Seminario
URBES, ARCHIMEDE, Censimento permanente
I Comuni verso l'uso statistico degli archivi amministrativi
e dei sistemi di integrazione delle fonti
18 giugno 2015, Terni

Uso statistico delle fonti amministrative:

la valutazione della qualità degli archivi e le basi integrate di microdati

Manlio Calzaroni – Istat, Direttore centrale delle rilevazioni censuarie e dei registri statistici

Istat

Prospettive da Eurostat

COMUNICAZIONE DELLA COMMISSIONE AL PARLAMENTO EUROPEO E AL CONSIGLIO

sul metodo di produzione delle statistiche UE: una visione per il prossimo decennio

Bruxelles, 10.8.2009

COM(2009) 404 definitivo

Nuove esigenze

In tutti i settori della statistica continua ad aumentare la necessità di informazioni. Con l'aumento della complessità e dell'interrelazione dei dati rilevati, crescono anche le esigenze degli utenti di disporre di dati **integrati e coerenti**. …
… su tematiche che riflettono diversi **fenomeni** di base **correlati e interdipendenti**.

Quindi il modello "stovepipe" in cui le statistiche nei diversi settori vengono prodotte in modo indipendente non è adatto a soddisfare le esigenze politiche di insiemi di dati integrati.

Nuovo modo di produrre

… le statistiche per settori specifici non sarebbero più prodotte in modo indipendente; sarebbero invece prodotte come **parti integrate in sistemi di produzione completi** [impostazione delle statistiche basata sull'idea di un magazzino di dati (*data warehouse*)] per gruppi di statistiche.

Questi sistemi sarebbero basati su una comune infrastruttura (tecnica); applicherebbero nella misura del possibile software standardizzato e utilizzerebbero tutte le fonti di dati disponibili (*statistiche e, soprattutto, amministrative*).

Nuovo problemi da affrontare

A tal fine occorre individuare come le informazioni da fonti diverse possono essere messe insieme e sfruttate per scopi diversi, ad es.:
mediante l'eliminazione di differenze metodologiche, uniformando le classificazioni statistiche, ecc.

Per ottimizzare l'efficienza gli Stati membri dovrebbero creare una rete di basi dati da cui sia possibile estrarre qualsiasi informazione pertinente.

Per ottenere questi risultati:
È indispensabile integrare microdati,
cioè dati di prevalente origine amministrativa

Il processo di modernizzazione dell'Istat

From multiple modes for surveys to multiple data sources for estimates

by Constance F. Citro – Statistics Canada

Register base statistics: Administrative data for statistical purposes

by Andres and Britt Wallgren– Statistics Sweden

Towards an integrated statistics programme for the post-2015 development agenda

by Geet Bruinooge – Statistics Denmark

Statistics 4.0 - **Are we at the edge of a new era for statistics?**

by Walter Radermacher – Eurostat

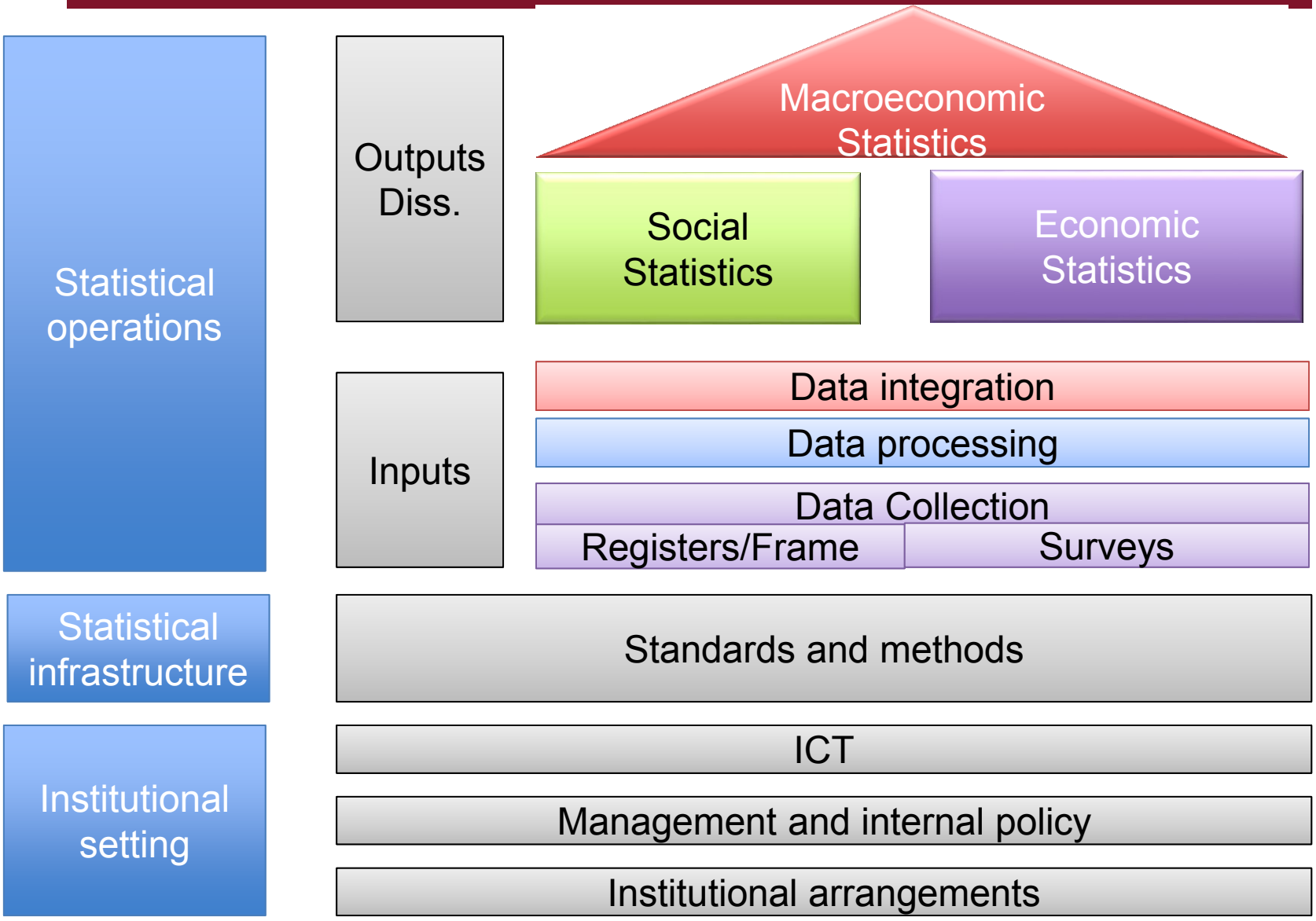
Il processo di modernizzazione dell'Istat

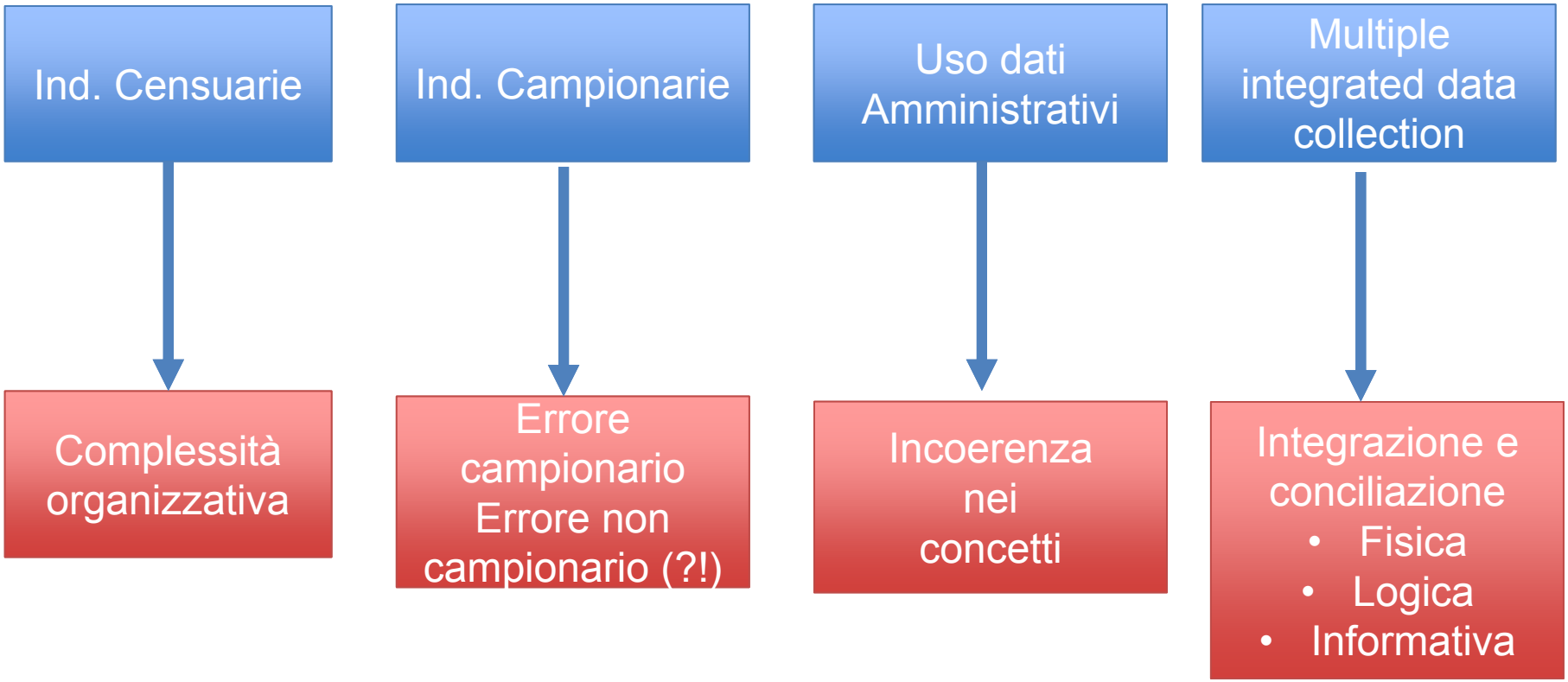
Business Architecture
Approccio "***per funzioni***" al processo statistico
Abbandono dei processi a "***silos***" per domini stat.

Incremento della trasversalità
Funzioni centralizzate a supporto di tutti i processi statistici

Riduzione dei costi
Uso ***massivo*** di dati ***non*** raccolti da indagine
Sfruttamento di tutte le informazioni disponibili per produrre statistiche "pubbliche"

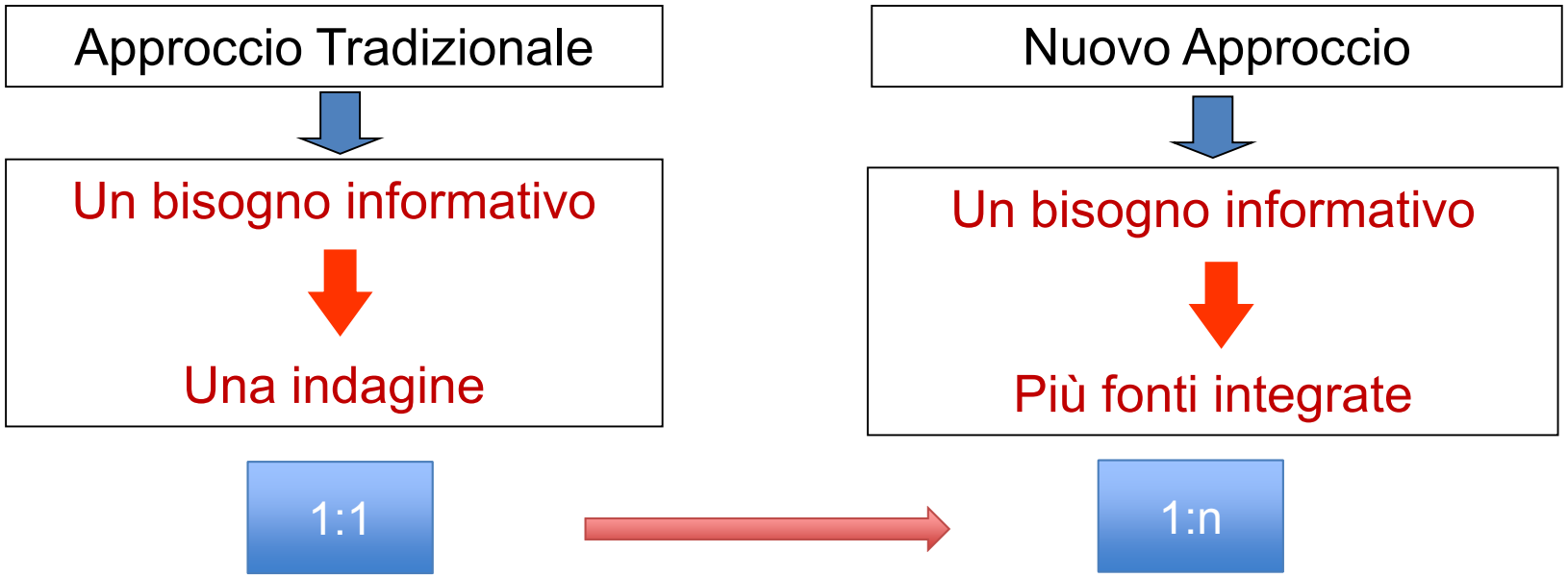
Il processo di modernizzazione dell'ISTAT (Geert Bruinooge)



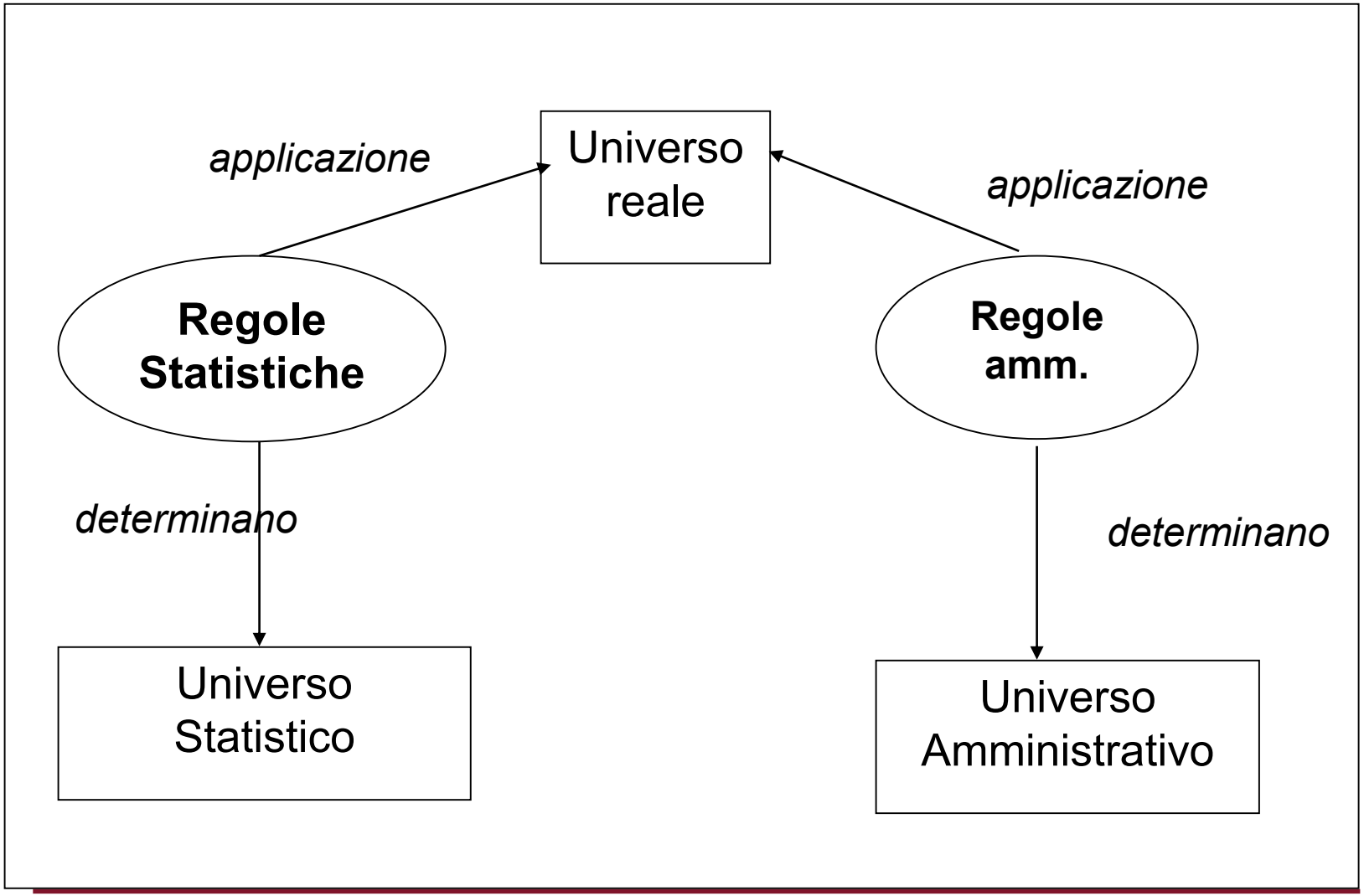

Ind. Censuarie
Ind. Campionarie
Uso dati Amministrativi
Multiple integrated data collection
Complessità organizzativa
Errore campionario
Errore non campionario (?!)
Incoerenza nei concetti
Integrazione e conciliazione
• Fisica
• Logica
• Informativa

Multiple Integrated Data Collection

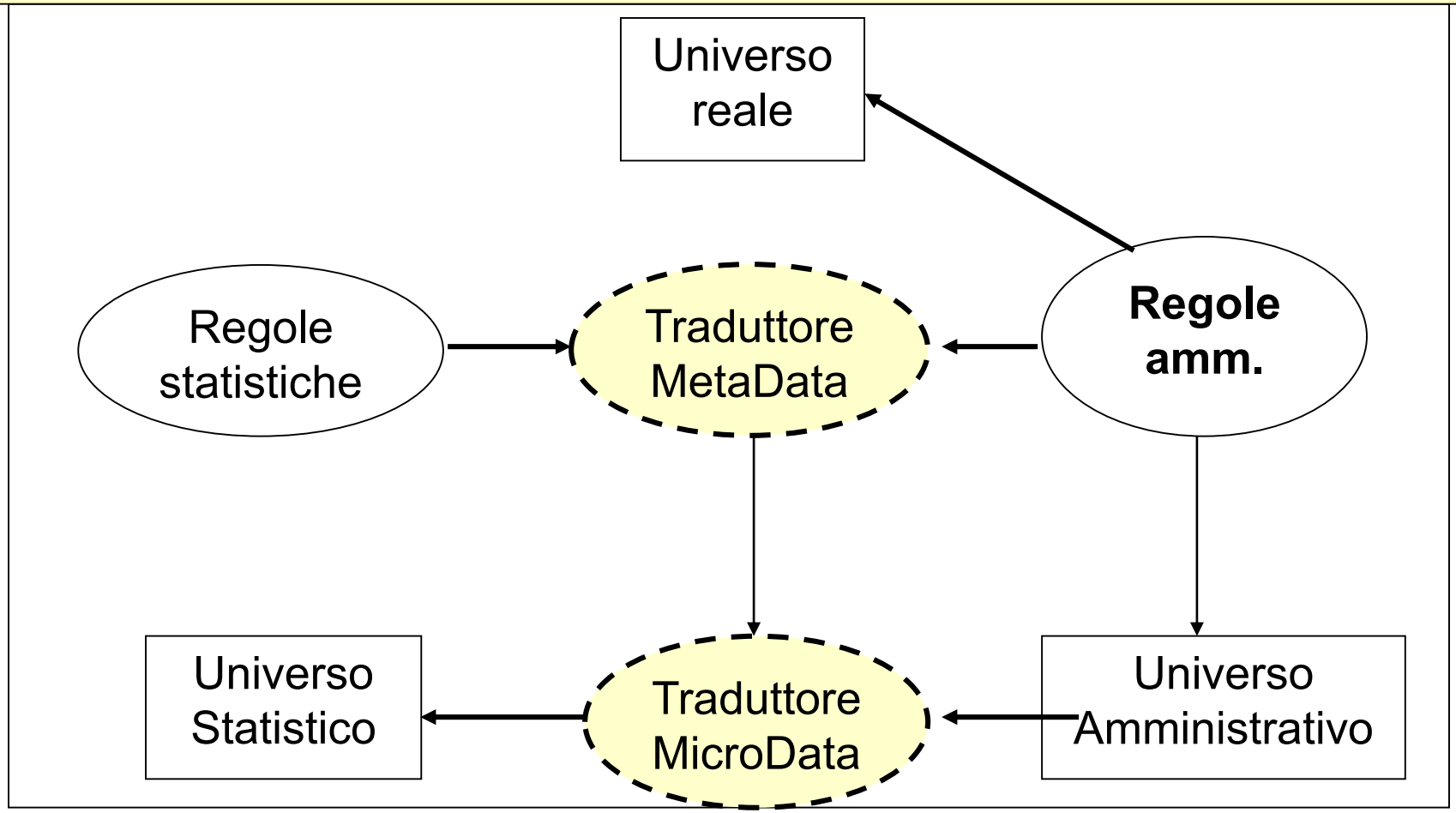
- Riduzione delle risorse finanziarie ed umane
- Riduzione del “fastidio statistico”
- Incremento, in quantità e qualità, delle informazioni statistiche richieste dagli utenti (nazionali e internazionali)
- Incremento della innovazione tecnologica e organizzativa
- Incremento di informazioni di natura differente (dichiarazioni, tracce digitali) disponibili.
- Nuove legislazioni, nazionali ed europee, che facilitano l’accesso da parte degli INS a dati non statistici



Cosa significa utilizzare fonti amministrative per fini statistici



Cosa significa utilizzare fonti amministrative per fini statistici



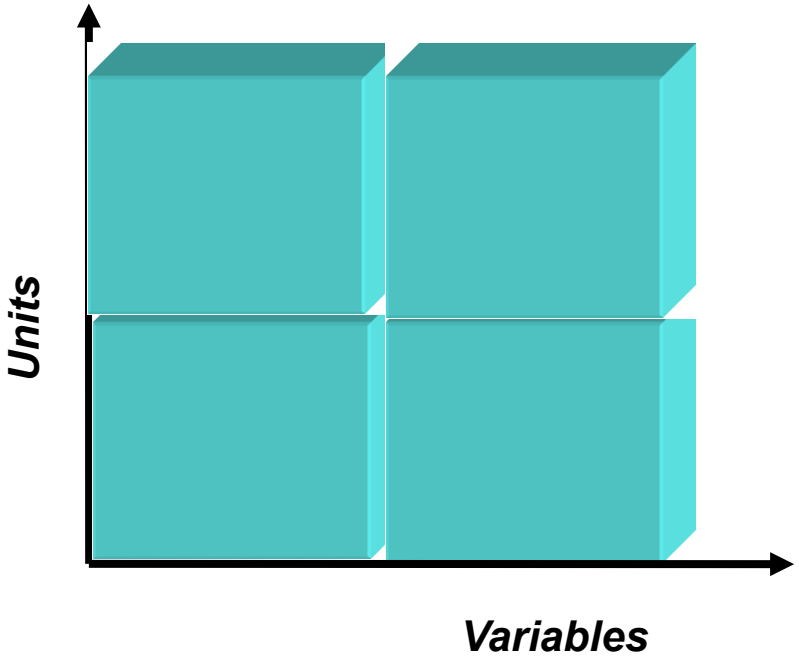
Qualità dei traduttori determina “Qualità statistica” dei dati amministrativi

Fonti amministrative utilizzate in Istat

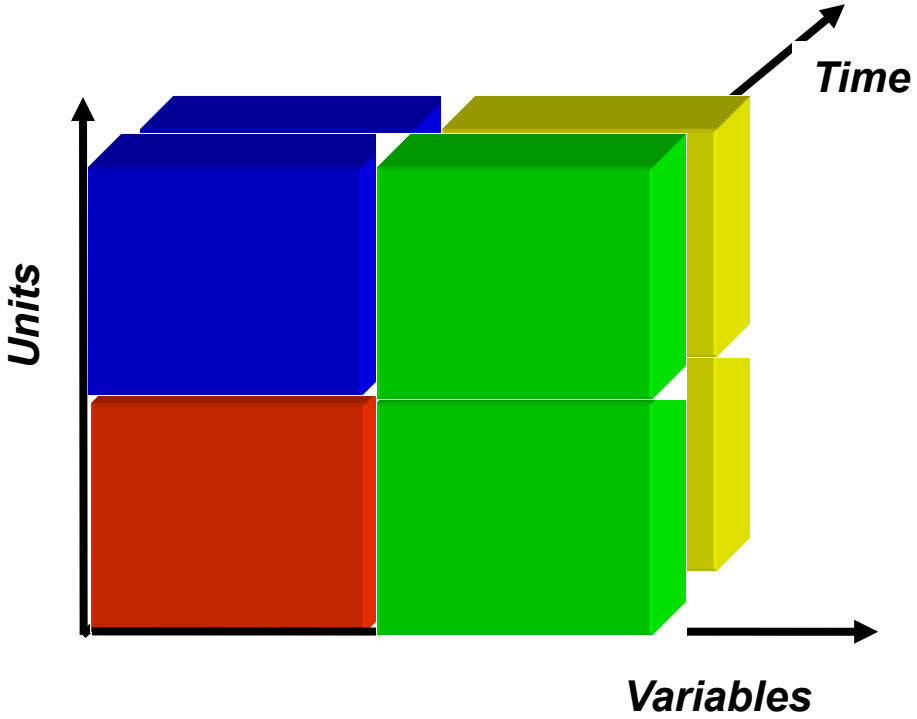
	2012	2013	2014
Enti fornitori	27	27	30
Forniture	199	261	278

Multiple Integrated Data Collection

Statistical Survey



Multiple Integrated Collection



MIDCS è un processo complesso caratterizzato dalle eterogeneità e variabilità (anche nel tempo) delle fonti utilizzate per la produzione di informazioni statistiche.

- **Integrazione Verticale**: per la stessa unità di analisi, le unità sono raccolte da differenti fonti
- **Integrazione orizzontale**: per ciascuna unità le variabili sono raccolte da differenti fonti

L'uso dell'MIDC modifica l'organizzazione, le tecnologie e le metodologie da adottare

MIDC: problematiche

- Difficoltà nell'integrazione fisica
- Le diverse fonti possono non essere disponibili in tempi diversi
- Possono utilizzare concetti/classificazioni non coerenti fra loro
- Possono utilizzare gli stessi concetti ma con visioni differenti (oggettivo/soggettivo)
- Contengono differenti tipologie di errori (non campionari/campionari)
- Possono contenere differenti livelli di qualità intrinseca

Processo produttivo complesso :

INDUSTRALIZZAZIONE/CENTRALIZZAZIONE

La risposta Istat

Sistema Integrato di Microdati - SIM

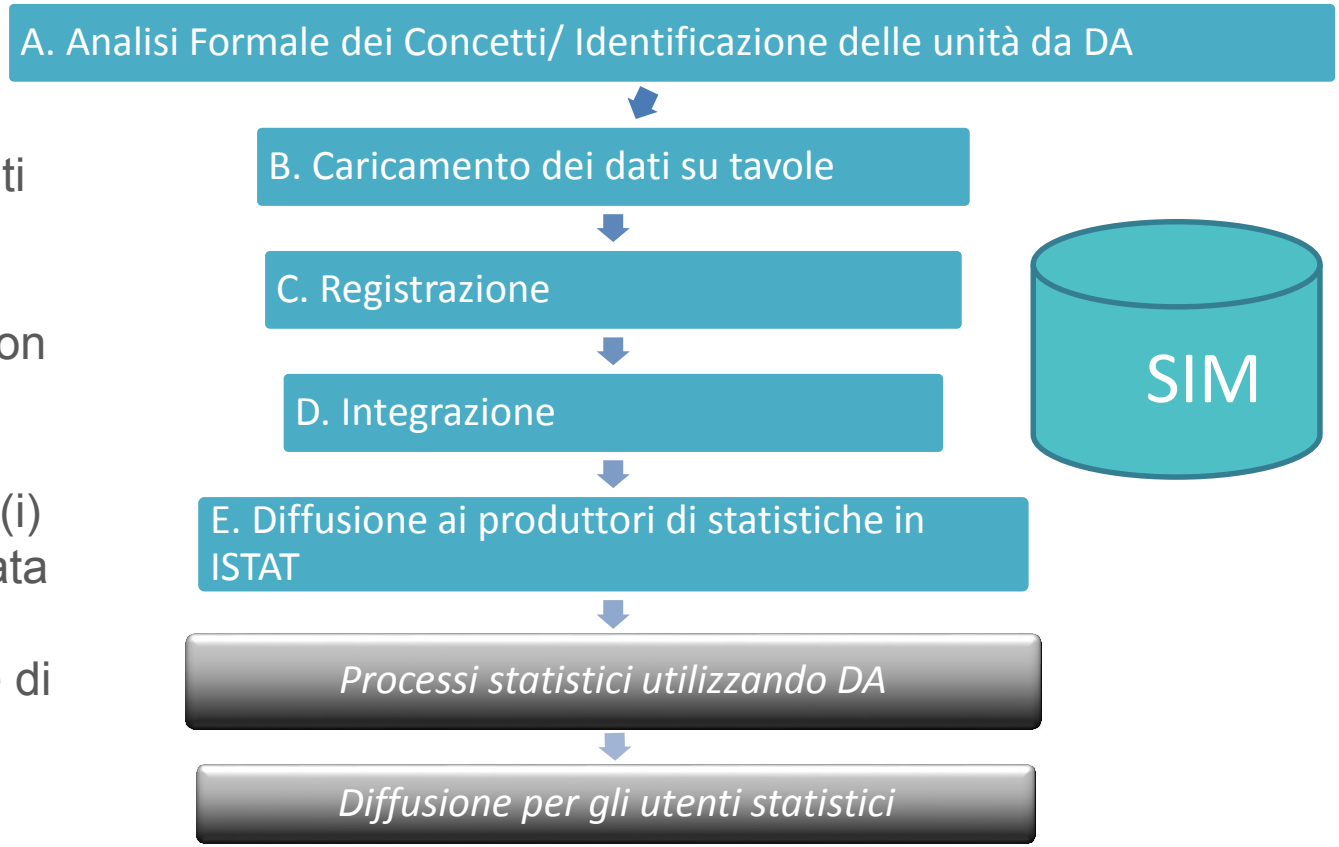
Def.: ***Archivio di microdati amministrativi e statistici integrati a supporto dei processi di produzione statistica***

Obiettivi

- Comune pretrattamento di Dati Amministrativi (DA)
- Conformità con le leggi sulla confidenzialità ed il collegamento di dati
- Uniformità di accesso ai DA per i produttori di statistiche
- Evitare duplicazioni di lavoro
- Comune descrizione di metadati e qualità dei DA

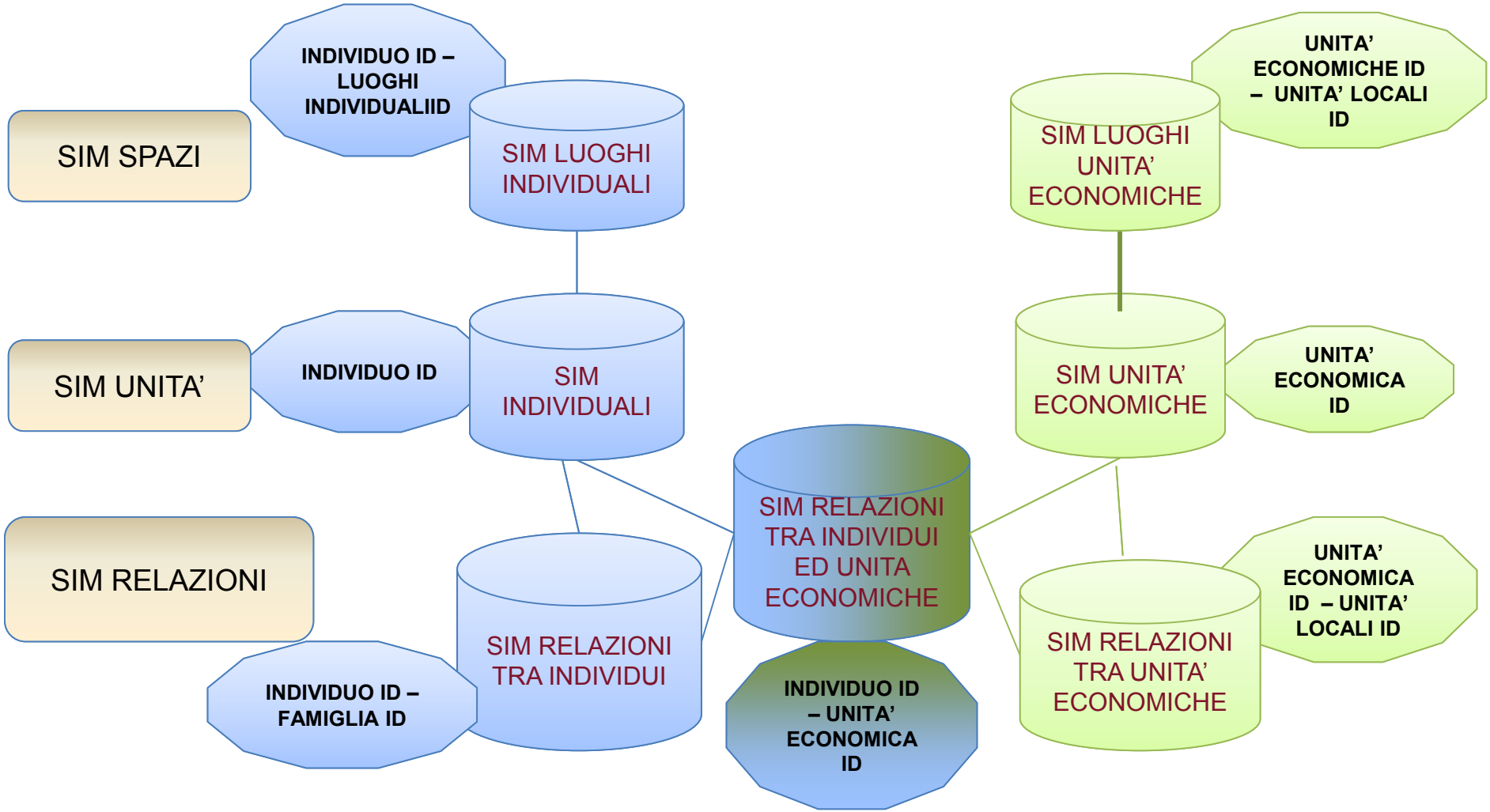
Funzioni del SIM

Integrazione riferita al processo di collegamento tra unità identificate in fonti diverse: individui, unità economiche, luoghi.
Ogni unità è identificata con un numero ID unico e stabile(nel tempo).
A seconda della variabile (i) di collegamento, è applicata un'adeguata strategia di collegamenti ed una serie di algoritmi.



L'integrazione alimenta lo sviluppo di Banche dati di l'integrazione di diversi sottosistemi di unità statistiche.
Le BD per l'integrazione sono contenitori di microdati utili per consentire una visione uniforme delle unità analizzate mostrando le informazioni disponibili nelle diverse fonti.

SIM: Il sottosistema



Dimensioni della qualità dei dati da indagine e dei DA

La misura della qualità dei dati amministrativi considerati come input del processo statistico si differenzia da quella dei dati di indagine output statistico.

Le dimensioni della qualità definite in ambito Eurostat per i dati statistici

Pertinenza, Accuratezza e attendibilità,
Tempestività e puntualità,
Coerenza e comparabilità, Accessibilità e chiarezza

non sono direttamente caratterizzanti la qualità statistica dei dati amministrativi.

Quindi:
Individuare le dimensioni specifiche della qualità a fini statistici di un dato amministrativo

La Ue ha sviluppato più studi che sono confluiti in uno schema condiviso - La Quality Report Card for Adm. data

La QRCA è un report condiviso a livello Ue (Blue-ETS) associato ai dati amministrativi

Descrive i principali aspetti della qualità dei DA

La documentazione della qualità dei dati del SIM avviene facendo riferimento alla QRCA (in corso di implementazione)

IPERDIMENSIONE	DIMENSIONE
	F1. Fonte dei DA
	F2. Rilevanza della fonte dei DA
METADATI	M1. Chiarezza e interpretabilità
Informazioni per la valutazione della qualità a livello concettuale e la descrizione delle procedure di acquisizione	M2. Comparabilità
	M3. Inalterabilità dei metadati
	M4. Acquisizione dei dati/trattamento
DATI	D1. Controlli tecnici
Indicatori di valutazione della qualità dei DA forniti	D2. Correlazione
	D3. Accuratezza
	D4. Completezza
	D5. Dimensione temporale

Obiettivo ->
automatizzare la produzione della QRCA creando inter-operabilità tra le fasi del processo di acquisizione

Fase di implementazione [1]

Dall'analisi concettuale dei dati amministrativi necessaria per il caricamento nelle tabelle Oracle del SIM si generano

-Indicatori della Dimensione dei Controlli tecnici (D1)
necessari per monitorare in modo tempestivo la fase di acquisizione dei dati (leggibilità dei file, conformità dei dati rispetto alla richiesta, data di acquisizione, ..)

- Individuazione degli oggetti/entità dell'archivio (M1) su cui calcolare gli indicatori di comparabilità, integrabilità,…

Fase di implementazione [2]

Dalla fase di Codifica degli oggetti (impresa, individuo, comune/provincia) si generano

- Indicatori di Integrabilità (D2-Dati):
  - *Comparabilità degli oggetti*
  - *Qualità delle variabili di linkage*
- Indicatori di Accuratezza (D3-Dati)
  - *Autenticità degli oggetti*
  - *Accuratezza delle variabili di classificazione (codici comune e provincia)*

Indicatori di Completezza (D4-Dati)
- *Copertura*
- *Valori mancanti degli identificativi*

Indicatori della Dimensione temporale (D5-Dati)
- *Dinamicità degli oggetti*

Esempio Laureati Miur

Nella fase di caricamento dei dati dei laureati del Miur si generano 5 *oggetti/entità* e quindi 5 tabelle :

Laureato,
Laurea,
Corso di studi universitario,
Facoltà,
Università.

Su cui calcolare gli indicatori.

Esempio Laureati Miur

La fase di Codifica dei laureati/individui consiste nell'assegnazione del codice individuo, unico in tutto il sistema.

Tale procedura genera
Indicatori di qualità delle variabili di linkage (D2.Integrabilità),
indicatori di autenticità degli oggetti (D3.Accuratezza).

Nel caso dell'esistenza di registri integrati in SIM, si generano gli ***Indicatori di copertura*** (D4.Completezza delle unità).

La registrazione della presenza dei codici individuo negli archivi amministrativi nel tempo genera
Indicatori della Dinamicità degli oggetti (D5.Dimensione temporale)

SIM – Quality Report Card for AD

D4. Completezza

Unità

Sottocopertura –
Sottocopertura per sottopopolazioni

Confronto con dati ufficiali prodotti dal Miur

Tipo di laurea	Miur dati ufficiali	Microdati	Sottocopertura	
	N	N	Diff	Diff %
Laurea Triennale (D.M. 509/99)	142.254	138.385	-3.869	-2,7
Laurea Triennale (D.M. 270/04)	26.484	25.939	-545	-2,1
Laurea Specialistica (D.M. 509/99)	51.297	50.407	-890	-1,7
Laurea Magistrale (D.M. 270/04)	35.244	34.826	-418	-1,2
Laurea a Ciclo Unico (D.M. 509/99)	13.533	13.093	-440	-3,3
Laurea Magistrale Ciclo Unico (D.M. 270/04)	13.346	11.367	-1.979	-14,8
Laurea Vecchio Ordinamento (antecedente D.M. 509/99)	16.647	15.588	-1.059	-6,4
Corsi di diploma universitario vecchio ordinamento (antecedente D.M. 509/99)	64	55	-9	-14,1
Scuole dirette a fini speciali vecchio ordinamento (antecedente D.M. 509/99)	3	1	-2	-66,7
Total	298.872	289.661	-9.211	-3,1

Esempio Report di Rilevanza

Dalla fase di acquisizione dei DA, si può generare in modo automatico un Report di Rilevanza per ciascuna fonte (F2. Rilevanza) che misura l'importanza della fonte, ad es., per la produzione Istat.
Si possono analizzare anche informazioni relative al tipo di uso (per frame di campionamento, per e&i, per produzione diretta delle statistiche), alla riduzione del response burden,…

Ente fornitore	Archivio	Numero richieste	Dip.to/ Dir.	PSN/ Accordo	Regolamenti Europei connessi	…
Inps	Archivio E-Mens	8	5	8	5	
Unioncamere-Infocamere	Dati dei bilanci delle società di capitali (XBRL)	5	5	5	2	
Agenzia delle Entrate	Modelli Unico	8	5	7	2	

Viste le prospettive che indicano come obiettivo

lo sviluppo di sistemi integrati di microdati
sostanzialmente amministrativi

Gestire la qualità statistica di questi dati significa:

1. Gestione centralizzata e coordinata della acquisizione e archiviazione dei dati – *Comitato e SIM Repository unico*

2. Definizione di uno schema unico di analisi di qualità statistica - *QRCA è un primo approccio disponibile da migliorare e adattare alla realtà nazionale*

3. Costruire e sviluppare indicatori specifici per le singole dimensioni individuate - *da sviluppare*

L'integrazione

Unità: integrazione fisica

Riconoscimento dello steso **oggetto** in più fonti e nel tempo

Variabili: integrazione logica

Riconoscimento dello stesso **contenuto semantico** in più fonti e nel tempo

Variabili: integrazione informativa

Riconoscimento della **coerenza sintattica** fra informazioni desumibili da più fonti

Integrazione fisica

A - Utilizzo di una chiave univoca

Codice Fiscale / Record Linkage deterministico

B - Utilizzo dei contenuti di caratteri identificati

- Persone fisiche - Nome e Cognome da solo o in combinazione con altri caratteri «discriminanti»: data di nascita, nazionalità, indirizzo di residenza, sesso,….
- Persone giuridiche – Denominazione da sola o in combinazione con altri caratteri «discriminanti»: attività economica, forma giuridica, dimensione, localizzazione,…

Analisi Testuale / Record Linkage probabilistico

L'utilizzo del Codice Fiscale

- Presente in tutte le fonti amministrative con un alto tasso di copertura
- E' lo «strumento migliore» per identificare uno stesso oggetto (persona fisica/persona giuridica) in più fonti.

Problematiche

- Non assume la caratteristica di codice identificativo universale
- Essendo un codice «parlante» si possono generare duplicazioni di codici (stesso CF per differenti individui), che se pur risolte a livello di Anagrafe Tributaria, possono non essere recepite, o recepite in ritardo da altri Enti.
- Gli individui non si referenziano in tutti gli ambiti nella stessa maniera.
- Non in tutte le culture è riconosciuta una strutturazione in termini di cognome e nome o c'è un attenzione particolare al momento di nascita (la misurazione del tempo non è universale!)
- Vengono fornite differenti strutture identificative ad enti differenti
- Date di nascita generiche (si predilige il primo giorno dell'anno)
- Duplicazioni nei nomi, particolarmente rilevante per individui nati in alcuni paesi esteri: SING (India), FERDINANDO (Sri Lanka)

Soluzione

Utilizzo congiunto del Codice Fiscale

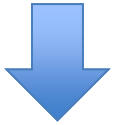
- con l'analisi testuale dei caratteri identificativi

e/o

- con tecniche di record linkage probabilistico

- Sistemi di classificazione (localizzazione, professioni, attività economica, tipologia di contratto,….)
  - Differenti tempistiche nell'aggiornamento di uno stesso sistema di classificazione
  - Differenti sistemi di classificazione

Privilegiare la descrizione rispetto a strumenti di decodifica



Analisi testuale

- Variabili numeriche
  - Riconoscimento della stessa variabile in due fonti – differente etichettatura con uguale contenuto informativo
  - Differente etichettatura nel tempo per una stessa fonte

- Processo difficilmente automatizzabile (necessità di analisi puntuale da parte di esperti)
  - Analisi testuale delle etichette
  - Analisi del contenuti informativo (analisi delle distribuzioni/ordini di grandezza dei valori)

Esempio di base dati integrata

Asia occupazione

Un archivio LEED – Linked Employer Employee Database

Integra 15 fonti

Tre punti di vista:

- Impresa
- Lavoratore
- Rapporto di lavoro

Integrazione informative: Il sistema Infor. sul mercato del lavoro

ASIA- Occupazione

Altre fonti amministrative

Base statistica micro per l'occupazione settore business

- ✓ E_mens/DMAG/CIGPD
- ✓ Artig./Commerc.
- ✓ ENPALS/INPGI(?)
- ✓ INAIL
- ✓ PARA_INPS
- ✓ CCIAA SOCI/PERSONE
- ✓ Unico quadro RH

Previdenziali

- ✓ Auton. Agricoltura
- ✓ Domestici
- ✓ INPDAP

NON Previdenziali

- ✓ 770 (CU)
- ✓ Cedolini stipendiali
- ✓ MIUR Lav. Univ.
- ✓ MIUR Lav. Scuola
- ✓ Dichiarazione dei redditi

Base informativa unica per l'occupazione

Integrazione informative: Il sistema Infor. sul mercato del lavoro

Base informativa unica per l'occupazione

riconciliazione dei disallineamenti delle fonti amministrative nei confronti dei regolamenti statistici

Dati Statistici

Censimento della popolazione

SISTEMA INFORMATIVO SUL MECATO DEL LAVORO

Correzione da modello dei dati amministrativi
Superamento della tradizionale dicotomia delle analisi economiche e sociali

Grazie per l’attenzione