Workshop

USO DEGLI ARCHIVI DI DATI CORRENTI PER L'EPIDEMIOLOGIA: IL SISTEMA DI **BABELE**

Tecniche di record-linkage e privacy

Nicola Caranci

Agenzia Sanitaria e Sociale, Regione Emilia-Romagna

Gruppo di lavoro AIE "Sistemi informativi e Privacy"

Premessa: un esempio di collegamento tra i flussi sanitari

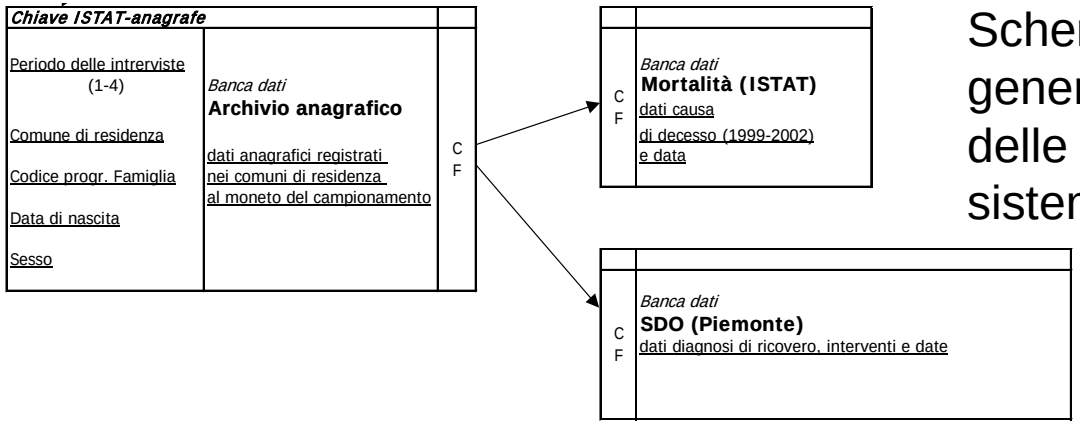
INTRODUZIONE DELL'IDENTIFICATIVO PERSONALE ANONIMO E REGOLAMENTO SULLA PRIVACY IN **EMILIA-ROMAGNA - SISEPS**

- Seguendo la L. 196/2003*, si è introdotto negli archivi contenenti dati sensibili un identificativo personale numerico anonimo (PROG_PAZ), in sostituzione dei dati anagrafici. E' un identificativo personale anonimo, comune a tutte le banche dati (NB: nei flussi SDO e Hospice il nuovo identificativo sostituisce quello precedente, introducendo un aumento dei ricoveri ripetuti valutato mediamente inferiore allo 0,5%)

- **Per coloro che possono accedere ai dati di dettaglio, è possibile ricostruire ed analizzare i percorsi assistenziali nel tempo, in tutto rispetto delle normative vigenti**

* Tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali

Collegamento dei dati sanitari con altri archivi analitici 1/2



Schema del *data base* relazionale generato dall'integrazione delle banche dati che compongono il sistema "campionario longitudinale"

Record linkage dei decessi tramite **CF**:

procedura di passi in successione con chiavi con potere discriminante decrescente:

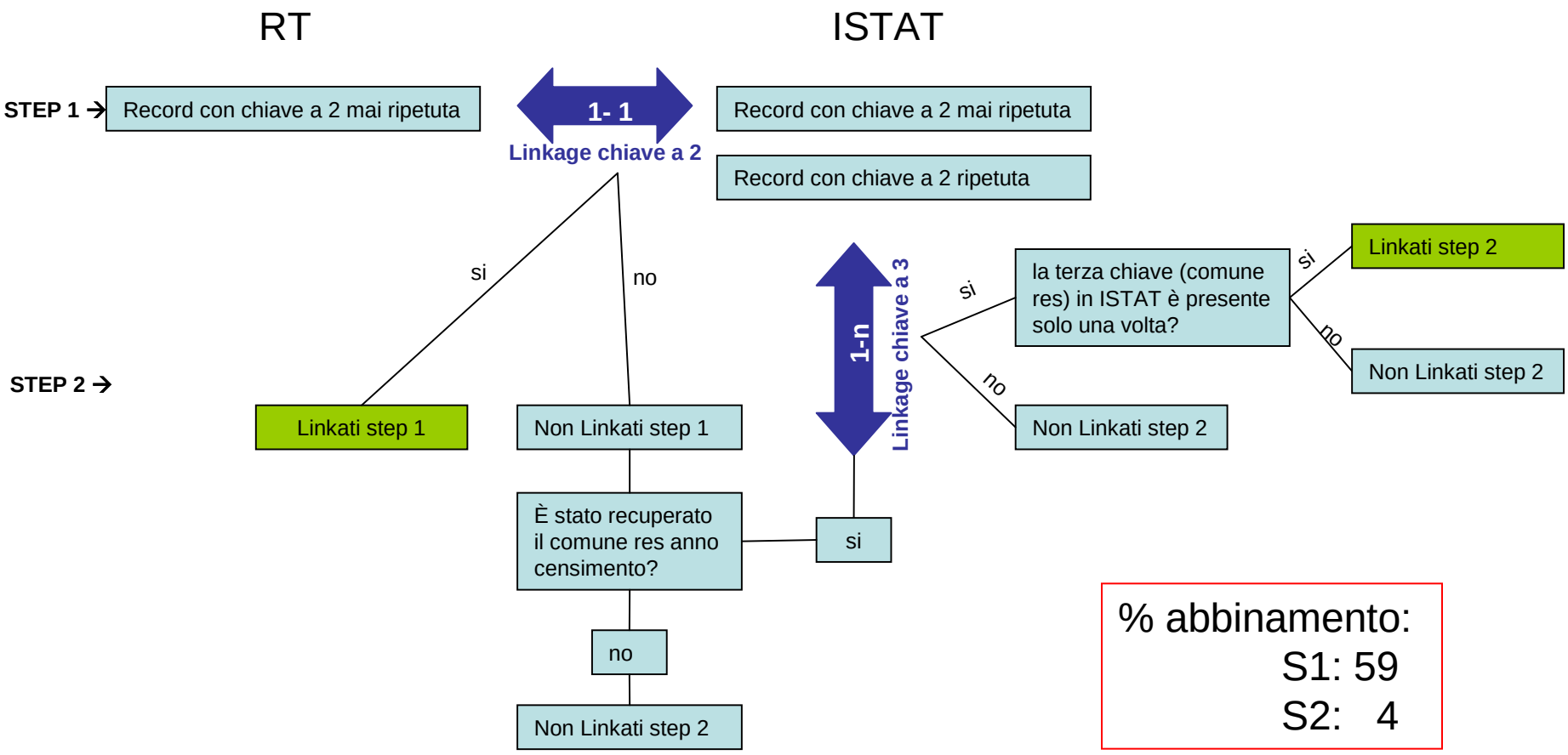
% abbinamento=97 (ipotetico dell'atteso)

1999-2002

n. key	chiave	Pattern	Linked	% sul tot. Dei candidati
0	pseudo CF completo	ABC XYZ 999 H 01 L219 M	2629	66.36
1	senza sesso	ABC XYZ 999 H 01 L219 -	12	0.30
2	senza comune	ABC XYZ 999 H 01 ---- M	372	9.39
3	senza giorno	ABC XYZ 999 H – L219 M	33	0.83
4	senza mese	ABC XYZ 999 – 01 L219 M	25	0.63
5	senza anno	ABC XYZ --- H 01 L219 M	32	0.81
6	senza nome	ABC --- 999 H 01 L219 M	79	1.99
7	senza cognome	--- XYZ 999 H 01 L219 M	46	1.16
8	senza sesso comune	ABC XYZ 999 H 01 ---- -	1	0.03
9	senza sesso giorno	ABC XYZ 999 H – L219 -	3	0.08
10	senza sesso mese	ABC XYZ 999 – 01 L219 -	0	0.00
11	senza sesso anno	ABC XYZ --- H 01 L219 -	10	0.25
12	senza sesso nome	ABC --- 999 H 01 L219 -	32	0.81
13	senza sesso cognome	--- XYZ 999 H 01 L219 -	37	0.93
14	senza comune giorno	ABC XYZ 999 H -- ---- M	215	5.43
15	senza comune mese	ABC XYZ 999 – 01 ---- M	2	0.05
16	senza comune anno	ABC XYZ --- H 01 ---- M	239	6.03
17	senza comune nome	ABC --- 999 H 01 ---- M	13	0.33
18	senza comune cognome	--- XYZ 999 H 01 ---- M	0	0.00
19	senza mese giorno	ABC XYZ 999 - -- L219 M	9	0.23
20	senza anno giorno	ABC XYZ --- H – L219 M	17	0.43
21	senza nome giorno	ABC --- 999 H – L219 M	1	0.03
22	senza cognome giorno	--- XYZ 999 H – L219 M	4	0.10
23	senza anno mese	ABC XYZ --- - 01 L219 M	94	2.37

Collegamento dei dati sanitari con altri archivi analitici 2/2

LINKAGE A 2 STEP



% abbinamento:
S1: 59
S2: 4

CHIAVE A 2 :data nascita + comune nascita
CHIAVE A 3 :data nascita + comune nascita + comune residenza anno censimento

Collegamento dei dati sanitari con altri archivi aggregati 1/2

Dati sanitari

(SDO, Mortalità…)

X%

Dati nominativi

Anagrafe Comune

Dati anagrafici individuali nominativi

Sezione di cens.

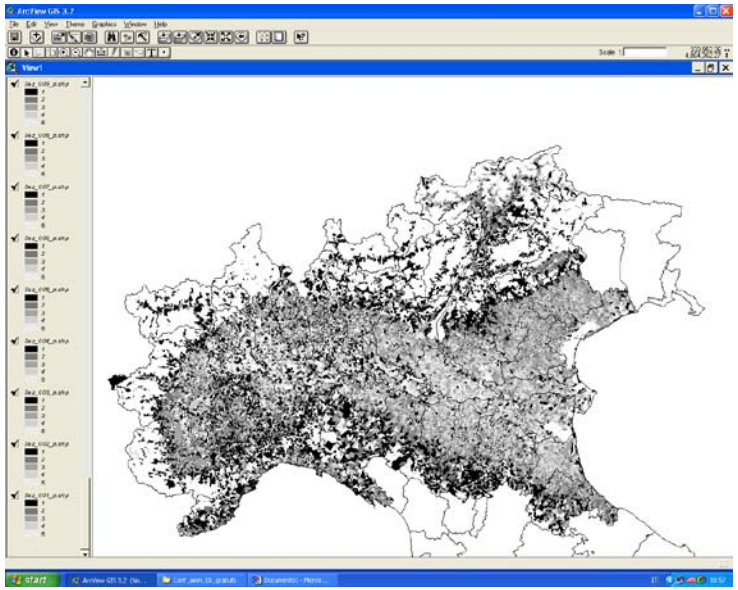
dei residenti,

o preferibilmente:

Georeferenziazione

ISTAT

(Censimento 2001)

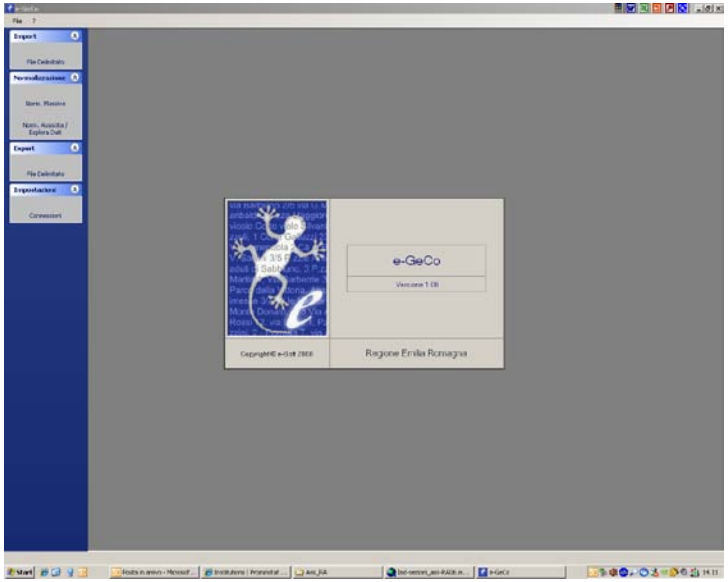


Frequenze per sezione

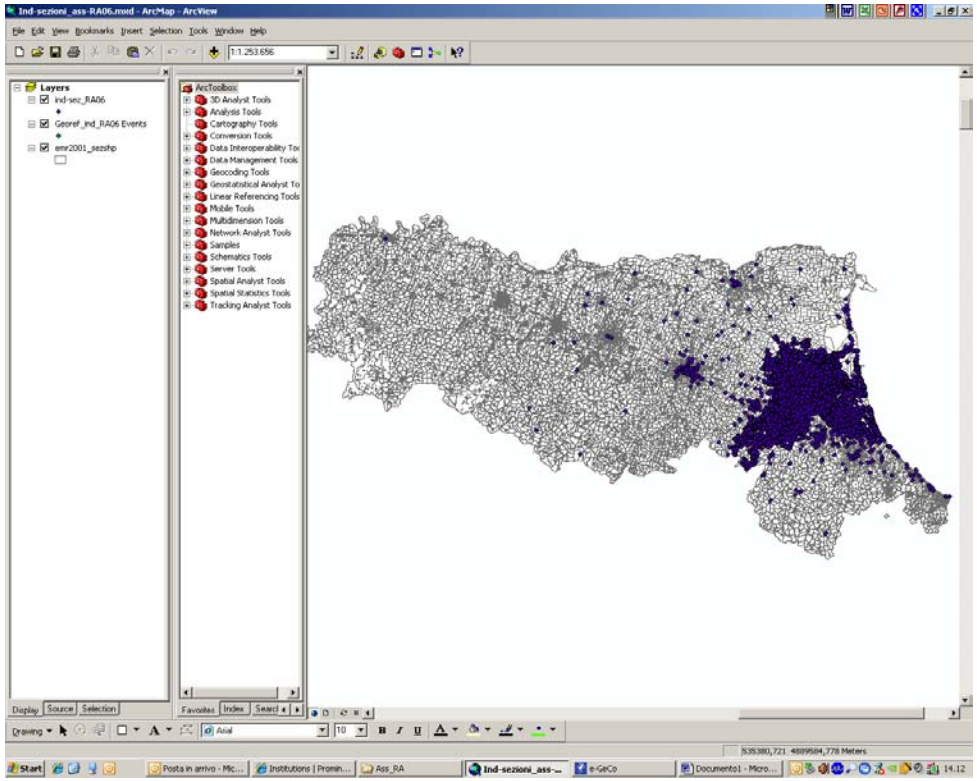
Indicatori sullo stato socio-demografico (es.: indice di deprivazione)

Collegamento dei dati sanitari con altri archivi aggregati 2/2

Georeferenziazione dell'*Anagrafe degli Assisti* nell'ASL di Ravenna



I passo: normalizzazione e attribuzione delle coordinate spaziali degli indirizzi (comune, toponimo e n° civico). L'uso del programma **eGeCo** (stradario del 2007-2009) consente di georeferenziare il 90% di 310.302 assistiti



II passo: *Join* **spaziale** delle coordinate assegnate agli indirizzi con la cartografia (poligoni delle sezioni di censimento 2001).

L'attribuzione della zona geografica avviene, in questo caso, con qualche approssimazione. Es.: disallineamento dell'informazione del comune (116) nell'1 per mille (301 indirizzi), corrispondente ad un errore di circa 3 metri

Per concludere, alcune considerazioni “iniziali”:

- Disponiamo di ampie fonti di dati, sia di natura epidemiologica che socio-demografica ed ecologia
- la mole delle informazione a disposizione cresce rapidamente e rimane sproporzionatamente inferiore la capacità di mettere a frutto informazioni spesso già esistenti
- i vincoli che derivano dalla normativa per la protezione dei dati personali impone, in Italia (e sempre più anche in Europa...), un livello di garanzia del cittadino che talvolta è di ostacolo alla ricerca (in particolare a quella epidemiologica)

- una domanda può essere: **quanto si adegua l’apparato istituzionale e le metodologie per non trasformare gli alti livelli di garanzia in ostacoli per attività di rilevanza e utilità pubbliche?**

- una prima risposta può essere: **le soluzioni tecnologiche e tecniche crescono a loro volta.** Ma ad oggi sono molto eterogenee; ricerca di un linguaggio comune per **BABELE**

Grazie per l’attenzione

ncaranci@regione.emilia-romagna.it