Dynamic anonymous index for confidential data

Guillermo Navarro-Arribas[1] Daniel Abril[2] Vicenç Torra[2]

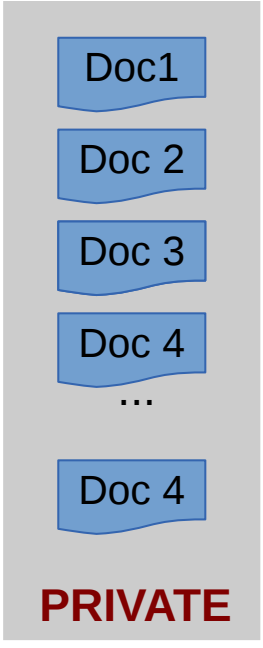
[1]Dep. Enginyeria de la Informació i de les Comunicacions (DEIC),
Universitat Autònoma de Barcelona (UAB).

[2]Institut d'Investigació en Intelligència Artificial (IIIA),
Consejo Superior de Investigaciones Científicas (CSIC).

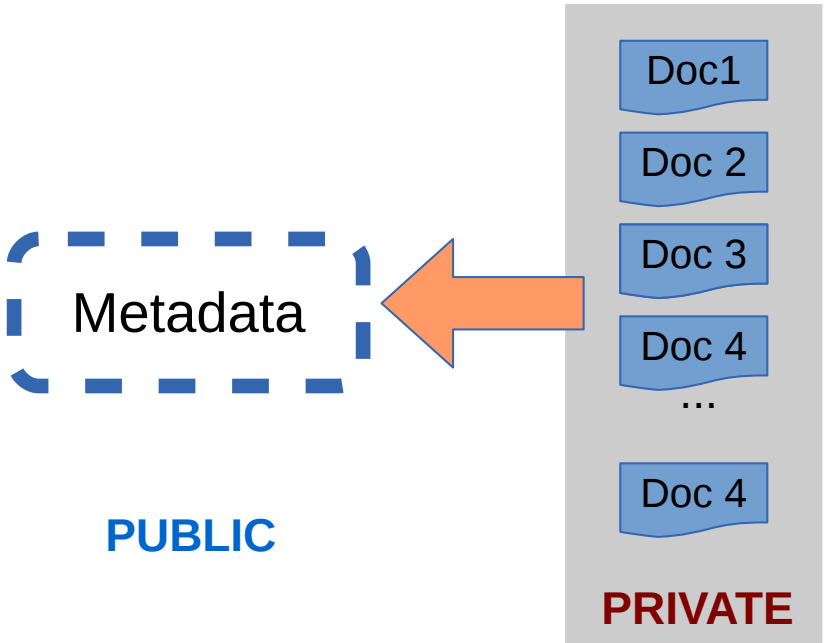
September 12, 2013

Scenario

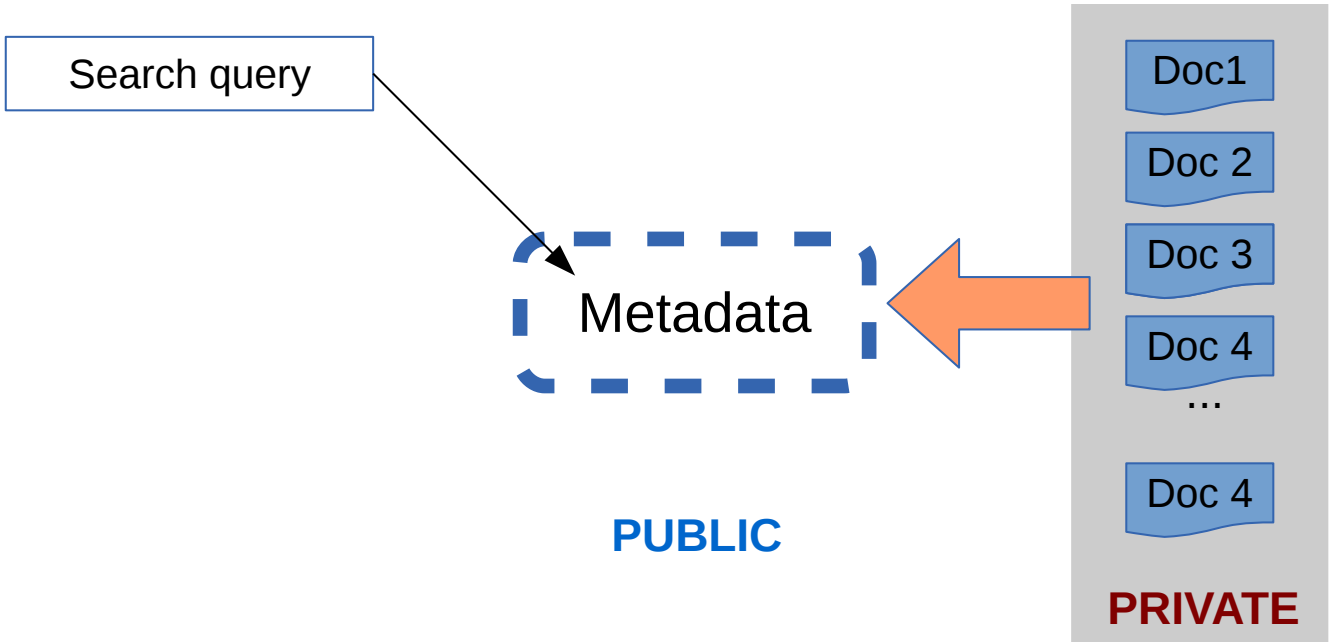
Scenario



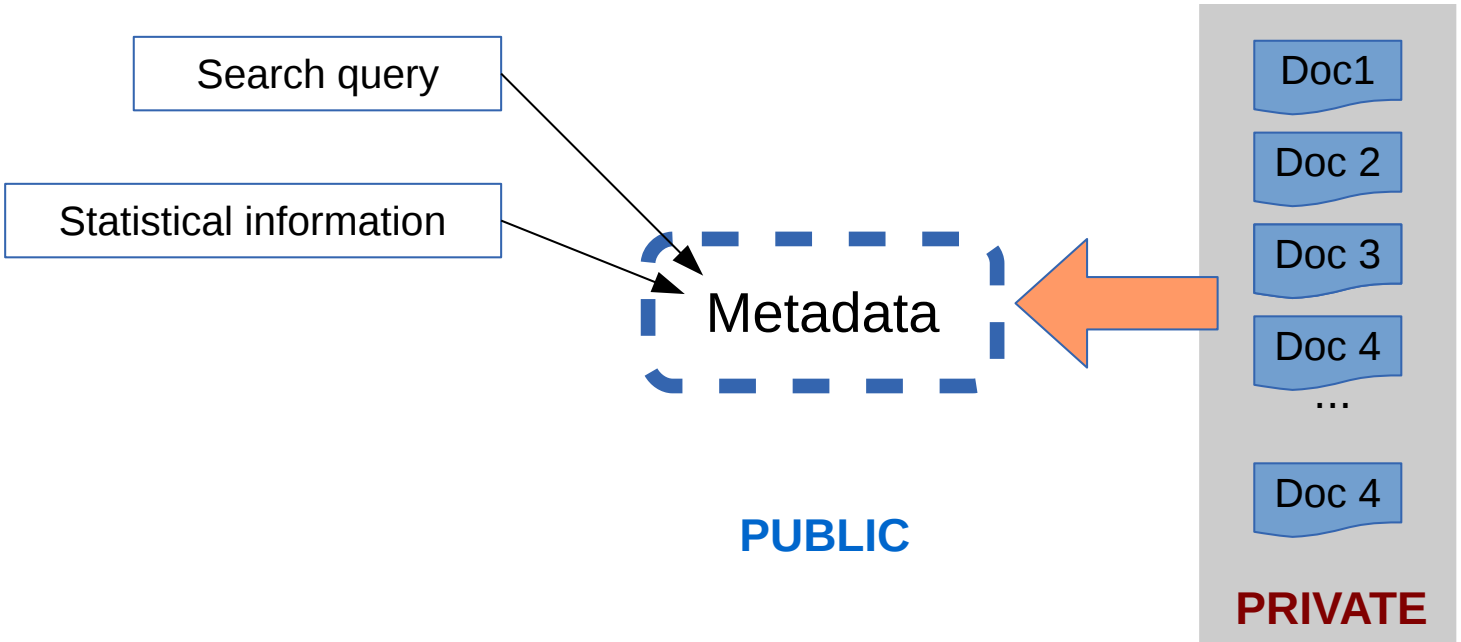
Scenario



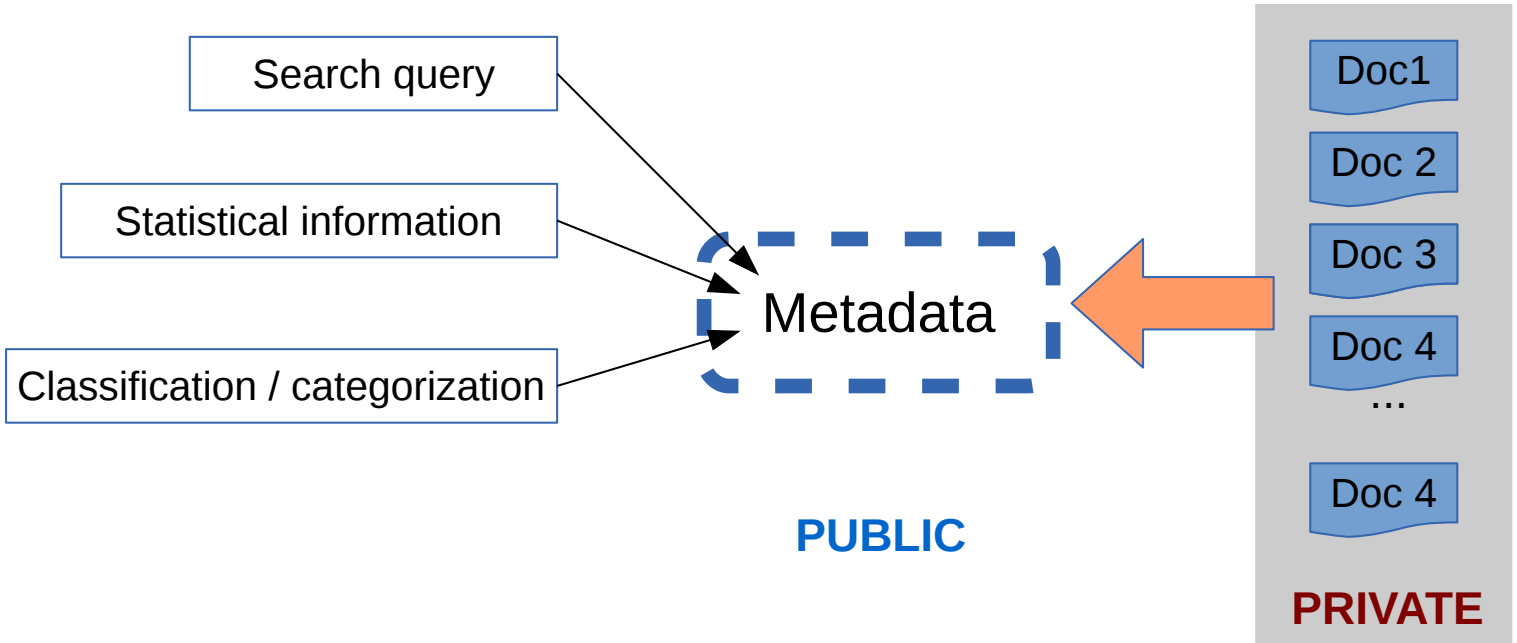
Scenario


Search query
Metadata
Doc1
Doc 2
Doc 3
Doc 4
...
Doc 4
PUBLIC
PRIVATE

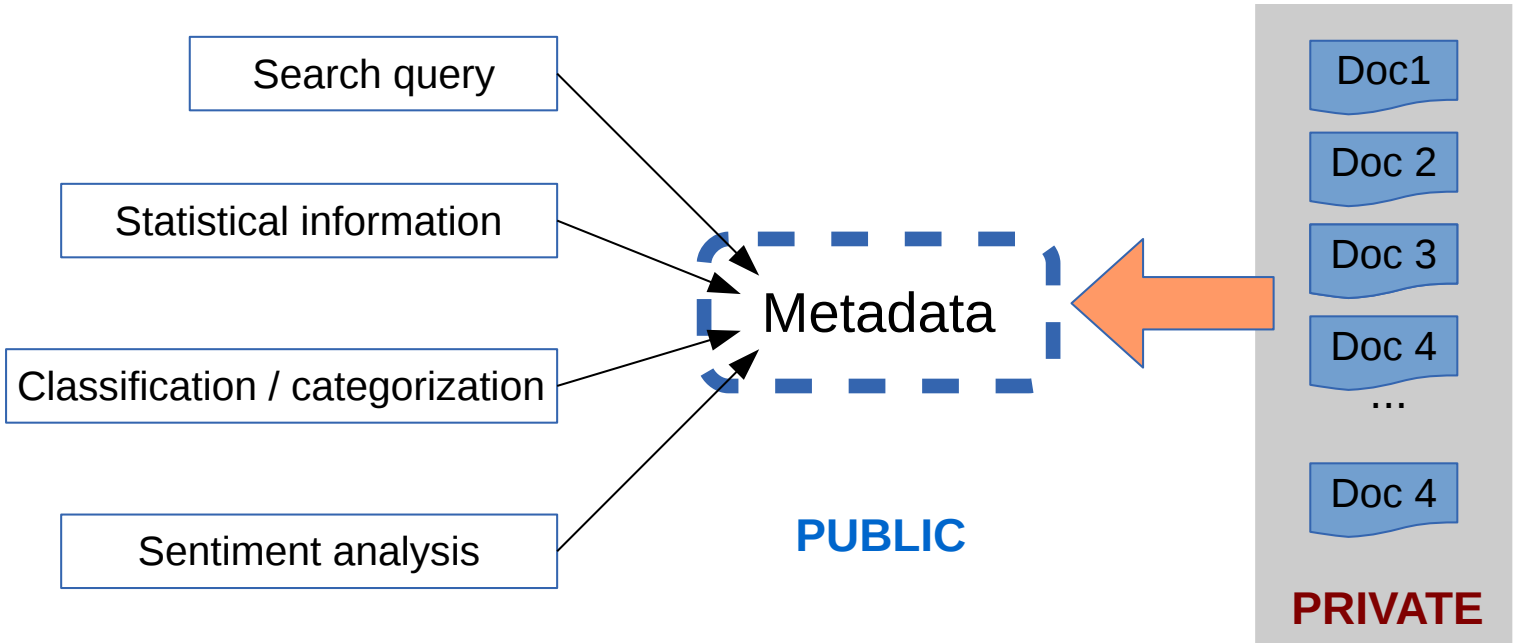
Scenario


Search query
Statistical information
Metadata
Doc1
Doc 2
Doc 3
Doc 4
...
Doc 4
PUBLIC
PRIVATE

Scenario


Search query
Statistical information
Classification / categorization
Metadata
Doc1
Doc 2
Doc 3
Doc 4
...
Doc 4
PUBLIC
PRIVATE

Scenario


Search query
Statistical information
Classification / categorization
Sentiment analysis
Metadata
PUBLIC
Doc1
Doc 2
Doc 3
Doc 4
...
Doc 4
PRIVATE

Metadata $\Rightarrow$ VSM

- **Vector Space Model (VSM)** as metadata.

	$term_1$	$term_2$	...	$term_M$
d_1	$w_{1,1}$	$w_{2,1}$	...	$w_{M,1}$
d_2	$w_{1,2}$	$w_{2,2}$	...	$w_{M,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_n	$w_{1,n}$	$w_{2,n}$	...	$w_{M,n}$

- w is a frequency based weight (term frequency, tf-idf, ...)

$\rightarrow$ VSM is treated as a microdata file for anonymization.

Protection of VSM

- We apply custom **microaggregation** to document vectors:

Microaggregation: commonly used in Statistical Disclosure Control

- **Partition**: on clusters of size $\geq k$.
- **Aggregation**: each vector is replaced by its centroid.

- Privacy is provided per document! (document anonymity)
- There are (at least) k **indistinguishable documents**.

Example of microaggregated VSM

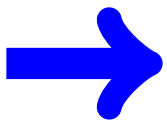
A (very) naive example:

Doc	"bullrun"
d_1	10
d_2	0
d_3	5
d_4	30
d_5	40

Example of microaggregated VSM

A (very) naive example:

Doc	"bullrun"
d_1	10
d_2	0
d_3	5
d_4	30
d_5	40



Doc	"bullrun"
d_1	5
d_2	5
d_3	5
d_4	35
d_5	35

This is a 2-anonymous VSM!

Dynamic VSM $\rightarrow$ inference problem

- Different protections of the same data:

Doc	t
d_1	1.0
d_2	2.0
d_3	3.0
d_4	4.0
d_5	5.0

(a) Original T

Doc	t
d_1	2.0
d_2	2.0
d_3	2.0
d_4	4.5
d_5	4.5

(b) T_1

Doc	t
d_1	1.5
d_2	1.5
d_3	4.0
d_4	4.0
d_5	4.0

(c) T_2

Dynamic VSM $\rightarrow$ inference problem

- Different protections of the same data:

Doc	t
d_1	1.0
d_2	2.0
d_3	3.0
d_4	4.0
d_5	5.0

(d) Original T

Doc	t
d_1	2.0
d_2	2.0
d_3	2.0
d_4	4.5
d_5	4.5

(e) T_1

Doc	t
d_1	1.5
d_2	1.5
d_3	4.0
d_4	4.0
d_5	4.0

(f) T_2

- Inference of d_3 by intersection of clusters.
- Deletions and insertions in protected data can lead to the same problem!

Dynamic VSM: Deletion and insertion

A very conservative approach:

- Initial static microaggregation of the VSM, then
- **Insertion**: find closest cluster and add the element to it.
- **Deletion**: delete element, if cluster is $< k$ add remaining elements to closest cluster.

the centroid is never re-computed

Example of deletion and insertion

1. Original and initial protection with $k = 2$.

Doc	t
d_1	1.0
d_2	2.0
d_3	4.0
d_4	5.0
d_5	8.0
d_5	9.0

Doc	t
d_1	1.5
d_2	1.5
d_3	4.5
d_4	4.5
d_5	8.5
d_5	8.5

Example of deletion and insertion

1. Original and initial protection with $k = 2$.
2. Add document $d_6, t = 10$.

Doc	t
d_1	1.0
d_2	2.0
d_3	4.0
d_4	5.0
d_5	8.0
d_5	9.0

Doc	t
d_1	1.5
d_2	1.5
d_3	4.5
d_4	4.5
d_5	8.5
d_5	8.5

Doc	t
d_1	1.5
d_2	1.5
d_3	4.5
d_4	4.5
d_5	8.5
d_5	8.5
d_6	8.5

Example of deletion and insertion

1. Original and initial protection with $k = 2$.
2. Add document $d_6, t = 10$.
3. Delete document d_3.

Doc	t
d_1	1.0
d_2	2.0
d_3	4.0
d_4	5.0
d_5	8.0
d_5	9.0

Doc	t
d_1	1.5
d_2	1.5
d_3	4.5
d_4	4.5
d_5	8.5
d_5	8.5

Doc	t
d_1	1.5
d_2	1.5
d_3	4.5
d_4	4.5
d_5	8.5
d_5	8.5
d_6	8.5

Doc	t
d_1	1.5
d_2	1.5
d_4	1.5
d_5	8.5
d_5	8.5

Summary

- Addition and deletion are very conservative.
- Easy and fast, but
- can lead to high information loss
  - OK for scenario with big set on initial anonymization, low k, and few additions/deletions.
- Does not have inference problems (no formal proof).

Thanks

Dynamic anonymous index for confidential data

Guillermo Navarro-Arribas[1] Daniel Abril[2] Vicenç Torra[2]

[1]Dep. Enginyeria de la Informació i de les Comunicacions (DEIC),
Universitat Autònoma de Barcelona (UAB).

[2]Institut d'Investigació en Intelligència Artificial (IIIA),
Consejo Superior de Investigaciones Científicas (CSIC).

September 12, 2013