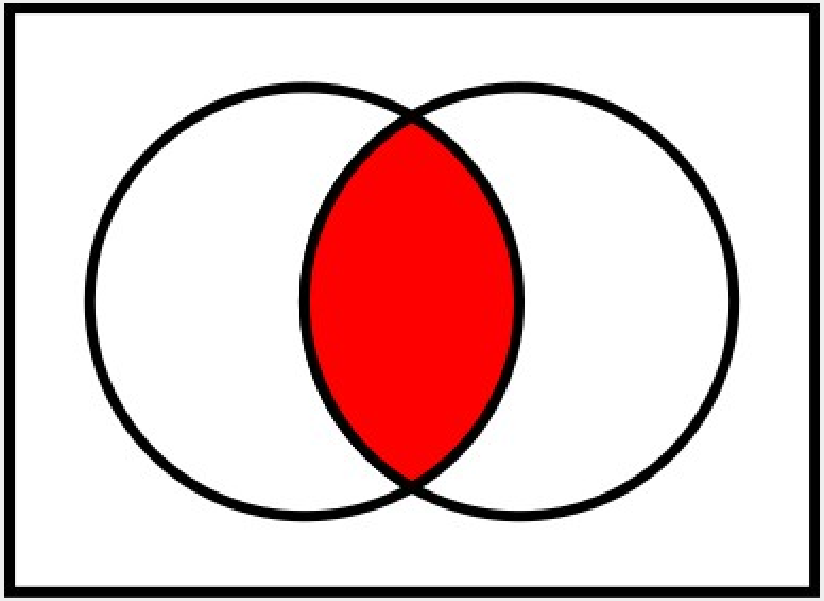
Private Set Intersection

A cryptographic technique that allows two parties to compute the intersection of their sets without revealing anything except the intersection

Matej Kovačič
Matjaž Rihtar

Private Set Intersection

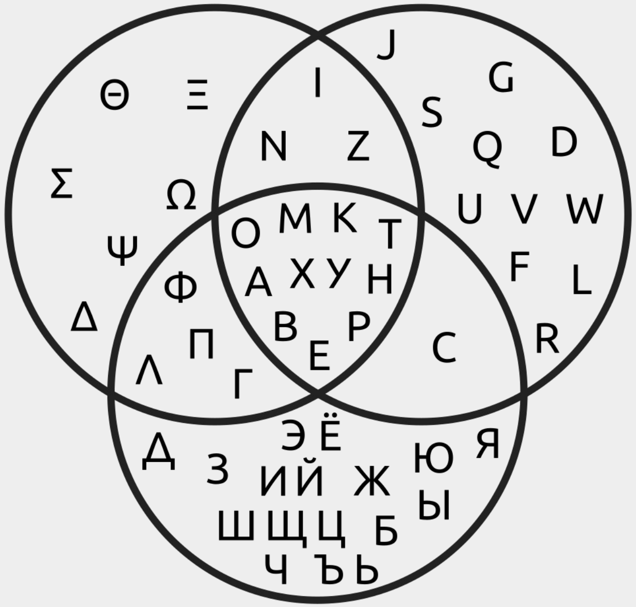
In a private set intersection (PSI) protocol two parties jointly compute the intersection of their private input sets.



A client and a server jointly compute the intersection of their private input sets in a manner that at the end the client learns the intersection and the server learns nothing (one-way PSI) or both learn the intersection (mutual PSI).

Private Set Intersection

Such protocols are especially useful whenever one or both parties (who do not fully trust each other) must compute an intersection of their respective data sets (i.e. only the required minimum amount of information should be disclosed).



Intersections of the Greek, English and Russian alphabet, considering only the shapes of the letters and ignoring their pronunciation.

Examples of use

1. A government agency needs to make sure that employees of its industrial contractor have no criminal records. Government does not want to disclose list of all convicted felons, and contractor does not want to disclose their full list of employees. However, both would like to know the intersection, if any.

2. Federal tax authority wants to learn whether any suspected tax evaders have accounts with a certain foreign bank and, if so, obtain their account records. The bank is forbidden of wholesale disclosure of their account holders and the tax authority can not reveal its list of suspects to a foreign bank.

Source: Emiliano De Cristofaro, Gene Tsudik. 2009. Practical Private Set Intersection Protocols with Linear Computational and Bandwidth Complexity. Cryptology ePrint Archive, Report 2009/491. <https://eprint.iacr.org/2009/491>.

Examples of use

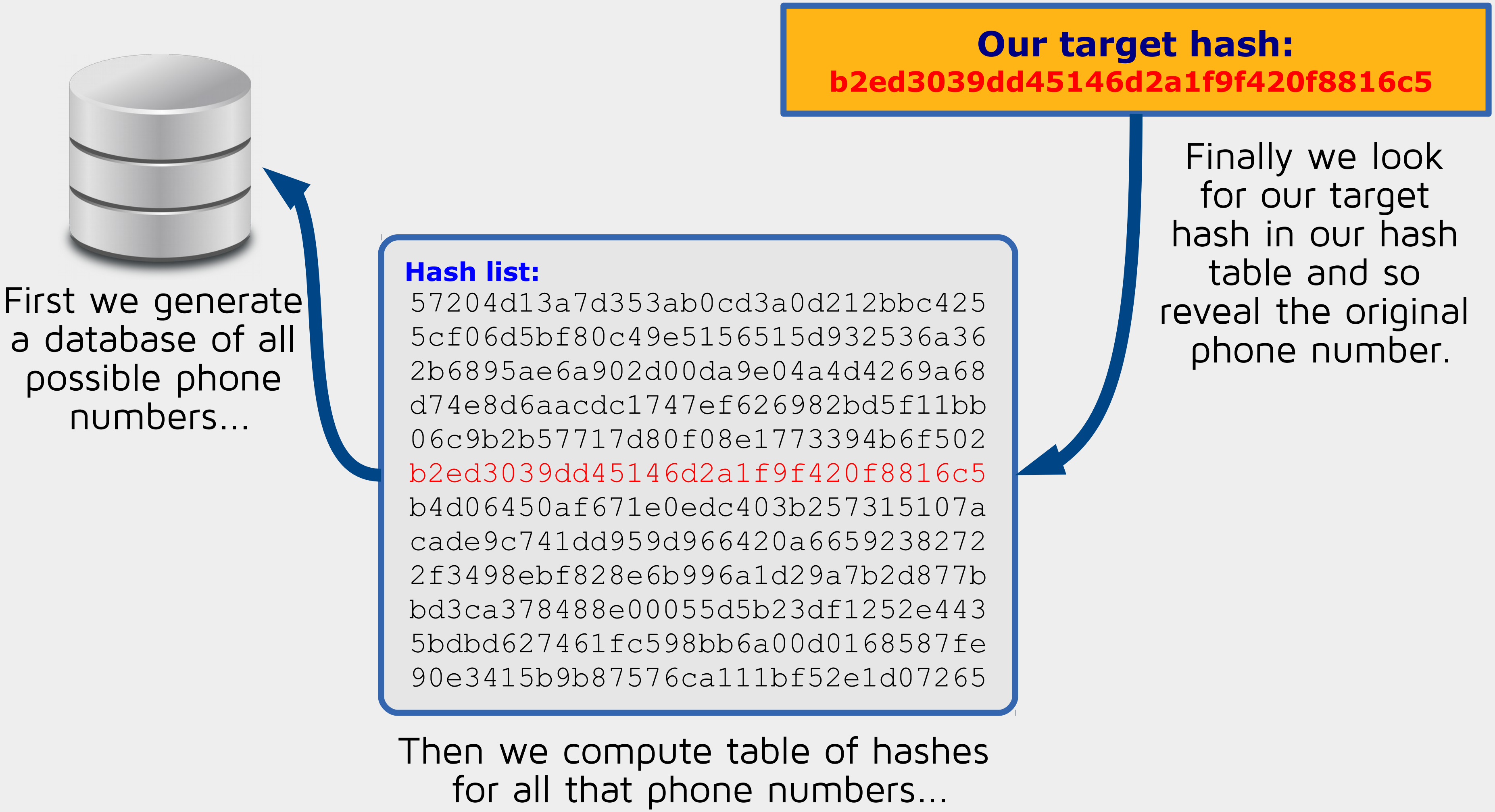
3. Two real estate companies would like to identify homeowners who are double-dealing. That means they have signed *exclusive* contracts *with both companies* to assist them in selling their properties. None of them would like to reveal the other party list of their customers.

4. Two national law enforcement bodies want to compare their respective databases of terrorist suspects. National privacy laws prevent them from revealing bulk data, however, by treaty, they are allowed to share information on suspects of common interest.

Source: Emiliano De Cristofaro, Gene Tsudik. 2009. Practical Private Set Intersection Protocols with Linear Computational and Bandwidth Complexity. Cryptology ePrint Archive, Report 2009/491. <https://eprint.iacr.org/2009/491>.

Why not hashes?

Hashes are supposed to be one-way functions, but this is not always true...



Our target hash:
b2ed3039dd45146d2a1f9f420f8816c5

First we generate a database of all possible phone numbers...

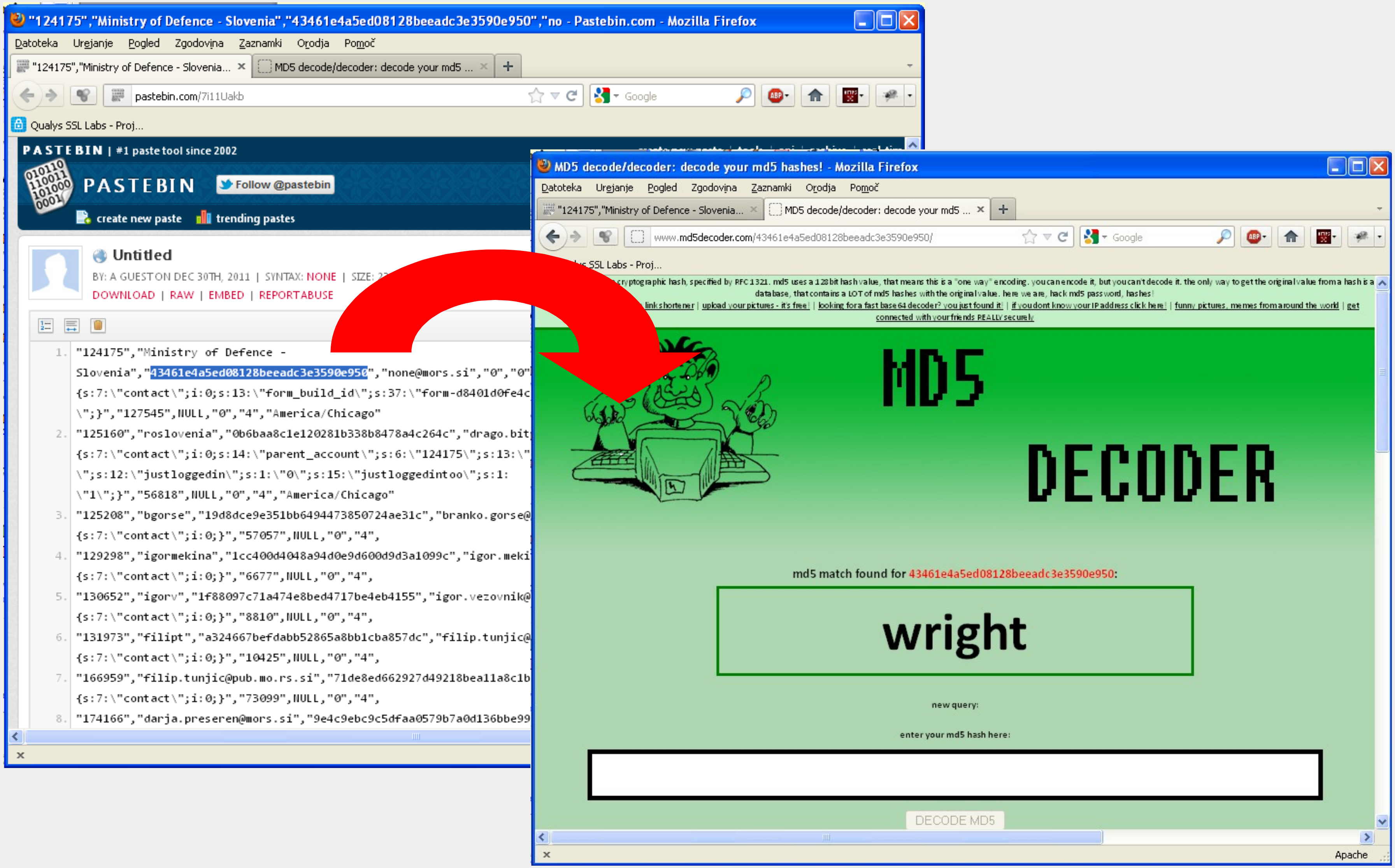
Hash list:

```
57204d13a7d353ab0cd3a0d212bbc425
5cf06d5bf80c49e5156515d932536a36
2b6895ae6a902d00da9e04a4d4269a68
d74e8d6aacdc1747ef626982bd5f11bb
06c9b2b57717d80f08e1773394b6f502
b2ed3039dd45146d2a1f9f420f8816c5
b4d06450af671e0edc403b257315107a
cade9c741dd959d966420a6659238272
2f3498ebf828e6b996a1d29a7b2d877b
bd3ca378488e00055d5b23df1252e443
5bdbd627461fc598bb6a00d0168587fe
90e3415b9b87576ca111bf52e1d07265
```

Then we compute table of hashes for all that phone numbers...

Finally we look for our target hash in our hash table and so reveal the original phone number.

Why not hashes?



"124175","Ministry of Defence - Slovenia","43461e4a5ed08128beeadc3e3590e950","no - Pastebin.com - Mozilla Firefox

Datoteka Urejanje Pogled Zgodovina Zaznamki Orodja Pomoč
"124175","Ministry of Defence - Slovenia...
MD5 decode/decoder: decode your md5 ...
pastebin.com/7i11Uakb
Google
Qualys SSL Labs - Proj...
PASTEBIN | #1 paste tool since 2002
PASTEBIN
Follow @pastebin
create new paste
trending pastes

Untitled

BY: A GUEST ON DEC 30TH, 2011 | SYNTAX: NONE | SIZE: 2

DOWNLOAD | RAW | EMBED | REPORTABUSE

```
"124175","Ministry of Defence -
Slovenia","43461e4a5ed08128beeadc3e3590e950","none@mors.si","0","0"
{s:7:\"contact\";i:0;s:13:\"form_build_id\";s:37:\"form-d8401d0fe4c
\";}","127545",NULL,"0","4","America/Chicago"
"125160","roslovenia","0b6baa8c1e120281b338b8478a4c264c","drago.bit
{s:7:\"contact\";i:0;s:14:\"parent_account\";s:6:\"124175\";s:13:\"
\";s:12:\"justloggedin\";s:1:\"0\";s:15:\"justloggedintoo\";s:1:
\"1\";}","56818",NULL,"0","4","America/Chicago"
"125208","bgorse","19d8dce9e351bb6494473850724ae31c","branko.gorse@
{s:7:\"contact\";i:0;}","57057",NULL,"0","4",
"129298","igormekina","1cc400d4048a94d0e9d600d9d3a1099c","igor.meki
{s:7:\"contact\";i:0;}","6677",NULL,"0","4",
"130652","igorv","1f88097c71a474e8bed4717be4eb4155","igor.vezovnik@
{s:7:\"contact\";i:0;}","8810",NULL,"0","4",
"131973","filipt","a324667befdabb52865a8bb1cba857dc","filip.tunjic@
{s:7:\"contact\";i:0;}","10425",NULL,"0","4",
"166959","filip.tunjic@pub.mo.rs.si","71de8ed662927d49218bea11a8c1b
{s:7:\"contact\";i:0;}","73099",NULL,"0","4",
"174166","darja.preseren@mors.si","9e4c9ebc9c5dfaa0579b7a0d136bbe99
```

MD5 decode/decoder: decode your md5 hashes! - Mozilla Firefox

Datoteka Urejanje Pogled Zgodovina Zaznamki Orodja Pomoč
"124175","Ministry of Defence - Slovenia...
MD5 decode/decoder: decode your md5 ...
www.md5decoder.com/43461e4a5ed08128beeadc3e3590e950/
Google
...s SSL Labs - Proj...

...cryptographic hash, specified by RFC 1321. md5 uses a 128bit hash value, that means this is a "one way" encoding. you can encode it, but you can't decode it. the only way to get the original value from a hash is a database, that contains a LOT of md5 hashes with the original value. here we are, hack md5 password, hashes!

link shortener | upload your pictures - it's free! | looking for a fast base64 decoder? you just found it! | if you dont know your IP address click here! | funny pictures, memes from around the world | get connected with your friends REALLY securely

md5 match found for 43461e4a5ed08128beeadc3e3590e950:

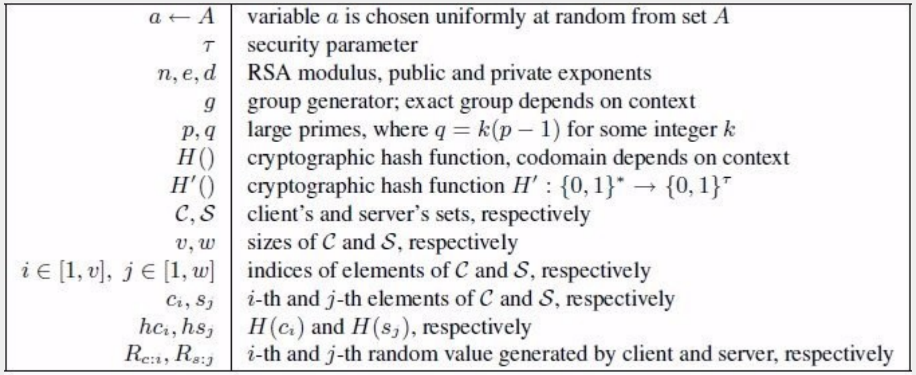
wright

new query:

enter your md5 hash here:

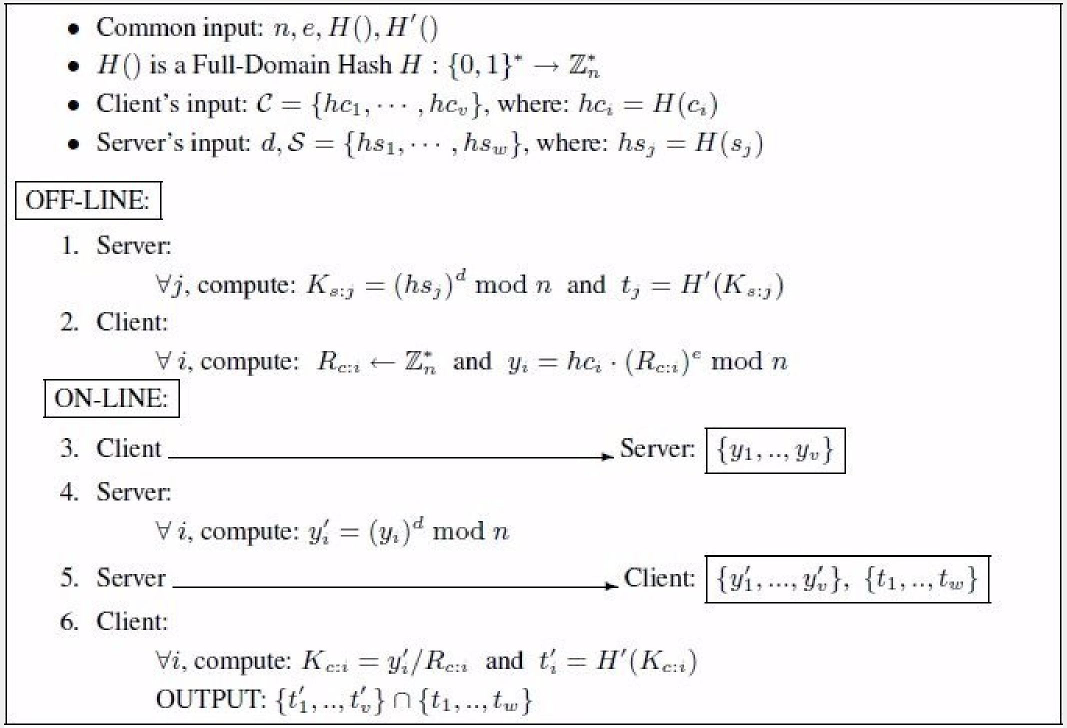
On the Internet we can find lists of precomputed hashes...

How does PSI work?



$a \leftarrow A$	variable a is chosen uniformly at random from set A
τ	security parameter
n, e, d	RSA modulus, public and private exponents
g	group generator; exact group depends on context
p, q	large primes, where $q = k(p-1)$ for some integer k
$H()$	cryptographic hash function, codomain depends on context
$H'()$	cryptographic hash function $H' : \{0,1\}^* \rightarrow \{0,1\}^\tau$
$\mathcal{C}, \mathcal{S}$	client's and server's sets, respectively
v, w	sizes of $\mathcal{C}$ and $\mathcal{S}$, respectively
$i \in [1, v],\ j \in [1, w]$	indices of elements of $\mathcal{C}$ and $\mathcal{S}$, respectively
c_i, s_j	i-th and j-th elements of $\mathcal{C}$ and $\mathcal{S}$, respectively
hc_i, hs_j	$H(c_i)$ and $H(s_j)$, respectively
$R_{c:i}, R_{s:j}$	i-th and j-th random value generated by client and server, respectively

Table 1: Notation



- Common input: $n, e, H(), H'()$
- $H()$ is a Full-Domain Hash $H : \{0,1\}^* \rightarrow \mathbb{Z}_n^*$
- Client's input: $\mathcal{C} = \{hc_1, \cdots, hc_v\}$, where: $hc_i = H(c_i)$
- Server's input: $d, \mathcal{S} = \{hs_1, \cdots, hs_w\}$, where: $hs_j = H(s_j)$

OFF-LINE:

1. Server:
   $\forall j$, compute: $K_{s:j} = (hs_j)^d \bmod n$ and $t_j = H'(K_{s:j})$
2. Client:
   $\forall i$, compute: $R_{c:i} \leftarrow \mathbb{Z}_n^*$ and $y_i = hc_i \cdot (R_{c:i})^e \bmod n$

ON-LINE:

3. Client $\longrightarrow$ Server: $\boxed{\{y_1, .., y_v\}}$
4. Server:
   $\forall i$, compute: $y_i' = (y_i)^d \bmod n$
5. Server $\longrightarrow$ Client: $\boxed{\{y_1', ..., y_v'\},\ \{t_1, .., t_w\}}$
6. Client:
   $\forall i$, compute: $K_{c:i} = y_i' / R_{c:i}$ and $t_i' = H'(K_{c:i})$
   OUTPUT: $\{t_1', .., t_v'\} \cap \{t_1, .., t_w\}$

Figure 4: Blind RSA-based PSI Protocol with linear complexity

PSI exchange, which is a functionality in secure multi-party computation (also known as privacy-preserving computation) **is much more secure than exchange with hashes**, because it uses encryption.

Source: Emiliano De Cristofaro, Gene Tsudik. 2009. Practical Private Set Intersection Protocols with Linear Computational and Bandwidth Complexity. Cryptology ePrint Archive, Report 2009/491. <https://eprint.iacr.org/2009/491>.

How does PSI work?

Server takes its data, computes hash (SHA-256), convert this to integer, makes RSA decode (mathematical operation which is a part of RSA decode procedure), and computes new hash (SHA-256).

data → hash(s_data) → s_number1 → (s_number1)$^{\wedge d}$ mod n → s_number2 → hash(s_number2) → s_number3

Client computes hashes (SHA-256) for all data.

c_data → hash(c_data) → c_number

Then client for each data element generates random number (c_rand).

Now client has: c_data :: c_number :: c_rand

How does PSI work?

Then client multiplies each data element (c_data) with RSA encoded random number. Random number c_rand is RSA encoded with server's public key. {RSA(c_rand) * c_data mod n}. → This result is then sent to server.

For each received data server then computes RSA decode. Decoded client data and server's final hash of its data (s_number3) are sent to client. ←

Finally client takes (his) RSA decoded client data from server and divides them with c_rand number and computes a hash. This hash is then compared with server's data (s_number3).

If both hashes are the same, the data are the same, so there is intersection between the data on server and a client.

Transliteration

Another problem is that different languages has different writing system.

Transliteration is utilized when a word or phrase must be “transmitted” to a language with a different writing system.

For the we took ICAO (The International Civil Aviation Organization) transliteration tables, however we improved them a little, so our system now uses “extended ICAO” transliteration standard.

Example (lines containing # are ICAO original standard, lines containing ## are our “extended ICAO”):

- 0x00DD: 'Y', # Y with acute
- 0x00DE: 'TH', ## Thorn (Iceland)
- 0x00DF: 'ss', # sharp s (Germany)
- 0x0136: 'K', # K with cedilla
- 0x0137: 'k', ## k with cedilla
- 0x0138: 'k', ## kra
- 0x0139: 'L', # L with acute

Transliteration

Our “extended ICAO” supports all small and some additional upper case and additional cyrillic characters.

Our procedure is that we first perform transliteration and then convert characters to uppercase.

Example of a problem with ICAO standard (one character could be transliterated to differend latin characters – depending of an original language):

- 0x0413: 'G',   # GHE (or G, except Belorussian and Serbian H)

Suggestions:

- If you are using small caps, you can leave them – our system will automatically transcode them to uppercase.
- If you have cyrillic characters in original, leave them in cyrillic.

Transliteration

Example of transliteration:

Matej Kovačič → MATEJ KOVACIC

Matjaž Rihtar → MATJAZ RIHTAR

Permutations

Asumption is that one person can have from 2 to 5 names (first name + up to 4 additional names or surnames).

This means from 2 to 394 possible permutations.

Example: "Kyle Reese Sergeant, born 1. 1. 2003".

All possible permutations:

SERGEANT REESE KYLE 2003-01-01
REESE SERGEANT KYLE 2003-01-01
SERGEANT REESE 2003-01-01
REESE SERGEANT 2003-01-01
REESE KYLE SERGEANT 2003-01-01
SERGEANT KYLE 2003-01-01
KYLE SERGEANT 2003-01-01
SERGEANT KYLE REESE 2003-01-01
KYLE REESE 2003-01-01
KYLE REESE SERGEANT 2003-01-01
REESE KYLE 2003-01-01
KYLE SERGEANT REESE 2003-01-01

Permutations

We can find even cases when one or more names are similar and not exactly the same. For example:

- country A has a person “SERGEANT REESE **KYLE** 2003-01-01”
- country B has a person “SERGEANT REESE **KYLLE** 2003-01-01”

PSI exchange will find a match, because one of the permutations in both countries is “SERGEANT REESE 2003-01-01” (name KYLE or KYLLE is neglected in this permutation).

However, please note, that birthday of both persons should always be the same. Otherwise, there is no match.

Technical requirements

PSI application is written in programming language Python.

Installation script provides all installation packages (Python 2.7, required Python modules and the application itself).

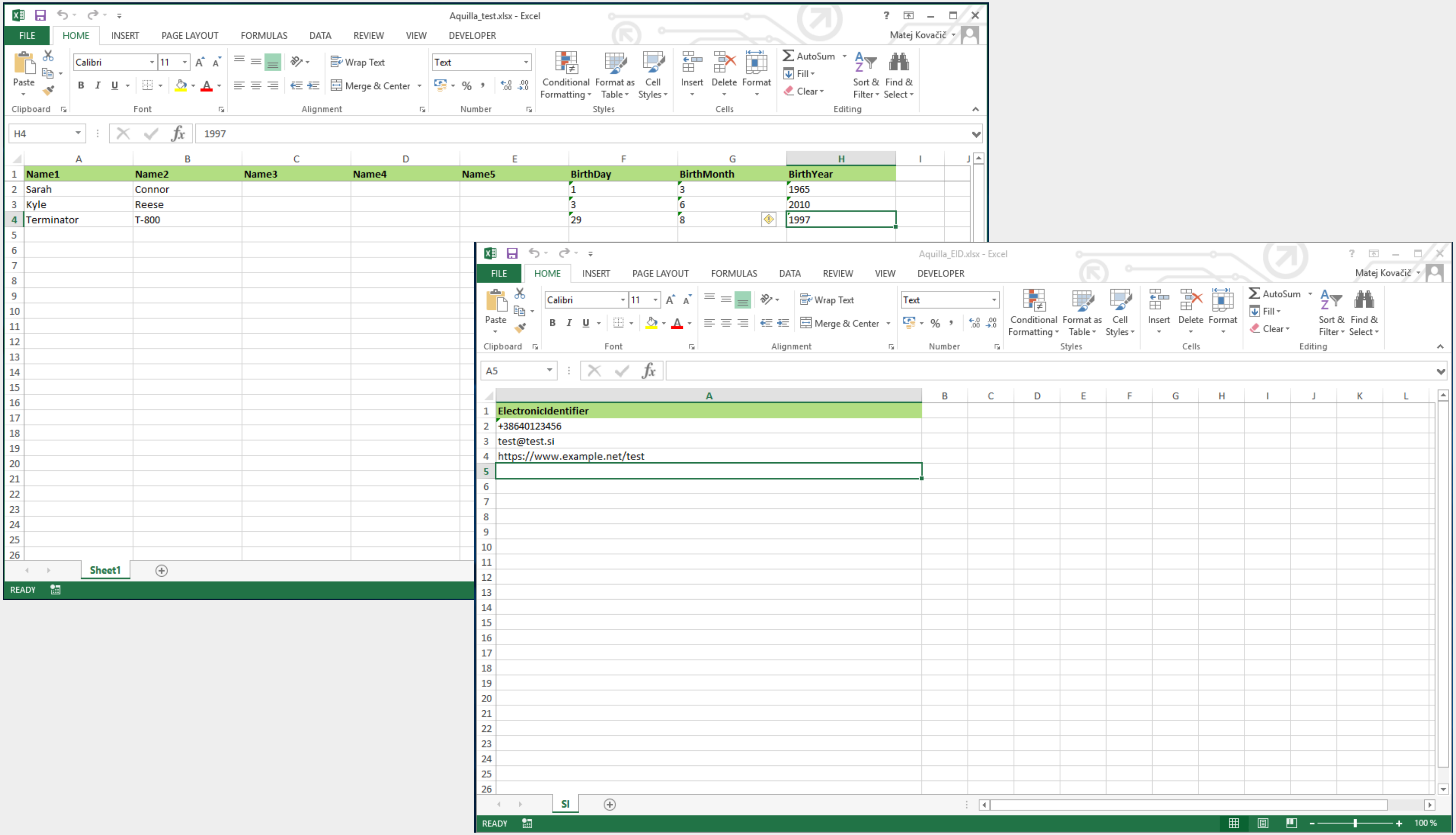
Application runs on operating system Windows 10 (minimum Windows 7), but it could be ported to other platforms as well.

Other needed software:

- for importing the data from Excell files, Microsoft Office 2010 or more recent version is required;
- web browser (suggested is the latest version of Chrome or Firefox).

Technical requirements

Data could be imported via CSV or Excell (XLSX) file with predefined set of fields.



Aquilla_test.xlsx - Excel

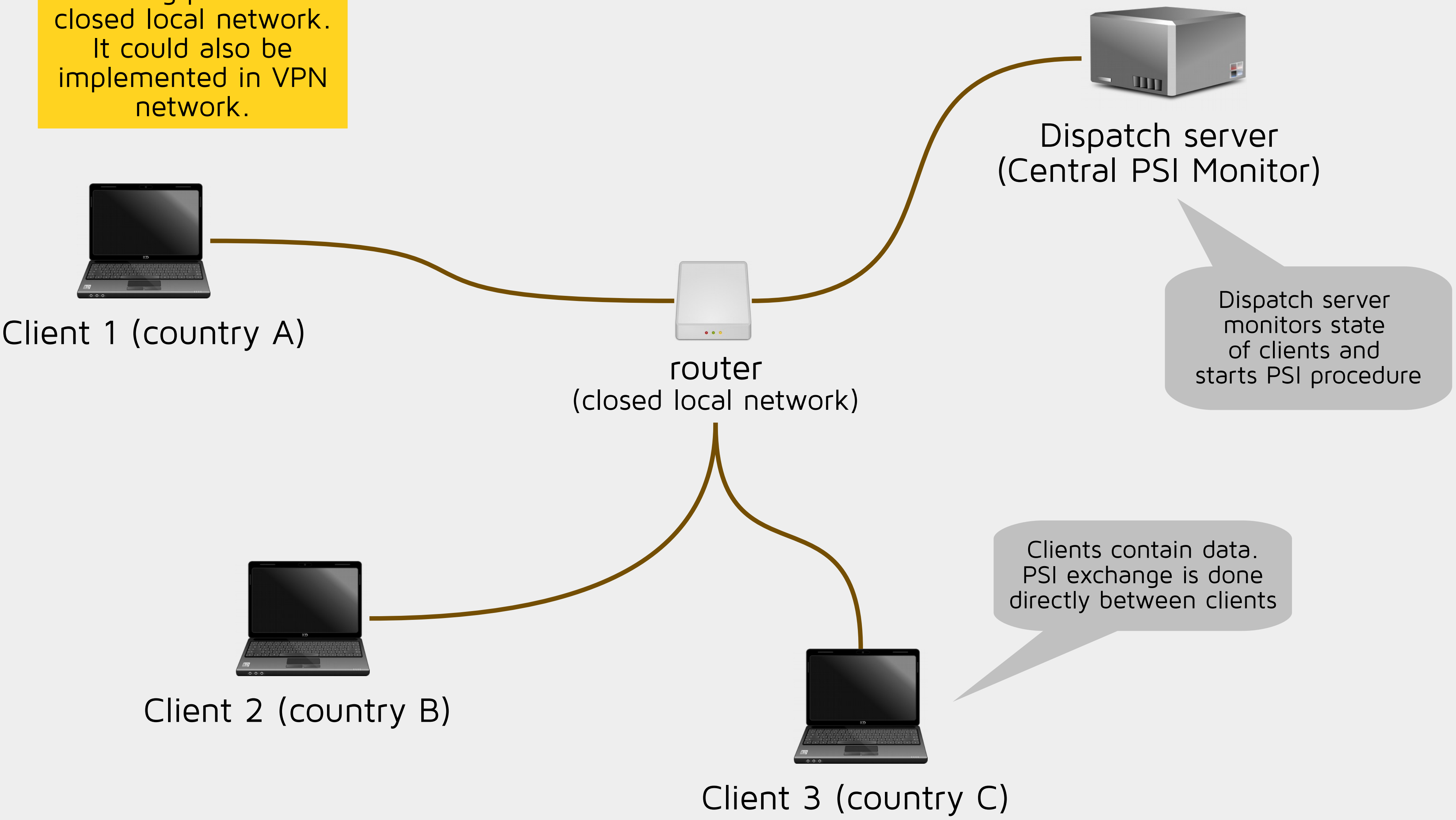
Name1	Name2	Name3	Name4	Name5	BirthDay	BirthMonth	BirthYear
Sarah	Connor				1	3	1965
Kyle	Reese				3	6	2010
Terminator	T-800				29	8	1997

Aquilla_EID.xlsx - Excel

ElectronicIdentifier
+38640123456
test@test.si
https://www.example.net/test

Hardware setup

PSI exchange is taking place in closed local network. It could also be implemented in VPN network.



Dispatch server
(Central PSI Monitor)

Dispatch server monitors state of clients and starts PSI procedure

Client 1 (country A)

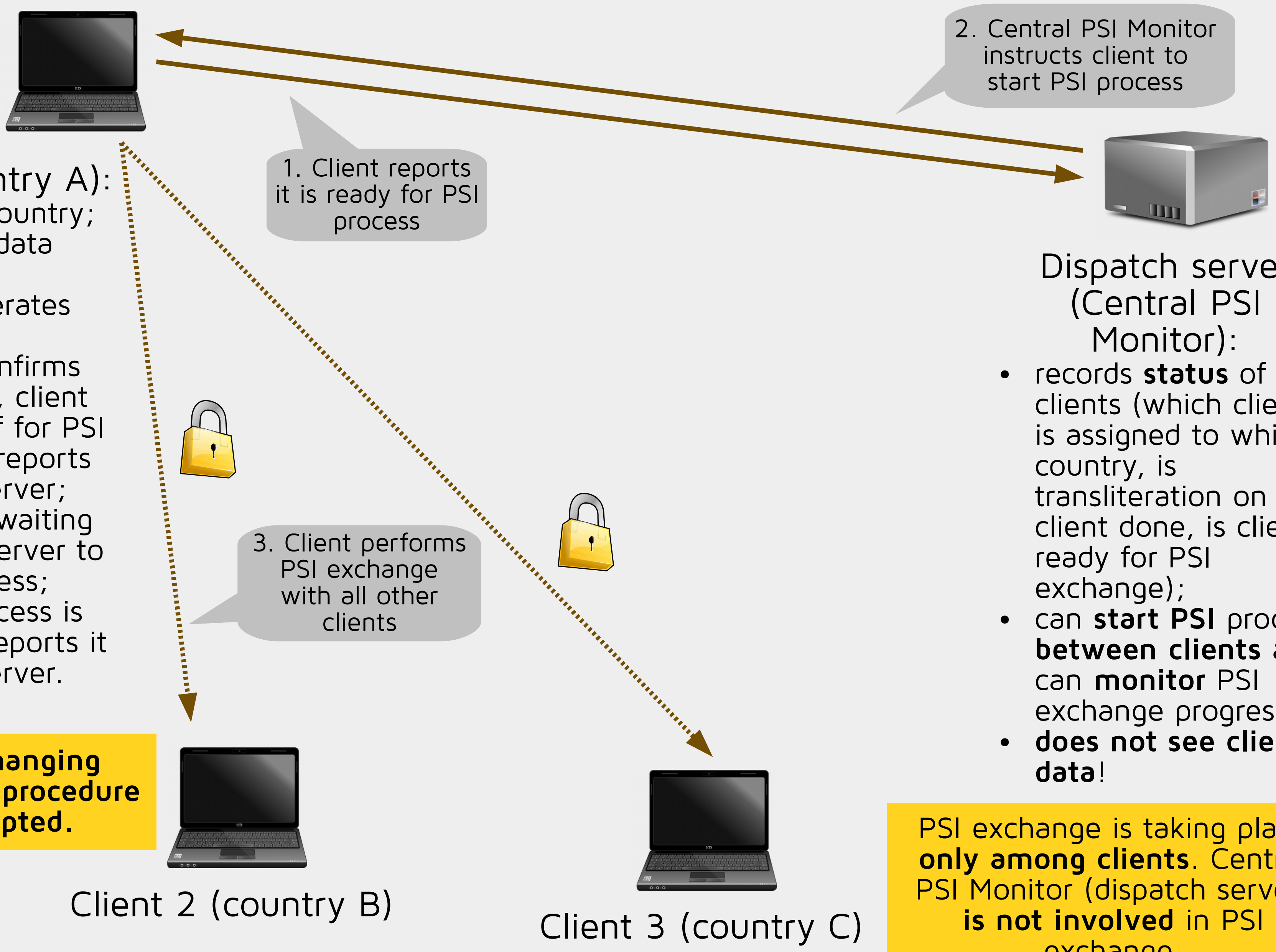
router
(closed local network)

Clients contain data. PSI exchange is done directly between clients

Client 2 (country B)

Client 3 (country C)

Data flow



Client 1 (country A):

- user selects country;
- user uploads data (XLSX/CSV);
- client transliterates data;
- when user confirms (s)he is ready, client prepares itself for PSI excange and reports to dispatch server;
- client is now waiting for dispatch server to start PSI process;
- when PSI process is done, client reports it to dispatch server.

1. Client reports it is ready for PSI process

2. Central PSI Monitor instructs client to start PSI process

3. Client performs PSI exchange with all other clients

Dispatch server (Central PSI Monitor):

- records **status** of clients (which client is assigned to which country, is transliteration on client done, is client ready for PSI exchange);
- can **start PSI** process **between clients** and can **monitor** PSI exchange progress;
- **does not see client data**!

Client 2 (country B)

Client 3 (country C)

Data exchanging through PSI procedure is encrypted.

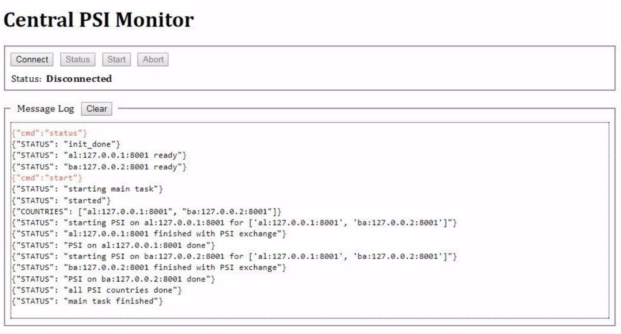
PSI exchange is taking place **only among clients**. Central PSI Monitor (dispatch server) **is not involved** in PSI exchange.

PSI application

Application consists of a three parts. First part is Central PSI Monitor, which monitors statuses of clients and starts PSI procedure.

Client application is divided into two parts. One is used to exchange the data about the persons and the other to exchange so called electronic identificator data (phone numbers, licence plates, e-mail addreses social network ID’s,...).

Each country can monitor its own status.



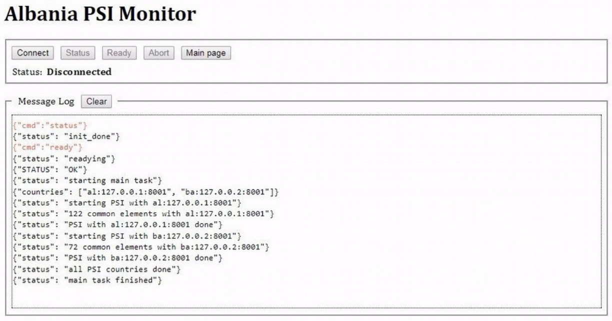
Central PSI Monitor

Status: **Disconnected**

Message Log

```
{"cmd":"status"}
{"STATUS": "init_done"}
{"STATUS": "al:127.0.0.1:8001 ready"}
{"STATUS": "ba:127.0.0.2:8001 ready"}
{"cmd":"start"}
{"STATUS": "starting main task"}
{"STATUS": "started"}
{"COUNTRIES": ["al:127.0.0.1:8001", "ba:127.0.0.2:8001"]}
{"STATUS": "starting PSI on al:127.0.0.1:8001 for ['al:127.0.0.1:8001', 'ba:127.0.0.2:8001']"}
{"STATUS": "al:127.0.0.1:8001 finished with PSI exchange"}
{"STATUS": "PSI on al:127.0.0.1:8001 done"}
{"STATUS": "starting PSI on ba:127.0.0.2:8001 for ['al:127.0.0.1:8001', 'ba:127.0.0.2:8001']"}
{"STATUS": "ba:127.0.0.2:8001 finished with PSI exchange"}
{"STATUS": "PSI on ba:127.0.0.2:8001 done"}
{"STATUS": "all PSI countries done"}
{"STATUS": "main task finished"}
```

Central PSI Monitor at dispatch server is monitoring country’s status and PSI progress.



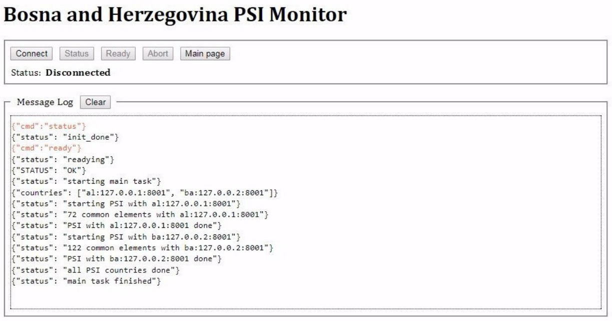
Albania PSI Monitor

Status: **Disconnected**

Message Log

```
{"cmd":"status"}
{"status": "init_done"}
{"cmd":"ready"}
{"status": "readying"}
{"STATUS": "OK"}
{"status": "starting main task"}
{"countries": ["al:127.0.0.1:8001", "ba:127.0.0.2:8001"]}
{"status": "starting PSI with al:127.0.0.1:8001"}
{"status": "122 common elements with al:127.0.0.1:8001"}
{"status": "PSI with al:127.0.0.1:8001 done"}
{"status": "starting PSI with ba:127.0.0.2:8001"}
{"status": "72 common elements with ba:127.0.0.2:8001"}
{"status": "PSI with ba:127.0.0.2:8001 done"}
{"status": "all PSI countries done"}
{"status": "main task finished"}
```

PSI exchange is taking place **directly between two countries** (peer-to-peer).



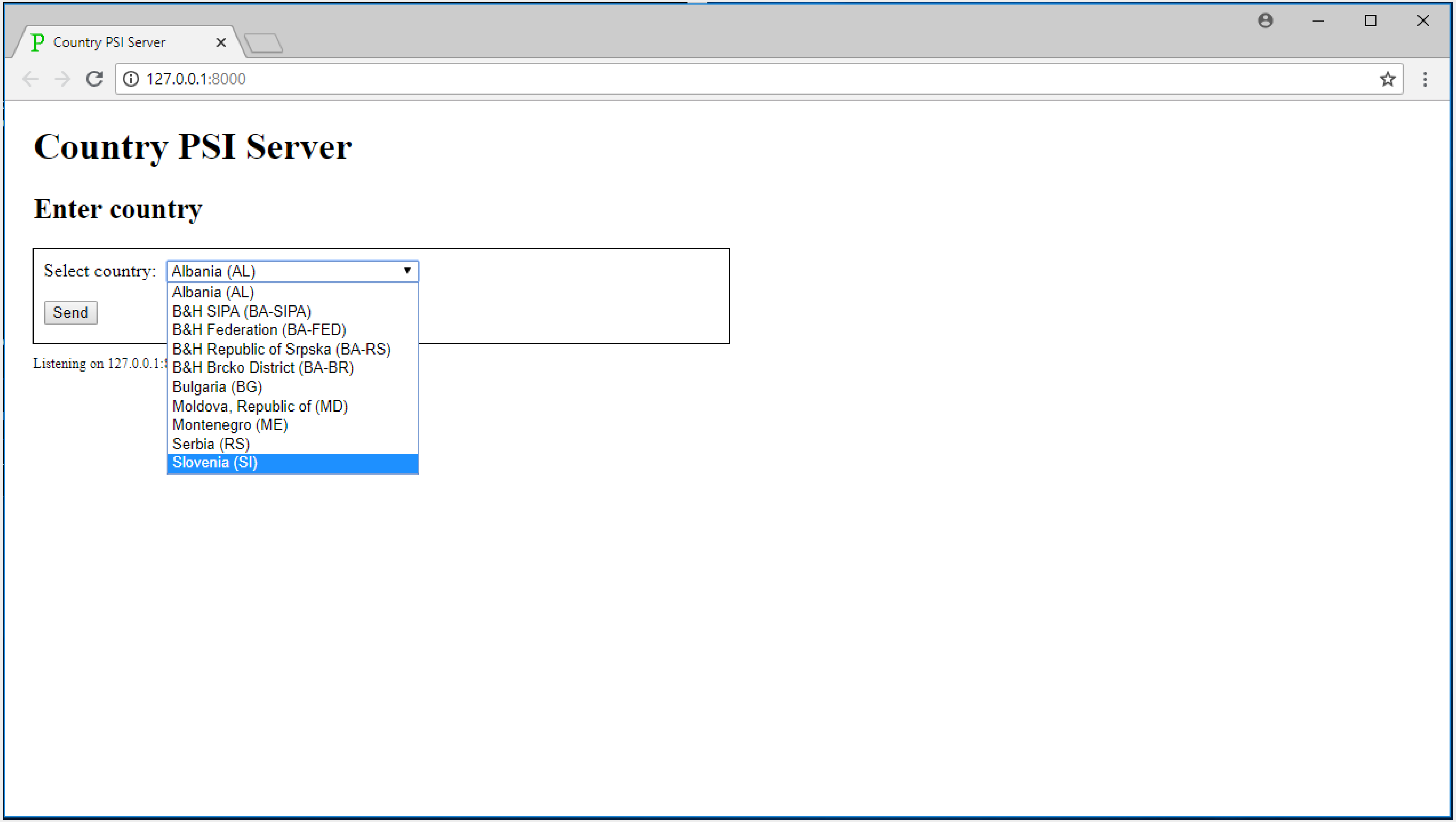
Bosna and Herzegovina PSI Monitor

Status: **Disconnected**

Message Log

```
{"cmd":"status"}
{"status": "init_done"}
{"cmd":"ready"}
{"status": "readying"}
{"STATUS": "OK"}
{"status": "starting main task"}
{"countries": ["al:127.0.0.1:8001", "ba:127.0.0.2:8001"]}
{"status": "starting PSI with al:127.0.0.1:8001"}
{"status": "72 common elements with al:127.0.0.1:8001"}
{"status": "PSI with al:127.0.0.1:8001 done"}
{"status": "starting PSI with ba:127.0.0.2:8001"}
{"status": "122 common elements with ba:127.0.0.2:8001"}
{"status": "PSI with ba:127.0.0.2:8001 done"}
{"status": "all PSI countries done"}
{"status": "main task finished"}
```

PSI application



Country PSI Server

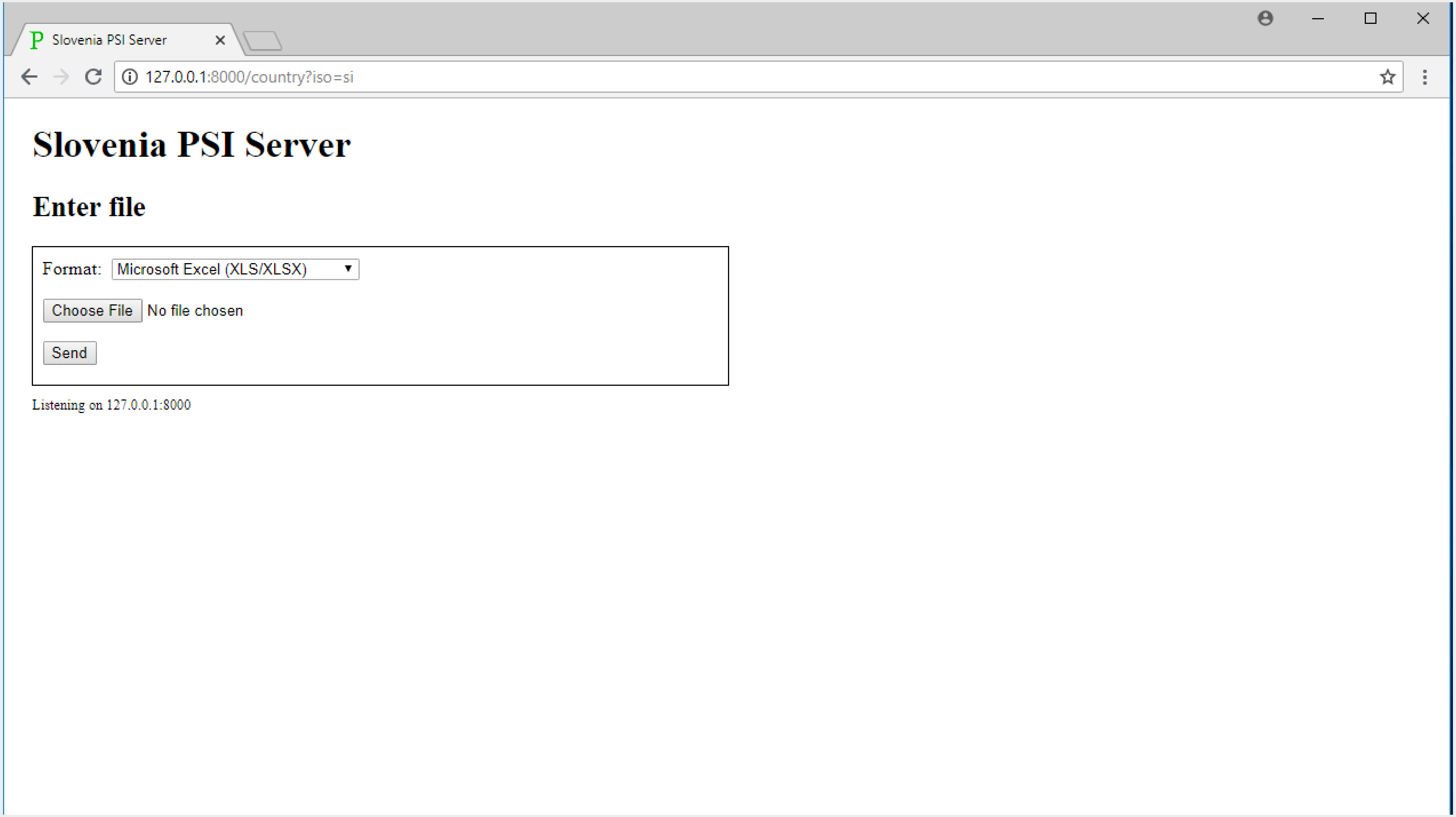
Enter country

Select country: Albania (AL)

Listening on 127.0.0.1:

Step 1: Select your country.

PSI application



Slovenia PSI Server

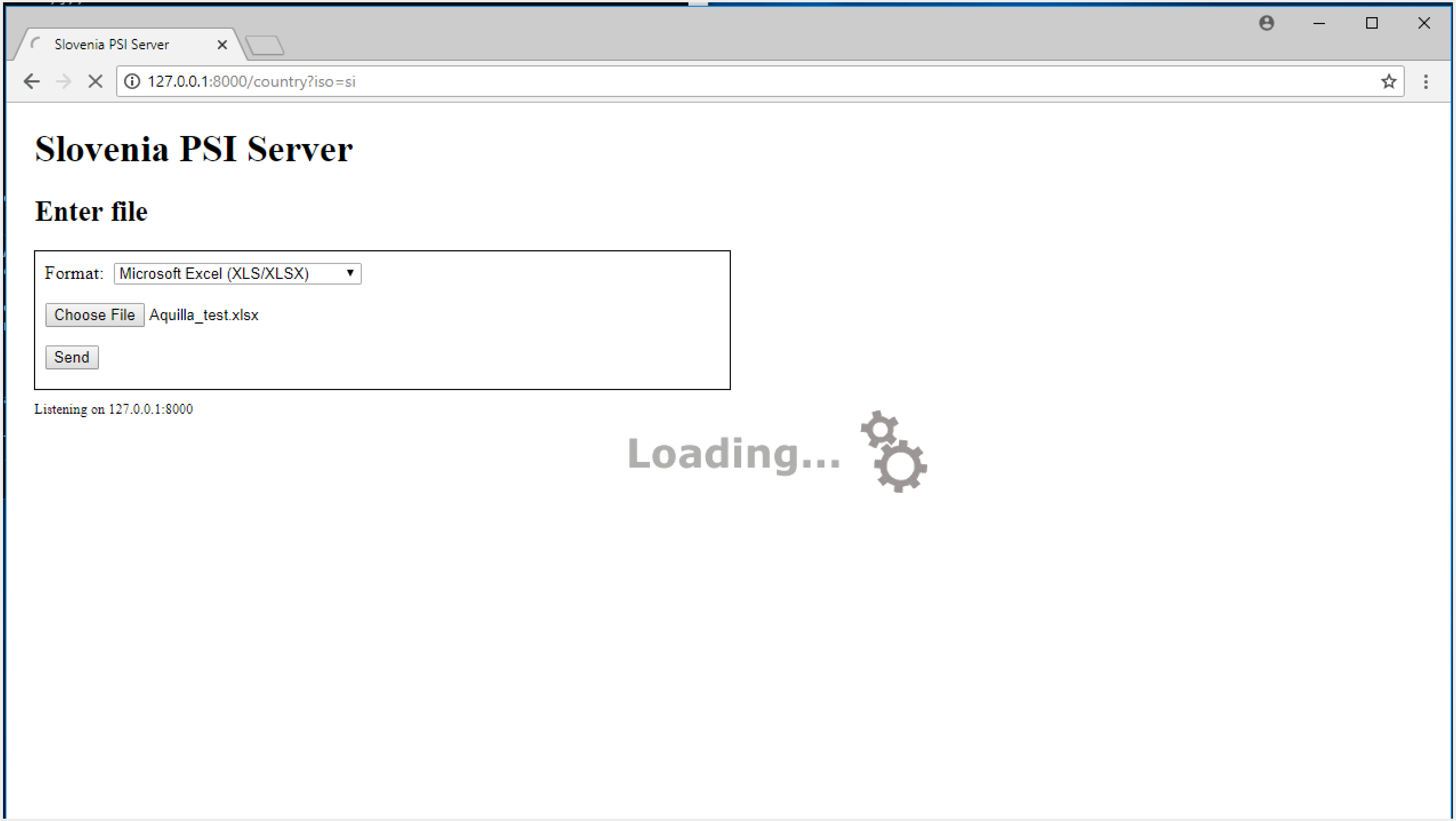
Enter file

Format: Microsoft Excel (XLS/XLSX)

Listening on 127.0.0.1:8000

Step 2: Select your file with data. You can upload Excel or CSV file.

PSI application



Slovenia PSI Server

127.0.0.1:8000/country?iso=si

Slovenia PSI Server

Enter file

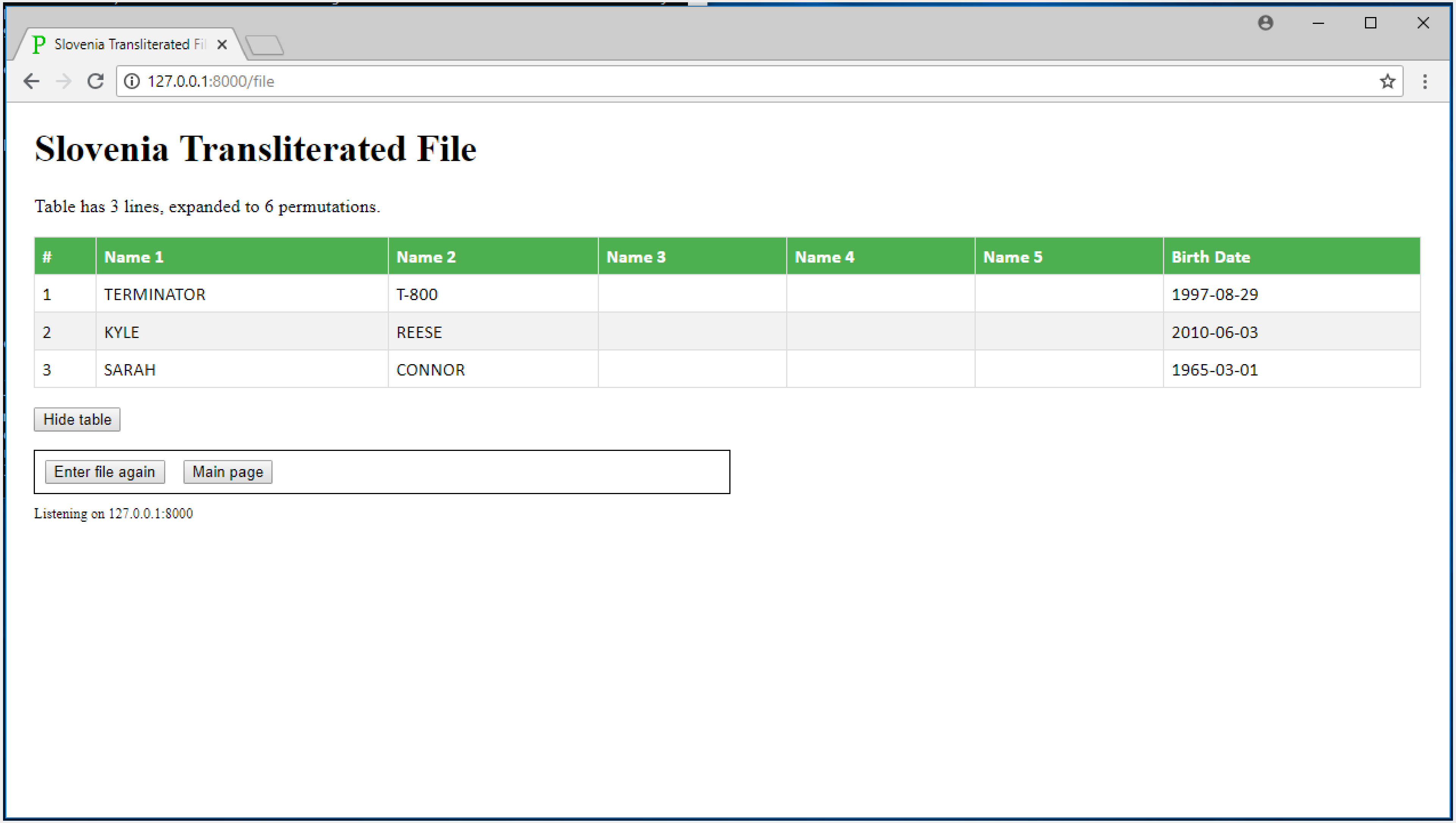
Format: Microsoft Excel (XLS/XLSX)

Aquilla_test.xlsx

Listening on 127.0.0.1:8000

Step 3: File is uploaded to the application. Transliteration is done and permutations are computed.

PSI application



Slovenia Transliterated File

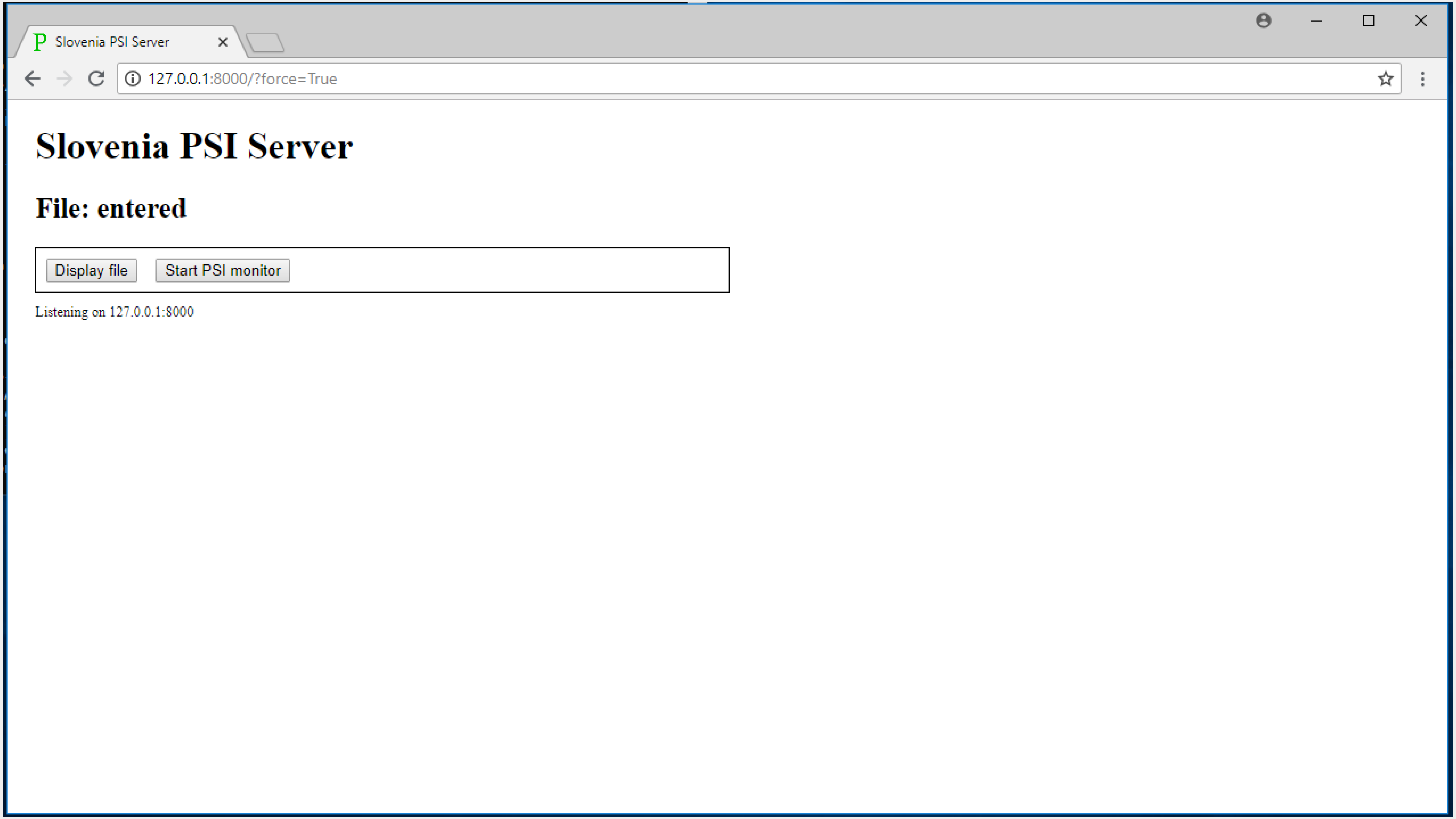
Table has 3 lines, expanded to 6 permutations.

#	Name 1	Name 2	Name 3	Name 4	Name 5	Birth Date
1	TERMINATOR	T-800				1997-08-29
2	KYLE	REESE				2010-06-03
3	SARAH	CONNOR				1965-03-01

Listening on 127.0.0.1:8000

Step 4: You can rewiev the data.

PSI application



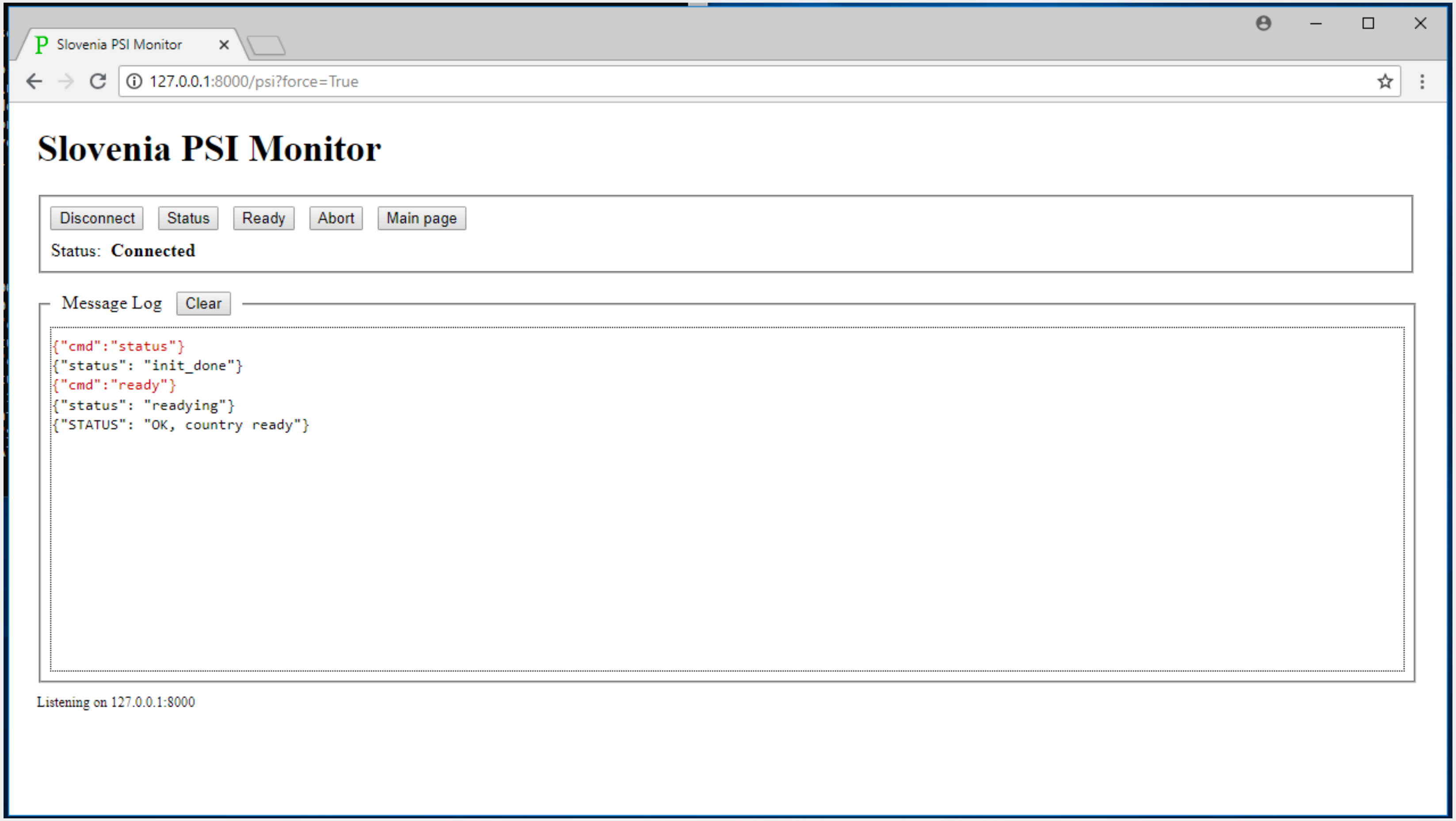
Slovenia PSI Server

File: entered

Listening on 127.0.0.1:8000

Step 5: Click to “Main page” and then “Start PSI monitor”.

PSI application



Slovenia PSI Monitor

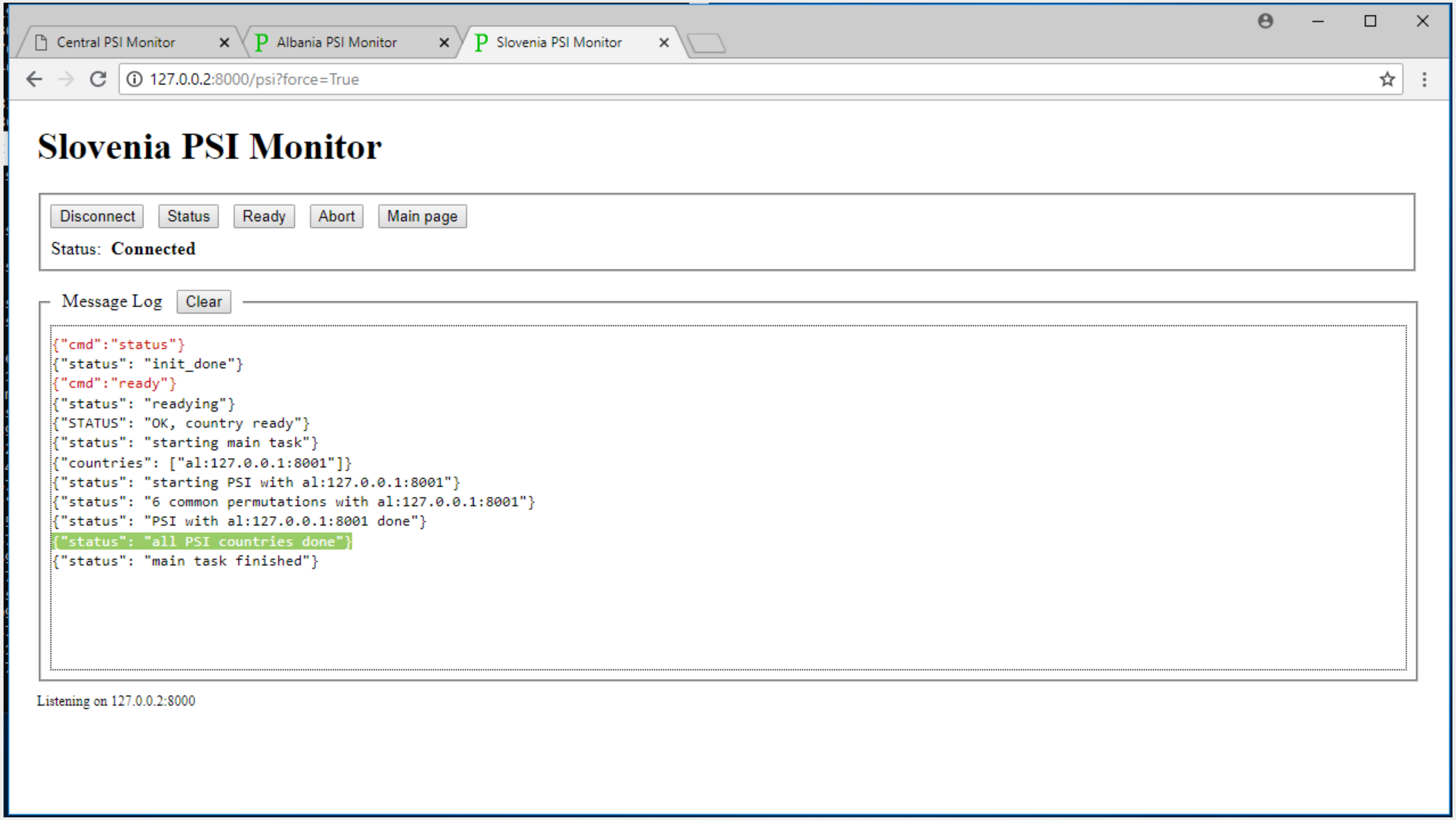
Status: **Connected**

```
{"cmd":"status"}
{"status": "init_done"}
{"cmd":"ready"}
{"status": "readying"}
{"STATUS": "OK, country ready"}
```

Listening on 127.0.0.1:8000

Step 6: Now application generates RSA keys. After that, you can click to “Connect”. Your client will be connected to Central PSI Monitor. Then you can click “Ready” button. This will inform the Central PSI Monitor that your country is ready for PSI exchange (this will be indicated by message: {"STATUS": "OK, country ready"}).

PSI application



Slovenia PSI Monitor

Status: **Connected**

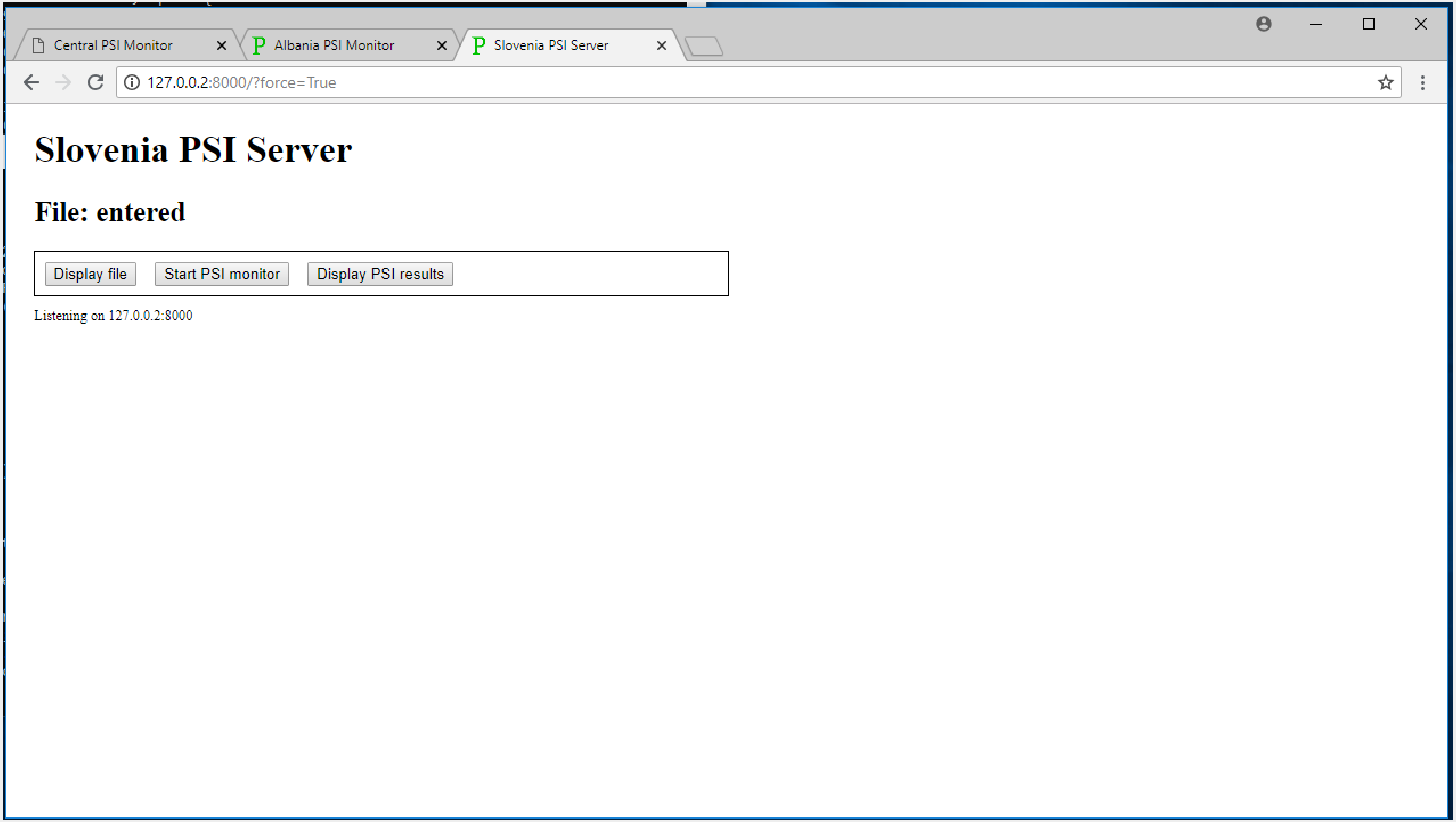
Message Log

```
{"cmd":"status"}
{"status": "init_done"}
{"cmd":"ready"}
{"status": "readying"}
{"STATUS": "OK, country ready"}
{"status": "starting main task"}
{"countries": ["al:127.0.0.1:8001"]}
{"status": "starting PSI with al:127.0.0.1:8001"}
{"status": "6 common permutations with al:127.0.0.1:8001"}
{"status": "PSI with al:127.0.0.1:8001 done"}
{"status": "all PSI countries done"}
{"status": "main task finished"}
```

Listening on 127.0.0.2:8000

Step 7: When all clients are ready, operator of the Central PSI Monitor can start PSI exchange among countries. Central PSI Monitor will then indicate which countries are doing PSI exchange and when the procedure is finished.

PSI application



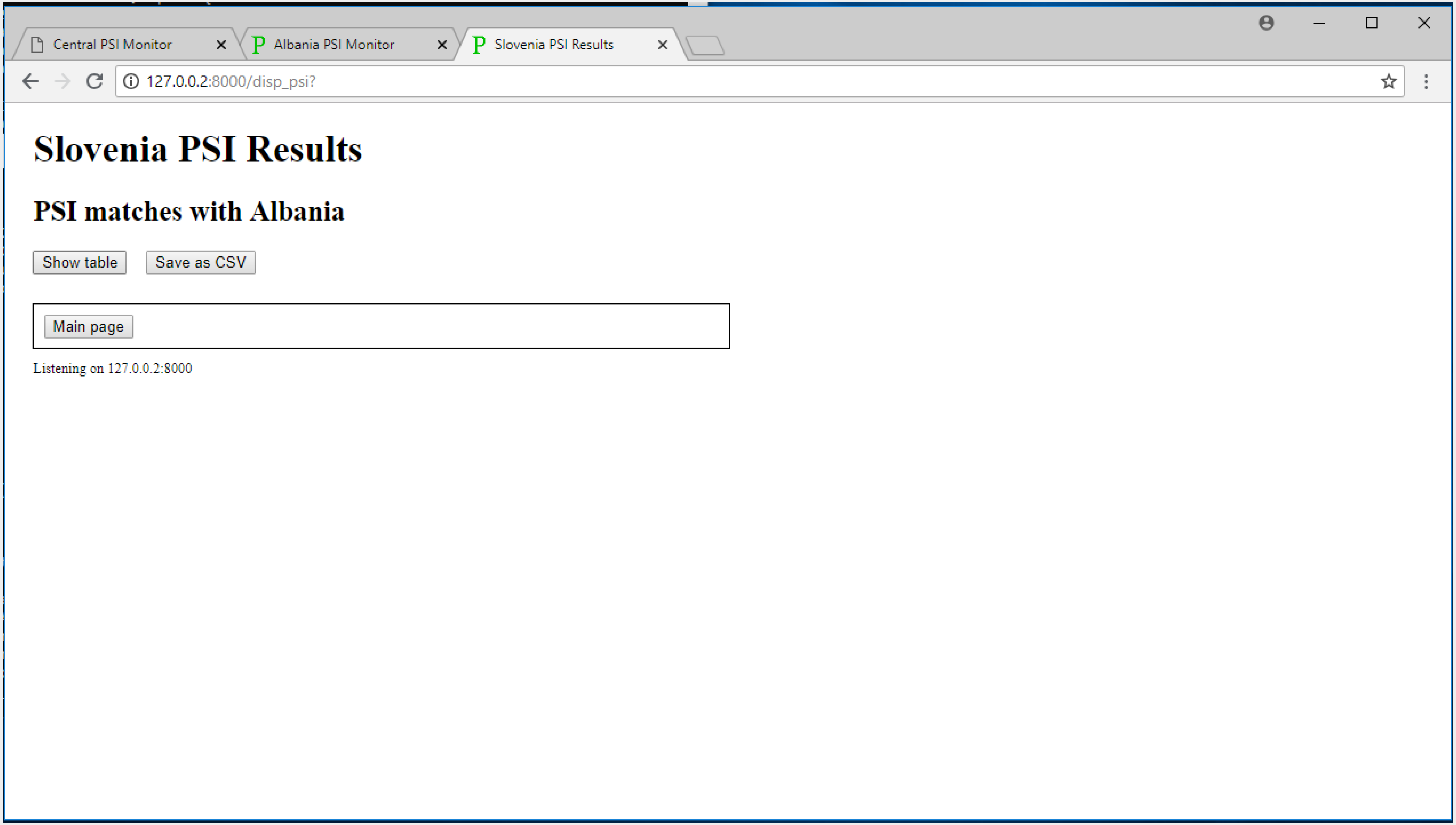
Slovenia PSI Server

File: entered

Listening on 127.0.0.2:8000

Step 8: Now you can return to main screen and check PSI results.

PSI application



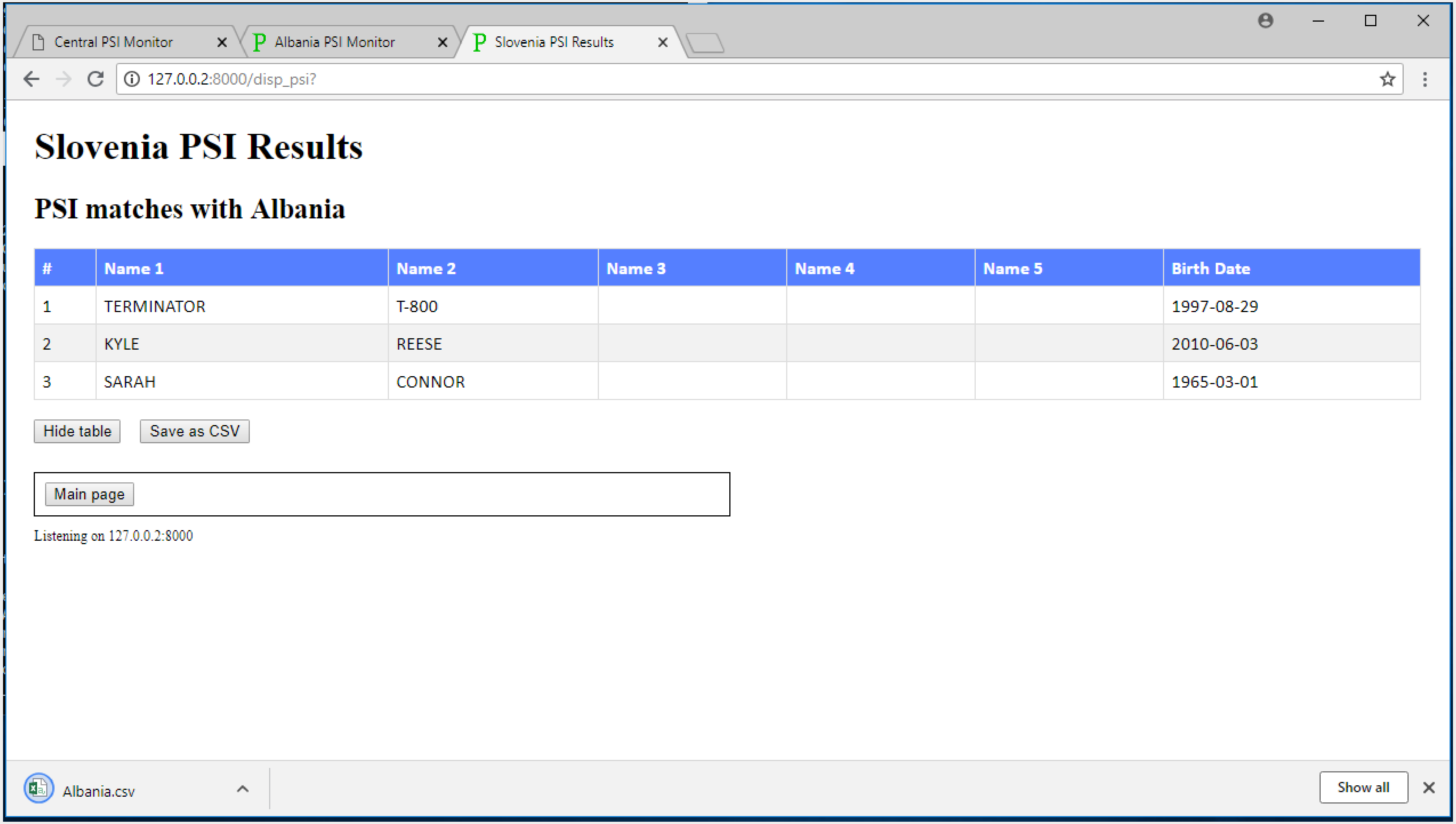
Slovenia PSI Results

PSI matches with Albania

Listening on 127.0.0.2:8000

Step 9: You can see with which clients (countries) you have common results.

PSI application



Slovenia PSI Results

PSI matches with Albania

#	Name 1	Name 2	Name 3	Name 4	Name 5	Birth Date
1	TERMINATOR	T-800				1997-08-29
2	KYLE	REESE				2010-06-03
3	SARAH	CONNOR				1965-03-01

Listening on 127.0.0.2:8000

Step 10: You can review the results (common records in both databases) and save the results in CSV file.

Discussion

matej.kovacic@telefoncek.si
matjaz@eunet.si