SRI vrea să cumpere softul “Big Brother”. Pentru eGuvernare.

Explicațiile unui caiet de sarcini tehnic de pe SEAP (nr. anunț de participare 168752/17.07.2016)[1]

#NuSupravegherii

Proiectul SRI denumit SII Analytics, deși identificat ca un proiect pentru promovarea eGuvernării și de combatere a fraudelor (“*prevenție, detectare și luare de măsuri pentru reducerea redundanței plăților în zona publică, prevenirea fraudei și abuzurilor și creșterea eficienței în actul guvernamental;*” - pag. 5, paragraful 3), este un proiect cu potențial de supraveghere generalizată a întregii populații a României – un software Big Brother – cu nicio măsură de limitare a accesului SRI sau al altor instituții publice la datele personale colectate și integrate în acest sistem.

Din analiza caietului de sarcini, dincolo de problema generică a competenței SRI de a implementa un astfel de proiect, identificăm trei zone mari de probleme punctuale:

1. Interceptarea comunicațiilor

1.1. Ce prevede caietul de sarcini al licitației?

Caietul de sarcini prevede hardware și software pentru scop precis de interceptare a traficului pentru monitorizare comunicațiilor – trafic și conținut.

La paginile 10 și 11, la punctul 1.2, SRI-ul vrea să achiziționeze 4 dispozitive pentru interceptarea traficului de Internet (1.2 Echipamentele de tip Traffic Access Point (TAP) - 4 buc.). La paginile 11 și 12, la punctul 1.3, descoperim că SRI-ul mai vrea să achiziționeze și o „Soluție de monitorizare a utilizării resurselor de comunicare colaborativă”.

Aceasta înseamnă un software, și hardware-ul necesar pentru acesta, care ia traficul interceptat de dispozitivele de care vorbeam mai devreme și îl analizează pentru a monitoriza comunicația desfășurată prin servicii și programe de instant messaging (cum ar fi Yahoo Messenger, Facebook Messenger, Skype etc.), email sau orice alte mecanisme similare de comunicație electronică.

Text original “*identificarea protocoalelor specifice de comunicare colaborativă în tehnologie Web, inclusiv a mesageriei electronice "instant" și a celei tradiționale (de tip email și asimilată), precum și a tehnologiilor conexe de schimb de informație nestructurată, indiferent dacă serviciile de acest tip vor fi accesate și utilizate prin aplicații dedicate, sau prin pagini Web, în instanțe de tip browser”*.

Din analiza textului caietului de sarcini, rezultă că vorbim inclusiv de accesul la conținut conform scopului *(„identificarea, decodarea și salvarea, la cerere, precum și vizualizarea artefactelor asociate (inclusiv a fișierelor text și multimedia atașate)”, „analiza și retenția selectivă a sesiunilor de comunicare respective” și merge până la „reconstrucția și vizualizarea contextului comunicării [...] pentru identificarea rapidă a secvenței operațiilor și, respectiv, a relațiilor între participanți”.)*

1.2. Care este problema?

Aceste dispozitive și programe nu au nici un rost în cadrul proiectului de Big Data sau e-

1 http://media.hotnews.ro/media_server1/document-2016-07-18-21166594-0-caiet-sarcini-sri.pdf

Guvernare, pentru datele transmise legal – acestea **au un rol doar pentru interceptarea comunicațiilor** fără acordul utilizatorilor.

Mai mult în mod normal supravegherea comunicațiilor se poate face doar în contextul stabilit de Codul de Procedură Penală, deci cu autorizarea judecătorului, altfel aceste activități sunt ilegale. Unde va instala SRI-ul aceste echipamente, ce date doresc să intercepteze și cu ce autoritate vor analiza și stoca datele și informațiile rezultate în urma acestei monitorizări?

1.3. Care este scenariul cel mai negru?

Dacă un astfel de dispozitiv este instalat oriunde la un furnizor de acces Internet (ISP) sau într-un nod de interconectare (gen RoNIX), acesta ar putea intercepta tot traficul de Internet. Prin capacitatea de 10GB/s înmulțit cu patru dispozitive s-ar putea intercepta aproximativ tot traficul din România prin RoNIX.

Prin software-ul de analiză, se poate identifica un anumit tip de conținut – de exemplu:

- email-uri trimise de la tribunalulbucuresti.ro
- email-uri cu fișier video atașat de dimensiune între 10MB și 25MB
- convorbiri pe Facebook/Yahoo chat între X si Y (s-ar putea ca întreg conținutul să fie mai greu de depistat, dar cel puțin faptul că 2 persoane au discutat se poate identifica)

2. Analiză comportamentală

2.1. Ce scrie în caietul de sarcini?

Proiectul vrea să aducă împreună surse de date, care înseamnă:

- baze de date – cam orice informație structurată tabelar pusă la dispoziție de către "instituțiile beneficiare" - e.g. ANAF, MAI, CNAS etc. - pct 2.19.3.1. - *"cel puțin 28.000 de tabele cu cel puțin 500.000 de coloane şi cel puțin 35.000.000.000 de înregistrări;"* (pag. 95, secțiunea 2.19.3.1.1 Depozite de date comune, primul punct)

"Nivelul surse de date cuprinde depozitele de date comune, așa cum au fost ele definite în cadrul proiectului, ca totalitatea depozitelor de date structurate, semi-structurate şi nestructurate – formă consolidată a datelor furnizate de instituțiile cu rol de furnizor în cadrul proiectului "SII INFRASTRUCTURĂ" - fără a se limita la acestea, precum şi seturile de date individuale, așa cum au fost ele definite în cadrul proiectului, ca totalitatea seturilor de date semi-structurate şi nestructurate, aflate în posesia utilizatorilor finali." (pag. 95, secțiunea 2.19.3.1 Nivel surse de date)

- alte documente nestructurate (formate acceptate - *"formatele PDF, HTML, HTM, XHTML, PPT, PPTX, DOC, DOCX, RTF, XLS, XLSX, CSV, XML, JSON şi TXT"*) (pag. 96, paragraful 3)

Toată aceste baze de date corelate vor putea fi căutate în orice mod – de la căutări simple (text, arii geografice) la chestiuni mai complexe (bazate pe analiza legăturilor dintre persoane, obiecte (mașini, telefoane etc.), evenimente sau orice alt fel de informație, plasarea în timp și spațiu a acestor legături și chiar estimarea comportamentului țintelor pornind de la aceste informații și analize.

Astfel se vrea ca sistemul să poată da *„răspunsuri la întrebări de tipul:*

- *A. care sunt societăţile comerciale ai căror acţionari sau administratori figurează la mai mult de N societăţi comerciale şi au cazier judiciar?*
- *B. care sunt societăţile comerciale înfiinţate între anii AAAA-AAAA, cu numărul mediu de angajaţi în perioada BBBB-BBBB mai mic decât N şi cu rambursări de TVA mai mari de S lei?*
- *C. care sunt societăţile comerciale cu cifra medie / angajat mai mare decât media + P% din abaterea medie standard a codului CAEN X?*
- ***D. care sunt persoanele fizice al căror venit personal declarat pe perioada AAAA - AAAA şi creditele bancare contractate pe perioada BBBB - BBBB este cu P% mai mic decât numărul de metri pătraţi de teren / clădiri * V şi auto de valoare minim S lei, dobândite în perioada CCCC – CCCC;" (pag. 123-124)***

Mai mult, caietul de sarcini precizează că datele colectate nu vor fi șterse probabil niciodată ("*vor fi stocate pe termen indefinit numai la nivelul componentelor offline ale sistemului.*" - pag 144)

Cel puțin 45 de persoane vor fi utilizatori finali /formatori, fiind instruiți în sistem, dar probabil în jur de 1000 de persoane vor avea acces la sistem.

Se cer 750 de terminale mobile cu acces la sistem și un număr nespecificat de terminale fixe. Se cere „să asigure o medie a timpului de răspuns de sub 5 secunde, în condițiile în care 500 de utilizatori concurenți vor efectua regăsiri într-un interval de 60 minute" pentru toate tipurile de căutări (pag. 154).

2.2. Care este problema?

În primul rând colectarea datelor în acest mod este ilegală. Cum deja a precizat Curtea Europeană de Justiție în cazul Bara vs CNAS și ANAF[2] în interpretarea Legii 677/2001 privind protecția datelor cu caracter personal, o instituție publică nu poate da datele colectate unei alte instituții publice doar pentru că "ar putea să fie de folos", ci trebuie să se bazeze fie pe consimțământ, fie pe informarea persoanei vizate dacă există restul de limitări conform legii.

Datele de mai sus – punctul D ne arată că cel puțin următoarele baze de date ar putea fi cumulate:

- declarații către ANAF cu privire la veniturile personale;
- credite bancare acordate;
- bunuri imobile cumpărate;
- automobile achiziționate.

Practic SRI își creează un sistem (unde s-ar putea ca și alții să aibă acces conform descrierii generale) unde **ar putea căuta orice informație despre orice persoană fizică care apare într-o bază de date a autorităților publice**, nerespectând în mod flagrant legislația privind protecția datelor cu caracter personal și neaplicând nicio măsură de protecție a acestor date.

2.3. Care este scenariul cel mai negru?

În varianta soft scenariul cel mai negru ar însemna faptul că o persoană ce are acces la sistem îl folosește în interes propriu – de exemplu:

2 Vezi http://curia.europa.eu/juris/document/document.jsf?docid=168943&doclang=EN

- vecinul tău dintr-o dată pare că știe prețul cu care ai cumpărat apartamentul sau ca mai ai 2 mașini pe numele tău; sau
- polițistul care te oprește află instantaneu că ai cumpărat 4 mașini de-al lungul vieții; sau
- se rulează o cerere de tipul – care este persoana care a vândut un apartament cel mai ieftin/cel mai scump în cartierul Militari; sau
- o firmă cunoscută ca "apropiată serviciilor" va găsi întotdeauna elemente pentru a șicana alte firme concurente.

Am făcut aceste scenarii exclusiv pe baza informațiilor despre bazele de date ce rezultă din caietul de sarcini. Dacă la aceste informații se alătură și alte baze de date – gen cele accesate prin cardul electronic de sănătate sau informațiile dorite recent de ANAF cu privire la plățile cu cardul, problemele pot fi multi mai complicate.

În varianta hard, ar fi creat automat câte un director (folder) pentru fiecare CNP identificat în vreo bază de date, astfel încât SRI și ceilalți care au acces la baza de date pot afla orice informații despre orice persoană dacă doresc.

Acesta este adevăratul Big Brother, pentru că poți știi tot despre toți!

2.4. Dar totuşi, nu ar avea voie statul să caute în propriile baze de date pentru a depista corupții?

Ba da, dar în condiții în care să respecte drepturile cetățeanului la viață privată.

Sistemele de e-guvernare ne pot ușura viața, dar cetățenii trebuie să aibă încredere faptul că instituțiile publice colectează informațiile într-un scop bine precizat și nu pentru orice alt scop viitor.

De aceea **interoperabilitatea datelor personale în zona de eGuvernare** trebuie să se bazeze pe **transparență** (cetățeanul să poată vedea în orice moment cine, când și de ce cineva a accesat datele lui) și **respectarea principiilor** colectării datelor cu caracter personal (limitarea scopului, bază legală, necesitate și proporționalitate, perioada de păstrare, respectarea drepturilor subiecților, notificarea încălcărilor de securitate, analiză de impact a datelor personale, angajarea unui responsabil privind protecția datelor personale etc.).

Vezi detalii în opinia Grupului de Lucru al Articolului 29[3] care oferă suficiente informații referitoare la principiile care trebuie să stea la baza transferului de date: Guidelines for Member States on the criteria to ensure compliance with data protection requirements in the context of the automatic exchange of personal data for tax purposes adopted on 15 December 2015.[4]

De asemenea, **accesul la <u>toate</u> datele personale dintr-o bază de date ar trebui de principiu interzis**, cu posibilitatea unor excepții pentru datele anonimizate sau pseudonimizate.

3. Recunoaştere facială

3 Grupul de Lucru al Articolului 29 este un grup format din totalitatea autorităților de protecție a datelor personale din Uniunea Europeană - http://ec.europa.eu/justice/data-protection/article-29/index_en.htm

4 http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2015/wp234_en.pdf

3.1. Ce scrie în caietul de sarcini?

Sistemul are inclusă și o componentă de recunoaștere facială (adică de recunoaștere a unei persoane dintr-o poză) care să poată opera cu o bază de date de 3 milioane de imagini – pct. 2.20.

Textul include și o serie de imagini de referință, parte a „datelor salvate în depozitele de date nestructurate”. Conform documentului „În prezent, există imagini de referință alb-negru și color, cu o dimensiune medie de 16KB, ce ocupă cel puțin 950 GB spațiu pe disk (…) O persoană poate avea asociate mai multe imagini de referință. **În aceste imagini, subiectul nu prezintă ocluzii sau variații ale aspectului și expresiei faciale, respectiv ale rotației capului față de poziția frontală.** Rezoluția nativă este de minim 200 x 240 pixeli.” (pag. 130, secțiunea 2.20.1 Imagini de referință)

Din definiția de mai sus noi estimăm că aceste poze nu pot fi decât cele din buletine, dar și cele din pașapoarte (atât expirate, cât și actuale).

Un calcul minimal din spațiul ocupat pe disc ne arată că probabil vorbim în jur de 50-60 de milioane de imagini, care ar fi considerate imagini de referință.

3.2. Care este problema?

În primul rând, soluția de recunoaștere facială (face recognition) nu are vreo legătură cu niciunul din scopurile proiectului – de prevenire și combatere a fraudei, e clar că scopul trebuie să fie altul.

În al doilea rând, este neclar în ce condiții SRI are acces la pozele din pașapoarte și buletine și cu ce scop le folosește.

În al treilea rând, se pare că imaginile de referință sunt doar începutul, ele urmând să creeze o bază de date cât mai largă. Vezi pagina 130, secțiunea 2.20.1 Imagini de referință, paragraful 2 - „Numărul și varietatea depozitelor de date comune este prevăzut să crească în continuare”– deci nu se va limita la pozele din buletine.

3.3. Care este scenariul cel mai negru?

Prin corelarea acestor informații cu cele din analiza comportamentală se pot identifica nu doar imaginile persoanelor, ci și alte detalii despre ele din restul bazelor de date.

De exemplu:

- SRI își face o aplicație astfel încât să „citească” în mod automat Facebook, astfel încât să integreze date de acolo în baza de date de recunoaștere facială. De exemplu, dacă aveți o poză reală la profilul de Facebook, chiar dacă sub alt nume sau un nume incomplet sau comun, SRI poate să conexeze restul de informații din bazele de date proprii cu acel cont de Facebook, pentru a ști că dvs. sunteți acel Ion Popescu pe care îl urmărește.

- SRI/Jandarmeria filmează (și preia apoi) sau fotografiază 5000 de participanți la un protest din Piața Universității. Înregistrarea foto/video poate fi, după aceea, procesată și apoi analizată cu software-ul de recunoaștere facială pentru a identifica exact ce persoane au participat pentru a fi contactați ulterior și amendați.