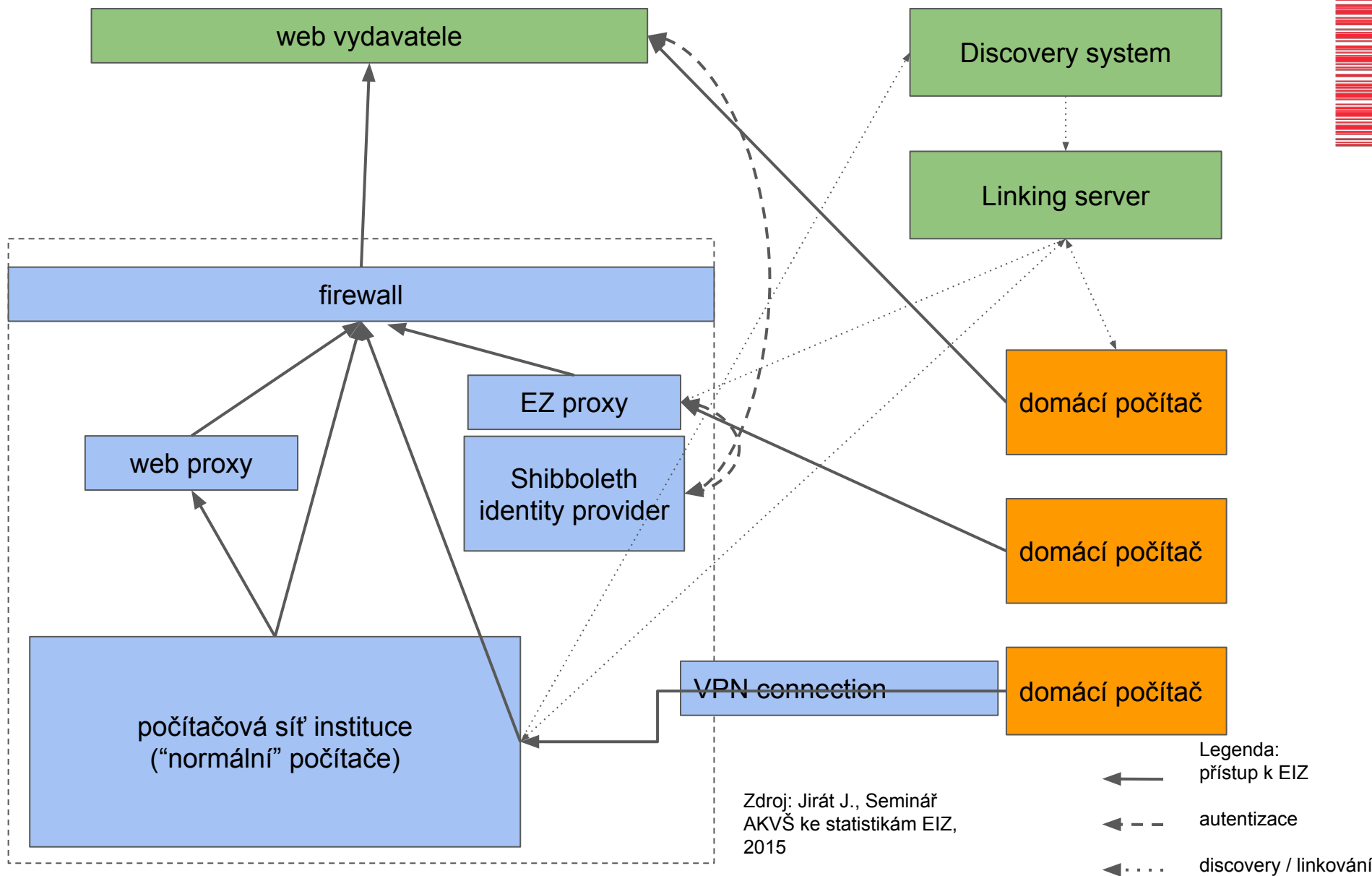
Jiří Jirát (CzechELib/VŠCHT/AKVŠ)

- Co a jak měřit
- Standard COUNTER, novinky v COUNTER 5
- Implementace v systému Celus
- Celus: demo

Poznámky

- Předpokládáme, že jste aspoň trochu slyšeli o COUNTER 4.
- Online seminář je zkrácená verze, v případě zájmu může CzechELib ve spolupráci s pracovní skupinou pro EIZ AKVŠ uspořádat detailní kontaktní seminář týkající se statistik využívanosti EIZ.
- Prezentace záměrně ponechává mnoho termínů v angličtině, aby uživatelé měli usnadněnou orientaci v datech a nástrojích používajících terminologii COUNTER 5.

Co a jak měřit



Srovnání

Místo měření	Formát dat	Obsah	Nevýhody	Podíl na celkovém provozu (odhad)
firewall	low-level	všechn provoz	na hranici zákona	> 99 % (ale díky Shibbolethu bude část provozu čím dál víc zcela mimo doménu instituce)
web proxy	low-level	všechn HTTP, HTTPS	dtto	30 – 40 % (podle instituce)
EZproxy	low-level	víceméně jen HTTP a HTTPS k EIZ	nelze jednoduše zjistit, zda se jednalo o OA, free nebo placený obsah	< 5 % (podle instituce), ale např. Francie má centrální proxy
linking server	zpracovaná data	data k e-časopisům a e-knihám	údaje o OA, free, placeném obsahu jen na úrovni titulů	< 5 %
data od vyd. (poskytovatele)	zpracovaná data	detailní data k EIZ jednoho poskytovatele	věříme jim?	100 % (ale nemusí obsahovat údaje o čtenosti dokumentů v repozitářích)

COUNTER

Counting Online Usage of Networked Electronic Resources

<https://www.projectcounter.org/>

nezisková organizace podporovaná globální komunitou knihoven, vydavatelů a poskytovatelů, kteří přispívají k vývoji Code of Practice (pracovními skupinami a komunikací)

COUNTER – historie

	Publikován	Platnost
Release 1 of the Code of Practice for Journals and databases	January 2003	
Release 2 of the Code of Practice for Journals and databases	April 2005	
Release 1 of the Code of Practice for Books and Reference Works	March 2006	
Release 3 of the Code of Practice for Journals and databases	August 2008	
Release 4 of the Code of Practice for e-Resources	April 2012	
The COUNTER Code of Practice for Release 5 (pak ještě 5.0.1)	July 2017	účinný od ledna 2018, deadline pro implementaci 2019-02-28 (with support for January 2019 usage)
Připravuje se aktualizace - Release 5.1		

COUNTER-compliant vendor

- musí podstoupit roční nezávislý audit
 - tj. nestačí, že data jen vypadají jako COUNTER reporty
 - i vytvoření dat musí splňovat podmínky standardu
- seznam “COUNTER-compliant” poskytovatelů:
 - <https://www.projectcounter.org/about/register/>
 - pouze ti, kteří jsou zde uvedeni
- musí poskytovat reporty uvedené v aktuálně platném “Release”

Co přináší COUNTER 5

COUNTER 4 vs. COUNTER 5

- COUNTER 5 je oproti COUNTER 4 velkým skokem vpřed
- Zavádí množství atributů, které lze různě kombinovat => velká flexibilita v analýze dat

COUNTER 4

- pouze standardní předdefinované reporty
 - s pevně nastavenými a vybranými parametry (JR1, JR1GOA, ..., BR2, ...)
 - typy položek oddělené a separované každá ve svých reportech (časopisy, kapitoly knihy, celé knihy ...) atd.
- SUSHI*
 - XML formát pro jednotlivé typy reportů
 - typ API: SOAP

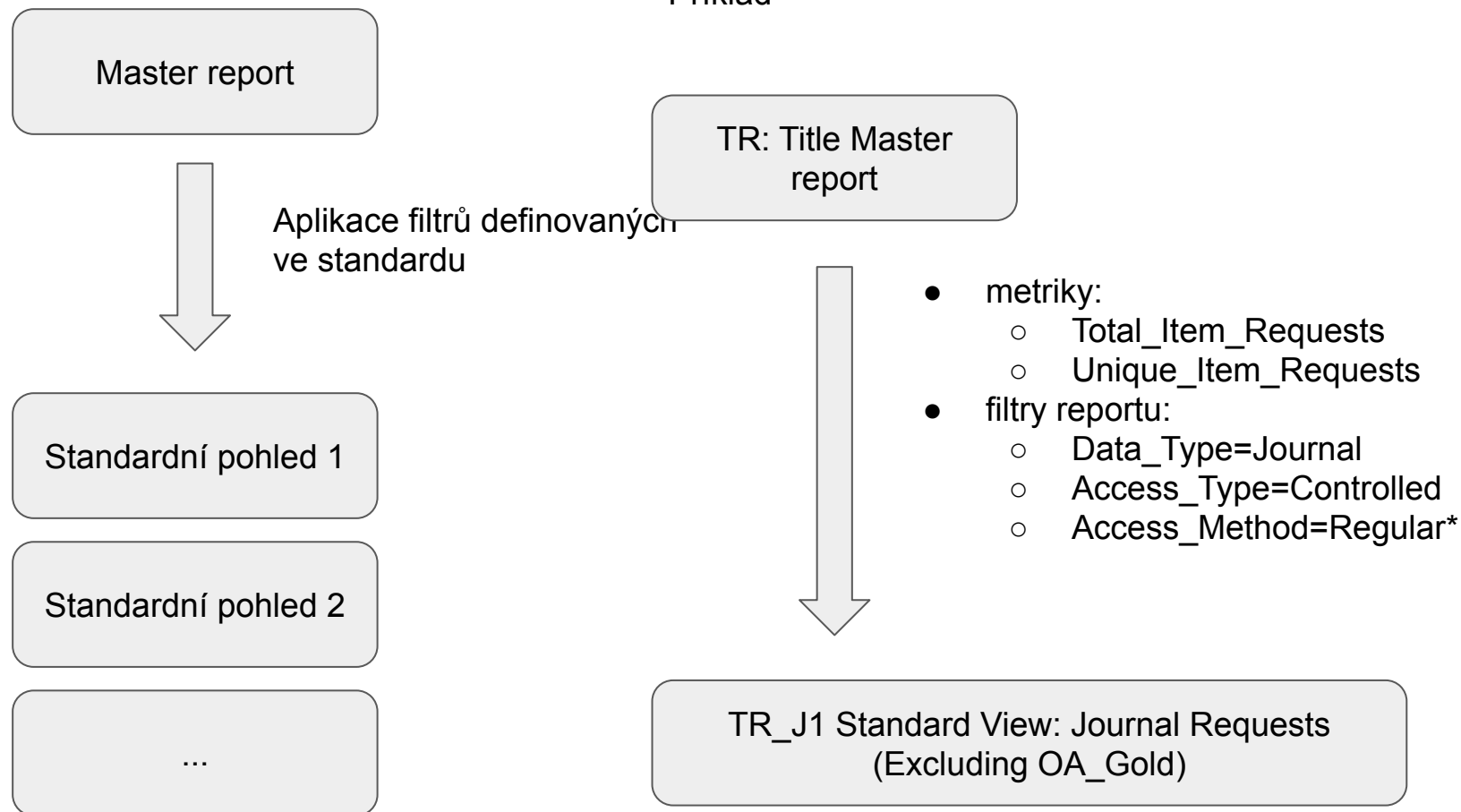
* SUSHI (Standardized Usage Statistics Harvesting Initiative) - automatické stahování

COUNTER 5

- Master reporty
 - PR (Platform Master Report), TR (Title...), DR (Database...), IR (Item...)
 - obsahují všechny relevantní metriky a atributy, všechny typy publikací
 - plně upravitelné aplikací filtrů a dalších konfigurací na...
- ... standardní pohledy (Standard Views)
 - reporty odvozené aplikací filtrů z Master reportů
- SUSHI
 - formát JSON
 - typ API: REST
- Nové atributy/hodnoty
 - Unique items, Unique titles
 - Investigations
 - Access Method, Access Type, ...
 - ...

Master reports a Standard views

Příklad



Ukázka dat (jedna položka - kniha)

▼ Report_Items:

▼ 0:

- ▶ Title: "Advances in Photoelectro...nt and Systems Analysis"

Data_Type: "Book"
Access_Method: "Regular"
YOP: "2018"
Access_Type: "Controlled"



atributy

▼ Item_ID:

▼ 0:
Type: "DOI"
Value: "10.1039/9781782629863"

▼ 1:
Type: "Proprietary"
Value: "rsc:9781782629252"

▼ 2:
Type: "Print_ISSN"
Value: "2044-0782"

▼ 3:
Type: "URI"
Value: "https://doi.org/10.1039/9781782629863"

Platform: "RSC eBooks"
Publisher: "Royal Society of Chemistry"

▶ Publisher_ID: [...]

▼ Performance:

▼ 0:
Period:
Begin_Date: "2020-03-01"
End_Date: "2020-03-31"



sledované období
(březen 2020)

Instance:

▼ 0:
Metric_Type: "Unique_Title_Investigations"
Count: 10
▼ 1:
Metric_Type: "Unique_Title_Requests"
Count: 10



metriky

Master reports a Standard Views

Master report (MR)		Standardní reporty (Standard views)
Zkratka	Název	
PR	Platform MR	PR_P1 (Platform-level usage summarized by Metric_Type)
DR	Database MR	DR_D1 (Database Search and Item Usage) DR_D2 (Database Access Denied)
TR	Title MR	TR_B1 (Book Requests (Excluding OA_Gold)) TR_B2 (Book Access Denied) TR_B3 (Book Usage by Access Type) TR_J1 (Journal Requests (Excluding OA_Gold)) TR_J2 (Journal Access Denied) TR_J3 (Journal Usage by Access Type) TR_J4 (Journal Requests by YOP (Excluding OA_Gold))
IR	Item MR	IR_A1 (Journal Article Requests) - jen pro repozitáře, nebo "Scholarly Collaboration Network" IR_M1 (Multimedia Item Requests)

- **Metric_Type (Metrika)**

- Hledání
 - Searches_Automated / Searches_Federated / Searches_Regular
- Využití
 - Total_Item_Investigations / Total_Item_Requests / Unique_Item_Investigations / Unique_Item_Requests / Unique_Title_Investigations / Unique_Title_Requests
- Odmítnutí
 - Limit_Exceeded / No_License

- **Data_Type (Typ dat)**

- Book / Journal / Report / Database / Other / ...

- **Access_Method (Způsob přístupu)**

- Regular / TDM

- **Access_Type (Typ přístupu)**

- Controlled / OA_Gold / Other_Free_To_Read

- **YOP (“Year of Publication” - Rok vydání)**

- **Section_Type**

- Chapter / Book / Article / Other / ...

Ne vždy (u všech reportů) jsou všechny atributy vyžadovány nebo dávají smysl

Hledání (Searches)

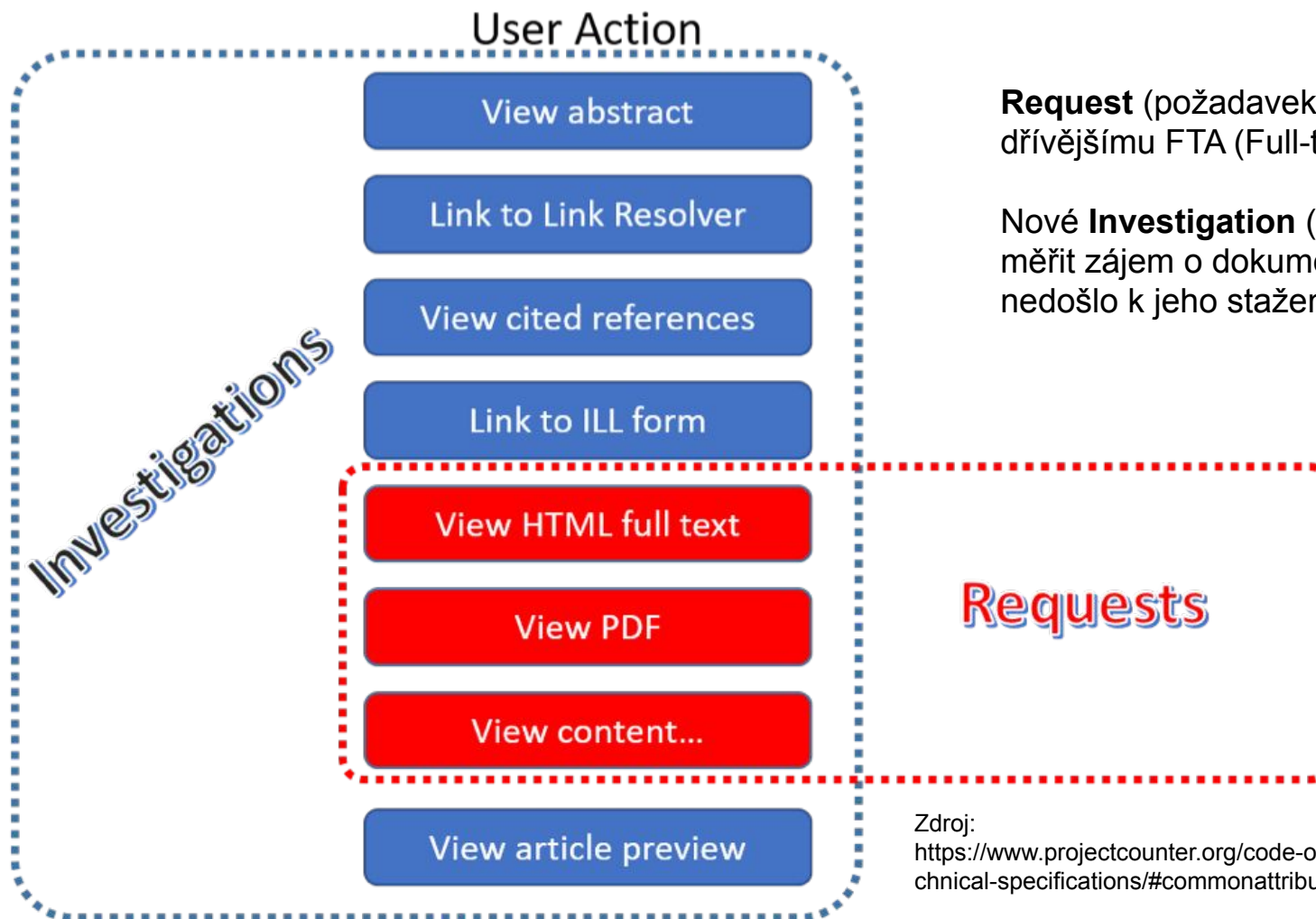
- **Searches_Regular (“normální” hledání)**
 - uživatelem položený intelektuální dotaz, typicky odeslání vyhledávacího formuláře na server
- **Searches_Automated (automatizované vyhledávání)**
 - hledání z discovery vrstvy nebo podobné technologie, kde je více databází prohledáváno simultánně jedním dotazem z uživ. rozhraní. Uživatel není zodpovědný za to, které databáze jsou prohledávány.
- **Searches_Federated (federované vyhledávání)**
 - umožňuje uživatelům hledat ve více databázích (i různých poskytovatelů) jedním dotazem z jednoho uživ. rozhraní. Uživatel není zodpovědný za to, které databáze jsou prohledávány.



*foceno před Covid-19, proto bez roušky

Female computer nerd working at her computer. Photography. Britannica ImageQuest, Encyclopædia Britannica, 25 May 2016.
quest.eb.com/search/132_1304093/1/132_1304093/cite. Accessed 2 Apr 2020.

Požadavky (Requests) vs. zkoumání (Investigations)



Request (požadavek) odpovídá víceméně dřívějšímu FTA (Full-text Article) download

Nové **Investigation** (zkoumání) umožňuje měřit zájem o dokument, i když nakonec nedošlo k jeho stažení/přečtení na platformě

Zdroj:
<https://www.projectcounter.org/code-of-practice-five-sections/3-0-technical-specifications/#commonattributes>

Requests

- **Celkové**
 - Total_Item_Requests
 - celková využívanost
- **Unikátní**
 - Unique_Item_Requests
 - užitečné zejména u časopisů - odstraňuje např. duplicitní započítání zobrazení HTML verze a pak stažení PDF verze
 - Unique_Title_Requests
 - využívanost titulu jako celku
 - užitečné pro knihy - srovnání napříč platformami (Total_Item_Requests se mohou mezi platformami hodně lišit, podle toho, v jakých kusech je čtenáři kniha servírována)

Investigations

- **Celkové**
 - Total_Item_Investigations
- **Unikátní**
 - Unique_Item_Investigations
 - Unique_Title_Investigations

Requests jsou podmnožinou Investigations!

Např. sečíst Requests a Investigations by bylo hrubé zkreslení!

Způsob přístupu (Access Method)

- **Regular**

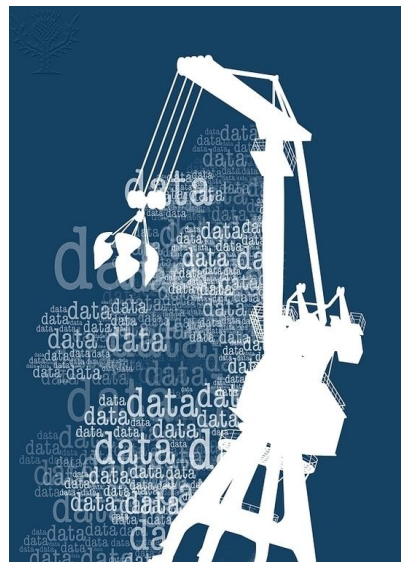
- odkazuje na aktivitu na platformě, kterou reprezentuje typického (živého) uživatele

- **TDM (nově)**

- přístup k obsahu a metadatům pro účely TDM (Text and data mining)
- typicky přístup přes dedikované API
- vyskytuje se pouze v Master reportu



*foceno před Covid-19, proto bez roušky



Grandmother and baby using a computer. Photograph. Britannica ImageQuest, Encyclopædia Britannica, 31 Aug 2017.

quest.eb.com/search/132_1512801/1/132_1512801/cite. Accessed 2 Apr 2020.

Data mining, artwork. Photograph. Britannica ImageQuest, Encyclopædia Britannica, 26 Mar 2018.

quest.eb.com/search/132_1527865/1/132_1527865/cite. Accessed 2 Apr 2020.

Odmítnuté přístupy (Access Denied)

Dává informaci o tom, proč byl přístup odmítnut

- **No_License**

- instituce vůbec nemá licenci k položce
- pozn.: když je uživatel místo na plný text přesměrován na abstrakt, tak se zároveň započte jako “Item_Investigation”

- **Limit_Exceeded**

- sice instituce licenci/e má, ale omezený počet, v některém okamžiku došlo k překročení počtu “křesel”

Jaký typ licence/omezení je na položce

- **Controlled**
 - “tradiční” licencovaný obsah
- **OA_Gold**
 - obsah přístupný v režimu Gold Open Access
- **OA_Delayed**
 - zatím neimplementován, plánován patrně v budoucí verzi 5.x
 - Gold OA obsah, ale se zpožděným uvolněním v tomto režimu
- **Other_Free_To_Read**
 - obsah sice volně dostupný (v ten okamžik), ale nesplňující definici Gold OA

- **Některé metriky jsou přímo podmnožinou jiných:**
 - Requests jsou podmnožinou Investigations, např.:
 - Total_Item_Requests je podmnožinou Total_Item_Investigations
 - jejich sečtení bude hrubým zkreslením
- **Jiné jsou odvozené z jiných:**
 - Unique_Title_Requests a Unique_Item_Requests z Total_Item_Requests
 - Unique_Item_Investigations z Total_Item_Investigations
 - opět - jejich sečtení nedává smysl!
- **Některé položky nepopisují chování “normálních” uživatelů:**
 - např. Access Method=TDM, nebo Searches_Automated
 - také nedává smysl je započítávat

Tedy pokud nechcete nesmyslně nafouknout čísla...

Implementace v systému Celus

- **Software plánován jako součást projektu CzechELib**

- Nikoli “in-house” vývoj, ale outsourcing přes veřejnou zakázku (výběrově řízení)
- Smlouva podepsána 7. 5. 2019, účinná od 9. 5. 2019
- Dodavatel: Big Dig Data s. r. o. (Praha)

- **Software dodán jako open source:**

- zdrojové kódy na GitHub
- licence - MIT License

- **Klíčové milníky**

- Analýza: 4 týdny, datum akceptace: 1. 7. 2019 (zpoždění kvůli nastavení prerekvizit na straně NTK)
- Implementace
 - do testu: 6 týdnů
 - do produkce: 4 týdny
 - obě podfáze akceptovány 31. 8. 2019
- Testovací fáze: 3 měsíce (na žádost CzechELib prodloužena z 30. 11. 2019 do 30. 1. 2020, aby bylo možno otestovat co nejvíce možných platform - tj. vč. ročních statistik z nových EIZ z vlny 2019+)

- **Cíl**

- systém, ve kterém budou agregovány statistiky pro všechny EIZ a instituce v rámci konsorcií
- možnost přidávat institucionální statistiky (EIZ mimo konsorcia)
- maximalizovat strojové zpracování

- **Požadavky (stručný souhrn z výzvy k VZ)**

- založeno na jazyku Python
- stahování SUSHI pro COUNTER-compliant poskytovatele
- možnost definovat vlastní report & nahrávání dat pro non-COUNTER EIZ

- **Krátké vývojové cykly (sprinty) - týdenní nebo čtrnáctidenní**

- Meetingy: CzechELib Licenční a administrativní jednotka (“product owner”) s vývojáři Big Dig Data
- Uploady nové verze

- **Výsledek**

- Jeden z mála (sic!) existujících fungujících systémů pro konsorciální statistiky na světě (UK/JISC - JUSP, FR/Couperin - EZPaarse, CZ/CzechELib - CELUS + komerční RedLink); projekt CC-PLUS stále ve vývoji

Napojení na další systémy

Z ERMS

Z české akademické federace
identit EduID.cz

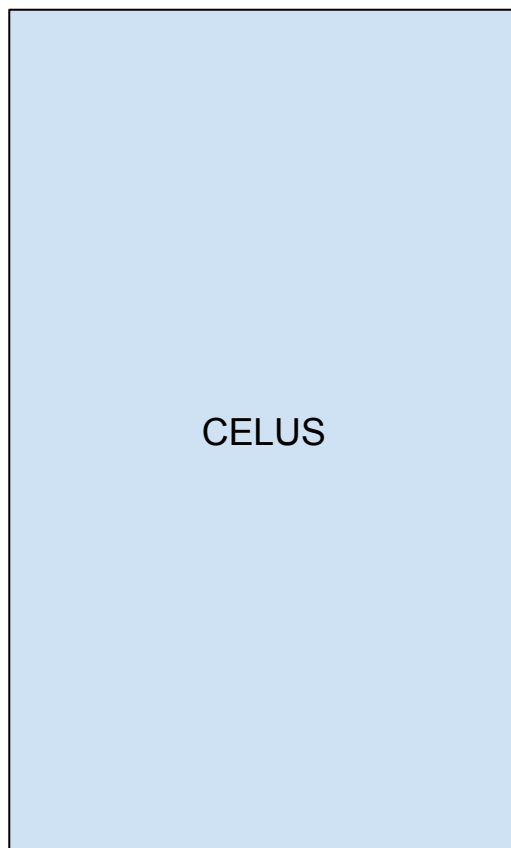
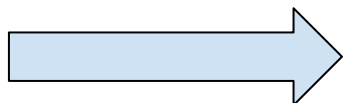
Autorizace - role (a oprávnění)



Institute



Platformy



Autentizace - Shibboleth



COUNTER a non-COUNTER reporty

● COP4 vs. COP5

- Zaměření na COP5 (rozhodnutí CzechELib)
- COP4 také zpracováván, ale žádný další vývoj nebo rozšíření nepožadovány (COP4 bude brzy “odstaven”, část vydavatelů již přes COP4 novější data neposílá)

● COP5

- stahují se pouze **plné reporty (full reports)**: (TR, DR, PR)
- umožňuje excelentní flexibilitu a možnosti filtrování
- vybrané standardní reporty (TR_J1, DR_D1, ...) jsou také nabízeny, ale jsou vytvořeny přímo v systému

● non-COUNTER

- možnost definovat vlastní strukturu reportů (se speciálními metrikami)
- nahrání dat s těmito metrikami ve formátu (TSV/CSV)
- CzechELib bude pro konsorcium nahrávat jednou ročně (typicky za předchozí kalendářní rok)

Dáváme přednost SUSHI



Sushi. Photography. Britannica ImageQuest, Encyclopædia Britannica, 25 May 2016.
quest.eb.com/search/156_2400334/1/156_2400334/cite. Accessed 29 Nov 2019.

COP = COUNTER Code of Practice

- **Konsorciální administrátor může snadno nastavit SUSHI linky / nahrát data pro jednotlivé členy**
- **Ukázalo se jako extrémně efektivní**
 - 90 % času stráveno nastavení prvního SUSHI linku pro daného poskytovatele, zbylé tucty institucí nastaveny v zanedbatelném čase
 - momentálně nastaveno přes 700 SUSHI linků
- **Zatím největší překážky**
 - neschopnost některých poskytovatelů nastavit práva pro správce konsorcia (možnost stahovat jednoduše statistiky jednotlivých institucí)
 - nebo nutnost nastavit další povolený IP rozsah pro stahování statistik
 - též technické problémy na straně poskytovatelů statistik (např. volný výklad a nesprávná implementace standardu) - CzechELib je “early adopter” u většiny platforem - bugreporty vydavatelům
- **Od jara 2020 bude CzechELib dodávat statistiky výhradně přes tento systém**
- **Nejčtenější/nejvíce odmítnuté tituly na úrovni konsorcia**

- Online **aktualizovaný přístup** ke statistikám (u COUNTER-compliant platform) **kdykoli**
- **Srovnání institucí** na platformě (anonymizované)
- **Nejčtenější/nejvíce odmítnuté přístupy** k titulům na úrovni instituce atd.
- U platform, kde jsou kromě e-časopisů i e-knihy, má instituce automaticky také **statistiky k e-knihám** (byť nejsou přes CzechELib předpláceny) a obráceně
 - Elsevier ScienceDirect, Wiley Online Library, SpringerLink, ACS, RSC, Thieme ...
- **Členové mají možnost nastavit SUSHI link a označit “mimo konsorcium”**
 - NTK a VŠCHT již nastavily maximum toho, co bylo možné
 - CzechELib může asistovat při nastavování (rozhodně doporučujeme - kontaktujte nás), obzvláště u nových platform je testování a nastavení prvního SUSHI linku netriviální
- Možnost **exportu** detailních dat pro další zpracování
 - Ale pozor: např. data pro ScienceDirect pro VŠCHT Praha:
 - CSV soubor: téměř 100 MB, přes 500 000 řádek (ale stahuje se jako ZIP soubor: cca 5 MB)

Role & oprávnění - detail

	Django admin (administrátor databáze)*	Konsorciální admin	Institucionální admin (správce inst. v ERMS)	Člen instituce (ale ne správce)
definování nastavení platforma-metrika / metrika-report**	globální nastavení	-	-	-
definování struktury dat pro (non-COUNTER) reporty**		-	-	-
SUSHI linky: zobrazit / vytvořit / smazat / zamknout / odemknout	*	pro jakoukoli instituci	jen pro svoji instituci (SUSHI: jen když je odemčeno)	-
nahrát non-COUNTER data	*			
zobrazit statistiky	*			jen pro svoji instituci

* může dělat téměř všechny low-level operace, včetně úplné katastrofy

** typicky úkony provedené jednou jako počáteční nastavení, bez nutnosti modifikovat po dalších několika let (pouze když poskytovatel změní strukturu dat)

Platformy s funkčním COP4 a/nebo COP5

- ACM Digital Library
- ACS
- AlexanderStreet
- Allen Press (single inst)
- AMA
- AMS
- ASCE Library (single inst)
- ASME (single inst)
- BioOne
- CUP
- DeGruyter
- Emerald Insight
 - čekáme na konsorc. admin
- GeoScienceWorld
- ICE Virtual Library (single inst)
- IEEE Xplore Digital Library
- IET Digital Library (single inst)
- IOPscience
- (J-STAGE) (single inst)
 - problémy s certifikátem na serveru vyd.
- JSTOR
- MathSciNet
- Nature.com
- NEJM
- NRC Research Press (single inst)
- OECD iLibrary
- OSA (single inst)
- OUP
- Ovid
- ProQuest
- ProQuest Ebook Central (individual SUSHIs)
- RSC
- Sage
- ScienceDirect
- Scitation
- Scopus
- SpringerLink
- Taylor & Francis Online
- Taylor and Francis ebooks (single inst)
 - nutno povolit IP range NTK
- Thieme (částečně)
- Web Of Science
- Wiley Online Library

(single inst) = SUSHI funguje, ale testováno jen s institucionálním účtem

Složitá diskuse s HighWire (AAAS, AACR) ohledně konsorc. admin

Pozn.: seznam je přibližný, situace se dynamicky vyvíjí, jak vydavatelé postupně implementují COUNTER 5

non-COUNTER reporty

- Např. následující platformy budou používat vlastní reporty:
 - Bisnode, Bookport, Brepolis, C.H.Beck, InCites, Knovel, Micromedex, Naxos, Reaxys, SciFinder, SciFinder-N, SciVal, UpToDate

non-COUNTER data

1 Data upload

SciFinder - custom

Interest defining metrics: Substance/Markush Reaction Reference CAPlus Reference Medline Commercial Source Regulatory

Standard dimensions: Metric Title

Dimensions specific to report:

Report type

SciFinder - custom

Data file to upload



Upload a file with tabular data in CSV fo

UPLOAD

Definice “Zájmu” a “Odmítnutého zájmu”

- **Účel a původ:**

- Mít (v ideálním případě) jednu vybranou metriku pro každou platformu pro kompaktní zobrazení a pro event. finanční kalkulace
- Základy definovány v “Metodice vyhodnocování statistik” (vytvořena ve spolupráci s pracovní skupinou AKVŠ)

- **Logika definice**

- Platforma - Report(y)
 - které reporty mají být brány v úvahu při výpočtu zájmu
 - např.: “ScienceDirect - TR” (Title Report v COP5) a “ScienceDirect - JR1” (Journal Report 1 v COP4)
- Report - Metrika
 - která metrika se bere v úvahu z daného reportu
 - např.: z reportu TR - Total_Item_Requests, z reportu JR1 - FT Article Requests
- rozhodnuto **neuvažovat** specifikace typu “Platforma - Report - Metrika” nebo dokonce “Institute - Platforma - Report - Metrika”

- **COP4 vs COP5**

- když jsou k dispozici obě verze COUNTER pro daný měsíc, vyšší verze má přednost (zabránění duplikování hodnot)

Definice “Zájmu” a “Odmítnutého zájmu”

- **Typy zájmu (a odmítnutého zájmu):**
 - Mají zabránit míchání “hrušek s jablky”
 - Tři typy:
 - full-text
 - search
 - other
 - Obecně
 - plnotextové platformy uvažují pouze typ využití “full-text” (žádné “investigations”, “searches”, ...)
 - citační/abstraktové databáze uvažují pouze typ “search”
 - V některých případech mají platformy nastaveny oba typy metrik (full-text i search), typicky agregátoři kombinující abstraktové i plnotextové databáze (EBSCOHost, ProQuest, Ovid), kde obě metriky mohou být zajímavé
- **Také námi definované reporty (non-COUNTER) mohou mít metriky, které se nezapočítávají do “zájmu”**

Shrnutí - očekávané “FAQ”

- **Kde systém najdu?**
 - <https://stats.czechelib.cz/>
- **Kdo se do něj dostane (autorizace)?**
 - Osoba, která je v ERMS přiřazena nějaké instituci
 - Má-li roli správce, může i editovat SUSHI
 - Nemá-li roli správce, může si jen zobrazovat statistiky
- **Co uvidím?**
 - Plná data jen pro svoji instituci
 - Ve srovnáních jsou ostatní instituce anonymizovány
- **Máme nějaké EIZ mimo CzechELib, můžeme stahovat statistiky přes Celus?**
 - Ano, pokud platforma umí SUSHI, můžete nastavit link i pro ni
 - Pokud neumí, můžete v definovaném formátu vkládat statistiky pomocí uploadu

<https://stats.czechelib.cz/>

- Ukázka administrace jedné instituce

event. (podle času)

- Ukázka konsorciální administrace

Děkujeme za pozornost

Děkujeme a nastavte svoje platformy, které nemáte přes CzechELib, rádi vám pomůžeme!

Těšíme se na kontaktní seminář, kde můžeme detailněji probrat celou problematiku.