Přílohy

Příloha č. 1: Kostra pro hloubkové rozhovory

Příloha č. 2: Hloubkové rozhovory

Příloha č. 1: Kostra hloubkových rozhovorů

1. V jakém médiu respondent pracuje?
2. Jakou má redakce daného média strukturu?
    a. Rozdělení rolí – působí v redakci jen novináři, nebo i techničtí pracovníci, programátoři/datoví analytici
        i. Pokud v redakci technici působí, jak moc je to důležité / pokud ne – měli by?
    b. Specializují se novináři v dané redakci na určité oblasti novinařiny?
    c. Navyšují se požadavky na jednotlivé novináře, aby uměli pracovat s technologiemi? Zvládají to? Je to žádoucí a podstatné?
3. Jak vypadá budoucnost datových leaků?
    a. Přibývá datových leaků? Jsou nějak specifické?
    b. Unikají větší objemy dat?
    c. Jaké to pro novináře a redakce prezentuje výzvy?
    d. V čem jsou datové úniky pozitivní a v čem negativní jev – vysvětlete.
4. Popis konkrétních leaků a redakčních rutin, novinářské praxe: Na kterých velkých data leaks se podíleli?
    a. Jak ta práce vypadala? Kdo a jak s daty pracoval? Bylo potřeba například využít externisty (analytiky, programátory)?
    b. Jaké má práce na velkých datových únicích výhody a nevýhody?
    c. Rozdíly mezi přístupem – jedná se spíše o datovou nebo investigativní novinařinu?

Příloha č. 2: Rozhovory

Jan Cibulka

Pro jaké médium pracuješ?

Pro online sekci Českého rozhlasu, iRozhlas, jako datový novinář.

Pracujete s datovými leaky?

Záleží na definici datového leaku. Pracujeme s leaklými informacemi, stejně jako všichni novináři. Ale v tomhle ohledu beru data jako strukturovaná, typicky tabulky, strojově čitelné datové struktury, a to s daty jak z Česka tak ze zahraničí. Ale nejsme součástí žádné velké novinářské organizace typu OCCRP nebo ICIJ, takže my s velkými úniky typu balíky dokumentů standartně nepracujeme. V posledních pěti letech jsem se setkal třeba s větším objemem uniklých emailů.

Je možné říct, že datová novinařina v Česku jde spíš směrem tabulek, dat, než právě směrem zpracovávání uniklých dokumentů?

Ano, určitě. Jednak je to pohodlnější, protože ve chvíli, kdy ta data mají nějakou strukturu je jejich zpracování přímočařejší. Další věc je, že Česko není náchylné k tomu, že by unikaly obrovské objemy dokumentů.

Takže jsi nikdy nepracoval na žádných „papers“?

Pracoval jsem na těch prvních, Panama Papers, ale jen v tom smyslu, že jsem pomáhal s nějakými technickými aspekty právě redakci investigace.cz, ale jako novinář jsem na nich nepracoval.

Myslíš, že je důležité, aby redakce měly datové analytiky, tech experty?

Určitě ano. Všechny velké vydavatelské domy mají své ajťáky, ale to nejsou lidé, jichž služby by mohli využívat novináři. Můžou vyvíjet redakční systémy, produkty pro klienty, starat se o počítače a telefony, ale nespolupracují s novináři, redakce je nemůže úkolovat. To vede k tomu, že ti lidé nejsou pružní – když se něco děje, nemůže redakce čekat na to, až vývojáři dokončí svůj projekt a budou se moci věnovat něčemu jinému. Zároveň není motivace tyhle světy propojovat. To si myslím, že je škoda – jednak proto, že ajťáci mohou svými znalostmi a zkušenostmi redakce obohatit, mohou jim něco naprogramovat, a tím nutně nemyslím nějakou vizualizaci na webu, ale třeba nějaký interní nástroj, který novinářům usnadní práci, zpracovat

data, roztřídit dokumenty… Další výhoda je, že ajťák může své technické znalosti nabídnout redaktorům, protože je to čím dál více potřeba. Celá veřejná správa se realizuje digitálně, a když tomu novináři nerozumí, nemohou efektivně popisovat problémy, které s sebou digitalizace přináší.

Je tedy lepší mít profesionální tech tým nebo by se novináři měli ty dovednosti učit?

Myslím, že oboje. Novináři by se tech věci rozhodně učit měli. Ale je samozřejmě i výhodné mít k ruce profesionály, kteří mají ty technologické znalosti hlubší. Ti lidé se mohou učit od sebe. A ani v malé redakci by to neměl dělat nebo umět jen jeden člověk.

Bude v budoucnu nutností, aby se novináři zároveň stali dataři a ajťáky?

Já si myslím, že už to nutnost je. Už jen proto, že v téhle oblasti leží spousta témat. A to, že jsou k tomu novináři slepí, bych přirovnal k tomu, když se vyhýbají ekonomickým tématům. Když třeba celá redakce nemá nikoho, kdo by dokázal popisovat „svět peněz“.

Takže by každá redakce měla mít odborníky specificky zaměřené na jejich oblast expertízy?

Rozhodně. Je hloupost si představovat, že je novinář schopen přeskakovat mezi tématy bez hlubší znalosti a nepodělat to. On to jen neví, že to podělal a pokud v té redakci těch lidí není víc, tak to nepozná nikdo a to je nebezpečné. Obecně se všechno lidské vědění prohlubuje, což nutně přináší potřebu hlubších specializací.

Takže se novináři budou muset naučit programovat?

Ano. Za pár let to bude nutností a ten, kdo to nebude umět, bude mít na trhu práce horší pozici.

Nestačí vědět, jak ty věci fungují a umět je předat tomu tech expertovi?

Je to možnost. Ale jestliže tomu ten novinář už rozumí, tak je jen krůček od toho do nějakého jednoduchého programování přeskočit a třeba ty základní úkoly už si dělat sám. Není to nutně tak, že novinář, který nebude umět programovat bude nepoužitelný, ale bude třeba jenom klopit ČTK, dělat nějaké zpravodajství. Ale nebude schopen hlubší analýzy a pokud ano, tak mu to bude dlouho trvat. To je neefektivní a vzhledem k tomu, jak jsou na tom média finančně špatně, je nutné hledět i na efektivitu.

Když ale bereme v potaz ty velké objemy dokumentů, jako Pandora Papers, stejně jsme ty dokumenty museli přečíst…

Ano, ale už jste je měli předpřipravené. Někdo z nich vytahal klíčová slova, z metadat vytahal

autory, to všechno pomůže novináři se zorientovat. Samozřejmě velká část práce bude vždycky sednout si na zadek a prostě si to přečíst, ale zase jsme u té efektivity. Navíc dokument leaky jsou čím dál větší, protože to technika dovoluje, a je nemyslitelné, že budeš terabyte dokumentů pročítat ručně.

Jsou datové leaky častější?

Rozhodně ano, vidíme to na různých Papers, jsou rozsáhlejší, týkají se větší části světa. Častější jsou i leaky kriminálního původu, které ale mají využití při novinářské práci – typicky přihlašovací údaje. Když se snažíš propojit nějaké firmy a zjistíš, že někdo používá ke všem účtům té firmy stejné heslo – třeba jméno manželky – pravděpodobně se bude jednat o stejného člověka. Tady si ale musí novináři dávat pozor, aby se pohybovali v hranách zákona.

Pak je tu otázka toho, nakolik je to etické?

Jistě, pokaždé je třeba uvážit nakolik je zveřejnění těch informací ve veřejném zájmu a kdy už ne.

Jsou ještě jiné důvody, proč bude laků přibývat než to, že je to technicky možné?

Ano, a dobrý příklad na to jsou spojené státy, kdy lidé viděli, že se těmi leaky dosáhlo něčeho dobrého. Typicky Snowden Manning, Assange, ti všichni vynesli informace a důkazy o státem prováděné kriminalitě. To znamená, že veřejný zájem na tom ta data zveřejnit byl extrémně silný. Když whistlebloweři vidí, že má jejich činnost nějaký reálný dopad, hluboký právní následky. A jistě, způsobilo to nějako škodu, ale zisky převažují.

Budou leaky větší?

Ano, opět kvůli technickým možnostem, prostě se dá těch dat víc sebrat. Taky se data budou uchovávat déle – a existuje maxima že čím déle informaci držím, tím je pravděpodobnější, že unikne. Takže když ji držím nekonečně dlouho, unikne 100%. Dokud tedy firmy nezačnou data „skartovat“, bude k únikům docházet. A tadyta datová nenasytnost se netýká jen korporátů, pro které je to byznys, ale začíná se týkat i států, typicky bezpečnostních složek. A bezpečnostní složky mají data často špatně zabezpečená, zaprvé kvůli nějaké aroganci a za druhé prostě proto, že na to nemají peníze.

Proč jsou leaky dobré?

Ne každý leak je dobrý. Když hacknou pojišťovnu a vytečou z ní informace o diagnózách pacientů, tak to dobré rozhodně není. I ten kriminální ekosystém ale může vypustit přínosná data, třeba z nějaké státní instituce, která dělá něco, co není správné a zatlouct to voličům. Nebo

je pozitivní únik dat z korporace, která má nějaký veřejná dopad, typicky z právní kanceláře, která poskytuje offshore služby. Musíme jednotlivé leaky posuzovat podle obsahu. Myslím, že konkrétně u leaků mají média důležitou roli gatekeepra. Například WikiLeaks, které byly veřejně přístupné, měly i po vydávání článků o datech v nich obsažených mnohem menší dosah a dopad než ty kauzy, se kterýma šli ti whistlebloweři rovnou do médií. To má dva důvody – zaprvé může whistleblower novinářům poradit, co je v datech nejdůležitější a za druhé, když mají novináři pocit, že na tom pracují exkluzivně, spíš tomu budou věnovat čas a energii.

Jaké prezentují leaky pro novináře výzvy, problémy?

Technické – umět to zpracovat. Bezpečnostní – někdo si pro ně může přijít, a tím nemyslím jen policii, někdo může na novináře zaútočit a to jak fyzicky, tak se je například snažit zdiskreditovat. Je třeba data zabezpečit, umět je bezpečně sdílet, zpracovávat. Pak samozřejmě bezpečnost zdroje. Pak jsou tam právní otázky, někdo se může chtít soudit. Etické otázky – ochrana soukromí osob, které sice v leaku jsou, ale není veřejný zájem o nich psát. Ochrana firem, daňové tajemství. Další věc je, že novináři nejsou od toho, aby pomáhali bezpečnostním složkám, ať to bezpečnostní složky myslí sebelépe. Takže i když chce finančák dát za uši politikovi, kterého jsme kritizovali, novináři tady nejsou od toho, aby veřejné správě pomáhali. A všechno je to náročné na prostředky.

Jak bys oddělil data leaky a dokumentové leaky?

I ty „papers“ a „files“ budou mít nějakou strukturovanou formu, minimálně třeba seznam dokumentů, kde se podle názvů dá vyčíst co asi jednotlivé dokumenty obsahují. Vy máte Aleph, který to umí zaindexovat, hledat klíčová slova, což jsou vlastně strukturovaná data, ačkoli ta struktura nebyla součástí původního leaku. Myslím, že pevně se ta hranice narýsovat nedá, ale jsou samozřejmě leaky, které vznikly jako balík dokumentů a které vznikly jako uniklá strukturovaná databáze. Příkladem datového úniku jsou počty nakažených během covidu, které stát nechtěl zveřejňovat, a tak museli novináři pracovat s leaky. Příklad dokument leaku jsou samozřejmě „papers“ a „files“

Stevan Dojčinović

Můžeme začít tím pro jaké médium pracuješ?

Pracuju pro jedno velké médium, pro nadnárodní novinářskou síť OCCRP, kde zastávám pozici regionálního šéfredaktora a zároveň šéfuju a jsem zakladatelem srbského Krik, kde pracuje 15 lidí. Dalo by se tedy říct, že pracuju pro jedno velké médium, které má dosah po celém světě a jedno malé. Pracuji pro ně najednou a mnohdy je ta práce propojená, například v případě Pandora Papers, Panama Papers, při těhle leacích, projektech, pracujeme společně. A já jsem ta spojka.

Jaké jsou role jednotlivých zaměstnanců KRIKu? Je někdo z nich specialista například na data?

Ne, všichni jsou novináři, nemáme rozdělené role, že by se někdo specializoval na jedno a jiný na druhé. Tyhle služby nám poskytuje OCCRP, které má celý tým rešeršistů – novinářů, kteří se soustředí jen na vyhledávání informací, vyhledávají pro ostatní novináře informace online, které si oni sami neumějí najít. Tohle jsou specialisti na analýzu dat. V Kriku je nás málo a každý děláme všechno. Nemáme role, specialisty. Jistě, někteří se lépe vyznají například v organizovaném zločinu, jiní v korupci, ale ve všeobecné rovině dokáže každý pokrýt jakékoli téma.

Vím, že jsi pracoval na několika velkých datových únicích – můžeš je vyjmenovat?

Panama Papers, Paradise Papers, FinCEN Files a Pandora Papers – ty byly pod taktovkou ICIJ. Pak jsme ale měli další v OCCRP – pamatuju si Lux Leaks (data v Lucemburska) a pak seznamy nemovitostí v Dubaji, na další si nevzpomenu. Celkem jsem pracoval minimálně na deseti leacích. Já se na to snažím nedívat optikou „další uniklé informace“, zajímá mě, co z nich vzejde za příběhy. Proto si je nepamatuju, ta zjištění pro mě nejsou leaky, pamatuju si příběhy, které jsme publikovali.

Co tedy děláte, když dostanete velký balík dat, jako třeba Pandora Papers? Jaký je postup?

Funguje to tak, že k samotnému leaku mají přístup naši editoři a ti je postahují. Konkrétně byl ten proces následující: já jsem měl přístup k datům, byl jsem součástí novinářů, kteří ten přístup dostali. Jen pro Srbsko bylo v Pandora Papers desítky tisíc dokumentů, a pak byly dokumenty pro jednotlivé společnosti, možná stovky tisíc. Rozhodl jsem se tedy dokumenty postupně začít stahovat a předával jsem je dvěma editorům a ti je prošli – ale neanalyzovali – soustředili se jen na jména lidí a názvy společností. Nešlo o to selektovat, která jména jsou důležitější. Byla to robotická práce. Pamatuju si, že jsem jeden den stáhl tisíc dokumentů a jednomu editorovi jsem

dal 500 a druhému taky. A oni měli na starosti vytvořit excelovou tabulku s názvy firem a jmény lidí. V podstatě dělali seznam. Seznam potom putoval k jednotlivým reportérům, těch je v Krik sedm, a ti začali o lidech a společnostech vyhledávat informace a ty nám zase předávali zpátky – tedy mně a mým editorům. Načež jsme se na základě informací, které reportéři našli, rozhodovali, kdo nás zajímá a ty entity v seznamu zvýrazňovali. A pak jsem stahoval další dokumenty a celý proces se opakoval. Zopakovali jsme to asi pětkrát.

Takže jsi do databáze zadal „Srbsko“ a stahoval všechno, co vyhodila.

Ano, to trvalo několik měsíců. Pak jsme měli zkompletovaný seznam doplněný o informace a mohli jsme se rozhodovat, kdo a jaké společnosti nás zajímají. Když jsme se rozhodli, která jména prozkoumáme, opět jsem si sednul k databázi a začal vyhledávat podle názvu firmy nebo jména osoby. To jsou další desítky až stovky dokumentů, které nemusí obsahovat klíčové slovo „Srbsko“, takže bychom je na poprvé nenašli. Tak jsem postahoval složky informací o jednotlivých entitách a dal je pak tomu novináři, který to zajímavé jméno nebo název našel. A ti je museli přečíst a analyzovat. Zároveň pro mě redaktoři sestavili seznam osob, například politiků nebo vlivných podnikatelů, kteří by pod klíčovým slovem „Srbsko“ vůbec nevyběhli, protože mají například francouzský pas a já prohledával tato jména. Reportéři mezitím analyzovali dokumenty a když to udělali, prezentovali nám, editorům, co našli. V té době jsem měl prohledaná potenciálně zajímavá jména a všechny tyhle informace jsme probírali a já spolu s editory jsme rozhodovali, které z nich by mohly být zajímavé pro čtenáře. Zúžili jsme to na deset případů, ze kterých jsme si mysleli, že by mohly být zajímavé texty. A tenhle výběr jsme zase předali reportérům, aby na těch kauzách už začali doopravdy pracovat. Tak to udělali a publikovali a je to.

V tomhle procesu tady nebyl nikdo, kdy by zastával funkci datového novináře nebo analytika?

Ne, všichni vyhledávali informace

Nebylo by to užitečné mít někoho takového?

Já si ale nemyslím, že by to nějak pomohlo, že by ten proces byl rychlejší. Tohle je práce, kterou musí dělat lidi, ne programy. Programy nezachytí podstatné informace. Ano, datoví analytici jsou dobří na analýzy, ale to jsou obecné informace. Když chcete skutečně dobře vystavit a pochopit ten příběh, musíte si tu složku přečíst a rozumět obsahu. Datová novinařina je dobrá na propojování informací. Při tomto typu práce jsem ale chtěl, aby si každý dokument přečetl

člověk. Nemusí to být důležitý případ, důležité jméno, nemusí z toho být žádný článek, ale je nutné, aby si lidi všechno přečetli. Měli jsme na to dostatek času, a proto jsem chtěl, aby byl každý dokument přečtený. Tohle může dělat jen novinář. Datoví novináři jsou dobří, možná by se nám jeden hodil, ale ne na tuhle práci. Navíc jsme ty obecné informace, tu datovou analýzu – kolik je v leaku společností, kolik peněz, lidí, zemí – měli od ICIJ, kteří celý projekt zastřešovali, ti mají k dispozici experty.

Stalo se vám někdy, že jste před tím, než se dala leaknutá data procházet ručně, potřebovali pomoc nějakého analytika / programátora, který by vám s nimi pomohl?

Ano, ale nebyli to externisti. Vlastně se to u velkých datových leaků děje běžně, jestliže je má k dispozici OCCRP, dostane se k nim nejprve jejich tým profesionálů, kteří nemají s novinařinou nic společného, jsou to analytici, v podstatě technická podpora. Podobně to funguje i v ICIJ. A právě tenhle tým nám mnohokrát pomohl porovnávat data. Například čas od času posíláme tomuhle tech týmu OCCRP jména politiků, třeba po volbách, když se mění pozice, a oni porovnávají náš seznam s daty z datových leaků a zjišťují, jestli v nich nefigurují. Navíc, když jste součástí OCCRP, máte k dispozici nástroje, které tohle vyhledávání a propojování usnadňují. Nakonec ale všechno skončí u lidí, kteří musí všechno přečíst. Pokud ale nemáme k dispozici dokumenty, ale jen nějaká syrová data, třeba seznamy, pak datové analytiky využijeme.

Myslíš si, že jsou čím dál větší požadavky na technické dovednosti novinářů?

Myslím, že není tak podstatné, aby se technické dovednosti učili přímo novináři, myslím, že je lepší prostě najmout nějakého odborníka, analytika, který pomůže například zorganizovat data. Obecně si myslím, že novináři nejsou dobří v organizování čehokoli, že jsou naopak velmi dezorganizovaní lidé. Vždycky tedy budu preferovat aby tu práci odvedl někdo, kdo se v tom skutečně vyzná, profesionál. To se myslím projevuje i v OCCRP, například lidé, kteří vytvářejí a spravují Aleph nejsou novináři, to jsou technici, ajťáci, analytici. Připadá mi, že každý druhý rok přijde na scénu nová dovednost, kterou by si novináři měli osvojit. A připadá mi, že už je toho moc. Takže to, že je někdo novinář a dělá práci novináře – což je analýza dokumentů, rozhovory, rešerše, psaní – není dost, měl by taky umět dobře fotit, natáčet, dělat podcasty, umět se prezentovat na sociálních sítích… Požadavky na novináře neustále rostou a myslím, že už je toho příliš. Proto si myslím, že je důležité mít k dispozici experty, kteří tu práci skutečně umí dělat, než to všechno házet na novináře. V redakci Kriku máme například zaměstnance,

kteří dělají jen fundraising, starají se o finance, účetnictví, a s novinařinou nemají v podstatě nic společného.

Myslíš si, že jsou leaky častější?

Ano, protože tak to funguje. Když lidé vidí, že někdo vypustil nějaká data, podněcuje je to k tomu, aby to udělali také. Je to skvělé. Čím víc se to bude dít, tím více jich bude. Je to jako lavina. Letos (v roce 2021) jsme za jeden rok měli hned dva velké leaky – Lux Leaks a Pandora Papers – jenom co jsme publikovali v naší organizaci, ale vím, že byly i další. Podporuje a přesvědčuje to whistleblowery.

Myslíš si tedy, že se to bude stupňovat, že budou leaky stále častější?

Ano.

Jsou uniklé informace čím dál tím větší, objemnější?

Ano, většinou to tak je, jednotlivé leaky obsahují více dokumentů. Rozhodně tomu tak je u Pandora Papers, ale ne vždycky to funguje tak, že ten další, mladší, je větší. Například Paradise Papers byly menší než Panama Papers. Nejsem si ale jistý, jestli to jde takhle analyzovat. Vždycky záleží na člověku, který data vypustí, na tom, k jakým objemům má přístup. Nemyslím si tedy, že to je úmysl, že se zvětšují, ale ve všeobecné rovině tomu tak je. A podle ICIJ je Pandora Papers zatím největší.

Naučil jste se v KRIK pracovat lépe s daty s každým dalším leakem?

Ano, pamatuju si, že Panama a Paradise Papers nebyly moc dobré. Začali jsme se zlepšovat až u FinCEN Files. Postupně jsme se stali mnohem více organizovaní ale i efektivnější v tom, jak jsme s daty pracovali. Myslím, že Pandoa Papers jsme zvládli opravdu dobře, jsem na to pyšný. Pamatuju si, že při Panama Papers to byl prostě chaos, všechno jsme dělali na poslední chvíli, byl to bolestivý proces. Pandora Papers jsme měli dobře naplánované, všechno do sebe zapadlo.

Jaké jsou klady a zápory práce na velkých datových leacích?

Krik není jen médium, které publikuje kauzy, ale děláme zprávy, publikujeme každý den. Kvůli Pandora Papers jsme s tímhle museli přestat. Asi měsíc, od začátku srpna (Pandora Papers začaly vycházet 5. září) jsme nevydali ani jeden kousek zpravodajství, což pro naše čtenáře muselo být divné. Prostě to nešlo. Blížil se konec projektu a všichni jsme chápali, že se musíme soustředit na něj, sto procent času. Prostě jsme denní publikování zamrazili. Myslím, že to byla chyba. Možná, že příště se připravíme ještě lépe.

Ublížilo vám to jako médiu?

Možná ano, ale srpen je všeobecně velmi špatný, co se týče čtenost. Horší byl začátek září, kdy jsme pořád nic nepublikovali a lidé se vracejí z dovolené a chtějí zase číst zprávy.

Jaké jsou klady a zápory práce s ostatními médii?

Já vidím jen pozitiva. Jsem vždycky pro mezinárodní spolupráci a nevidím na ní nic záporného. Vždycky jste silnější, když je vás víc, máte více informací, práce je efektivnější, má větší dosah. Jediný problém může nastat při dohadování deadlinů a dat vydání. Nadnárodní spolupráce je budoucností novinařiny, mělo by se to dělat častěji.

Pavla Holcová

Pro jaké médium pracuješ?

Jsem šéfredaktorkou investigace.cz a zároveň pracuji pro OCCRP.

Jakou má investigace.cz redakční strukturu? Jsou v ní specializovaní novináři?

V redakci máme specialisty, respektive se jednotlivá redaktoři profilují podle toho, co jim jde lépe. Máme lidi, kteří rozumí datům, máme lidi, kteří dokážou dobře pracovat s otevřenými zdroji, dělat rešerše nebo dokážou číst různé dokumenty – například právě dokumenty, které jsou součástí datových úniků –, máme lidi, kteří raději telefonují. Myslím si, že je to mix lidí, kteří se vhodně doplňují.

Zmiňuješ lidi – novináře, zaměstnance investigace.cz –, kteří rozumí datům – co přesně to znamená?

Tím myslím lidi, kteří umí například scrapeovat databáze, třeba i nějak základně programovat, jde o lidi, kteří mají nějaký vyšší level technických dovedností. Toho máme jednoho, datového analytika, který do investigace.cz přišel asi před rokem.

Myslíš si, že je nutné mít v redakci takového člověka? Technicky zdatného, datového analytika?

Nemyslím, že je to nutné, ale rozhodně je to užitečné a výhodné. Já datovou a investigativní žurnalistiku odlišuji, odděluji. Kdybychom dělali čistě investigativní žurnalistiku, tak bychom toho „tech člověka“ nebo dataře zas tak nepotřebovali.

Takže míra důležitosti mít „dataře“ v týmu se odvíjí od toho, kam směřuje dané médium? Jestli je to investigativní nebo datová novinařina?

Ne. Obecně si myslím, že každá větší redakce by „dataře“ mít měla. Přičemž zdůrazňuji výraz větší – my větší ještě nejsme. Myslím si, že je krátkozraké takového člověka z dlouhodobého hlediska v týmu nemít. Nemyslím si ale, že se celá žurnalistika bude ubírat směrem k datařině, rozhodně se ale investigativa s datařinou budou více prolínat, ale nemyslím si, že by někdy investigativní žurnalistika byla závislá na datařině, nebo na psaní kódu, nebo na scrapingu databází. Je to jedna z podmnožin novinařiny.

Jestliže tedy mají mít redakce datové analytiky, měli by to být novináři?

Pokud dokáží novináři správně expertům vysvětlit, co chtějí, aby pro ně udělali – což je

většinou dost složité – tak je možné mít tyto dvě profese, a tedy osoby, oddělené. Je ale samozřejmě výhodnější mít člověka, který přemýšlí jako novinář, což znamená že dokáže vymyslet ten příběh, a zároveň umí vymyslet metodologii, jak toho dosáhnout, jak se k těm datům a číslům dostat.

Myslíš, že se stane pro novináře nutností tyhle technické dovednosti mít? Že se je prostě budou muset naučit?

Ne. Nemyslím si, že to bude nutnost, ale myslím si, že budou specialisti, tedy někdo, kdo je novinář, umí psát a vidět za těmi daty příběh a zároveň vědět, kde ta data získat, a lidé, kteří tohle budou umět a velice žádaní.

Shrnula bych to takhle: když budou novináři rozumět tomu, jak práce s daty a s jejich zdroji funguje, a budou to umět předat expertům, není nutné, aby ty expertní schopnosti sami měli. Nejde to dělat tak, že novinář řekne: tady mám leak a je to 27 gigabytů. Musí umět říct: tady je leak, je to 27 gigabytů, jsou v něm tyhle a tyhle formáty, potřebuju ten objem dat pročistit, smazat soubory, které nejsou potřeba, vyextrahovat entity (jména lidí, názvy firem), je potřeba roztřídit dokumenty podle data vzniku na základě metadat… Zkrátka tomu technickému člověku vysvětlit, co má s tím leakem udělat. Protože ten expert většinou nedokáže odhadnout co novinář potřebuje.

Dokud investigace.cz datového analytika neměla, kdo s tímto druhem praxe pomáhal?

OCCRP. Tak to ostatně funguje doteď, u těch opravdu velkých datových leaků, které by jeden člověk zkrátka nezvládl zpracovat. Ale to zase trvá dlouho. OCCRP totiž tenhle servis poskytuje všem jeho členským centrům, takže se nemůže věnovat hned každému, kdo s nimi přijde s velkými objemy dat. Proto je dobé mít někoho, kdo ty menší úkony zvládne třeba během dvou dnů a na svém počítači. Jednotlivé leaky jsou často menší a je například jen potřeba vyčistit data tak aby byla čitelná, sjednotit formát…

Na kterých velkých datových leacích jsi pracovala?

Panama Papers, Pandora Papers, Kočnerova knižnica, Laundromaty, pašování a prodej zbraní do Sýrie.

Jak práce na velkých datových leacích vypadá?

Pokaždé jinak, záleží na projektu. Mezi Panama Papers a Pandora Papers urazili ICIJ obrovský kus cesty, obrovský vývoj. Úplně první databáze, která vznikla na prohlížení Panama Papers byla uživatelsky dost nepříjemná. Například se v ní nedalo filtrovat podle typu dokumentů, což

je důležité, protože je nutné rozlišit například podepsané a naskenované smlouvy, které mají většinou formát pdf. a třeba nástřely smluv, které mohou být ve Wordu. Zároveň jsou součástí dat často obrovské tabulky, které sice mohou být užitečné, ale není reálně možné je procházet celé. Další problém spočíval v tom, že některé formáty databáze vůbec neuměla načíst, třeba neuměla načíst emaily, nebo neuměla zobrazovat náhled dokumentů. Zkrátka byla ta první databáze dost neohrabaná, ta poslední, k Pandora Papers je mnohem jednodušší, příjemnější na používání. Zároveň, stejně jako u Pandora Papers byla novinářům k dispozici komunikační platforma. To je systém ICIJ, který principiálně funguje vždycky stejně, jen se technologicky vylepšuje. OCCRP má trochu jiný systém a na jednotlivých projektech taky pracuje méně lidí. Pokud projekt zahrnuje nějakou databázi dat, tak ta je pak prohledávatelná v Alephu, přičemž tam jsou nejen například obchodní rejstříky, ale i třeba dokumenty. Ke komunikaci a sdílení zjištění slouží u OCCRP Wiki a pak většinou existuje ještě nějaká skupina na Signalu, skrze kterou se novináři například domlouvají na pravidelných videohovorech.

Můžeš porovnat jak vypadala práce na Panama Papers a na Pandora Papers?

Obecně by se dalo říct, že Pandora byla mnohem jednodušší. Už jen z toho důvodu, že jsme jako novináři věděli, jak s těmi daty a informacemi pracovat. Nebylo to tak „tady máte dokumenty a hrajte si“, ale věděli jsme co hledat, znali vzorce chování, věděli, které smlouvy a dokumenty co znamenají, jak například offshorová schémata fungují. Tohle jsme u Panama Papers netušili. Také bylo lepší, že šlo o více leaků, tedy data z více společností, tudíž dávalo logický smysl, že se dokumenty nahrávaly do databáze postupně. Panama Papers se také nahrávaly postupně přestože to byl jeden leak, což znamenalo, že se člověk musel do databáze po čase vracet a začít s novým vyhledáváním i u informací, které už měl.

Dobře, a jak pak vypadala práce například na Pandora Papers v redakci investigace.cz?

Dostala jsem přístup do databáze s dokumenty a do komunikační platformy. Pak jsem požádala o přístup pro dva další kolegy. Ze začátku každého takového projektu je člověk nadšený a začne na první dobrou vyhledávat jména, která ho z nějakého důvodu zajímají, kauzy, na kterých už pracoval. Nutno podotknout, že na začátku jsme v databázi vyhledávaly jen my dvě, ne celá redakce. Je to vlastně jak když člověk dostane zabalený dáreček a začne zkoušet, co tam bude, co to dělá. A velmi rychle si uvědomí, že je těch informací strašně moc a je třeba k jejich procházení přistoupit nějak systematicky. V tomhle ICIJ také pomohlo, protože hned ze začátku pro nějakou základní orientaci vytvořili shrnující tabulky – například tabulky (beneficial owners) konečných vlastníků firem, ve kterých jsme mohli hledat podle trvalého pobytu v České republice. Další novinkou, kterou tahle databáze umožňovala, bylo dostopovat odkud

jednotlivé dokumenty pochází – tedy k nějaké složce – a zjistit, jaké další dokumenty ta složka obsahuje. Tohle u Panamy nešlo.

Takže první věc je hledat zajímavá jména, ale pak je třeba vytvořit systém…

Ano. Tady je třeba zdůraznit, že my jsme od každé té právní kanceláře – tudíž z každého leaku – měli seznam konečných vlastníků s českým pobytem. Jenže to zaprvé nejsou všichni Češi v Pandora Papers, už jenom proto, že ne každý Čech uvedl v dokumentech, že v Česku žije nebo to zkrátka neuvedl vůbec. A například je v tom seznamu hodně Rusů, kteří to naopak uvádějí. Reálně tohle byla tak třetina všech Čechů. Museli jsme tedy stejně dokumenty pomocí různých filtrů projít ručně. Tak vznikla excelová tabulka. Pak jsme mohli začít zběžně vyhledávat, co se nám tam vyskytuje za lidi – na každého byly tak 4 minuty se rozhodnout, jestli je bereme nebo ne, a začít přemýšlet nad tím, z čeho by mohly vzejít zajímavé příběhy a začali jsme načítat dokumenty. Tohle ale pořád dělali v podstatě jen dva lidé. Velmi brzy jsme zjistili, že je v Pandoře Andrej Babiš a že to je tak důležitá kauza, že se na ni musíme soustředit a že musíme obě načíst všechny dokumenty. U těch dalších kauz už jsem nečetla všechny dokumenty, tos dělala ty. A vydáváme je vlastně doteď.

Bylo k práci s uniklými daty potřeba využít externisty? Analytiky, techniky, IT lidi?

V případě Panama a Pandora Papers ne, protože se o technickou stránku staralo ICIJ. V případě Laundromatů taky ne, tam se o přípravu dokumentů staralo OCCRP. U Kočnerovy knižnice ano, částečně sice data zpracovávalo OCCRP, ale například na uživatelsky přívětivé zprostředkování zpráv přes různé aplikace, jsme měli externího technika, programátora.

Jaké má práce na datových únicích výhody a nevýhody?

Výhody jsou že to jsou často data, ke kterým se nemáme jak jinak dostat. My jako novináři sice můžeme mít představu o tom, jak něco funguje, ale nemáme pro to důkazy, nevíme, jaký to má rozsah. My často díky dokumentům, které získáme například z otevřených zdrojů můžeme vidět, že se něco děje, ale nevíme, co přesně. Tyhle úniky dokumentů a leaky obecně jsou něco, co nám umožňuje pochopit dění za oponou, to, co nemá být vidět, co je velmi uzavřený, tajnůstkářský svět. Je to pohled do světa, který je pro normální lidi naprosto nedosažitelný.

Nevýhody: je to úmorná, mravenčí práce a je nesmírně náročné koordinovat větší týmy novinářů. To je taky věc, která se u takhle velkých projektů musela postupně vyvinout – začínalo to tak, že na jeden projekt byl jeden koordinátor. To se ale nedá. Na Panama Papers pracovalo 400 lidí, na Pandora asi 600. Ale ono stačí, aby těch lidí bylo 15 – jsou z jiných zemí, z jiných médií, mají spolu spory, potřebují, aby to koordinátor řešil, rozhodoval…

Omezila práce na velkých datových leacích fungování redakce jako takové?

Částečně. Ale nebylo to tak, že bychom si mohli dovolit půl roku pracovat jen na Pandoře. Nemohli jsme přestat vydávat každý den jako je ve zvyku, už jenom proto, že by to bylo divné, že by bylo jasné, že se něco chystá. Vlastně jsme ani neměli novináře, kteří by se věnovali exkluzivně jen Pandoře, museli jsme to skloubit s normálním fungováním. Ke konci pak ano, tys dělala jen Pandoru. A poslední týdny před vydáváním byly plnou parou před a všichni.

Bude datových leaků přibývat?

Ano, a už se to děje. Protože lidé, kteří jsou frustrovaní z toho, že se něco děje, vidí do zákulisí moci, obcházení pravidel, vidí, že to co dělají novináři má reálný dopad a je to větší, než co by v těch věcech dokázal udělat stát.

Takže je to přímá úměra?

Ano, čím víc tím víc.

Budou leaky větší?

Pravděpodobně ano, protože je to technicky možné, vyvíjí se větší úložiště dat, data jsou v lepší kvalitě, skladují se ve větším objemu. Není to tak dlouho, co se považovalo za vrchol vyspělé technologie, když měl disk na počítači 3 giga a vešly se na ně dva filmy. Teď je naprosto běžné, že film, který si za chvilku stáhnete má 300 giga. Tady je ale potřeba rozlišit mezi velikostí a kvalitou informací. Větší neznamená nutně informačně nasycený.

Jaké prezentují datové leaky pro novináře výzvy, problémy?

Bezpečnost – jak zdroje tak novináře. Pochopení širšího kontextu – leaknuté informace samy o sobě nemusí nic znamenat; pochopení toho, že novinářská práce nespočívá v tom, že nám někdo něco pošeptá do ucha a my to obratem vysypeme na papír, že vidíme jen díl mozaiky a potřebujeme pochopit, co to znamená v širším celku. Samotný leak je třeba 10-15% kauzy, my musíme zjistit, co se dělo v dalších firmách, v životě toho člověka, vstup do politiky, organizovaný zločin… ten kontext je mnohem širší než jen popsat co se píše v dokumentech.

V čem je pozitivní, že bude leaků víc a budou větší?

Budeme lépe chápat, jak svět funguje.

Friedrich Lindenberg

Jak vznikl OCCRP datový tým? Kdo v něm působí?

OCCRP sis s tou myšlenkou nějakou dobu pohrávalo, dokonce najali člověka, který jim pomáhal scrapeovat webové stránky. Pak najali mě, a ještě jednoho člověka a vznikající tech tým jsme si rozdělili na dva sektory – jeden sektor jsou ajťáci, kteří spravují webové stránky a kteří mají na starosti bezpečnost – a druhá část, ta moje, se snažila vytvářet nástroje, které by pomáhaly investigativním novinářům, které umí například procházet data nebo jsme data aktivně vyhledávali, třeba ze zdrojů, které nutně nemusí každý znát. Součástí toho byl právě Aleph, který jsme vyvinuli v roce 2016, protože jsme na to dostali grant, což je software, v podstatě databáze, kam nahráváme všechna data, co máme k dispozici. Zároveň jsme chtěli usnadnit komunikaci a spolupráci mezi novináři, tak jsme vyvinuli Wiki, kde mohou společně pracovat na jednotlivých projektech a sdílet své poznatky.

Takže na jedné straně jste data sbírali a na druhé pomáhali členům, kteří přišli s daty, se kterými si nevědí rady?

Přesně tak. Bylo to třeba, protože OCCRP se velmi rychle rozvíjelo. Na začátku v něm bylo dohromady asi 8 lidí. Takže nebyl problém spolupracovat na osobní rovině, potkávat se a sdílet informace. Zároveň mohli novináři snadno pracovat s námi, někdo přišel, dal mi data a řekl specificky, co s nimi potřebuje dělat, udělal jsem to. Jenže teď má OCCRP na plný úvazek snad dvě stě lidí, a to nepočítám všechna členská centra, která mají svoje zaměstnance, což dohromady dává klidně tisíc lidí. A tech tým je stále tvořený asi osmi lidmi. Nejde tedy s každým pracovat individuálně a do hloubky na jeho datech, bylo třeba tohle pochytit softwarem, který může používat každý. Proto jsme se snažili vytvořit nástroje, které zpracovávání dat zautomatizují.

U každého leaku ale stejně ta data musí vidět člověk, nebo ne?

Ne nutně. Od začátku války na Ukrajině je těch leaků tolik, že to ani nejde. Software je užitečný, například na nějaké základní úrovni chápe, jaké existují druhy dokumentů, a na základě toho, že identifikuje video, pdf, power point, tak přistupuje k jejich zpracování, aby v nich bylo například jednoduché hledat podle klíčových slov. To je první krok, zpracovat data tak, aby byla prohledávatelná, a to je automatizovaný proces. Pak je samozřejmě otázka, jak v těch datech hledat zajímavé informace, z čeho se dá napsat příběh, vytvořit kontext.

Nakonec ale novináři musí ty dokumenty přečíst?

Ano, ale ještě před tím můžeme pomoci předvybrat ty, které stojí za to číst. Například když na začátku roku unikla data z ruského mediálního domu – od těch co dělají Russia Today a televizi Rossia 1 – bylo to 800 gigabitů emailů. To by nikdo nepřečetl. Takže nastává otázka, jak ta data zpracovat tak, aby novináři věděli, které číst první, kde jsou zajímavé informace. Když třeba řeknou – mám na to hodinu, co si mám přečíst, mám na to den, co si mám přečíst, a tak dále… A v tom si myslím, že jsme ještě dost špatní.

Jak by tohle vůbec fungovalo? Podle klíčových frází?

Mě se líbí teorie, že by se dala selektovat jména lidí, nebo jejich označení – třeba když bude ruská televize zmiňovat v emailu dceru Putina, rozhodně je to email, co stojí za přečtení.

Na kolika velkých datových leacích jsi pracoval? A na kterých?

Na to nemám odpověď. Něco mezi deseti a padesáti. Podle toho, jak definujeme velký data leak.

Když budeme mluvit o velkých „papers“ a „files“?

To je další věc, data z Pandora a Panama Papers OCCRP nikdy nedostalo, ICIJ jim je prostě nedalo. Je to nějaký konkurenční boj. ICIJ chtělo ta data mít pod kontrolou, takže v Alephu nejsou, protože jsme je prostě nedostali. Ale pracoval jsem třeba na Laundromatech, to sice nebyly tak velké objemy dat, ale výsledné kauzy byly obrovské. Taky si pamatuju zajímavý leak, kdy někdo hacknul banku a pak ty informace prostě vypustil na internet. Takže je mohl zpracovat kdokoli, ale nikdo na to neměl technologie, takže OCCRP založilo celou koalici na základě toho, že jsme měli prostředky ta data zaindexovat, aby se v nich dalo hledat. Vtipné je, že z toho nakonec nic moc nebylo, že ta banka nic zas tak hrozného nedělala.

Pracoval jsi i na Kočnerovo knižnici – jak to vypadalo?

To bylo zajímavé hlavně z hlediska velikosti materiálu. Nejprve na Slovensko zavolali mého kolegu, když se snažili získat přístup k bezpečnostním kamerám, což se nepovedlo. A pak nám prostě leaknuli celý policejní spis. Normálně jsou za velké datové leaky považovány už stovky gigabitů. Jen pro představu – gigabite je tisíc megabitů a Bible, což je opravdu velká kniha s malými písmenky, má 3 megabity. U Kočnerovy knižnice to bylo 70 terabitů dat. Jistě, byla tam videa, to to hodně nafoukne. První problém byl, jak ta data zkopírovat, protože kdybychom to dělali klasicky na disky, trvalo by to zhruba dva měsíce. Nakonec jsme tedy vytvořili deset minipočítačů, které data kopírovaly najednou a stihli jsme to za tři dny. Pak je potřeba data nahrát na internet, což by běžnou cestou také trvalo dlouho. Tenkrát jsme skrze poskytovatele

připojení z Prahy data nahráli připojením přímo do páteře internetu. A pak jsme ta data museli projít a identifikovat co je důležité a co ne. Něco málo jsme dokonce nahráli do Alephu, ale zbytek je přístupný jen skrze zabezpečené počítače v Bratislavě. Skvělé bylo, že nám tenkrát pomohl externí programátor, který zpracoval všechny chatovací zprávy ze všech aplikací do jedné, aby šly snadno procházet. To totiž Aleph neumí. Je skvělé kolik různých přístupů k tomu leaků můžete mít, podle toho, co zrovna chcete analyzovat.

Myslíš si, že je důležité, aby měly redakce své vlastní datové analytiky, tech tým? Ve smyslu, jestli je důležité, aby tyhle schopnosti měli novináři, nebo by měli raději využívat experty?

Myslím, že trocha technických schopností je pro novináře dobrá věc, děláte velmi specifickou a specializovanou disciplínu. Každý novinář se časem naučí, jak číst účetní závěrky, jak analyzovat smlouvy… Líbí se mi, že novináři zakládají koalice, že spolupracují – konkrétně na těch velkých data leacích, protože informace v nich obsažené jsou relevantní pro mnoho zemí, takže dává smysl, aby na nich pracovali ti novináři, jejichž zemí se to týká. A je skvělé, že OCCRP a ICIJ agregují tyhle technické dovednosti a nástroje a nabízí je, protože nejen že to je náročné, ale je to neuvěřitelně drahé. Jen udržování Alephu v provozu stálo do dnešního dne 1,5 milionu dolarů.

Do jaké míry se u datových leaků dá mluvit o datové novinařině a co už je prostě investigativní novinařina?

Byl jsem teď na jedné data journalism konferenci, kde různá média prezentovala své kauzy. A ty nejlepší vznikají právě díky tomu, že se tyhle dva obory propojují, že spolu technici a investigativci spolupracují. Myslím, že OCCRP vlastně datovou novinařinu moc dělat neumí. Pro jejich kauzy jsou data odrazovým můstkem, který pak zasazují do nějakého širšího kontextu. „Setřiďte nám to a my si už pak poradíme“, technici a novináři spolu vlastně moc nespolupracují, OCCRP je v tomhle směru neumí propojit. Kdyby na projektech spolupracovali už od začátku, myslím, že by se nich daly vykřesat skvělé věci. Možná je to ale také tím, že kauzy OCCRP jsou vlastně dost standardizované – zkorumpovaný politik tady a támhle. Takže odpověď je ano, redakce by měly mít analytiky a tech týmy.

Je datových leaků čím dál tím víc?

Ano. Zaprvé je nutné vzít v potaz rozšíření ransomware (kradení nebo blokování dat za účelem vydírání – pozn.), stal se z toho velmi výnosný byznys. Dřív se lidé nabourávali třeba do státních institucí, protože to byla sranda, teď je to byznys. Zároveň je všude tolik dat, všechno

se nahrává na internet, do systémů, které nejsou aktualizované, takže je velmi jednoduché se k nim dostat. Zároveň s těmi technologiemi lidé často neumí pracovat, takže se může stát, že někdo něco zveřejní omylem, a stačí vědět, kde to najít.

Dobře, ale platit hackerům za data asi není úplně etické…

Někdo to dělá, vím, že XX médium podplácí úředníky z Ruska, aby jim vynášeli informace. Myslím, že OCCRP tohle nedělá. V Alephu je ale hodně databází, které unikly z Ruska – když tam padl komunismus, neexistovala sjednocená data o obyvatelích, aby například banky věděly, že jim mohou půjčovat peníze. Takže banky, pojišťovny, instituce poskytující půjčky, všechny sbíraly obrovská množství dat a tenkrát se někdo rozhodl všechna tahle data koupit a vynést na dark web. A odtud je získalo OCCRP.

Existují nějaké jiné důvody, kromě toho, že technologie zkrátka úniky dat umožňují, proč leaků přibývá?

Myslím, že z whistleblowerství se stala v podstatě kultura, trend. Je to zkrátka cool vynášet data. A samozřejmě je to užitečné. Pamatuju si na jednoho člověka, který měl k dispozici počítač plný zajímavých dat z Deutschebank a rozhodl se, že sejme Donalda Trumpa. Tak s těmi daty šel za novináři s tím, že jim dá podklady k tomu, aby ho zdiskreditovali. A na tom počítači byly skvělé věci, třeba různé podivné obchody z Kazachstánem. Ale nic o Trumpovi. A ten whistleblower byl nakonec hrozně nešťastný. Pro mě je tohle skvělý příklad toho, jak se z whistleblowerství stal nějaký kulturní fenomén, že je to mnohdy performativní, spíš než užitečné.

Jsou leaky také čím dál větší?

Ano, rozhodně. Zároveň si ale myslím, že si lidé uvědomují, že velikost se nerovná relevanci. Je důležité nad těmi leaky přemýšlet, než do nich investujete čas – říct si, jasně, tohle je obrovské množství dokumentů, ale stojí nám to za ty náklady?

Takže je potřeba nejdříve zjistit, jestli jsou ta data vůbec zajímavá?

Je to začarovaný kruh. Jsou zajímavá? Nevíme, musíme je nejdřív prozkoumat. Pamatuju si na jeden leak, taky spojený s Deutschebank a data utekla z britské firmy, která zakládala offshory. Znělo to skvěle, jako další Panama Papares. Tak jsme data zaindexovali, připravili k prohledávání, dali dohromady špičkový tým novinářů a čekali, co tam najdou za obrovské praní peněz… a ono nic. Měli tam nějaké podivné machinace, jakože se třeba snažili udělat z Kamerunu marihuanovou plantáž, ale žádné velké praní peněz nebo jiné porušování zákona

jsme nenašli. Takže jedni z nejlepších novinářů světa zjistili, že analyzují nějaké nepodstatné podvodníčky.

Kromě technických požadavků, jaké jsou další výzvy, problémy, které datové leaky představují?

Rozhodně ověřování toho, jestli jsou to reálné dokumenty a jestli s nimi někdo nemanipuloval. V největším rozkvětu ruských hackerů, kteří šířili informace do Evropy jsme narazili na případy, kdy ta data upravovali. Další problém je například dekódování konverzací – měli jsme nějaký ruský leak, kde se ti lidé bavili o skříních a různých kusech nábytku z Ikei, což byly názvy pro raketomety. Takže když někdo mluvil o posílání 20 skříní na Donbas, tak nám došlo, že to asi nejsou truhláři. Stejně tak u Kočnerovy knižnice, prostě si musíte vytvořit slovník. Další věc je etická stránka věci – myslím, že bychom s daty měli zacházet s úctou, často máme v dokumentech spousty osobních informací. Může se taky stát, že jsou v datech fotky dětí, což je obrovský právní problém, protože dětskou pornografii nesmí mít ani novináři pro legitimní účely. U Kočnerovy knižnice jsme taky měli dost intimní fotky, které by se asi neměly dostávat ven. Další problém je samozřejmě otázka kolaborace, jak lidi přimět, aby spolu sdíleli informace. Pak je tu otázka dlouhodobého využití těch dat – novináři jsou nastavení vnímat svět podle projektů. Když je jeden hotový, přijde další. Jenže co když je příběh 27 propojený s příběhem 51? To je věc, nad kterou hodně přemýšlím, jak pomocí dat propojovat ty jednotlivé projekty, sjednocovat informace. Třeba o tom člověku, který vás momentálně zajímá, už psali před pěti lety v Černé hoře? Když bylo OCCRP malé, mohli se jednou za rok scházet, dát si pivo a povídat si o tom, jakého mafiána nebo politika mají zrovna v hledáčku.

Proč jsou datové leaky pozitivní?

Myslím, že i když se vlády snaží být transparentní, je to pořád národně-centrické. Data leaky přinášejí globální transparentnost. Najednou se nabídne pohled do toho, jak tyhle nadnárodní šedé a černé ekonomiky fungují. Vidíme, jak jsou životy lidí v malé africké vesnici ovlivňovány tím, co se děje ve Švýcarsku.

A proč jsou negativní?

Jak už jsem říkal, ransomware jako dobrý byznys. Zároveň je tu otázka ochrany soukromí, protože leaky se můžou dotknout každého. Jsme nadšení, že utekla data z ruské mediální organizace, která sice nefunguje tak, jak by média měla, ale pořád jsou to média. Je jen otázka času, kdy se to stane nám. Zároveň je to otázka nálepkování – když teď leaknou data z nějaké organizace, automaticky předpokládáme, že je zlá, špatná. Ale třeba z těch nejzlejších data

neutíkají. Novináři se musejí rozhodovat, čemu budou věnovat pozornost, jestli společnosti, ze které data unikla, ale vlastně nedělá nic až tak hrozného, nebo společnosti, která je čisté zlo, ale data z ní ještě neunikla.

Szabolcs Panyi

Můžeš popsat redakci Direkt36 a jestli jsou v ní specializovaní novináři?

Jsme malé médium a děláme jen dlouhodobé kauzy, takže ne zpravodajství. Je nás 8. většina našich kauz se točí kolem toho, jak se obohacuje rodina premiéra Viktora Orbána, témata spojená s Ruskem – například dostavba jedených bloků v elektrárně a kontrakty s projektem spojené – ruský vliv na maďarskou vládu, a někdy se účastníme nadnárodních spoluprací jako jsou Pandora Papers, nebo Pegasus Project.

Máte v redakci někoho, kdo se specializuje na analýzu dat nebo je jinak technologicky zdatný?

Bohužel ne, bylo by to skvělé.

Myslíš si, že je důležité někoho takového mít?

Ano, byl by to pro nás naprosto klíčový posun. Myslím, že časem někoho takového najmeme, protože obecně redakce roste, ale momentálně se spíš soustředíme na to, abychom měli dostatek obsahu a novinářů, kteří dokáží pracovat na investigativních projektech. Myslím, že je to také částečně způsobeno tím, jak těžké je v Maďarsku získat jakékoli informace, nemáme moc otevřených zdrojů a FOI žádosti jsou většinou zamítány. Nemáme tím pádem k dispozici moc dat, na kterých bychom naše kauzy mohli stavět. V tomhle ohledu dost spoléháme na whistleblowery, na lidi, kteří nám poskytnou dokumenty nebo nám ty informace převypráví, ale že bychom stavěli na nějakých číslech nebo tabulkách, to ne.

Takže většinou pracujete s uniklými dokumenty spíše než s excelovými tabulkami?

Ano, třeba se smlouvami, kontrakty, interní korespondencí mezi nebo vně jednotlivých ministerstev, emaily…

Myslíš si, že je nutné, aby se novináři učili datovou analýzu nebo jiné technické dovednosti? Nebo je lepší mít v redakci profesionála?

Myslím si, že je efektivnější mít v týmu nebo po ruce experta, protože jen než se to lidé naučí… a taky vůbec se o datovou novinařinu zajímat je specifické, nemyslím, že novináři jsou techničtí lidé. když soudím podle sebe – nejsem matematik. Bylo by to pro mě bolestné se tyhle věci učit. Samozřejmě že nějaké základní schopnosti, jak s daty pracovat mám, ale je pro mě mnohem jednodušší mít někoho, kdo má pro ty technické věci cítění, rozumí jim a pomůže mi. Ať už jde o nějaké triky v excelu nebo o vizualizaci dat.

Na jakých velkých datových leacích jsi se podílel?

Jen na jednom – na projektu Pegasus. Byl jsem zároveň jedním z novinářů, kterým hacknuli telefon. Součástí této kauzy byla organizace Frobidden Stories a také OCCRP. Tenkrát jsme měli k dispozici uniklý seznam telefonních čísel a pak ještě pár dalších věcí, o kterých ale nesmím mluvit. Byla to telefonní čísla, která měla být cílem špionážního softwaru Pegasus. Obecně to ale všechno byla čísla. Seznam telefonních čísel obsahoval více než 50 tisíc položek a náš úkol byl propojit ta telefonní čísla s jejich majiteli, lidmi, kteří je používali. Měli jsme k dispozici databázi, o které vlastně můžu říct jen to, že čísla mají samozřejmě specifickou podobu pro různé země, takže časem se čísla rozdělila podle zemí.

Dobře, když nemůžeme mluvit o tom, co je obsahem databáze, můžeš mi nastínit, jak práce s daty fungovala z technické stránky?

Největší problém tkvěl v tom, o jak citlivá data se jedná. Tím, že Pegasus napadal telefony prakticky bez jakéhokoli přispění majitele telefonu – ve smyslu že nebylo třeba klidnou na žádný odkaz – nemohli jsme telefony tak úplně používat. Takže vůbec přihlášení do databáze, kterou jsme používali, byl šílený proces. Každý musel dodržovat všemožná opatření – jak se přihlašovat, z jakého zařízení, odkud, přes speciální účet s různými hesly. Když jsme pak projekt publikovali, což už je něco přes rok, koordinátoři projektu – Forbidden Stories – se rozhodli, že musí z databáze vyhodit novináře z „rizikových zemí“. Pro jistotu, kdyby například do redakce přišla policie, nebo je někdo hacknul. Takže například mě. Když jsem chtěl k databázi přístup, musel jsem požádat někoho z „bezpečných zemí“, aby mi vyhledal konkrétní informace. Celý projekt Pegasus ale vlastně nevisel na datech, jistě byla na začátku, ale ten příběh jsme museli postavit novinářskou prací, museli jsme dohledat spoustu informací. Nabyla to tedy klasická datová novinařina, kdy například taháte čísla z nějakých databází. Další součástí celého projektu byla analýza napadených telefonů, s čímž nám pomáhala Amnesty International, která k tomu účelu vyvinula speciální nástroj. Nejdřív jsme museli potenciální cíle přesvědčit, aby si telefony nechaly analyzovat. Pak Amnesty provedlo technickou analýzu a postupně mám dávalo zprávy, které osahovaly informace o tom kdy byly telefony napadnuty, na jak dlouho a kolik dat bylo z telefonů ukradeno. Právě díky této analýze jsme mohli s jistotu říct, že šlo o Pegasus, protože telefony napadal velmi specifickou mezerou v zabezpečení iMessages. Do telefonů přišla neviditelná iMessage od falešného účtu a když jsme viděli, že stejné falešné účty komunikovaly s určitými lidmi, bylo jasné, že se jedná o stejný systém a je jedno, že jej používaly bezpečností agentury z celého světa.

S jakými zdroji tedy pracujete, když nemáte k dispozici otevřená data? S whistleblowery?

Ano, většinou se jedná o interní emaily nebo diplomatickou komunikaci, které nám někdo poskytne. Každý leak je jiný – někdy je to digitální kopie třeba nějakého dokumentu, což je skvělé, protože se snadno ověřuje jejich pravost. Někdy získám dokumenty, které mohou mít nějaké identifikační znaky, kterým třeba nemusím nutně rozumět, a tak se rozhodnu je raději nezveřejňovat, abych ochránil zdroj. Někdy můžu jen parafrázovat část toho leaku, ne ho přímo citovat. Pro jistotu nezveřejňuji dokumenty v originální podobě, ale udělám screenshot, abych zamezil tomu, že uniknou nějaká metadata, nebo dokument vytisknu a pak ho vyfotím. Maďarská vláda velmi stojí o to lidi, kteří ty informace vypouští najít a v nejlepším případě jen vyhodit. Například se může stát, že je jeden dokument poslaný více příjemcům a my si nemůžeme být jistí, jestli se jedná přesně o ten samý dokument, nebo má nějaké drobné odlišnosti určené právě k tomu, aby bylo jednodušší dostopovat odkud pochází. Samozřejmě musím být opatrný také kvůli tomu, že mě někdo může sledovat – přeci jen už se to stalo. To, jak s dokumenty a obecně s informacemi nakládám je extrémně složité, vlastně mě to dost zatěžuje. Všechno skladuji offline, vytištěné, píšu si poznámky na papír a leaknuté informace z veškerého digitálu mažu co nejrychleji. Tohle ale funguje jen u malých projektů, na kterých pracuju sám.

Potřebovali jste někdy pomoc datového analytika nebo tech experta?

Nevybavuju si, že bychom někoho takového potřebovali. Zároveň by to bylo složité z bezpečnostních důvodů. Většinou nechci používat nikoho externího, částečně kvůli bezpečnosti, ale také proto, že ten proces je už tak pomalý, a ještě kdybych si představil, že musím koordinovat někoho zvenčí.

Myslíš si, že leaků přibývá?

Ano, jednoduše proto, že je mnohem jednodušší kopírovat a skladovat informace, máme k tomu velmi dobře dostupné technologie, třeba chytré telefony. Je to přirozená součást technologického vývoje. A je to dobře. Před dvaceti lety byste musela ty dokumenty doslova propašovat v krabici, dneska stačí foťák nebo disk.

Takže jsou leaky pozitivní?

Ano.

Mají také nějakou negativní stránku?

Ano, vždycky je tu otázka kredibility těch informací. Nevím o tom, že by se to stalo, ale neviděl bych to jako nereálné, kdyby se vláda snažila šířit mezi novináře falešné informace, aby pak

mohla ukázat, že jsme nevěrohodní. A proti tomu se toho moc dělat nedá. Když vám v poště přistane nějaká bomba, je většinou nemožné se jít doptat dalšího zdroje, jestli je to pravda, je nerealistické že by si novinář našel druhého whistleblowera, který mu poskytne stejné bombové informace.

Jsou leaky čím dál větší?

Ano, dá se to tak říct. Jen když vezmeme Pandora Papers, ty byly obrovské a zároveň přinesly obrovské množství informací. Navíc vzhledem k tomu, že soukromé i státní instituce ukládají čím dál tím více informací online, je to logický vývoj. Je snazší se k těm informacím dostat a bude se to dít stále častěji.

Miranda Patrucić

Můžeme začít tím, že krátce vysvětlíte, pro jaké médium pracujete?

Pracuji pro OCCRP už asi deset let, a v podstatě jsem součástí jejich pevného jádra, nepracuju pro žádné jiné médium, ani jsem jiné médium nezaložila. Formálně by se moje pozice dala nazvat šéfredaktor pro jihovýchodní Asii a Kavkaz. Mám pod sebou tým 14 lidí z různých zemí.

Jak váš tým vypadá? Jsou v něm nějací specialisté, specificky zaměření novináři?

Ne, nic takového. Máme 4 stálé redaktory, se kterými pracuji už dlouho, ale ani jeden z nich nemá přímo specializaci, všichni dělají všechno. Samozřejmě mají každý specifický set schopností a dovedností, které se stále rozvíjejí, ale musí být schopní pracovat na jakémkoli tématu, ačkoli mají logicky nějaké preference, když si mohou vybírat.

Je práce vašeho týmu o to náročnější, že zkoumá dění v ne úplně demokratických zemích?

Jistě, například nemají k dispozici žádná veřejná data – nebo k nim mají jen velmi omezený přístup. Proto se musí naučit mnohem víc dovedností, než například my v Evropě. Oni nemůžou prostě dojít na úřad a na něco se zeptat.

Nehodilo by se tedy mít v týmu nějakého experta na data? Analytika, ajťáka?

Nikoho takového nemáme a vlastně nepotřebujeme – od toho je tu OCCRP. Ti mají celé datové oddělení, mají lidi na rešerše…

Takže když se bavíme o robotickém zpracovávání velkého objemu dat, prostě byste zavolali OCCRP?

Přesně tak.

Je takové datové oddělení pro redakci (nebo v tomto případě organizaci) důležité?

Ano, extrémně. Dělají toho spoustu – od scrapeování přes zpracovávání velkých úniků dat, dali dohromady celou databázi (Aleph), se kterou denně pracujeme… Rozhodně jsme na nich závislí, protože data jsou jedním z nejdůležitějších zdrojů informací, který máme jako novináři k dispozici. Například jenom k Ázerbájdžánu máme k dispozici tolik dat, že bychom si vystačili na nějakou tu dobu, kdyby se nestalo nic nového, je v nich tolik příběhů.

Myslíte si, že by si práci s daty měli osvojit novináři? Že by měli umět například scrapeovat, kódovat…?

Osobně si myslím, že ne. Prostě nemůžete být expert na všechno. Když to vezmu podle sebe –

když jsem se potřebovala naučit scrapeovat, prostě jsem si to vygooglila a stáhla jsem si nějaký add-on, kde jsem pak zadala „stáhni tyhle odkazy“ a ono se tak stalo, ale například nevím, jak pracovat s Pythonem. Na jednu stranu je jistě přínosné, aby se novináři neustále vzdělávali a naučili se, co je možné dělat, jaké jsou možnosti… Ale že by to museli například umět všichni v jednom týmu, to rozhodně ne, protože pak nebudou mít třeba jiné schopnosti, které jsou klíčové k dělání dobré investigativy. Z pragmatického hlediska to nedává smysl.

Myslíte si tedy, že je dobré mít přehled, ale místo toho, aby se s obtížnějšími technickými úkony „patlali“ novináři, je lepší je zadat někomu, kdo se tím živí…

Ano, v podstatě ano. Je dobré vědět, co se s jakými daty dá udělat, ale je praktičtější prostě zavolat data týmu, který to udělá za mě. Tím nechci říct, že by se novináři neměli učit nové věci, jen si nemyslím, že by to mělo být standardem, že by se mělo očekávat, že novináři jsou zároveň IT experti.

Máte pocit, že se postupem času zvyšuje požadavek na novináře, aby měli vyšší úroveň technických dovedností?

Ano i ne. Vzpomínám si na velký boom datové žurnalistiky a v tomhle smyslu myslím, že redakce nabíraly lidi, kteří uměli například dělat krásné vizualizace. Prostě to bylo trendy. Ale co se týká investigativní novinařiny obecně, nemyslím si, že by byl požadavek být „datař“. Jasně, hodí se, když to například někdo v týmu umí, ale zdaleka není třeba, aby to byla většina.

Myslíte, že se datové úniky vyskytují častěji?

Obecně ano, a bude jich více. Jako novináři máme k dispozici čím dál více dat, vidíme více uniklých informací, scrapeovaných registrů, hacků. Je to způsob, jak lidé, kteří nesouhlasí s fungováním světa reagují. Myslím, že se to stává novým standardem. Pokud se nějaká společnost nechová korektně, dělá něco nelegálního, dá se očekávat, že z ní někdo vytáhne data. Navíc je předávání informací čím dál jednodušší, dřív, když jste chtěli vynést informace, museli jste to udělat doslova – nějak propašovat ven krabici dokumentů. Teď stačí mít disk. Taky je více příležitostí zneužít toho, že lidé nemají technologie zvládnuté, mají například slabé heslo do emailu.

Dají se jednotlivé datové úniky nějak specifikovat?

Myslím si, že jsou různé trendy v tom, jaká data zrovna unikají. Byly tady leaky americké armády – jako Wiki Leaks, pak jsme měli úniky typu Laundromat (informace o globálních sítích na praní peněz – pozn. red.), těch bylo hned několik po sobě, a pak jsme pracovali na těch

nejznámějších – datech uniklých z korporací, z offshore právních kanceláří a podobných institucí, jako jsou Panama Papers a teď Pandora. Myslím si, že na určité úrovni se dá mluvit o trendech – čím více se o tématu mluví, tím je pravděpodobnější, že o něm budou unikat další data. Problém je, že nikdy nevíte, kdy ten další leak přijde – a jestli vůbec.

A co na základě institucí, ze kterých data unikají?

Neexistuje společnost, která by byla proti leakům imunní. Měli jsme data uniklá z vládních institucí, z bank a privátních firem, od poskytovatelů telekomunikačních služeb… Rozdíl je v tom, jak velké objemy dat to jsou. Může jít o pár papírů, třeba o nějakou důležitou zprávu, a jindy mohou uniknout celé databáze klientů, interní komunikace, transakce…

Dalo by se říct, že se objemy uniklých dat kontinuálně zvětšují?

Ano, zvětšují, ale teď mluvíme jen o těch, ke kterým už reportéři získali přístup. Kdo ví jaká kvanta dat ještě existují – co já vím, jestli někde nejsou data celých vlád. Pokusů o hackování institucí je neustále plno, třeba někde existují data, která ještě vůbec neznáme. Ale v kontextu leaků, se kterými pracovalo ICIJ a OCCRP, ano, zvětšují se, ale ne nutně kontinuálně. Pak se nabízí otázka, zda nemají menší leaky větší relevanci, zda jde vůbec ty úniky posuzovat na základě objemu, a ne například na základě společenského dopadu. Například Pegasus Project (data uniklá ze špionážní firmy která ukázala zneužívání jejího systému – pozn. red.) byl v podstatě jen seznam čísel, ale byl to enormně důležitý leak. Na druhou stranu můžete mít terabity dat, která jsou ale z většiny k ničemu.

Takže to není otázka objemu dat, ale jejich signifikance…

Ano, a také toho, jak těžké muselo být ta data získat. Vynést ten seznam čísel z Pegasu muselo být extrémně riskantní, na rozdíl od toho, že se například nebouráte do serverů právnické firmy, která o tom nemá ani páru.

Je to dobře, že uniká více dat? Že máme více „leakových“ projektů?

Ano, myslím, že ano. Když se podíváme na příklad offshorů – všichni víme, že jsou špatné. Ale kdyby nebylo uniklých dat, nemůžeme ukázat kolik lidí do toho systému přispívá. Leaky jsou důležité, protože odhalují netransparentnost tam, kde by správně měla být naprostá transparentnost.

Na kterých leacích jste pracovala?

Na všech, aspoň myslím. Panama, Pandora, Swiss Leaks, Pegasus, Laundromats, nevím, jestli byly ještě další velké OCCRP + ICIJ projekty, ale dělala jsem je všechny.

Můžete mi popsat, jak ta práce vypadá?

Typicky zavolají, nebo nás zkrátka nějak informují – mě konkrétně bude informovat někdo z OCCRP, pak získáš přístup, typicky přes nějakou zabezpečenou platformu, kam ta data nahráli (před tím, je někdo zprocesoval). Já konkrétně bych tedy hledala informace o zemích ze svého regionu a snažila se zjistit, jestli jsou tam nějaká vodítka, nějaké relevantní informace, a pak na základě toho, co najdu, tak začnu na něčem pracovat. Samotný únik dat ale v mém případě bývá tak 10% celé story, zapojím do ní i další zdroje, vždycky se snažím vytvořit nějakou větší story.

Jak pro vás začíná práce na leaku? Vyhledávání jmen?

Ne nutně, vzhledem k tomu, že nejsem ze zemí centrální Asie, neznám všechny relevantní jména, takže začnu tím, že vyhledám třeba Ázerbájdžán a kouknu, co na mě vyběhne a pak si proklepnu i lidi, co tam vypadnou. Ale třeba Alijevy kontroluju vždycky na první dobrou, to je klasika. Je třeba nad tím přemýšlet, například co porovnávat, jaké databáze a seznamy, protože to někdy nedává smysl – třeba v Pandoře kde máš milion dokumentů Word Searches a podobný kraviny.

Co se děje dál? Seznam lidí, co vás zajímají a prohledávání?

Ano, v případě Pandory jsme na to šli systematicky – vytvořili jsme excell, kam jsme nacpali všechna jména a názvy firem, co jsme v dokumentech našli, a snažili se zjistit, co jsou zač. Když jsme pak zjistili, že nás nějaké jméno nebo firma zajímá, tak jsme se vrátili k datům a začali je studovat.

Může se ale stát, že tam sice máme jméno, ale žádné relevantní dokumenty…

Ano. Dokonce to tak je většinou, ten leak je nedostatečný, neřekne ti celou story.

Museli jste využít služeb nějakých externistů například ke zpracování dat?

Vlastně ne. Konkrétně v případě Pandory vlastně stačila moje expertíza – follow the money – takže samotná data se mi zpracovávala dobře. Ale pro můj tým, který s daty o offshore firmách nikdy nepracoval, to bylo extrémně náročné. Nevěděli si s tím rady. A já jsem jim s tímhle nebyla schopna pomoci, protože když máte k jedné firmě 700 dokumentů, tak je zkrátka někdo musí projít, i když relevantních je jen asi 20 – a na tohle já zkrátka nemám čas.

Nebylo by tedy užitečné mít třeba nějakého externistu, který by pomáhal s procházením dat nějakou formou automatizace?

Ne, protože u dat o firmách je automatizované procházení dat vlastně k ničemu. Samozřejmě to pomůže ze začátku se samotným zpracováním toho leaku, a následně je užitečné vytahat statistická data – kolik bylo v leaku dokumentů, kolik zemí… Nakonec ale stejně skončíte u ruční práce, protože je potřeba otevřít každý dokument a podívat se, co v něm je, vytvořit časové osy… V tomhle už automatizace nepomůže. To se týká Pandory, ale například u Laundromatů byla práce datového týmu naprosto klíčová, to oni pospojovali jednotlivé transakce a vysvětlili nám, která firma kam posílala kolik peněz… Tenhle druh analýzy by žádný reportér nikdy nedokázal udělat.

A jaká jsou pozitiva a negativa práce na velkých datových únicích pro jednotlivá média?

Negativum je rozhodně že zabírají tolik času. Vznikají díky nim úžasné příběhy, ale ta data v podstatě monopolizují veškerý čas, veškerý prostor pracovat na jiných kauzách. Vlastně si člověk musí vybrat, které příběhy jsou pro něj důležitější, mezi loajalitou k partnerům, kteří s ním na projektech pracují a mezi loajalitou k tomu, co je třeba publikovat. To není dobré ani špatné. Ale zkrátka vždycky vám něco uteče.