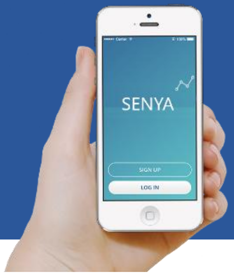
# Intelligent Data Discovery & Classification



the promise of understood Customer data…

## Getting a handle on Personal Data!

You should document what Personal Data you hold, where it came from and who you share it with. Many reasons why, the essence of your Digital efforts, Customer 360 and now GDPR. GDPR requires you to maintain records of your processing activities. For example, if you have inaccurate personal data and have shared this with another organisation, you will have to tell the other organisation about the inaccuracy so it can correct its own records. You won't be able to do this **unless you know what personal data you hold,** where it came from and who you share it with.

The SENYA Intelligent Data Discovery & Classification Template provides you with a free Template already containing the base classification structures to identify and understand the Personal Data your organization holds.

**PII (Personally Identifiable Information)** is a term mostly used in the US with Personal Data the **European equivalent**, albeit different, for purposes of this document they are used interchangeably. As we speak the line between them are blurring.

# What is Personal Data?

The National Institute of Standards and Technology (NIST) provides the following definition of PII:

PII is **any information** about an **individual** maintained by an agency, including (1) any information that can be **used to distinguish or trace** an **individual's identity**, such as **name**, **social security number,** date and place of birth, mother's maiden name, or biometric records; and (2) any **other information** that is **linked or linkable** to an **individual**, such as medical, educational, financial, and employment information.

Simplistically PII can be divided into two categories: linked information and linkable information.

**Linked information** is any piece of **personal information** that can be used to **identify an individual** and includes amongst others:

a) Full name
b) Home address
c) Email address
d) Social security number
e) Passport number
f) Driver's license number
g) Credit card numbers
h) Date of birth

**Linkable information**, on the other hand, is information that **on its own** may **not be able to identify a person**, but when **combined** with another piece of information **could identify**, trace, or locate a person, some examples:

a) Last Name
b) City, Post Code
c) Work Place

# Nub of the challenge, where is the PII data in my Organization?

Most people will say everywhere, some will say nowhere but we think we could safely conclude **somewhere.** So **how do we find it?** Many Vendors proclaim with automated software which by magic will scan your systems and give you a report. Some even go further and declare once identified will protect that data so that you and your Customers can sleep easily.

Nothing in our considered **opinion** is further from the truth. Yes, there are automated systems, yes, they can scan pockets of your Enterprise, yes, they can find certain elements of PII and yes, some specific elements can be protected. It is all the bits between the combined Yesses's that is the true conundrum. All Data and systems are not created equal. For example, show me where one of those solutions can scan the transactional 24x7 Banking systems for PII data. What however is possible is downstream sets of isolated data to be examined in a way similar to Virus scanning, again provided it met certain conditions.

## Our view; use the smarts of the Organization

Most organizations will quickly through a collection of SME tell you the major coarse grain Sources of Information. Typically, they associate them as Systems, Databases, FileStores, Employee PC's, Email Repositories etc.

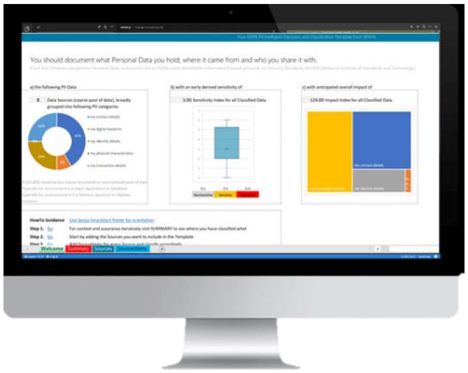
Once known this list or **SuperList** becomes the target for **finding personal data**. Preferably the SuperList is arranged into Business/Technical Domains and prioritized to maximize bang-for-buck.

For each **Source** a **method**/**tool** needs to be found to **determine** the internal **structure**/**schema** or in the case that there is absolutely no-structure(RAW) a way to scan the actual content. Many Sources will use the same method/tool so that one could get by with typically 3-5 discrete elementary detection tools covering most if not all proprietary Sources. These Tools ranged from expensive Commercial to highly sophisticated free Open Source and some decisions will be required to cater for what you are comfortable with. There is even some Tools to scan your entire Network to find Sources to ensure your Superlist is complete.

In this process SENYA offers the free **Capiible Template** which will allow you to easily pick-up these structures from above mentioned Elementary Scanners and **classify** them with **pre-build PII standards**. Lastly by arranging the Templates into a Document library like Sharepoint one quickly can orchestrate and record an Organization wide PII effort.

## What is in the Template?

1) **DASHBOARD View**, showing **an early assessment** (effectively whilst you classify) on what the sensitivity and impact is of the discovered PII Data. Philosophy is to empower the Classification teams for early decsion making instead of waiting for Reports, traditionally weeks/months after the exercise.



You should document what Personal Data you hold, where it came from and who you share it with.

a) the following PII Data

b) with an early derived sensitivity of

c) with anticipated overall impact of

3 Data Sources (coarse pool of data), broadly grouped into following PII categories

3.00 Sensitivity Index for all Classified Data

124.00 Impact Index for all Classified Data

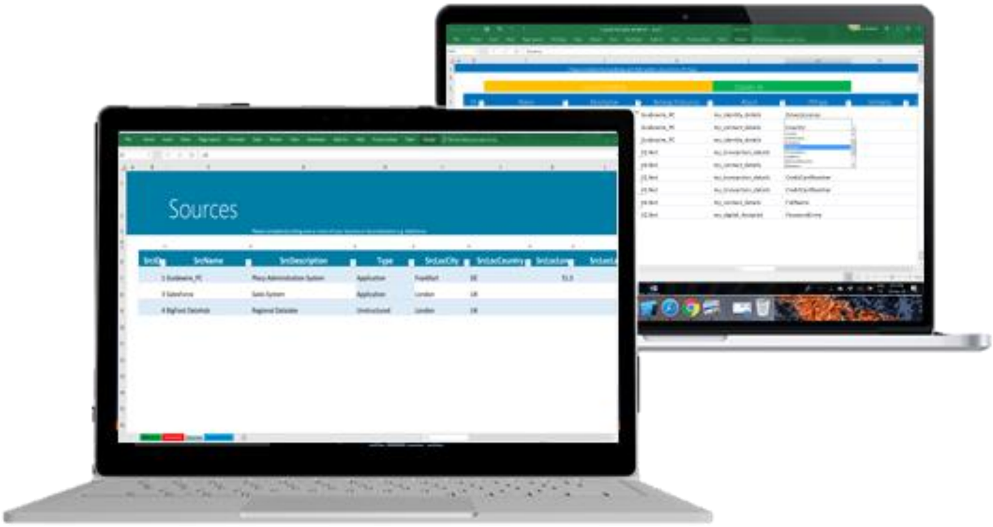
HowTo Guidance Use Senya Smartstart Poster for orientation

Step 1. Go For context and assurance iteratively visit SUMMARY to see where you have classified what

Step 2. Go Start by adding the Sources you want to include in the Template

2) **SUMMARY View**, showing frequency of **what and where your PII information** is. If you go for an approach where you only fill in one Source per Template the scope of this view will be limited to a single Source. If, however, you decide to complete multiple Sources in a single temple the scope will vary accordingly. For more comprehensive Views we recommend using our platform in conjunction with any of your favorite BI tools e.g. PowerBI, Tableau, MicroStrategy etc.

3) **Discovering Classifying** is a simple process whereby you add Sources, and foreach paste Source Elements. Thereafter you classify them from a drop-down picklist.

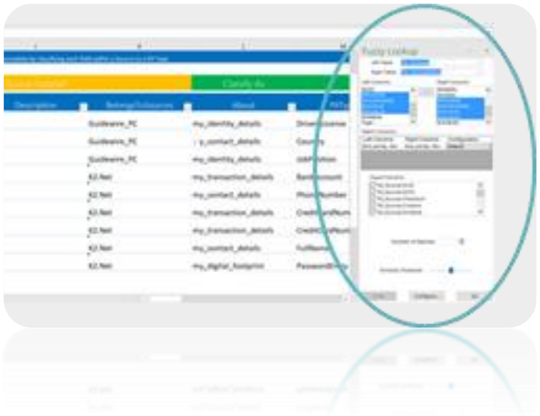


4) **FastStart Pack** containing some **guidance** on getting to the essence of your pII discovery strategy and using the Template.

Capible Template
Faststart Orientation
1
It is about (Personal) Data!!!
2
...dealing with multiple Sources of Structured & Unstructured Data
Your Company
Personal Data ???
3
It is not rocket science, start by building a strategy around seeing the overall pic and your current intent...
4
...get your ability to locate, detect and scan these Sources operational
b
...use the Capiible Template to arrange and smart Classify
c
7
...close the loop
d
8
...iteratively refine and cycle thru the prioritized list
a
d
9

5) **PowerQuery** to easily and **rapidly interrogate Sources yourself**. Amongst others caters for most structured formats like Databases, XML/JSON and various file formats. It also easily deals with all Office documents considered to be by many as unstructured. For true unstructured sources you will have to rely on one or more of the Elementary Scanners discussed earlier.

## Go from data to this-is personal-data in minutes – Fuzzy Logic



When you have 3 or more Sources to classify, applying automated cognitive technology to recognize personal data makes a lot of sense.  It will help you get there quicker, cheaper and typically with more precision.

Fuzzy Logic uses Jaccard similarity, loosely defined as the size of the set intersection divided by the size of the set union for two sets of objects. This is used to determine the degree of similarity between a data element and the target PII corpus.  Contrast this to most of the scanning tools that uses pattern recognition in the form of Regex expressions to find PII data in the actual data.

Senya is using Fuzzy Logic as a cognitive service inside or outside of Machine Learning.  It is not included as part of the free Template but is available as-a-service.  More info

## More questions about the SENYA Capiible solution

| | TEMPLATE | PLATFORM | SERVICE |
|---|---|---|---|
| Arrangement | You get the Template and you do all the discovery and classification yourself | In-Cloud platform that you subscribe to and use. | Senya provide turn-key service with negotiated Service Level Agreement and accompanying custom services |
| Cost | FREE | Monthly subscription or price per Domain | Flexible; to-be-negotiated |
| Usage | Unlimited use within your Organization bounded to person requesting Template | Limited to agreed Domain/s.  GDPR tag all data for agreed Domain/s | Limited to agreed Domain/s.  GDPR tag all data for agreed Domain/s |
| Term | Unlimited | Month by Month | Annual Subscription |
| GDPR | PII Standard List | PII Standard List | PII, SHI, SPI, Other |
| RecordKeeping | ○ | ● | ● |
| Fuzzy Logic/ML | ○ | ● | ● |
| FastStart Pack included | ● | ● | ● |
| Access Devices | Phone, Tablet, PC | Phone, Tablet, PC | Phone, Tablet, PC |

## Get your copy now!

Get your free copy of the Template now.  As always input to improve or your experiences in using it, would be most appreciated.

**Contact** |*info@senya.co.uk* | *www.senya.co.uk*