Direzione centrale per lo Sviluppo e il Coordinamento della Rete territoriale e del Sistan – DCSR

FRAMEWORK PER LA QUALITA' DEGLI ARCHIVI AMMINISTRATIVI

Seconda versione 2015

SOMMARIO

PARTE PRIMA: INTRODUZIONE AL FRAMEWORK PER LA QUALITA' DEGLI ARCHIVI AMMINISTRATIVI 4
CONSIDERAZIONI SULLA VALUTAZIONE DELLA QUALITA' DEGLI ARCHIVI AMMINISTRATIVI 4
La valutazione della qualità degli archivi amministrativi: perché un nuovo approccio 4
La valutazione della qualità delle fonti d'informazione nel mondo esterno alla statistica ufficiale 6
Il contenuto del presente Framework 7
DALL'INFORMAZIONE AMMINISTRATIVA ALL'INFORMAZIONE STATISTICA 10
Il ciclo di vita dell'informazione amministrativa 10
Dall'informazione amministrativa all'informazione statistica: i collettivi amministrativi d'interesse statistico contenuti in un archivio amministrativo 12
Dall'informazione amministrativa all'informazione statistica: gli utilizzi dell'informazione amministrativa 16
Dall'informazione amministrativa all'informazione statistica: le modalità temporali di osservazione 18
PARTE SECONDA: INDICATORI DI QUALITA' RELATIVI ALLA FONTE E ALLA SUA DOCUMENTAZIONE 20
Il Framework Istat di indicatori per la documentazione della qualità delle fonti amministrative 20
IPERDIMENSIONE FONTE: Indicatori 21
Dimensione 1. Fornitore 21
Dimensione 2. Rilevanza e usi 22
Dimensione 3. Privacy e sicurezza informatica 22
Dimensione 4. Disponibilità dell'archivio 22
IPERDIMENSIONE DOCUMENTAZIONE DEI DATI E DELLE PROCEDURE: Indicatori 23
Dimensione 1. Chiarezza e standardizzazione della documentazione dei dati 23
Dimensione 2. Confrontabilità della documentazione dei dati 24
Dimensione 3. Chiavi identificative univoche e chiavi di raccordo (esistenza, tipo e qualità) 24
Dimensione 4. Documentazione delle procedure di acquisizione e trattamento dei dati 25
PARTE TERZA: UN NUOVO APPROCCIO ALLA VALUTAZIONE DELLA QUALITA' DEI DATI DI FONTE AMMINISTRATIVA 26
INTRODUZIONE 26
Linee generali dell'approccio seguito: l'ontologia dell'archivio e gli errori possibili 26
L'ontologia dell'archivio e le diverse tipologie d'informazione 27
Le diverse tipologie d'informazione e gli errori possibili 28
I contenuti della PARTE TERZA del Framework 30
I CONTENUTI INFORMATIVI DELL'ARCHIVIO AMMINISTRATIVO 31
I contenuti informativi dell'archivio: diversi tipi di enunciati 31
I contenuti informativi dell'archivio: l'organizzazione degli enunciati in record 35
LA DINAMICA DI AGGIORNAMENTO DELL'ARCHIVIO AMMINISTRATIVO 38
I diversi tipi di aggiornamenti 38
I diversi tipi di aggiornamenti: le classi REG, ELIM e MOD 40
Le operazioni nelle classi REG ed ELIM: ingresso e uscita dai collettivi 41
Le operazioni nella classe MOD: modifiche relative al possesso di caratteristiche e relazioni 43
Le operazioni nelle classi REG ed ELIM: gli effetti degli eventi istantanei di ingresso e uscita 43
Le operazioni nelle classi REG ed ELIM: ESEMPI di successioni di ingressi e uscite per elementi dei collettivi di tipo popolazione o evento con durata 44
Le operazioni nella classe MOD: ESEMPI di aggiornamento delle caratteristiche e relazioni per elementi dei collettivi di tipo popolazione o evento con durata 46
Serie di record presenti in archivio come effetto della dinamica degli aggiornamenti, per ogni elemento di popolazione o evento con durata 47
I DIVERSI TIPI DI ERRORE 55

I diversi tipi di errore in generale 55
Gli errori possibili in occasione di operazioni nella classe *REG* 58
Gli errori possibili in occasione di operazioni nella classe *ELIM* 60
Gli errori possibili in occasione di operazioni nella classe *MOD* 62
Considerazioni su ERRORI DI COPERTURA, MANCATE RISPOSTE ed ERRORI DI ACCURATEZZA e sui tipi di errori più frequenti in pratica 64
Errori di duplicazione 70
LE POSSIBILITA' DI DIAGNOSI PER I DIVERSI TIPI DI ERRORE 73
Le tipologie di controlli applicabili 73
Controlli strutturali 73
Controlli iniziali di errori evidenti 73
Altri controlli 74
Le tipologie di controlli applicabili e gli identificativi 75
Tabelle di associazione tra tipologie di controlli e tipologie di errore 77
APPENDICE 1 Statistical Network on Administrative Data: methodologies for an integrated use of administrative data in the statistical process 89
Preliminary technical report on a generic quality assessment framework 89
Usage of Administrative Data Sources for Statistical Purposes 90
APPENDICE 2: Indicatori Blue–Ets 93
Introduzione 93
Tabella 1. Dimensioni descrittive della qualità statistica dei dati amministrativi 94
Tabella 2. Indicatori di Integrabilità 94
Tabella 3. Indicatori di Accuratezza 94
Tabella 4. Indicatori di Completezza 95
Tabella 5. Indicatori della dimensione temporale 95
INTEGRABILITA' 97
ACCURATEZZA 98
COMPLETEZZA 100
DIMENSIONE TEMPORALE 102
APPENDICE 3: Documentare l'Ontologia di un Archivio Amministrativo 103
Perché documentare l'ontologia di un archivio amministrativo 103
Documentare l'ontologia di un archivio amministrativo: i collettivi 104
Documentare l'ontologia di un archivio amministrativo: le caratteristiche 108
Documentare l'ontologia di un archivio amministrativo: gli identificativi 109
Documentare l'ontologia di un archivio amministrativo: le relazioni 110
Documentare l'ontologia di un archivio amministrativo: costruire nuove caratteristiche 111
Documentare l'ontologia di un archivio amministrativo: l'archivio come rete di relazioni tra collettivi 121
Documentare l'ontologia di un archivio amministrativo: i sottoinsiemi dei collettivi principali 121
Gli oggetti nell'ontologia dell'archivio e i relativi tipi di enunciati 124
Gli oggetti nell'ontologia dell'archivio e i relativi tipi di enunciati, riformulati tenendo conto dei riferimenti temporali 126
L'archivio amministrativo come strumento di raccolta di enunciati di appartenenza a collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi 128
L'archivio amministrativo come strumento di osservazione dell'estensione dei collettivi, delle caratteristiche e delle relazioni 130
La qualità dell'archivio come corretta osservazione dell'estensione dei collettivi, delle caratteristiche e delle relazioni 141

PARTE PRIMA: INTRODUZIONE AL FRAMEWORK PER LA QUALITA' DEGLI ARCHIVI AMMINISTRATIVI

CONSIDERAZIONI SULLA VALUTAZIONE DELLA QUALITA' DEGLI ARCHIVI AMMINISTRATIVI

La valutazione della qualità degli archivi amministrativi: perché un nuovo approccio

La riflessione condotta da alcuni anni sul tema dell'utilizzabilità dei dati amministrativi per finalità statistiche ha portato in un primo tempo a identificare come attività preliminare e fondamentale l'analisi della qualità degli archivi amministrativi che sono utilizzati da specifici processi statistici. Questo obiettivo viene spesso sinteticamente enunciato descrivendolo come valutazione della qualità dell'input.

Di recente, i lavori condotti nell'ambito dello Statistical Network on Administrative Data hanno segnalato l'esigenza di valutare la qualità degli archivi disponibili in quanto potenzialmente utilizzabili per finalità statistiche generiche, non ancora specificate, senza vincolarsi all'utilizzo da parte di un processo specifico. Questo è l'obiettivo perseguito dal presente Framework.

Il lavoro di specifica di criteri per la valutazione della qualità degli archivi amministrativi portato avanti in questi anni si è evoluto lungo due direttrici:

- specifica di un modello descrittivo della qualità degli archivi amministrativi, articolato in aree tematiche a diversi livelli di specificità: l'esempio al quale si rifà, in parte, il presente Framework è il modello proposto da Statistics Netherlands nel 2009 [1], articolato in tre iperdimensioni della qualità, ciascuna delle quali comprende dimensioni e relativi indicatori

- individuazione di indicatori di qualità sintetici per gli archivi amministrativi, formulati tanto come informazioni di tipo qualitativo quanto come indicatori numerici calcolabili a partire dai dati in archivio: un esempio è il lavoro svolto nell'ambito del progetto europeo BLUE-ETS.

Tuttavia, manca ancora una concettualizzazione che sia di guida per analizzare l'utilizzabilità degli archivi amministrativi in modo sistematico e approfondito e al tempo stesso realmente generale, cioè indipendente da specifici usi, e organizzare di conseguenza il Framework di criteri di valutazione proposto.

Un Framework per valutare la qualità di una classe di fonti d'informazione non si dovrebbe limitare a fornire un elenco organizzato di indicatori ma dovrebbe giustificarne la formulazione sulla base di un approccio concettuale che ne garantisca l'esaustività e la sistematicità, il fatto cioè che gli indicatori proposti esauriscano, almeno in linea di principio, tutti gli aspetti da considerare rilevanti rispetto all'obiettivo di valutare la qualità della fonte.

[1] P. Daas, S. Ossen,R. Vis-Visschers, J. Arends-Toth, "Checklist for the Quality evaluation of Administrative Data Sources", CBS 2009

Occorre inoltre sottolineare che in concreto l'utilizzo statistico dei dati di fonte amministrativa per lo studio di specifiche variabili associate a specifici collettivi comporta un'attività di analisi mirata della qualità che è oggi generalmente svolta informalmente, senza un adeguato supporto metodologico e procedurale.

Si tratta di un'attività di analisi paragonabile alla fase esplorativa delle procedure di controllo e correzione che sono correntemente condotte sui dati d'indagine. In essa la qualità della specifica collezione di dati che si intende utilizzare è valutata in base a diversi vincoli logici come per le indagini ma con la necessità, a differenza della gran parte delle rilevazioni, di prendere in considerazione vincoli logici di tipo dinamico (longitudinale) tra eventi e tra variabili. In questa attività il supporto degli esperti della fonte amministrativa è necessario tanto per ricevere prime indicazioni sulla qualità, utili a orientare l'analisi, quanto per la specifica dei diversi generi di vincoli.

Un insieme di indicatori, anche se certamente utile per sintetizzare i risultati di tale attività, non è certamente sufficiente a guidarla.

E' necessario definire un più articolato Protocollo di guida alla valutazione della qualità dei dati di fonte amministrativa, generalizzabile in prospettiva a ogni tipo di fonte non statistica, che fornisca criteri standard e condivisi per valutare la qualità con la quale ogni data fonte d'informazione osserva specifici collettivi e specifiche variabili.

Per essere utilizzabile in ogni contesto, un tale Protocollo dovrà essere necessariamente articolato per generi di informazioni da valutare, generi di errori possibili, tipi di vincoli logici specificabili e di attività di controllo applicabili.

Nel caso delle indagini statistiche, così come delle elaborazioni statistiche successive e in generale dell'informazione trattata da processi statistici, sono le procedure di raccolta e utilizzo dell'informazione a suggerire una struttura concettuale di guida per enucleare le esigenze e i metodi di valutazione della qualità dell'informazione prodotta.

Nel caso particolare delle indagini un modello che consenta di descrivere le fasi dell'indagine come processo di osservazione offre una concettualizzazione adatta a individuare le dimensioni della qualità dell'indagine e organizzarle in un Framework. Per la valutazione di dettaglio della qualità di osservazione di specifici collettivi e variabili è inoltre disponibile una vasta offerta di tecniche di controllo e correzione.

Per definire le dimensioni della qualità degli archivi amministrativi, è necessario definire una concettualizzazione solida, che non può però poggiare su un'analoga caratterizzazione del processo di formazione del dato amministrativo.

Tale processo costituisce infatti per molti aspetti una "scatola nera" al di fuori del controllo dello statistico utilizzatore. Essendo orientato a fini amministrativi e non direttamente di conoscenza e misura dei fenomeni è costituito da procedure, spesso regolate da norme, nelle quali eventuali interventi e controlli mirano a fini operativi e non conoscitivi.

Inoltre i suoi esiti sono nella maggior parte dei casi fortemente influenzati dalle attitudini e peculiarità dei fornitori dell'informazione e in parte anche degli organismi che la raccolgono, in

una misura che è spesso non facilmente valutabile da parte di questi stessi organismi e comunque misurabile solo a posteriori, dando luogo in particolare a un peso degli errori sistematici certamente maggiore rispetto a ciò che si riscontra per le indagini.

Nel caso degli archivi amministrativi quindi occorre sviluppare una riflessione apposita sull'oggetto dell'attività di valutazione di qualità, con l'obiettivo di arrivare a definire il necessario Protocollo di guida alla valutazione della qualità dei dati di fonte amministrativa e in generale di garantire una struttura sistematica al Framework per la qualità definito.

Non potendosi basare, come per le indagini, sulla proceduralità del processo di formazione del dato è necessario ancorare i criteri di valutazione della qualità della fonte direttamente ai suoi contenuti informativi, cioè alla sua ontologia, descritta in termini di collettivi e variabili osservate.

Infine, l'analisi empirica puntuale di tutti i possibili errori finalizzata a definire un Protocollo di guida alla valutazione della qualità dei dati di fonte amministrativa costituisce il necessario presupposto pratico-operativo e documentativo per perseguire un obiettivo metodologico di più ampia portata: definire un modello teorico generale dell'errore per le fonti d'informazione non statistiche, adatto tra l'altro a consentire il controllo simultaneo di qualità su dati provenienti da più fonti, mediante l'assegnazione di probabilità di errore.

La valutazione della qualità delle fonti d'informazione nel mondo esterno alla statistica ufficiale

Altre considerazioni che rendono indispensabile sviluppare un nuovo approccio alla valutazione della qualità dei dati prodotti da fonti non statistiche riguardano, da un lato, l'evoluzione attualmente in corso verso un crescente utilizzo di una varietà di tali fonti e, dall'altro, la necessità di tenere conto delle tradizioni di analisi della qualità dei dati che si sono andate consolidando nel mondo esterno alla statistica ufficiale, in particolare in ambito informatico.

Per ciò che riguarda il primo punto, è immediato osservare che i Framework per la diagnosi della qualità delle fonti amministrative oggi più diffusi nella statistica ufficiale sono spesso basati su una specifica di classi di errore ispirata a quella utilizzata per le indagini e puntano direttamente a individuare una struttura di indicatori di qualità che estenda e adatti al contesto degli archivi amministrativi gli indicatori di qualità utilizzati per le indagini.

Questo approccio mostra i suoi limiti nel momento in cui si considera seriamente l'attuale tendenza verso l'utilizzo crescente di un gran numero di fonti d'informazione con caratteristiche diversificate, dagli archivi amministrativi ai big data. E' necessario un rovesciamento di prospettiva, che veda le indagini statistiche come un caso particolare in un più generale universo delle fonti d'informazione disponibili e utilizzabili a scopo statistico.

Le diverse fonti d'informazione attuano processi di osservazione della realtà e raccolta di informazione che comportano propri meccanismi di generazione dell'errore, tutti attualmente ancora da descrivere su un piano pratico e da formalizzare su un piano teorico che sia il più generale possibile, in modo da ricomprendere come casi particolari la valutazione della qualità dei dati

provenienti da indagini, da archivi amministrativi, da fonti caratterizzabili in base al paradigma dei big data.

Gli ambiti di progettazione, gestione e utilizzo di tutte queste fonti non statistiche coincidono con l'ambito della progettazione e gestione di database, che è di tradizionale interesse per la ricerca e la pratica informatica e al quale possono fare riferimento gli esperti degli archivi presso gli enti gestori.

In questo campo esiste un'ampia letteratura sulla qualità dei dati, che prescinde dall'uso statistico ma non dall'uso dei dati a supporto delle decisioni da parte del management, un contesto d'uso che è molto vicino a quello statistico e con esso può condividere metodi e tecniche.

Un Framework per la qualità degli archivi amministrativi a supporto dell'uso statistico non può ignorare questa realtà: per essere adottato deve poter essere compreso, apprezzato e utilizzato anche da questo diverso punto di vista. Anche per questo è importante una solida motivazione concettuale degli indicatori proposti, basata su paradigmi concettuali familiari all'informatica e agli informatici, anche se funzionali all'uso statistico.

Il contenuto del presente Framework

In linea con le precedenti considerazioni, il presente Framework adotta una concettualizzazione che àncora le diverse esigenze di valutazione della qualità ai diversi tipi di informazione gestita negli archivi amministrativi, enucleati mediante il riferimento diretto ai diversi tipi di oggetti osservati, che costituiscono l'ontologia d'archivio.

Tale concettualizzazione è stata definita in tre passi:

a) descrivendo in modo il più generale possibile il ciclo di vita del dato amministrativo, in modo da enucleare gli oggetti fondamentali ai quali è riferita l'informazione gestita negli archivi amministrativi: soggetti ed eventi amministrativi e loro proprietà;

b) riconducendo tali oggetti fondamentali ai concetti fondamentali alla base dell'uso statistico dell'informazione: collettivi d'interesse statistico e loro proprietà;

c) enucleando le specifiche esigenze di valutazione della qualità per i collettivi amministrativi d'interesse statistico e per le loro proprietà, che costituiscono l'ontologia d'archivio.

Nel **successivo paragrafo** della presente **PARTE PRIMA** si illustrano i passi a) e b), delineando così gli aspetti fondamentali della concettualizzazione adottata.

In particolare nel paragrafo seguente si discute l'articolazione del ciclo di vita del dato amministrativo e si presentano i concetti fondamentali utili a descriverlo. Nei successivi paragrafi tali concetti sono analizzati e reinterpretati da un punto di vista statistico, introducendo una prima concettualizzazione degli oggetti di riferimento dell'informazione statistica che si basa sull'individuazione di diversi tipi di collettivi amministrativi d'interesse statistico e delle proprietà loro attribuibili.

L'intero Framework rappresenta poi uno sviluppo del compito descritto al punto c).

Si tratta del compito metodologicamente più impegnativo, che ha richiesto:

- la definizione dell'ontologia degli archivi amministrativi
- una definizione dettagliata degli errori ancorata alle ontologie degli archivi, definita in termini di errata identificazione, errata inclusione o errata esclusione di elementi
- la disamina, che richiede un continuo sviluppo, dei diversi tipi di controlli applicabili alle diverse tipologie di errori individuate, con l'utilizzo di vincoli anche longitudinali.

Generalmente i modelli descrittivi della qualità degli archivi amministrativi finora proposti operano una distinzione tra la qualità generale di una fonte amministrativa, che determina l'utilizzabilità statistica della fonte nel suo complesso e può essere diagnosticata mediante indicatori sintetici, e la qualità dei dati prodotti dalla fonte amministrativa considerata, che dev'essere studiata con indicatori più specifici, spesso ottenuti mediante elaborazioni sui dati stessi.

In particolare molti di questi modelli, incluso il modello oggi adottato in Istat, adottano lo schema proposto da Statistics Netherlands che individua tre componenti principali della qualità di un archivio amministrativo, identificate nelle tre iperdimensioni Fonte, Documentazione, Dati.

La prima componente attiene a tutte le caratteristiche che riguardano nel suo complesso la Fonte, laddove si considerano l'autorità responsabile, la rilevanza, la disponibilità dell'archivio. La componente Documentazione riguarda invece la presenza e l'adeguatezza dei metadati che descrivono i dati. Infine la componente Dati riguarda la qualità intrinseca dei dati dell'archivio.

Per quanto detto l'approccio proposto ha impatto soprattutto sulla diagnosi della qualità dei dati, rispetto alla specifica della qualità generale della fonte amministrativa e, quindi, in termini di componenti della qualità, ha impatto soprattutto sulla componente Dati.

Di conseguenza per ciò che riguarda le componenti Fonte e Documentazione il presente Framework si attiene largamente a quanto già definito all'interno del Framework Istat per la valutazione della qualità degli archivi amministrativi mentre per quel che riguarda la componente Dati se ne discosta, proprio perché mira a definire, più che un sistema di indicatori, un più articolato Protocollo di guida alla valutazione della qualità dei dati di fonte amministrativa.

Comunque il lavoro di analisi delle determinanti della qualità dei dati che si intende condurre dovrà anche alla fine suggerire un insieme di indicatori sintetici che comprenda e arricchisca il sistema di indicatori della qualità dei dati già utilizzato in Istat, anche mediante l'investigazione di ulteriori dimensioni della qualità, ad esempio l'articolazione della qualità dei dati lungo la dimensione temporale.

Nella successiva **PARTE SECONDA** sono preliminarmente presentati gli indicatori Istat relativi alle componenti Fonte e Documentazione, perlopiù di tipo qualitativo, confermando indicatori già proposti in letteratura, in particolare nel progetto Blue-ETS, e suggerendone di nuovi.

Il nuovo approccio proposto è dettagliato nella **PARTE TERZA**, nella quale viene prima introdotta una specifica semi-formale, basata sull'ontologia d'archivio, dei contenuti informativi di un archivio amministrativo e della loro dinamica di aggiornamento, sono poi individuate le diverse categorie di errori possibili e le loro combinazioni nel modo più generale, come possibilità teoriche, delineando anche le possibilità di diagnosi per tali categorie di errori

La **PARTE QUARTA** del Framework (in corso di revisione) approfondisce le determinanti degli *errori di copertura dei collettivi*, e dei connessi *errori sugli identificativi*, anche con riferimento ai

diversi tipi di collettivi osservati in pratica dai diversi tipi di archivi amministrativi, ed elenca tutti i possibili errori di questo tipo.

La **PARTE QUINTA** del Framework (in corso di stesura) approfondisce le determinanti degli *errori di accuratezza relativi alle caratteristiche e alle relazioni osservate per gli elementi dei collettivi* ed elenca tutti i possibili errori di questo tipo. Analizza inoltre i possibili *errori nel matching* tra informazioni riferite allo stesso elemento del collettivo.

L'analisi condotta nella PARTE QUARTA e QUINTA del Framework prende in esplicita considerazione gli aspetti legati al riferimento temporale delle informazioni da accettare o accettate in archivio, in modo da tenere conto adeguatamente della natura continua nel tempo della raccolta di informazioni che è tipica degli archivi amministrativi e delle fonti loro simili nonché della sua influenza sugli errori possibili in concreto.

DALL'INFORMAZIONE AMMINISTRATIVA ALL'INFORMAZIONE STATISTICA

Il ciclo di vita dell'informazione amministrativa

In un articolo su Survey methodology del 1987, J. B. Brackstone di Statistics Canada distingue le seguenti categorie di registrazioni amministrative:

- Registrazioni mantenute per regolare il flusso di beni e persone tra le frontiere: import ed export, immigrazione ed emigrazione;
- Registrazioni risultanti dall'esigenza legale di registrare particolari eventi: nascite, morti, matrimoni, divorzi, fusioni di imprese, licenze;
- Registrazioni necessarie per amministrare benefici e obbligazioni come tasse, indennità di disoccupazione, pensioni, assicurazioni sulla vita, trasferimenti alle famiglie;
- Registrazioni necessarie per amministrare istituzioni pubbliche come scuole, università, istituzioni della sanità pubblica, tribunali, prigioni;
- Registrazioni provenienti dalla regolazione dell'attività economica, relativa ad esempio a trasporti, banche, telecomunicazioni, anche regolazione dell'offerta o del prezzo di alcuni beni;
- Registrazioni provenienti dall'offerta di servizi come elettricità, telefoni, acqua.

Nello stesso articolo, a questa disamina delle diverse categorie di registrazioni amministrative basata sul loro obiettivo Brackstone aggiunge poi alcune osservazioni molto preliminari sulla qualità dell'informazione raccolta.

Anzitutto osserva che l'obiettivo della registrazione influenza la qualità della registrazione stessa: quando le registrazioni servono a scopi di regolazione la loro qualità è in generale variabile e dipende dal grado di rispetto delle leggi nei diversi contesti, mentre nei casi in cui la registrazione è necessaria per avviare attività o compiere operazioni, oppure per ottenere benefici, si può in generale assumere che la registrazione sia di buona qualità.

Infine osserva che le definizioni delle entità oggetto di registrazione dipendono dallo scopo amministrativo della registrazione e non sempre sono utili a fini statistici.

Sulla base di queste osservazioni di Brackstone, si può delineare una descrizione molto generale del ciclo di vita dell'informazione amministrativa.

L'attività amministrativa può essere:

- di regolazione della vita collettiva o dell'attività economica;
- di amministrazione di benefici e obblighi;
- di erogazione di servizi immateriali (istruzione, sanità, sicurezza…) o materiali.

Si hanno comunque sempre:

- da una parte, un'istituzione o una rete di istituzioni che esercitano una specifica attività amministrativa, articolata in procedimenti amministrativi basati su norme;

- dall'altra, le due popolazioni di base sulle quali in generale si esercita l'attività amministrativa, che sono:
    - la popolazione delle persone che vivono sul territorio amministrato, con le loro aggregazioni: famiglie, convivenze;
    - la popolazione degli organismi che svolgono attività economica sul territorio amministrato: imprese e organismi non profit in senso lato, con le loro componenti funzionali e territoriali e le loro aggregazioni.

Oltre che su tali popolazioni di base, l'attività amministrativa si può esercitare direttamente sul territorio amministrato: si pensi ai piani paesaggistici, alla manutenzione di registri relativi a patrimoni naturalistici, alla prevenzione antincendi.

Le norme che regolano una data attività amministrativa determinano il sottoinsieme delle due popolazioni di base, o la porzione di territorio, che è specificamente coinvolto nell'attività.

Nell'esercizio di una data attività amministrativa, in genere un elemento generico di una delle due popolazioni di base, che in base alla norma fa parte del sottoinsieme specificamente coinvolto nell'attività, entra in contatto con una delle istituzioni che esercitano l'attività: da questo momento l'istituzione attiva una raccolta continua di quella informazione che è necessaria all'esercizio dell'attività, fino a quando l'elemento non cessa di far parte del sottoinsieme delle due popolazioni di base coinvolto nell'attività (rimanendo però spesso registrato in archivio).

Il ciclo di vita dell'informazione amministrativa si può quindi sinteticamente descrivere come segue.

Una specifica *attività amministrativa* è articolata in *procedure amministrative* basate su norme.

I *soggetti* coinvolti nell'attività, generalmente coincidenti con i soggetti normativamente definiti, sono da una parte le istituzioni che esercitano l'attività, dall'altra quegli specifici sottoinsiemi delle due popolazioni di base, la popolazione delle persone e la popolazione degli organismi che svolgono attività economica, sui quali si esercita l'attività.

Nel caso di attività amministrative rivolte direttamente al territorio i soggetti sono unicamente le istituzioni che esercitano l'attività.

Gli *oggetti* dell'attività sono *specifici eventi, fatti, accadimenti riferiti ai soggetti su cui si esercita l'attività*, inclusa l'erogazione di servizi ad essi diretti, o relativi direttamente a un territorio.

In molte specifiche attività tali eventi, fatti, accadimenti sono *dichiarati* dai soggetti di riferimento o da altri soggetti in loro nome o per loro conto. In altre attività, più facilmente quelle orientate all'erogazione di servizi, tali eventi, fatti, accadimenti sono oggetto di *registrazione diretta*.

Le *informazioni* gestite per lo svolgimento dell'attività riguardano *le proprietà dei soggetti e degli oggetti dell'attività, vale a dire le loro caratteristiche e le relazioni che le legano***,** e sono raccolte dalle istituzioni che esercitano l'attività.

Precisamente, una data istituzione che esercita la specifica attività amministrativa raccoglie informazioni su quegli elementi dei sottoinsiemi delle popolazioni di base rilevanti per l'attività con i quali viene a contatto, per tutto il tempo per cui rimangono tali, e sui relativi eventi, fatti, accadimenti, ma può anche raccogliere informazioni relative alle altre istituzioni che esercitano l'attività ad un livello differente (ad esempio il MIUR gestisce informazioni sugli studenti ed anche sugli atenei). Nel caso di un'attività amministrativa rivolta direttamente al territorio l'istituzione che la esercita raccoglie sul territorio tutte le informazioni necessarie all'attività specifica.

Si può delineare una prima definizione molto generale della qualità di questa informazione.

Un giacimento di informazioni costituito ai fini dell'esercizio di una specifica attività amministrativa deve assicurare la *rappresentazione esatta delle proprietà dei soggetti e degli oggetti dell'attività in qualsiasi momento d'osservazione*: a questo scopo deve assicurare la *registrazione immediata e accurata* di tutti i cambiamenti di tali proprietà.

Dall'informazione amministrativa all'informazione statistica: i collettivi amministrativi d'interesse statistico contenuti in un archivio amministrativo

Primo passo per l'utilizzo statistico di un giacimento di informazione amministrativa è la caratterizzazione del suo contenuto in termini di *collettivi d'interesse statistico* e loro proprietà.

Per individuare correttamente tali collettivi occorre premettere alcune considerazioni relative alle modalità con le quali può essere organizzato l'esercizio di un'attività amministrativa.

Una specifica attività amministrativa può essere esercitata da una singola istituzione oppure da una rete di istituzioni. In questa seconda eventualità è importante distinguere il caso in cui le istituzioni cooperano su un unico livello rispetto al caso in cui la rete di istituzioni è organizzata su più livelli gerarchici. E' questo il caso ad esempio in cui il servizio dell'istruzione pubblica è diretto da un ente centrale come il MIUR ed è poi erogato da altri enti, con livelli di autonomia diversi, come le scuole e le università.

Nel primo caso le istituzioni collegate in rete gestiscono tutte giacimenti d'informazione che sono direttamente riferiti ai soggetti su cui si esercita l'attività, cioè ai sottoinsiemi delle popolazioni di base rilevanti per l'attività.

Nel secondo caso invece le istituzioni o i gruppi di istituzioni a ogni livello gestiscono giacimenti d'informazione relativi all'attività che possono contenere tanto informazioni dirette sui soggetti ai quali è indirizzato il servizio o, in generale, in relazione ai quali si esercita l'attività (nel nostro esempio sono gli studenti), quanto informazioni sulle istituzioni che concorrono ad esercitare l'attività a un livello sottostante (nel nostro esempio sono le scuole o gli atenei).

Fatta questa premessa, con riferimento alla specifica dell'ontologia d'archivio (si veda l'apposita Appendice) si può affermare che:

- i soggetti su cui si esercita l'attività, cioè i sottoinsiemi delle popolazioni di base rilevanti per l'attività, corrispondono a collettivi d'interesse statistico di tipo popolazione (esempio Studenti);

- nel caso che l'attività sia esercitata da una rete di istituzioni organizzata in livelli, alcuni soggetti che esercitano l'attività possono anch'essi corrispondere a collettivi d'interesse statistico di tipo popolazione (esempio Scuole, Atenei);

- gli oggetti dell'attività corrispondono a collettivi d'interesse statistico di tipo evento, con durata (esempio Carriera dello studente, Degenza, Rapporto di lavoro) o istantaneo (esempio Immatricolazione, Esame, Ricovero, Avvio rapporto di lavoro); si incontrano spesso eventi di tipo associativo, cioè riferiti a due o più elementi di collettivi di tipo popolazione (ad esempio Rapporto di lavoro).

Chiamiamo questi collettivi, corrispondenti ai soggetti e agli oggetti di un'attività amministrativa, *collettivi amministrativi d'interesse statistico*.

E' interessante osservare che il territorio può essere considerato come un particolare tipo di collettivo, che evidentemente non è costituito da un insieme di elementi distinti ma ha comunque una proprio estensione che è continua, e può essere discretizzata fissando un'unità di misura.

I collettivi amministrativi d'interesse statistico possono avere come proprietà caratteristiche qualitative o quantitative o relazioni con altri collettivi (si veda sempre l'Appendice dedicata all'ontologia degli archivi amministrativi).

Le caratteristiche possedute dagli elementi dei collettivi amministrativi d'interesse statistico sono viste dallo statistico come *variabili.*

Il territorio in particolare ha associate variabili che sono oggetto di operazioni di misura anziché di registrazione discreta elemento per elemento. Eventi direttamente riferiti al territorio possono inoltre essere oggetto dell'attività amministrativa.

Se le relazioni tra collettivi sono correttamente documentate, possono essere sfruttate dallo statistico per costruire nuove variabili. Ad esempio si può attribuire ad un evento di tipo immatricolazione la residenza dello studente immatricolato, ad uno studente immatricolato la data di immatricolazione, ad un esame il sesso dello studente che l'ha sostenuto, ad uno studente il numero totale di esami sostenuti.

Dal punto di vista statistico, quindi, in generale un giacimento di informazione amministrativa contiene informazioni su uno o più collettivi di tipo popolazione, corrispondenti ai soggetti dell'attività amministrativa, e su uno o più collettivi ad essi connessi, di tipo evento, corrispondenti agli oggetti dell'attività amministrativa, e sulle loro caratteristiche e relazioni alle quali possono corrispondere variabili d'interesse statistico. Può contenere informazioni relative direttamente a caratteristiche del territorio o a eventi direttamente connessi al territorio con le loro caratteristiche.

Per uniformarci ad un uso corrente chiamiamo d'ora in poi *archivi amministrativi* i giacimenti di informazione amministrativa così descritti, in termini di collettivi d'interesse statistico.

Per quanto detto su come si attiva e viene condotta la registrazione delle informazioni in tali archivi, possiamo affermare che *gli archivi amministrativi sono strutturati come registri relativi a uno o più collettivi di tipo popolazione corrispondenti ai soggetti dell'attività amministrativa (* (o direttamente al territorio) *e a collettivi ad essi legati di tipo evento***,** registri che rilevano con continuità i seguenti tipi di informazioni:

- tutti gli elementi di ciascun collettivo di tipo evento istantaneo, i quali sono legati a elementi dei collettivi di tipo popolazione (ad esempio Immatricolazione, Iscrizione, Esame, Laurea, Avvio rapporto di Lavoro, Ricovero ospedaliero), con le loro proprietà al momento in cui occorrono;
- tutti gli elementi di ciascun collettivo di tipo popolazione (ad esempio Studente, Ateneo, Lavoratore, Datore di lavoro, Degente), con le loro proprietà - caratteristiche e relazioni con altri collettivi - al loro ingresso nel registro;
- per ogni elemento di ciascun collettivo di tipo popolazione, tutti i cambiamenti relativi alle proprietà, fino alla sua eventuale uscita dal registro;
- tutti gli elementi di ciascun collettivo di tipo evento con durata, i quali sono legati ad elementi dei collettivi di tipo popolazione (ad esempio Carriera dello studente, Rapporto di lavoro, Degenza), con le loro proprietà al momento in cui iniziano;
- per ogni elemento di ciascun collettivo di tipo evento con durata, tutti i cambiamenti relativi alle proprietà, fino alla sua fine.

C'è una prima differenza pratica tra indagini e archivi amministrativi, relativa alla raccolta di informazioni sugli eventi. In un'indagine statistica nella maggior parte dei casi le informazioni relative agli eventi sono raccolte sfruttando il legame concettuale che lega sempre questi tipi di collettivi a collettivi di tipo popolazione, osservandole sugli elementi di una, o talvolta più, popolazioni che costituiscono le unità d'analisi principali dell'indagine.

Per ciò che riguarda gli archivi amministrativi, invece, questi possono essere riferiti, come le indagini, a collettivi di tipo popolazione (corrispondenti ai soggetti su cui si esercita l'attività amministrativa), ma perlopiù, in particolare nel caso di attività di erogazione di servizi, gli archivi amministrativi sono strutturati come registri relativi direttamente ad uno o più collettivi di tipo evento (ad esempio ricoveri in casa di cura, telefonate), che rilevano con continuità l'occorrenza degli elementi di tali collettivi, con le loro caratteristiche e relazioni con elementi di altri collettivi, tra le quali ci sono le relazioni con il loro elemento di riferimento in un collettivo di tipo popolazione (ad esempio, per i ricoveri in casa di cura, le persone).

In base alle precedenti considerazioni, possiamo precisare una definizione generale per la qualità dei dati contenuti negli archivi amministrativi, riferendola esplicitamente ai collettivi amministrativi d'interesse statistico che ne definiscono l'ontologia.

Un archivio amministrativo deve assicurare:

- la registrazione accurata e immediata di tutti gli elementi di ciascun collettivo di tipo evento istantaneo, i quali sono legati a elementi dei collettivi di tipo popolazione, con le loro proprietà al momento in cui occorrono;

- la registrazione accurata e immediata di tutti gli elementi di ciascun collettivo di tipo popolazione, con le loro proprietà - caratteristiche e relazioni con altri collettivi - al loro ingresso nel registro;
- per ogni elemento di ciascun collettivo di tipo popolazione, la registrazione accurata e immediata di tutti i cambiamenti relativi alle proprietà, caratteristiche e relazioni con altri collettivi, fino alla sua eventuale uscita dal registro;
- la registrazione accurata e immediata di tutti gli elementi di ciascun collettivo di tipo evento con durata, che sono legati ad elementi dei collettivi di tipo popolazione, con le loro proprietà al momento in cui iniziano;
- per ogni elemento di ciascun collettivo di tipo evento con durata, la registrazione accurata e immediata di tutti i cambiamenti relativi alle proprietà, fino alla sua fine.

Questa definizione è relativa alla qualità con la quale l'archivio amministrativo registra l'intero ciclo di vita di tutti i soggetti e oggetti coinvolti nell'attività amministrativa ed è quindi diacronica (o in altre parole longitudinale).

Si noti che in essa la qualità della registrazione è riconducibile in ultima analisi a:

- la corretta copertura dei collettivi degli eventi istantanei (corrispondenti a oggetti dell'archivio), i quali sono legati a elementi dei collettivi di tipo popolazione e si distinguono in:
    - collettivi di eventi istantanei legati a elementi dei collettivi di tipo popolazione che determinano l'ingresso e l'uscita dell'elemento legato dal collettivo di tipo popolazione, ad esempio Immatricolazione, Ricovero ospedaliero e/o l'inizio o la fine di eventi con durata;
    - altri collettivi di eventi istantanei legati a elementi dei collettivi di tipo popolazione, ad esempio Esame
    - particolari collettivi che raccolgono i cambiamenti di stato, cioè gli eventi corrispondenti all'acquisizione e alla perdita oppure al cambiamento delle proprietà – caratteristiche e relazioni – degli elementi dei collettivi di tipo popolazione o evento con durata: questi eventi solitamente non hanno interesse statistico diretto (a meno di particolari indagini longitudinali) ma solo per il cambiamento che determinano
- la corretta osservazione delle proprietà – caratteristiche e relazioni - degli elementi dei collettivi di eventi istantanei
- la corretta copertura dei collettivi di tipo popolazione (corrispondenti a soggetti dell'archivio) o evento con durata (corrispondenti ad oggetti dell'archivio), la quale dipende dalla corretta copertura dei collettivi degli eventi istantanei di ingresso/inizio o uscita/fine, e dalla corretta osservazione delle loro proprietà
- la corretta osservazione delle proprietà – caratteristiche e relazioni - degli elementi dei collettivi di tipo popolazione o evento con durata in ogni momento della loro esistenza, la quale dipende anche dalla corretta copertura dei collettivi che raccolgono i cambiamenti di stato.

Per quanto sopra le definizioni della copertura dei collettivi sono riferite all'intera storia dell'archivio amministrativo, storia risultante dalla sua alimentazione che è continua.

In linea di principio, infatti, in un archivio amministrativo ogni collettivo registra tutti gli elementi occorsi dalla data di costituzione dell'archivio fino alla data dell'ultimo aggiornamento

in ordine di tempo, ciascuno con il proprio momento (se istantaneo) o periodo di riferimento, ed è quindi definito in modo longitudinale.

La corretta copertura è la corretta registrazione di tali elementi con i loro riferimenti temporali.

Come discusso nell'ultimo paragrafo, dedicato alle modalità temporali di osservazione, l'utilizzo statistico dell'informazione gestita nell'archivio amministrativo comporta l'estrazione di una sezione dell'archivio, più o meno longitudinale in funzione delle esigenze conoscitive, ma comunque perlopiù ancorata a uno specifico momento o periodo di osservazione.

Da questo punto di vista è spesso operativamente sufficiente una definizione dei vari aspetti della qualità ancorata a un singolo momento o periodo di osservazione; è bene però tenere presente che questa definizione è sempre derivata da quella qui sopra specificata, basata sulla corretta registrazione dell'intero ciclo di vita di ogni elemento registrato nell'archivio, in termini degli eventi istantanei di ingresso/inizio e uscita/fine, e dei cambiamenti delle caratteristiche o relazioni.

Dall'informazione amministrativa all'informazione statistica: gli utilizzi dell'informazione amministrativa

Una classificazione degli utilizzi dell'informazione di origine amministrativa ancorata alle fasi della produzione del dato statistico è quella proposta nell'ambito dello Statistical Network on Administrative Data (si veda l'Appendice 1):

1. Creazione e gestione di registri e liste (survey frames): comprende la creazione di registri così come di liste ausiliarie alle indagini come le liste di campionamento, le liste di rilevazione

2. Definizione del disegno campionario

3. Sostituzione della rilevazione diretta, anche mediante integrazione di diverse fonti riferite allo stesso collettivo

4. Controllo e correzione (ad esempio mediante uso di dati ausiliari per l'imputazione)

5. Stime indirette (anche mediante stime da modello, predizione) e ponderazione

6. Tabulazione diretta

7. Validazione (di dati da indagine o di altri archivi) e confronto di dati

L'utilizzo di un archivio amministrativo, sia diretto che in funzione ausiliaria a un'indagine, comporta comunque un ulteriore passo, cioè il confronto tra la definizione dei collettivi amministrativi d'interesse statistico che lo descrivono e l'unità d'analisi, o le unità d'analisi, che sono caratteristiche del processo statistico che lo utilizza o che comunque, nel caso di registri, sono di interesse per gli statistici.

L'unità d'analisi di un processo statistico è un collettivo di elementi osservabili che è stato scelto come collettivo d'interesse statistico per il dato processo, delimitandolo mediante una definizione rispondente agli scopi dello studio.

In un processo statistico è sempre definito un collettivo principale d'interesse che nella maggior parte dei casi è di tipo popolazione e costituisce uno specifico sottoinsieme, individuato dalla definizione, delle due popolazioni di base, la popolazione delle persone e la popolazione degli organismi che svolgono attività economica. Questo collettivo costituisce l'unità d'analisi principale del processo, alla quale ad esempio riferire il piano di campionamento.

Possono essere individuati altri collettivi ad esso connessi, di tipo popolazione o evento (si pensi ad esempio alle indagini multiscopo). Tutti questi collettivi costituiscono le unità d'analisi del processo statistico.

Proprio per permettere il confronto tra le unità d'analisi dei processi statistici (indagini, elaborazioni, statistiche da fonte amministrativa organizzata) e le unità di riferimento dei registri, da una parte, e i contenuti degli archivi amministrativi, dall'altra, abbiamo introdotto nel precedente paragrafo il concetto di collettivo amministrativo d'interesse statistico, presentandone una prima tipizzazione il più possibile aderente a quella tipica dei processi statistici.

Una più dettagliata discussione dei contenuti informativi degli archivi mirata a permettere il confronto con i contenuti dei processi statistici è presentata nell'apposita Appendice dedicata all'ontologia degli archivi amministrativi.

Nel caso di un processo statistico che intenda utilizzare un archivio amministrativo, a seconda dei risultati del confronto tra le unità d'analisi del processo statistico, da una parte, e i collettivi amministrativi d'interesse statistico dell'archivio amministrativo che si intende utilizzare, dall'altra, si avrà che l'unità d'analisi potrà coincidere con uno dei collettivi amministrativi d'interesse statistico che caratterizzano l'archivio amministrativo utilizzato oppure potrà essere da questi derivata, mediante elaborazioni semplici, come l'estrazione di un sottoinsieme, o più complesse, tra le quali anche il linkage esatto o probabilistico di elementi appartenenti allo stesso collettivo registrati in archivi diversi.

Analoghe considerazioni valgono per il confronto tra le proprietà dei collettivi amministrativi d'interesse statistico utilizzati e le variabili obiettivo dei processi statistici utilizzatori o gestite dai registri.

Anzitutto in generale l'archivio amministrativo utilizza sue proprie specifiche classificazioni per registrare le caratteristiche osservate sui singoli elementi dei collettivi amministrativi d'interesse statistico.

Affinché queste caratteristiche diano luogo a variabili utilizzabili a scopo statistico è in primo luogo necessario che tali classificazioni siano qualitativamente adeguate allo scopo.

In particolare, per le caratteristiche più generali e utilizzate nella gran parte degli archivi (ad esempio luogo nascita e di residenza, livello di istruzione, condizione professionale per le persone, forma giuridica, attività economica per gli organismi che svolgono attività economica) tali classificazioni devono essere standard o riportabili a standard.

Come avviene per i collettivi poi, le variabili utili a scopo statistico possono riferirsi direttamente a caratteristiche registrate dall'archivio per gli elementi dei diversi collettivi o essere costruite combinando opportunamente le caratteristiche e anche le relazioni registrate, anche mediante operatori logici e aritmetici.

Dall'informazione amministrativa all'informazione statistica: le modalità temporali di osservazione

Le modalità temporali di osservazione dei collettivi d'interesse statistico sono una componente importante di un'indagine o di un'elaborazione statistica, che concorre a definirla come specifico contesto d'osservazione.

In generale si osserva che:

- spesso l'indagine ha come obiettivo l'istantanea dello stato di tutti gli elementi di un collettivo di tipo popolazione al tempo t, intendendo per stato il possesso di proprietà in quel momento, cioè di caratteristiche e relazioni con altri elementi (esempio: appartenenza di un componente a una famiglia al 30 ottobre 2012, professione di un componente la famiglia al 30 ottobre 2012, residenza di una famiglia al 30 ottobre 2012);

- per osservare gli eventi istantanei d'altra parte occorre sempre definire un periodo di osservazione (esempio: ricoveri ospedalieri di un componente la famiglia tra 30 ottobre 2011 e 30 ottobre 2012); nelle indagini che osservano tanto le caratteristiche di una popolazione al tempo t quanto gli eventi relativi agli elementi della popolazione si definisce in genere un periodo di osservazione per gli eventi ancorato al momento t d'osservazione per lo stato (esempio, ricoveri ospedalieri nell'ultimo anno);

- gli eventi con durata (ad esempio Carriera dello studente, Degenza) hanno caratteristiche intermedie che possono suggerire scelte diverse a seconda di diversi fattori, tra i quali: la durata effettiva che può essere consistente o trascurabile; il fatto che, come le popolazioni, se hanno una durata consistente i loro elementi possono cambiare proprietà nel corso del loro ciclo di vita; infine, soprattutto, l'obiettivo dell'osservazione.

Ad esempio in una Multiscopo si possono rilevare le vacanze effettuate da una famiglia in un anno, come se le vacanze fossero eventi istantanei anche se hanno una durata, o all'opposto si possono rilevare le proprietà al tempo t delle degenze ospedaliere che sono in corso in quel momento trascurandone la storia; è facile osservare che se un evento con durata può cambiare le sue proprietà nel corso del proprio ciclo di vita considerarlo come un evento istantaneo comporta specifiche scelte, come non rilevare le proprietà che cambiano o fissare per esse un riferimento temporale, rilevandole ad esempio al momento d'inizio o al momento di fine.

Infine un'indagine può essere specificamente mirata a studiare proprio i cambiamenti delle proprietà possedute dagli elementi di una popolazione (è il caso delle indagini longitudinali), ad esempio il cambiamento di professione, il cambiamento di famiglia d'appartenenza: come si è osservato nel paragrafo precedente questi eventi di cambiamento sono anch'essi eventi istantanei e quindi per queste indagini si definisce un periodo di osservazione, oltre a modalità d'osservazione che consentano di costruire un effetto di continuità d'osservazione.

Ciò suggerisce che gli archivi amministrativi possano costituire un patrimonio utile per ogni osservazione dei fenomeni di tipo longitudinale, in quanto come si è detto risultanti da un'osservazione continua che in linea di principio, a meno di considerazioni di qualità, rispecchia esattamente "quello che succede" agli elementi dei propri collettivi di riferimento.

In alternativa, e più frequentemente, gli archivi amministrativi saranno utilizzati per ricavare stime ancorate a riferimenti temporali definiti in analogia a quelli più comunemente utilizzati

nelle indagini, quindi per osservare lo stato di una popolazione al tempo t o gli eventi intercorsi in un periodo.

In questo caso per ciascuno dei collettivi amministrativi d'interesse sarà definito un momento o un periodo d'osservazione, con gli stessi criteri utilizzati nell'osservazione diretta brevemente discussi qui sopra, effettuando così una particolare "estrazione" rispetto all'informazione raccolta con continuità dall'archivio.

Ad esempio da un archivio sugli studenti universitari si potrà scegliere di estrarre informazioni di stato, ad esempio la residenza dello studente al 30 ottobre 2012, il corso di laurea d'appartenenza dello studente al 30 ottobre 2012 o informazioni relative agli eventi occorsi in un periodo come le iscrizioni a corsi di laurea tra 1 settembre 2011 e 1 settembre 2012.

In effetti da un archivio amministrativo si possono estrarre informazioni a diverso "grado di longitudinalità" a seconda delle esigenze, andando dall'elaborazione di informazioni ancorate a momenti dati t, come nelle indagini, fino, come caso estremo, a elaborare con continuità le informazioni relative all'intero ciclo di vita degli elementi delle popolazioni d'interesse.

PARTE SECONDA: INDICATORI DI QUALITA' RELATIVI ALLA FONTE E ALLA SUA DOCUMENTAZIONE

Il Framework Istat di indicatori per la documentazione della qualità delle fonti amministrative

Nell'ambito della statistica ufficiale, molta parte della riflessione dedicata alla valutazione della qualità dell'informazione di fonte amministrativa è indirizzata verso la definizione di sistemi strutturati di indicatori di qualità che mirano a documentare tutti gli aspetti della qualità, partendo dalla usabilità della fonte e arrivando fino alla qualità dei dati prodotti, quest'ultima spesso definita facendo ricorso a concetti utilizzati per le indagini, come la copertura dei collettivi e la precisione di misura delle variabili.

In particolare Statistics Netherlands [2] propone un sistema di indicatori qualitativi e quantitativi mirato a valutare la qualità delle fonti amministrative in quanto input di specifici processi statistici che le utilizzano, che organizza gli indicatori in accordo a un sistema di iperdimensioni e dimensioni, precisamente in tre iperdimensioni della qualità, ciascuna delle quali comprende dimensioni e relativi indicatori. Le iperdimensioni identificate sono Fonte, Documentazione, Dati.

La prima iperdimensione attiene a tutte le caratteristiche che riguardano nel suo complesso la Fonte, laddove si intende l'autorità responsabile e le modalità di fornitura del dato amministrativo. L'iperdimensione relativa alla Documentazione riguarda invece i metadati strutturali e referenziali che descrivono i dati. Infine l'iperdimensione dei Dati riguarda la qualità intrinseca dei dati dell'archivio.

Per la Fonte (o Documentazione della fonte), le dimensioni identificate sono: 1. Fornitore, 2. Rilevanza e usi, 3. Privacy e sicurezza, 4. Disponibilità.

Per la Documentazione dei dati e delle procedure le dimensioni sono: 1. Chiarezza e standardizzazione della documentazione dei dati, 2. Confrontabilità della documentazione dei dati, 3. Chiavi identificative univoche e chiavi di raccordo (esistenza, tipo e qualità), 4. Documentazione delle procedure di acquisizione e trattamento dei dati.

Infine per i Dati le dimensioni sono: 1. Identificazione, Copertura che include Sovracopertura e Sottocopertura, Integrabilità effettiva (Linkabilità), 2. Accuratezza (che include valori mancanti, errori di misura, inconsistenze e valori dubbi), 3. Dimensione Temporale (è da valutare l'effettiva necessità di una dimensione separata)

Sulla base di tale modello l'Istat ha definito un proprio Framework di indicatori per la documentazione della qualità delle fonti amministrative ai fini dell'utilizzo statistico, indipendentemente da ogni specifico processo statistico utilizzatore.

Non tutte le dimensioni proposte nel suddetto modello vengono incorporate nel modello italiano, ma solo quelle coerenti con gli obiettivi specifici, ossia fornire indicatori per descrivere e valutare l'utilizzabilità e la qualità della fonte per finalità statistiche generiche, senza prendere in considerazione utilizzi specifici.

[2] P. Daas, S. Ossen,R. Vis-Visschers, J. Arends-Toth, "Checklist for the Quality evaluation of Administrative Data Sources", CBS 2009

Alcune dimensioni vengono modificate in relazione al diverso contesto (per es. non si può parlare di "fornitura", e questa dimensione viene sostituita da una di "disponibilità"). Inoltre, a volte queste vengono sostanziate da indicatori, altre volte solo da indicazioni di tipo qualitativo e in generale metadati descrittivi. Per quanto riguarda gli indicatori, questi sono in parte sovrapponibili a quelli definiti in Blue-Ets, in particolare quegli indicatori che in questo progetto attengono alla qualità dell'input, anche se organizzati secondo una logica più consona al modello delle iperdimensioni e dimensioni adottato.

Per quanto osservato nell'Introduzione, di seguito si presentano solo gli indicatori del Framework Istat relativi alle dimensioni Fonte e Documentazione.

IPERDIMENSIONE FONTE: Indicatori

Dimensione 1. Fornitore

I metadati da acquisire e manutenere includono:

1.1. Nome archivio

1.2. Versione e data dell'archivio

1.3. Denominazione dell'ente titolare della fonte

1.4. Responsabile dell'AA interno all'autorità: Nome e struttura di appartenenza

1.5. Referente interno all'autorità: Nome e struttura di appartenenza

1.6. Altri enti coinvolti nella gestione dell'archivio

1.7. Ambito tematico di riferimento (uno o più fenomeni di pertinenza)

1.8. Riferimenti normativi

1.9. Contesto, ossia descrizione delle procedure amministrative che, da una parte, implementano la norma e quindi ne sono influenzate e, dall'altra, determinano le modalità di alimentazione dell'archivio, con particolare riferimento a:

- Carattere della registrazione (obbligatoria o facoltativa)
- Dimensione temporale: periodicità dell'aggiornamento e riferimento temporale dei dati amministrativi contenuti nell'archivio.

  Questo aspetto può essere diverso per i diversi collettivi di tipo evento che caratterizzano l'archivio. La dimensione temporale deve essere documentata per i principali tra questi collettivi, vale a dire quelli di maggior interesse per lo statistico.

  In una successiva versione del Framework potranno essere definiti criteri di prevalenza o rilevanza finalizzati alla costruzione di un indicatore sintetico capace di descrivere la

dimensione temporale per l'archivio nel suo complesso, tenendo conto dell'insieme dei collettivi di tipo evento che lo caratterizzano.

- Dimensione spaziale: omogeneità della normativa e delle procedure sul territorio

  Si può supporre infatti che, laddove la gestione e manutenzione dell'AA risieda presso un'autorità univocamente e chiaramente identificabile e vi sia omogeneità nella normativa e nelle procedure di alimentazione dell'archivio, il contesto per l'uso attuale e potenziale dell'archivio sia maggiormente favorevole rispetto a situazioni in cui vi siano responsabilità sparse e regolamentazione eterogenea sul territorio.

Molte delle informazioni qui elencate sono desumibili dalle Istruttorie condotte dal Nucleo Tecnico di supporto alla Commissione sulla modulistica amministrativa dell'Istat

Dimensione 2. Rilevanza e usi

2.1. Utilizzi aggiuntivi dell'archivio da parte dell'ente (supporto alle decisioni, supporto a indagini statistiche, elaborazioni statistiche,…)

2.2. Utilizzatori per finalità statistiche e descrizione sintetica delle statistiche prodotte

2.3. Usi potenziali che l'autorità responsabile identifica come possibili

Infatti, se l'archivio è già utilizzato per finalità statistiche, questo può riflettere un terreno maggiormente favorevole per ulteriori usi. In questa dimensione si documentano gli usi attuali internamente all'Ente proprietario e da parte di altri utilizzatori e gli usi potenziali che l'ente proprietario vede come possibili.

Anche le informazioni qui elencate sono desumibili dalle Istruttorie condotte dal Nucleo Tecnico di supporto alla Commissione sulla modulistica amministrativa dell'Istat

Dimensione 3. Privacy e sicurezza informatica

3.1. Normativa di riferimento per il trattamento dei dati sensibili presso l'ente

Dimensione 4. Disponibilità dell'archivio

4.1. Esistenza di rilasci dell'archivio per usi statistici

4.2. Formato tecnico e Standard di trasmissione per rilasci già effettuati

4.3. Date e periodicità del possibile rilascio

4.4. Tempestività del rilascio: data di possibile disponibilità archivio – periodo riferimento dati archivio (ultimo giorno).

Poiché, come precisato al punto 1.9, esistono diversi periodi di riferimento dei dati per diversi eventi dell'archivio, anche in questo caso occorre documentare la tempestività del rilascio per i principali tra questi collettivi, vale a dire quelli di maggior interesse per lo statistico.

In una successiva versione del Framework potranno essere definiti criteri di prevalenza o rilevanza finalizzati alla costruzione di un indicatore sintetico capace di descrivere la tempestività del rilascio per l'archivio nel suo complesso, tenendo conto dell'insieme dei collettivi di tipo evento che lo caratterizzano.

4.5. Accessibilità dell'archivio: eventuali procedure e condizioni per accedere all'AA

Si ipotizza che, se l'archivio è già rilasciato per usi statistici, la sua utilizzabilità sia maggiore e, inoltre, che l'ente abbia interesse a utilizzare condizioni e formati tecnici già adottati. Inoltre è importante iniziare ad esplorare la dimensione temporale attraverso gli indicatori 4.3. e 4.4.

IPERDIMENSIONE DOCUMENTAZIONE DEI DATI E DELLE PROCEDURE: Indicatori

Dimensione 1. Chiarezza e standardizzazione della documentazione dei dati

2.1.Disponibilità di documentazione del contenuto dell'archivio almeno per i principali collettivi e le principali caratteristiche/variabili. Ciò implica la disponibilità delle definizioni, o delle condizioni necessarie di appartenenza, associate ai principali collettivi e, quando pertinente, alle principali variabili, e la specifica delle classificazioni utilizzate per osservare le principali variabili.

2.2. Grado relativo di completezza della documentazione rispetto alla totalità dei collettivi, caratteristiche e variabili, classificazioni.

2.3. Descrizione dell'eventuale modello concettuale utilizzato per specificare la documentazione.

2.4. Grado di aggiornamento della documentazione a fronte di modifiche nella struttura e contenuto dell'archivio.

Dimensione 2. Confrontabilità della documentazione dei dati

2.5. Accuratezza della documentazione delle definizioni dei collettivi di riferimento dell'archivio: in particolare, livello di dettaglio fino al quale sono specificati tutti i sottocasi inclusi ed esclusi in tali definizioni.

2.6. Disponibilità di documentazione che permetta di valutare l'effettiva corrispondenza o raccordabilità con i principali collettivi, variabili e classificazioni ufficiali.

Dimensione 3. Chiavi identificative univoche e chiavi di raccordo (esistenza, tipo e qualità)

3.1. Esistenza delle chiavi identificative e documentazione della loro struttura. Segnalare se nell'archivio esistono collettivi di tipo popolazione o evento per i quali non esiste una chiave identificativa univoca o non è documentata la sua struttura, assegnare un valore sintetico all'indicatore sulla base del numero di tali collettivi e della loro rilevanza

3.2. Qualità delle chiavi identificative: giudizio generale sulla qualità della chiave identificativa (frequenza della scorrettezza sintattica o semantica, mancata stabilità temporale) sia nel ruolo di identificazione che in quello di raccordo.

Questo aspetto può essere diifferenziato per i diversi collettivi di tipo popolazione o evento che caratterizzano l'archivio. La qualità delle chiavi deve essere documentata per i principali tra questi collettivi, vale a dire quelli di maggior interesse per lo statistico.

In una successiva versione del Framework potranno essere definiti criteri di prevalenza o rilevanza finalizzati alla costruzione di un indicatore sintetico capace di descrivere la qualità delle chiavi identificative per l'archivio nel suo complesso, tenendo conto dell'insieme dei collettivi che lo caratterizzano.

3.3. Grado di coincidenza o mappabilità delle chiavi identificative utilizzate con chiavi utilizzate in altri archivi (codici fiscali, partite iva ….): giudizio generale sulla coincidenza o mappabilità della chiave.

Questo aspetto può essere differenziato per i diversi collettivi di tipo popolazione o evento che caratterizzano l'archivio. La coincidenza o mappabilità delle chiavi identificative con chiavi utilizzate in altri archivi deve essere documentata per i principali tra questi collettivi, vale a dire quelli di maggior interesse per lo statistico.

In una successiva versione del Framework potranno essere definiti criteri di prevalenza o rilevanza finalizzati alla costruzione di un indicatore sintetico capace di descrivere la coincidenza o mappabilità delle chiavi identificative con chiavi utilizzate in altri archivi per l'archivio nel suo complesso, tenendo conto dell'insieme dei collettivi che lo caratterizzano.

L’esistenza di chiavi identificative univoche e condivise da altri archivi, permettendo la linkabilità dei dati contenuti nell’AA con dati presenti in altri archivi, è condizione favorevole per l’utilizzabilità dell’AA a fini statistici.

Dimensione 4. Documentazione delle procedure di acquisizione e trattamento dei dati

4.1.Disponibilità di una specifica delle procedure di acquisizione e trattamento dei dati dell’archivio

La descrizione dei controlli che vengono messi in atto da parte dell’ente amministrativo in fase di acquisizione del dato amministrativo e delle verifiche ed eventuali correzioni dei dati amministrativi mancanti o incongruenti sono aspetti fondamentali per comprendere l’utilizzabilità dell’archivio a fini statistici.

PARTE TERZA: UN NUOVO APPROCCIO ALLA VALUTAZIONE DELLA QUALITA' DEI DATI DI FONTE AMMINISTRATIVA

INTRODUZIONE

Linee generali dell'approccio seguito: l'ontologia dell'archivio e gli errori possibili

Si è detto che le diverse fonti d'informazione attuano processi di osservazione della realtà e raccolta di informazione che comportano propri meccanismi di generazione dell'errore, tutti fino ad oggi ancora da descrivere su un piano pratico e formalizzare su un piano teorico.

Un primo passo in questa direzione è la descrizione sistematica da un punto di vista pratico, che sia il più generale possibile, delle *diverse tipologie di errore* che possono influenzare la qualità dell'informazione prodotta da una fonte d'informazione generica, il cui processo d'osservazione cioè non sia pianificato per la statistica, come avviene per le indagini.

Nel nostro approccio si individuano le diverse tipologie di errore possibili a partire da una specifica concettuale delle diverse *tipologie di informazione* che possono caratterizzare la generica fonte d'informazione, in particolare un archivio amministrativo. Tale specifica concettuale a sua volta si basa su una disamina di quali sono le *componenti tipiche dell'ontologia di una fonte di informazione*.

Precisamente, l'idea alla base del lavoro di definizione e classificazione degli errori illustrato nel presente documento è quella di *definire rigorosamente, per ogni informazione acquisita da una fonte, le condizioni di errore, in termini di acquisizione di informazione falsa o mancata acquisizione di informazione vera*. Ciò richiede due operazioni di concettualizzazione:

- anzitutto distinguere le diverse tipologie di informazioni acquisite da una fonte, in particolare da un archivio amministrativo
- per ogni tipologia, definire rigorosamente le nozioni di acquisizione di informazione falsa o mancata acquisizione di informazione vera, cioè i possibili errori.

La prima operazione di concettualizzazione si attua *definendo un'ontologia per la fonte d'informazione, in particolare l'archivio amministrativo*, basata su una serie di criteri per individuare e distinguere le diverse entità osservate, vale a dire gli *oggetti dell'archivio* intesi in senso lato, comprendendo soggetti e oggetti dell'attività amministrativa, con le loro proprietà. Questi criteri congiuntamente costituiscono in pratica un modello concettuale ispirato a quelli comunemente utilizzati in informatica, ma orientato all'utilizzo statistico dell'informazione.

La seconda operazione di concettualizzazione consiste nel *fare in modo che per tutte le informazioni riferibili ad ogni oggetto osservato si possa valutare una condizione di verità o falsità*, in modo da poter distinguere le diverse eventualità: acquisizione di informazione vera, acquisizione di informazione falsa o mancata acquisizione di informazione vera, e quindi poter ragionare sulla diagnosticabilità delle ultime due eventualità, che configurano errori.

Di seguito riassumiamo in breve queste due operazioni di concettualizzazione. Tutti i concetti qui introdotti sono estesamente illustrati nel capitolo seguente.

L'ontologia dell'archivio e le diverse tipologie d'informazione

L'analisi condotta nella PRIMA PARTE suggerisce quali sono le *componenti tipiche dell'ontologia di una fonte di informazione* e cioè, in generale:

- collettivi di tipo popolazione o di tipo evento, istantaneo o con durata
- caratteristiche associate agli elementi dei collettivi (che danno luogo a variabili una volta considerate da un punto di vista statistico), ciascuna delle quali utilizza:
    - una classificazione costituita da diverse modalità, se di classificazione
    - un dominio/range costituito da diversi valori, se numerica
- relazioni 1-n (oppure 1-1) [3] tra due elementi di collettivi (due diversi collettivi o lo stesso collettivo) [4].

Spesso una fonte d'informazione osserva un certo numero di *collettivi principali* e, per ogni collettivo principale, una *gerarchia di collettivi sottoinsieme* sui quali vengono osservate specifiche caratteristiche e relazioni.

Nei capitoli seguenti il contenuto dell'archivio e gli errori possibili sono analizzati con riferimento ai soli collettivi principali, i concetti introdotti sono comunque facilmente estendibili al caso di collettivi sottoinsieme, come sarà mostrato in un apposito paragrafo nella PARTE QUINTA.

Le diverse *tipologie di informazione* che nel corso del tempo vengono raccolte da una fonte, in particolare da un archivio amministrativo, sono allora descrivibili come *diverse tipologie di enunciati relativi a tali componenti dell'ontologia della fonte e riferiti a specifici elementi*, e quindi:

- enunciati che asseriscono l'appartenenza di un elemento ad un collettivo
- enunciati che asseriscono l'esistenza di un'associazione tra un elemento di un collettivo e una modalità, o valore, tra quelli assumibili per una particolare caratteristica, in quanto appartenenti alla classificazione o al dominio/range utilizzato per la caratteristica
- enunciati che asseriscono l'esistenza di una relazione tra due elementi di collettivi.

E' importante sottolineare che tale concettualizzazione, poiché vede le informazioni raccolte come enunciati relativi a singoli elementi o a loro associazioni, presuppone un contesto che prendendo a prestito la terminologia statistica si può definire discreto, nel quale cioè si osservano lo stato e le caratteristiche di singoli elementi piuttosto che misurare gli stati di un'entità continua.

Questa concettualizzazione non si presta perciò a caratterizzare le informazioni riferite direttamente al territorio il quale, come precedentemente osservato, è naturalmente descrivibile come un collettivo con un'estensione continua, non costituita di singoli elementi, a meno di non adottare opportune convenzioni.

Si potrebbe osservare peraltro che tutte le informazioni riferite ad elementi di popolazioni e relativi eventi possono essere concettualizzate come misure riferite al territorio su cui le popolazioni insistono.

Di fatto le informazioni relative a una stessa realtà osservata si possono concettualizzare in entrambi i modi a seconda delle esigenze. L'ipotesi di un contesto d'osservazione costituito di elementi

[3] Nel nostro approccio le relazioni m-n tra due elementi e le relazioni tra più elementi sono considerate come particolari collettivi, di tipo associativo.
[4] In una relazione 1-n i due collettivi, o lo stesso, assumono i due ruoli distinti di dominio e codominio della relazione: ogni elemento del dominio può essere legato a un singolo elemento del codominio, ogni elemento del codominio può essere legato a più elementi del dominio.

singolarmente osservabili è un presupposto comunemente assunto tanto nella gestione di archivi amministrativi quanto nella statistica ufficiale.

Nella concettualizzazione proposta ci si attiene perciò a questo presupposto, rimandando ad un'apposita trattazione l'esame delle conseguenze differenziali che comporta l'assumere un contesto di osservazione discreto piuttosto che continuo, nel senso che abbiamo precedentemente attribuito a queste espressioni.

Le diverse tipologie di enunciati sono quindi riferite a *singoli elementi opportunamente identificati*, precisamente:

- gli enunciati che asseriscono l'appartenenza di un elemento ad un collettivo sono *riferiti a*:
  - *un singolo elemento del collettivo* individuato da un proprio *identificativo univoco*
- gli enunciati che asseriscono l'associazione tra un elemento di un collettivo e una modalità tra quelle assumibili per una particolare caratteristica sono *riferiti a*:
  - *un singolo elemento del collettivo* individuato da un proprio *identificativo univoco*
  - *una modalità, o valore, appartenente alla classificazione o al dominio/range utilizzato per la caratteristica,* anch'esso individuato da un *proprio identificativo univoco*
- gli enunciati che asseriscono l'esistenza di una relazione tra due elementi di collettivi (due diversi collettivi o lo stesso collettivo) sono riferiti ai due elementi legati, ciascuno individuato da un proprio identificativo univoco, precisamente sono *riferiti a*:
  - *un singolo elemento del collettivo dominio* individuato dal *proprio codice identificativo*
  - *un singolo elemento del collettivo codominio* individuato dal *proprio codice identificativo* utilizzato come *codice di raccordo* [5].

Una discussione approfondita della concettualizzazione utilizzata è presentata nell'Appendice 3. "Documentare l'ontologia di un archivio amministrativo".

Le diverse tipologie d'informazione e gli errori possibili

Gli errori possibili sono quindi definiti come *errori relativi all'accettazione nella fonte dei diversi tipi di enunciati riferiti a specifici elementi e relativi all'appartenenza ad un collettivo, al possesso di una caratteristica, all'esistenza di una relazione*, quindi essenzialmente:

- accettazione di enunciati falsi
- mancata accettazione di enunciati veri
- ma anche accettazione ripetuta di enunciati (duplicazione).

Per ciò che riguarda gli *errori sugli identificativi*, l'assegnazione di identificativi agli elementi dei collettivi di tipo popolazione, in quanto operazione connessa alla registrazione in archivio di nuovi elementi, è comunemente considerata in molti Framework esistenti una possibile causa di errore che si combina con gli errori connessi all'inclusione del nuovo elemento in archivio.

La precedente disamina dei diversi tipi di enunciati registrabili e del ruolo degli identificativi nella loro formulazione suggerisce di considerare in modo più sistematico le possibilità di errori sugli

[5] Operativamente l'informazione sul possesso di una relazione è registrata con riferimento ad un elemento del collettivo dominio, individuando l'elemento del codominio legato mediante il suo identificativo utilizzato in funzione di codice di raccordo.

identificativi, esplicitando i possibili errori sugli identificativi degli elementi dei collettivi di qualsiasi tipo, nonché sugli identificativi quando vengono utilizzati come codici di raccordo e anche, in teoria, sugli identificativi utilizzati per distinguere le diverse modalità, o valori, assumibili da una caratteristica.

E' uso poi distinguere tra errori sintattici e semantici sugli identificativi.

Gli *errori sintattici* riguardano la presenza effettiva o l'esatta costruzione dell'identificativo e sono sempre evidenti e diagnosticabili: è il modo in cui poi vengono risolti che può influenzare la qualità della fonte.

Gli *errori semantici* sono errori relativi all'effettiva capacità di identificare. Infatti per tutti i tipi di identificativo si può dire che:

- un identificativo deve permettere di individuare univocamente l'elemento (che può essere l'elemento di un collettivo, una modalità, un valore) cui è attribuito

quindi gli errori semantici sugli identificativi sono quelli che compromettono questo requisito, cioè:

a) nel sistema di identificazione adottato, alcuni elementi non hanno assegnato un proprio identificativo
b) nel sistema di identificazione adottato, ci sono identificativi condivisi tra più elementi
c) nel sistema di identificazione adottato, ci sono identificativi per elementi non esistenti
d) nel sistema di identificazione adottato, ci sono elementi che hanno un doppio identificativo
e) nel sistema di identificazione adottato, ci sono elementi che ricevono identificativi diversi in momenti diversi (analogo al precedente).

E' immediato osservare che per gli *identificativi delle modalità, o valori*:

- non esiste un errore semantico, in quanto le modalità e i valori appartengono a un insieme prefissato a priori di elementi pre-identificati, appunto la classificazione o il dominio/range utilizzato
- il relativo errore sintattico consiste nell'uso di una modalità o di un valore non previsto, ed è comunemente diagnosticato come modalità non prevista per la classificazione utilizzata, o valore fuori range.

Vanno poi indagate *le diverse possibilità di combinazione tra errori relativi ai diversi tipi di enunciati ed errori di identificazione*, precisamente:

- per gli *enunciati che asseriscono l'appartenenza di un elemento ad un collettivo*, le possibili combinazioni tra errori di accettazione nella fonte dell'enunciato ed errori sintattici o semantici di identificazione dell'elemento
- per gli *enunciati che asseriscono l'associazione tra un elemento di un collettivo e una modalità, o valore*, tra quelli assumibili per una particolare caratteristica, le possibili combinazioni tra errori di accettazione nella fonte dell'enunciato ed errori sintattici o semantici di identificazione dell'elemento nonché errori sintattici di identificazione della modalità, o valore (modalità non prevista o valore fuori range)
- per gli *enunciati che asseriscono l'esistenza di una particolare relazione tra due elementi di collettivi* (due diversi collettivi o lo stesso collettivo) le possibili combinazioni tra errori di accettazione nella fonte dell'enunciato ed errori sintattici o semantici di identificazione sia dell'elemento di riferimento del dominio sia dell'elemento di riferimento del codominio (in

questo secondo caso, si tratta dell'errore nel codice identificativo utilizzato come codice di raccordo).

Infine, tutti i tipi di enunciati fin qui introdotti hanno un *riferimento temporale*, precisamente:

- un istante t_i per gli enunciati relativi ad elementi dei collettivi di eventi istantanei
- una durata $t_i - t$ oppure $t_i - t_j$, per gli enunciati relativi ad elementi dei collettivi di tipo popolazione o agli eventi con durata, i quali possono essere riferiti ad un periodo aperto $t_i - t$ o chiuso $t_i - t_j$..

Un'ulteriore fonte di errore è originata dall'eventuale differenza tra i riferimenti temporali effettivi degli enunciati e i riferimenti temporali per essi registrati in archivio.

I contenuti della PARTE TERZA del Framework

L'approccio che è stato esposto per grandi linee può apparire astratto a prima vista ma, proprio per questo, offre un fondamento analitico più efficace e generalizzato all'analisi dei diversi aspetti della qualità della generica fonte d'informazione, anche perché si presta a una formalizzazione logica che è nella tradizione della probabilità e della statistica, anche se spesso non esplicitata nei testi più scolastici.

Per sfruttare al massimo questo approccio, nella presente PARTE TERZA del Framework viene prima introdotta una specifica semi-formale, basata sull'ontologia d'archivio, dei contenuti informativi di un archivio amministrativo e della loro dinamica di aggiornamento; sono poi individuate le diverse categorie di errori possibili e le loro combinazioni nel modo più generale, come possibilità teoriche, delineando anche le possibilità di diagnosi per tali categorie di errori.

Nel capitolo I CONTENUTI INFORMATIVI DELL'ARCHIVIO AMMINISTRATIVO viene sommariamente introdotta la specifica dell'ontologia di un archivio amministrativo (illustrata più in dettaglio nell'apposita Appendice), specificando poi i diversi tipi di enunciati accettabili in archivio e la loro organizzazione in record.

Nel capitolo LA DINAMICA DI AGGIORNAMENTO DELL'ARCHIVIO AMMINISTRATIVO viene descritta la dinamica di aggiornamento dell'archivio, in termini di accettazione e modifica di record.

Nel capitolo I DIVERSI TIPI DI ERRORE sono individuate le diverse categorie di errori possibili e le loro combinazioni.

Nel capitolo LE POSSIBILITA' DI DIAGNOSI PER I DIVERSI TIPI DI ERRORE vengono sinteticamente delineate le diverse possibilità di diagnosi per ogni tipo di errore.

I CONTENUTI INFORMATIVI DELL'ARCHIVIO AMMINISTRATIVO

I contenuti informativi dell'archivio: diversi tipi di enunciati

In sintesi, gli oggetti che definiscono l'*ontologia di un archivio* sono:

- collettivi di tipo popolazione o di tipo evento, istantaneo o con durata
- caratteristiche associate agli elementi dei collettivi (che danno luogo a variabili una volta considerate da un punto di vista statistico), ciascuna delle quali utilizza:
  - una classificazione costituita da diverse modalità, se di classificazione
  - un dominio/range costituito da diversi valori, se numerica
- relazioni 1-n (oppure 1-1) [6] tra due elementi di collettivi (due diversi collettivi o lo stesso collettivo) [7].

Quindi gli oggetti che definiscono l'ontologia di un archivio sono riconducibili a insiemi di elementi, oppure a relazioni tra elementi appartenenti a tali insiemi. In particolare:

- i collettivi di tipo popolazione o evento - come Studenti, Immatricolazioni - sono concettualizzati come insiemi;
- le caratteristiche degli elementi - come Sesso, Residenza - sono concettualizzate come relazioni che legano un elemento di un insieme a una modalità di una classificazione o a un valore in un dominio numerico;
- le relazioni tra elementi di insiemi sono concettualizzate ovviamente come relazioni, di tipo funzionale.

Ciò significa che le informazioni correntemente acquisite dall'archivio sono sempre interpretabili come affermazioni circa l'appartenenza di un elemento ad un insieme, o di una associazione tra elementi ad una relazione, e l'errore riguarda la verità o falsità di tali affermazioni di appartenenza**.**

Ad esempio, una singola persona di nome Rossi può in un dato momento appartenere effettivamente o non appartenere al collettivo Studenti. A fronte di questa situazione reale, l'archivio presenta un errore se registra l'informazione che Rossi appartiene al collettivo Studenti quando ciò non è vero, oppure non registra l'informazione che Rossi appartiene al collettivo Studenti quando ciò è vero.

Analogamente, uno studente di nome Rossi può in un dato momento avere effettivamente o non avere la residenza a Roma: vale a dire, la coppia (Rossi, Roma) può in un dato momento appartenere effettivamente o non appartenere alla relazione Residenza che lega gli elementi del collettivo Studenti alle modalità della classificazione dei comuni italiani. A fronte di questa situazione reale, l'archivio presenta un errore se registra l'informazione che la coppia (Rossi, Roma) appartiene alla relazione Residenza quando ciò non è vero, oppure non registra l'informazione che la coppia (Rossi, Roma) appartiene alla relazione Residenza quando ciò è vero.

Per descrivere e ragionare sulle situazioni di errore, associamo ad ogni *oggetto osservato* da un archivio *un tipo di enunciato di appartenenza*, cioè un'affermazione di appartenenza potenzialmente attribuibile a elementi osservabili, che quando viene riferita ad un elemento

[6] Nel nostro approccio le relazioni m-n tra due elementi e le relazioni tra più elementi sono considerate come particolari collettivi, di tipo associativo.
[7] In una relazione 1-n i due collettivi, o lo stesso, assumono i due ruoli distinti di dominio e codominio della relazione: ogni elemento del dominio può essere legato a un singolo elemento del codominio, ogni elemento del codominio può essere legato a più elementi del dominio.

osservabile dà luogo ad un singolo enunciato di appartenenza effettivo, il quale poi può essere vero oppure falso.

Quindi ad esempio associamo al collettivo degli Studenti il tipo di enunciato Studente (x): si tratta di un'affermazione "aperta", che dà luogo a un'affermazione effettiva, cioè a un enunciato vero o falso, ponendo al posto della x l'identificativo di un elemento osservabile. Data ad esempio la persona Rossi, l'enunciato Studente (Rossi) può essere in un dato momento vero o falso.

Analogamente, associamo alla relazione Residenza che lega gli elementi del collettivo Studenti alle modalità della classificazione dei comuni italiani il tipo di enunciato Residenza (x y): si tratta di un'affermazione "aperta", che dà luogo a un'affermazione effettiva, cioè a un enunciato vero o falso, ponendo al posto della x l'identificativo di un elemento osservabile e al posto della y una modalità di una classificazione o un valore numerico. Dati ad esempio lo studente Rossi e la modalità Roma appartenente alla classificazione dei comuni italiani, l'enunciato Residenza (Rossi, Roma) può essere in un dato momento vero o falso.

Le diverse *tipologie di informazione* che nel corso del tempo vengono raccolte da una fonte, in particolare da un archivio amministrativo, sono proprio tali enunciati di appartenenza riferiti a singoli elementi osservabili (eventi o unità di popolazione), ad esempio Studente (Rossi), Residenza (Rossi, Roma).

Per questo possiamo dire che una fonte d'informazione accetta e gestisce *diverse tipologie di enunciati relativi ai collettivi, alle caratteristiche, alle relazioni componenti l'ontologia della fonte e riferiti a specifici elementi osservati*, precisamente:

- enunciati che asseriscono l'appartenenza di un elemento (un evento o un'unità di popolazione) ad un collettivo
- enunciati che asseriscono l'esistenza di un'associazione tra un elemento di un collettivo (un evento o un'unità di popolazione) e una modalità, o valore, tra quelli assumibili per una particolare caratteristica, in quanto appartenenti alla classificazione o al dominio/range utilizzato per la caratteristica
- enunciati che asseriscono l'esistenza di una relazione tra due elementi di collettivi (eventi o unità di popolazione).

Si è detto che è caratteristico delle fonti non statistiche raccogliere con continuità nel tempo le informazioni relative agli elementi osservabili.

In generale una fonte non statistica mira a gestire informazioni sullo stato delle unità osservate in ciascun momento d'osservazione, riferendo quindi tutte le informazioni relative agli eventi istantanei osservati al loro momento d'occorrenza e aggiornando le informazioni relative alle unità delle popolazioni e agli eventi con durata per tutto il tempo della loro appartenenza allo specifico collettivo caratteristico dell'archivio.

Per questo tutti i tipi di enunciati accettati e gestiti in un archivio amministrativo hanno un *riferimento temporale*, precisamente:

- un istante t_i per gli enunciati relativi a elementi dei collettivi di eventi istantanei
- una durata $t_i - t$ oppure $t_i - t_j$, per gli enunciati relativi ad elementi dei collettivi di tipo popolazione o agli eventi con durata, enunciati che possono essere riferiti ad un periodo aperto $t_i - t$ o chiuso $t_i - t_j$.

Per quanto detto i tipi di enunciati d'interesse statistico accettati e gestiti in un archivio amministrativo possono essere in generale di tre tipi:

A) enunciati di appartenenza di un elemento u_i ad un collettivo di tipo popolazione o evento (istantaneo o con durata) A, riferiti ad un momento o un periodo, precisamente:

> *$A(u_i, t_i)$* per i collettivi di tipo evento istantaneo, dove e $u_{i\ i}$ è l'identificativo dell'evento istantaneo appartenente al collettivo *A*, e $t_{i\ i}$ identifica il momento di occorrenza
>
> *$A(u_i, t_i - t)$* per i collettivi di tipo popolazione o evento con durata, dove u_i è l'identificativo dell'unità o evento con durata appartenente al collettivo *A*, t_i identifica il momento di ingresso dell'unità nel collettivo, *t* identifica un momento di uscita dell'unità dal collettivo ancora indefinito
>
> *$A(u_i, t_i - t_j)$* per i collettivi di tipo popolazione o evento con durata, dove u_i è l'identificativo dell'unità o evento con durata appartenente al collettivo *A* , t_i identifica il momento di ingresso dell'unità nel collettivo, t_j identifica il momento di uscita dell'unità dal collettivo.

Esempi (nei quali l'identificativo di un evento è indicato come $evento_i$): Studente (Verdi, $t_i - t$), Avvio rapporto di lavoro (avvio rapporto di $lavoro_i$, t_i), Ricovero ($ricovero_i$, t_i), Rapporto di lavoro (rapporto di $lavoro_i$, $t_i - t$), Degenza ($degenza_i$, $t_i - t$);

B) enunciati riguardanti il possesso da parte di un'unità u_i di un collettivo di tipo popolazione o evento, istantaneo o con durata, di una specifica modalità c_i di una classificazione o di uno specifico valore c_i di un dominio numerico per una certa caratteristica B, corrispondente all'appartenenza della coppia *(u_i, c_i)* alla caratteristica B, ad uno specifico momento o per un periodo, precisamente:

> *$B(u_i, c_i, t_i)$* per i collettivi di tipo evento istantaneo, dove u_i, è l'identificativo dell'evento istantaneo, c_i è l'identificativo di una modalità appartenente alla classificazione utilizzata per la caratteristica o di un valore appartenente al dominio numerico della caratteristica, t_i identifica il momento di occorrenza dell'evento istantaneo
>
> *$B(u_i, c_i, t_i - t)$* per i collettivi di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo, c_i è l'identificativo di una modalità appartenente alla classificazione utilizzata per la caratteristica o di un valore appartenente al dominio numerico della caratteristica, t_i identifica il momento di acquisizione della modalità c_i, *t* identifica un momento nel quale la modalità o valore c_i sarà sostituito da un'altra modalità o valore, momento che è ancora indefinito
>
> *$B(u_i, c_i, t_i - t_j)$* per i collettivi di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo, c_i è l'identificativo di una modalità appartenente alla classificazione utilizzata per la caratteristica o di un valore appartenente al dominio numerico della caratteristica, t_i identifica il momento di acquisizione della modalità c_i, mentre t_j identifica il momento nel quale la modalità o valore c_i è stato sostituito da un'altra modalità o valore.

Esempi (nei quali l'identificativo di un evento è indicato come $evento_i$): Classe di età (Verdi, 20-25, $t_i - t$), Sesso(Verdi, femmina, $t_i - t$), Residenza (Rossi, Roma, $t_i - t$), Valore della produzione (Fiat, 100000 euro, $t_i - t$), Motivo ricovero ($ricovero_i$, incidente, t_i), Spesa per degenza ($degenza_i$, 550 euro, $t_i - t$);

C) enunciati riguardanti l'esistenza di una particolare relazione tra due elementi u_i, u_j appartenenti a due collettivi di tipo popolazione o evento (o anche allo stesso collettivo) corrispondente all'appartenenza della coppia *(u_i, u_j)* alla relazione C [8], ad uno specifico momento o per un periodo, precisamente:

> *C(u_i, u_j, t_i)* se il collettivo nel ruolo di dominio è di tipo evento istantaneo, dove u_i, è l'identificativo dell'unità appartenente al collettivo dominio, u_j è l'identificativo dell'unità legata appartenente al collettivo codominio, e t_i identifica il momento di occorrenza dell'evento istantaneo
>
> *C(u_i, u_j, $t_i - t$)* se il collettivo nel ruolo di dominio è di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo dominio, u_j, è l'identificativo dell'unità legata appartenente al collettivo codominio, t_i identifica il momento di acquisizione del legame con l'elemento u_j,, t identifica un momento nel quale l'elemento legato u_j, sarà sostituito da un altro elemento legato, momento che è ancora indefinito
>
> *C(u_i, u_j, $t_i - t_j$)* se il collettivo nel ruolo di dominio è di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo dominio, u_j, è l'identificativo dell'unità legata appartenente al collettivo codominio, t_i identifica il momento di acquisizione del legame con l'elemento u_j,, t_j identifica il momento nel quale l'elemento legato u_j, è stato sostituito da un altro elemento.

Esempi (nei quali l'identificativo di un evento è indicato come $evento_i$): Azienda agricola è condotta da Conduttore (Le querce, Bianchi, $t_i - t$), Unità locale appartiene Impresa (Sede Torino, Fiat, $t_i - t$), Immatricolazione inizia Studente ($immatricolazione_i$, Verdi, t_i), Acquisizione crediti riguarda Studente ($esame_i$, Verdi, t_i), Avvio rapporto di lavoro riguarda Lavoratore (avvio rapporto di $lavoro_i$, Bianchi, t_i), Avvio rapporto di lavoro riguarda Datore di lavoro (avvio rapporto di $lavoro_i$, Neri, t_i), Avvio rapporto di lavoro inizia Rapporto di lavoro (avvio rapporto di $lavoro_i$, rapporto di $lavoro_i$, t_i), Rapporto di lavoro riguarda Lavoratore (rapporto di $lavoro_i$, Bianchi, $t_i - t$), Rapporto di lavoro riguarda Datore di lavoro (rapporto di $lavoro_i$, Neri, $t_i - t$).

In conclusione si può dire quanto segue.

L'*ontologia di una fonte d'informazione* è descritta dai seguenti *oggetti*:

- una serie di collettivi di tipo popolazione o di tipo evento, istantaneo o con durata, A^1,…, A^k, ….A^m
- una serie di caratteristiche , B_1 ,…, B_h, ..., B_n, (che danno luogo a variabili una volta considerate da un punto di vista statistico), ciascuna delle quali utilizza una classificazione costituita da diverse modalità, se di classificazione, un dominio/range costituito da diversi valori, se numerica
- una serie di relazioni 1-n (oppure 1-1) [9] tra due elementi di collettivi (due diversi collettivi o lo stesso collettivo) [10], , C_1,…, C_l, ..., C_r.

[8] Come illustrato nell'Appendice dedicata all'ontologia di un archivio amministrativo, le relazioni C sono sempre di tipo funzionale, in quanto legano un elemento di un collettivo codominio a più elementi di un collettivo dominio, o in particolare ad uno solo. Una relazione m-n, non funzionale, tra elementi di due o più collettivi è vista come un ulteriore collettivo, precisamente un evento di tipo associativo legato funzionalmente a tali collettivi, ad esempio Rapporto di lavoro è un evento di tipo associativo che è legato mediante due relazioni C, nel ruolo di dominio, ai due collettivi Lavoratore e Datore di lavoro nel ruolo di codomini.

[9] Nel nostro approccio come si è detto le relazioni m-n tra due elementi e le relazioni tra più elementi sono considerate come particolari collettivi, di tipo associativo.

e dalla *rete di associazioni che lega questi oggetti*, stabilendo su quali collettivi è osservata ciascuna caratteristica e quali sono i collettivi dominio e codominio per ogni relazione [11].

In generale ad ogni collettivo A^k è associata una serie di caratteristiche e una serie di relazioni, tenendo presente che in linea generale una caratteristica o relazione può essere associata a più collettivi, anche se ciò nella pratica avviene per poche caratteristiche o relazioni, ad esempio per i riferimenti territoriali. Ciò premesso usiamo in seguito le notazioni B_h^k, C_l^k quando occorre ricordare che la caratteristica B_h o la relazione C_l è associata al collettivo A^k.

L'informazione accettata e gestita dalla fonte è quindi in generale costituita dalle seguenti collezioni di enunciati:

- per ciascun collettivo A^k, con $k = 1, \ldots m$, una collezione di *enunciati di appartenenza* $A^k(u_i, t_i)$ oppure $A^k(u_i, t_i - t)$ o $A^k(u_i, t_i - t_j)$ secondo il tipo di collettivo, dove u_i identifica un generico evento o unità osservata, t_i, t_j rappresentano momenti del tempo
- per ciascuna caratteristica B_h, con $h = 1, \ldots n$, una collezione di *enunciati riguardanti il possesso di una specifica modalità o valore*, $B_h(u_i, c_i, t_i)$ oppure $B_h(u_i, c_i, t_i - t)$ o $B_h(u_i, c_i, t_i - t_j)$ secondo il tipo di collettivo, dove u_i identifica un generico evento o unità osservata, c_i rappresenta una generica modalità o valore osservato, t_i, t_j rappresentano momenti del tempo
- per ciascuna caratteristica C_l, con $l = 1, \ldots r$, una collezione di *enunciati riguardanti l'esistenza di una particolare relazione tra due elementi*, $C_l(u_i, u_j, t_i)$ oppure $C_l(u_i, u_j, t_i - t)$ o $C_l(u_i, u_j, t_i - t_j)$ secondo il tipo di collettivo, dove u_i identifica un generico evento o unità osservata, u_j rappresenta un generico evento o unità osservata legato, t_i, t_j rappresentano momenti del tempo

I contenuti informativi dell'archivio: l'organizzazione degli enunciati in record

Oggi l'informazione raccolta da un archivio amministrativo è nella maggior parte dei casi gestita in appositi database relazionali. L'ontologia dell'archivio può fungere anche da modello concettuale che guida l'organizzazione logico-fisica dell'informazione all'interno del database.

Secondo la teoria della progettazione dei database relazionali, in linea teorica ad ogni collettivo corrisponderà una tabella composta di tante righe quanti sono gli elementi appartenenti al collettivo. Ogni riga della tabella è un *record* e contiene tutte le informazioni relative ad un elemento del collettivo, vale a dire l'identificativo dell'elemento, le modalità e i valori per ciascuna caratteristica e, per ciascuna relazione, i codici di raccordo con gli elementi legati della stessa tabella o di un'altra tabella [12] .

Tuttavia nella pratica se l'archivio amministrativo, come avviene nella maggior parte dei casi, conserva uno storico delle entrate e delle uscite di elementi dai collettivi ed eventualmente anche delle variazioni relative alle caratteristiche e alle relazioni, l'informazione relativa agli elementi di un collettivo potrà trovarsi fisicamente dislocata in tabelle diverse tra loro connesse.

Il termine *record* è utilizzato anche nella letteratura sui controlli di qualità per denotare il raggruppamento di tutte le informazioni relative ad un elemento di un collettivo, ai fini, ad esempio, della specifica dei vincoli di incompatibilità tra modalità o valori.

---

[10] In una relazione 1-n i due collettivi, o lo stesso, assumono i due ruoli distinti di dominio e codominio della relazione: ogni elemento del dominio può essere legato a un singolo elemento del codominio, ogni elemento del codominio può essere legato a più elementi del dominio.

[11] Un'ontologia può essere specificata mediante appositi linguaggi basati sulla logica e dotati di una semantica formale.

[12] In termini informatici, l'identificativo è la chiave primaria del record, i codici di raccordo sono sue chiavi esterne.

Utilizzeremo di seguito il termine conformemente a questa seconda tradizione, ciò che comporta specificarne il significato ad un livello concettuale piuttosto che logico-fisico, indicando quindi con il termine *record* un gruppo di informazioni relative allo stesso elemento di un collettivo, a prescindere dalla loro organizzazione fisica. Un record così inteso può anche raggruppare informazioni dislocate in tabelle diverse, non corrispondendo quindi esattamente ad un singolo record in una tabella relazionale.

Poiché nel nostro approccio un'informazione è un enunciato, si può dire che:

- un *RECORD raggruppa enunciati relativi ad ogni evento o unità u_i appartenente ad uno specifico collettivo a partire da un momento iniziale t_i* ed eventualmente fino a un momento finale t_j.

Ai fini della successiva descrizione delle modalità di aggiornamento dell'archivio specifichiamo meglio come è composto un record.

Per comodità anziché utilizzare la notazione generica A^k per ogni tipo di collettivo indichiamo con E^k, P^k, D^k il generico collettivo di tipo evento istantaneo, popolazione, evento con durata rispettivamente, e con:

- $E^k(e_i, t_i)$ gli enunciati di appartenenza riferiti al collettivo di eventi istantanei E^k
- $P^k(p_i, t_i - t)$ o $P^k(p_i, t_i - t_j)$ gli enunciati di appartenenza con periodo aperto o chiuso riferiti al collettivo di tipo popolazione P^k
- $D^k(d_i, t_i - t)$ o $D^k(d_i, t_i - t_j)$ gli enunciati di appartenenza con periodo aperto o chiuso riferiti al collettivo di eventi con durata D^k

Utilizzando poi le notazioni B_h^k, C_l^k per ricordare che la caratteristica B_h o la relazione C_l è associata al collettivo E^k, o P^k, o D^k, per quanto detto possiamo assumere che:

- a ogni collettivo E^k, o P^k, o D^k è associata una serie di caratteristiche $B_1^k,\ldots, B_h^k,\ldots B_{n_k}^k$ e una serie di relazioni $C_1^k,\ldots, C_l^k,\ldots C_{r_k}^k$,

e inoltre indichiamo con:

- $B_h^k(e_i, c_i, t_i)$ gli enunciati relativi al possesso della caratteristica B_h per gli elementi di un collettivo E^k
- $B_h^k(p_i, c_i, t_i - t)$ o $B_h^k(p_i, c_i, t_i - t_j)$ gli enunciati con periodo aperto o chiuso relativi al possesso della caratteristica B_h per gli elementi di un collettivo P^k
- $B_h^k(d_i, c_i, t_i - t)$ o $B_h^k(d_i, c_i, t_i - t_j)$ gli enunciati con periodo aperto o chiuso relativi al possesso della caratteristica B_h per gli elementi di un collettivo D^k
- $C_l^k(e_i, u_j, t_i)$ gli enunciati relativi al possesso di una relazione C_l per gli elementi del collettivo E^k
- $C_l^k(p_i, u_j, t_i - t)$, $C_l^k(p_i, u_j, t_i - t_j)$ gli enunciati con periodo aperto o chiuso relativi al possesso di una relazione C_l per gli elementi del collettivo P^k
- $C_l^k(d_i, u_j, t_i - t)$, $C_l^k(d_i, u_j, t_i - t_j)$ gli enunciati con periodo aperto o chiuso relativi al possesso di una relazione C_l per gli elementi del collettivo D^k.

Allora un *RECORD* $E^k[e_i, t_i]$, o $P^k[p_i, t_i\text{-}t]$, o $P^k[p_i, t_i\text{-}t_j]$, o $D^k[d_i, t_i\text{-}t]$, o $D^k[d_i, t_i\text{-}t_j]$, raggruppa un enunciato di appartenenza ad un collettivo E^k, o P^k, o D^k relativo ad un elemento e_i o p_i o d_i e a un momento t_i con una serie di enunciati di possesso di caratteristiche e relazioni associate a tale collettivo relativi allo stesso elemento. Quindi ad esempio:

- $E^k\,[e_{i,,}\,t_i] = [E^k\,(e_i,\,t_i),\,B_1^k\,(e_i,\,c_i,\,t_i),\ldots,\,B_h^k\,e_i,\,c_i,\,t_i),\ldots\,B_{n_k}^k\,(e_i,\,c_i,\,t_i),\,C_1^k\,(e_i,\,u_j\,t_i),\ldots,\,C_l^k\,(e_i,\,u_j,\,t_i),\ldots\,C_{r_k}^k\,(e_i,\,u_j,\,t_i)]$,
- $P^k\,[p_{i\,},t_i\text{-}t] = [P^k\,(p_i,\,t_i\text{-}t),\,B_1^k\,(p_i,\,c_i,\,t_i\text{-}t),\ldots,\,B_h^k\,(p_i,\,c_i,\,t_i\text{-}t),\ldots\,B_{n_k}^k\,(p_i,\,c_i,\,t_i\text{-}t),\,C_1^k\,(p_i,\,u_j,\,t_i\text{-}t),\ldots,\,C_l^k\,(p_i,\,u_j,\,t_i\text{-}t),\ldots\,C_{r_k}^k\,(p_i,\,u_j,\,t_i\text{-}t)]$
- $P^k\,[p_{i\,},t_i\text{-}\,t_j] = [P^k\,(p_i,\,t_i\text{-}t_j),\,B_1^k\,(p_i,\,c_i,\,t_i\text{-}t_j),\ldots,\,B_h^k\,(p_i,\,c_i,\,t_i\text{-}t_j),\ldots\,B_{n_k}^k\,(p_i,\,c_i,\,t_i\text{-}t_j),\,C_1^k\,(p_i,\,u_j,\,t_i\text{-}t_j),\ldots,\,C_l^k\,(p_i,\,u_j,\,t_i\text{-}t_j),\ldots\,C_{r_k}^k\,(p_i,\,u_j,\,t_i\text{-}t_j)]$
- $D^k\,[d_{i\,},t_i\text{-}t] = [D^k\,(d_i,\,t_i\text{-}t),\,B_1^k\,(d_i,\,c_i,\,t_i\text{-}t),\ldots,\,B_h^k\,(d_i,\,c_i,\,t_i\text{-}t),\ldots\,B_{n_k}^k\,(d_i,\,c_i,\,t_i\text{-}t),\,C_1^k\,(d_i,\,u_j,\,t_i\text{-}t),\ldots,\,C_l^k\,(d_i,\,u_j,\,t_i\text{-}t),\ldots\,C_{r_k}^k\,(d_i,\,u_j,\,t_i\text{-}t)]$
- $D^k\,[d_{i\,},t_i\text{-}\,t_j] = [D^k\,(d_i,\,t_i\text{-}t_j),\,B_1^k\,(d_i,\,c_i,\,t_i\text{-}t_j),\ldots,\,B_h^k\,(d_i,\,c_i,\,t_i\text{-}t_j),\ldots\,B_{n_k}^k\,(d_i,\,c_i,\,t_i\text{-}t_j),\,C_1^k\,(d_i,\,u_j,\,t_i\text{-}t_j),\ldots,\,C_l^k\,(d_i,\,u_j,\,t_i\text{-}t_j),\ldots\,C_{r_k}^k\,(d_i,\,u_j,\,t_i\text{-}t_j)]$

Per i vari tipi di record useremo anche le notazioni semplificate $E_i^k[t_i]$, $P_i^k[t_i\text{-}t]$,], $P_i^k[t_i\text{-}t]$, $P_i^k[t_i\text{-}t_j]$, $D_i^k\,[t_i\text{-}t]$, $D_i^k[t_i\text{-}t_j]$, dove l'indice *i* in basso è riferito all'elemento e_i o p_i o d_i.

Per i collettivi di tipo popolazione o evento con durata, al momento t_i del primo o unico ingresso di un elemento p_i o d_i in archivio viene creato un record con periodo aperto $P_i^k[t_i\text{-}t]$ o $D_i^k\,[t_i\text{-}t]$. In assenza di aggiornamenti di qualsiasi tipo, questo record viene sostituito da un record con periodo chiuso $P^k\,[p_{i\,},t_i\text{-}t_j]$ e $D^k\,[d_{i\,},t_i\text{-}t_j]$ al momento dell'uscita definitiva dell'elemento dal collettivo.

Di solito però l'informazione relativa agli elementi dei collettivi di tipo popolazione e anche, generalmente in misura minore, agli eventi con durata è soggetta ad aggiornamenti nel periodo che intercorre tra il primo o unico ingresso dell'elemento nel collettivo e l'uscita definitiva dal collettivo.

Un aggiornamento può essere un cambiamento relativo alle caratteristiche o relazioni possedute dall'elemento oppure, quando ciò è possibile, un'uscita temporanea dal collettivo o un nuovo ingresso nel collettivo successivo ad un'uscita temporanea. In seguito ad un aggiornamento si chiude il periodo di un record e un nuovo record con periodo aperto può essere creato contestualmente o successivamente, secondo il tipo di aggiornamento.

Di conseguenza, come sarà più chiaro in seguito, al momento attuale t_A per ogni elemento p_i o d_i si potrà avere in archivio una serie di record con periodi di validità chiusi successivi, ciascuno generato da uno dei successivi aggiornamenti intervenuti successivamente al momento di primo o unico ingresso dell'elemento in archivio e, se l'elemento p_i o d_i è ancora appartenente al collettivo nel momento t_A, un unico record con periodo aperto che inizia dal momento dell'ultimo aggiornamento.

I record con periodo di validità chiuso al momento t_j contengono uno o più enunciati che condividono il momento di chiusura t_j, assieme eventualmente ad altri enunciati aperti, mentre l'unico record aperto contiene solo enunciati aperti. In entrambi i casi i diversi enunciati che compongono il record possono avere associati periodi di validità con momenti d'inizio t_i differenti, tra i quali il più recente è il momento d'inizio di validità dell'intero record.

LA DINAMICA DI AGGIORNAMENTO DELL'ARCHIVIO AMMINISTRATIVO

I diversi tipi di aggiornamenti

Nella maggior parte dei casi un archivio amministrativo, a differenza di un'indagine, non osserva direttamente i collettivi di tipo popolazione. Tutte le informazioni relative ai collettivi di tipo popolazione e anche ai collettivi di tipo evento con durata vengono aggiornate a seguito di eventi istantanei.

Gli eventi istantanei descritti finora hanno proprie caratteristiche e rivestono di per sé interesse statistico (ad esempio immatricolazioni, ricoveri ospedalieri, avvio di un rapporto di lavoro).

Anche gli aggiornamenti delle caratteristiche o relazioni per gli elementi dei collettivi di tipo popolazione o gli eventi con durata (ad esempio il cambio di residenza di una persona, il cambio di tipologia di un rapporto di lavoro) si possono considerare come particolari specie di eventi istantanei che non hanno caratteristiche proprie e in generale non rivestono interesse di per sé, se non per studi longitudinali.

Come dettagliato meglio nel paragrafo successivo e nel capitolo dedicato alla copertura dei collettivi, l'ingresso o l'uscita dal collettivo di appartenenza di un elemento di un collettivo di tipo popolazione così come l'ingresso o l'uscita dal collettivo di appartenenza per un evento con durata sono provocati da specifici eventi istantanei di ingresso o uscita e, talvolta, da specifici cambiamenti di caratteristiche o relazioni o anche da entrambi i tipi di eventi in combinazione.

Ciò premesso vediamo quali sono le diverse operazioni di accettazione in archivio e di modifica delle informazioni, organizzate in record.

Anzitutto l'occorrenza di un evento istantaneo (ad esempio un'immatricolazione, un esame, l'avvio di un rapporto di lavoro) ad un momento t_i comporta l'accettazione in archivio di un nuovo record contenente l'enunciato di appartenenza e tutti gli enunciati di possesso di caratteristiche e relazioni relativi all'evento istantaneo, tutti riferiti al momento t_i.

L'occorrenza al momento t_i di un evento istantaneo che è di primo o unico ingresso per un elemento di un collettivo di tipo popolazione o un evento con durata (ad esempio l'immatricolazione per uno studente, l'avvio per un rapporto di lavoro) e/o l'occorrenza di uno specifico cambiamento di caratteristiche o relazioni che ha lo stesso effetto determinano:

- l'accettazione in archivio di un nuovo record relativo a un elemento di un collettivo di tipo popolazione o a un evento con durata, contenente l'enunciato di appartenenza e tutti gli enunciati di possesso di caratteristiche e relazioni relativi all'elemento della popolazione o all'evento con durata, tutti riferiti al periodo t_i-t.

Successivamente, a seguito di specifici eventi istantanei o e/o cambiamenti di specifiche caratteristiche o relazioni, si può avere ad un momento t_j un'uscita temporanea o definitiva dell'elemento dal collettivo di tipo popolazione o evento con durata (ad esempio nei casi di un evento di laurea per uno studente, un licenziamento per il rapporto di lavoro), che determina:

- la chiusura dell'enunciato di appartenenza che non è più attuale dal momento t_j, rappresentata mediante la chiusura del suo periodo di validità che diventa t_i-t_j, e la conseguente chiusura del periodo di validità del record che lo contiene, in quanto anch'esso diventa non più attuale dal momento t_j, e di tutti gli enunciati relativi al possesso di

caratteristiche e relazioni che lo compongono, che non possono essere più aggiornati in quanto l'elemento non è più osservato dall'archivio [13].

All'uscita temporanea (rara per gli eventi con durata), sempre a seguito di specifici eventi istantanei e/o cambiamenti di specifiche caratteristiche o relazioni, può seguire ad un successivo momento $t_{i'}$ un nuovo ingresso (ad esempio nel caso di una nuova iscrizione dopo una laurea, per cui una persona torna ad essere uno studente), che determina:

- l'accettazione di un nuovo record aperto contenente tutti enunciati con periodo di validità t_i-t con $t_i = t_{i'}$ e $t_{i'} > t_j$, se t_j era il momento della precedente uscita temporanea.

Un generico aggiornamento di una caratteristica o relazione (o più) relativa all'elemento di una popolazione o a un evento con durata (ad esempio la residenza di uno studente, il tipo di un rapporto di lavoro) al momento t_m provoca:

- la chiusura di un enunciato (o più) che non è più attuale dal momento t_m [14], rappresentata mediante la chiusura del suo periodo di validità che diventa t_i-t_j con $t_j = t_m$, e la conseguente chiusura del periodo di validità del record che lo contiene, in quanto anch'esso diventa non più attuale dal momento t_m, e contestualmente:
- l'accettazione di un nuovo enunciato con periodo di validità aperto t_i–t con $t_i = t_m$ e quindi di un record con lo stesso periodo di validità contenente il nuovo enunciato assieme agli altri enunciati aperti ancora validi.

Questa descrizione degli effetti degli aggiornamenti in termini di chiusura e apertura di record è puramente concettuale e comoda per condurre analisi della qualità, mentre è evidente che nella pratica della gestione dell'archivio il trattamento a livello fisico degli aggiornamenti non avviene solitamente in questo modo, anche perché le politiche di organizzazione e gestione degli archivi possono essere le più diverse.

Anzitutto più in generale gli archivi possono adottare politiche diverse rispetto alla conservazione degli enunciati non più attuali, eliminandoli fisicamente o meno. Dato che la maggioranza degli archivi amministrativi osserva i collettivi con continuità nel tempo, possiamo assumere che nella maggior parte dei casi le informazioni non più attuali siano conservate, cioè che l'archivio gestisca uno storico dell'appartenenza ai collettivi e in molti casi anche uno storico delle variazioni relative alle caratteristiche e relazioni.

Le soluzioni concretamente adottate a questo scopo potranno poi essere diverse, ad esempio associare periodi di validità agli elementi appartenenti ai collettivi, così come alle modalità o valori assunti nel corso del tempo per le diverse caratteristiche e ai diversi elementi legati nel corso del tempo per le relazioni, oppure in alternativa registrare la situazione osservata al momento del primo o unico ingresso di un elemento in archivio senza espliciti aggiornamenti successivi, dato che i cambiamenti successivi nell'appartenenza al collettivo o nel possesso di caratteristiche e relazioni sono comunque sempre ricostruibili dagli eventi istantanei che li hanno provocati, o qualsiasi altra soluzione pratica per memorizzare i cambiamenti.

In generale quando l'archivio è strutturato per gestire uno storico dell'appartenenza ai collettivi e/o delle variazioni relative alle caratteristiche e relazioni un unico record concettuale, come qui inteso,

[13] come sarà spiegato nell'apposito capitolo, cioè è vero solo per i collettivi principali dell'archivio, non per i collettivi sottoinsieme

[14] assumiamo per semplicità che sia sempre presente nel record un enunciato per ciascuna caratteristica o relazione, ciò significa che per le eventuali caratteristiche o relazioni opzionali, cioè per le quali la modalità o l'elemento legato può non essere osservato, si utilizza una modalità o un legame fittizio

non corrisponde in genere ad un unico record fisico memorizzato in una tabella relazionale ma è comunque ricostruibile .

I diversi tipi di aggiornamenti: le classi REG, ELIM e MOD

Di seguito descriviamo meglio le operazioni di accettazione e aggiornamento delle informazioni in archivio. E' utile distinguere tre classi di operazioni, che indichiamo con REG, ELIM e MOD.

Classe REG: con le operazioni in questa classe vengono accettati in archivio elementi di un collettivo con tutte le loro proprietà. Comprende tutte le operazioni di registrazione dell'ingresso di elementi nei collettivi di tipo evento istantaneo, popolazione, evento con durata, con contestuale registrazione delle caratteristiche e relazioni relative all'elemento.

Viene quindi accettato un intero nuovo record con periodo aperto t_i-*t*, definito come al precedente paragrafo.

Classe ELIM: con le operazioni in questa classe vengono eliminati elementi di un collettivo di tipo popolazione o evento con durata. Comprende tutte le operazioni di eliminazione di elementi dai collettivi di tipo popolazione o evento con durata, chiudendo temporaneamente o definitivamente il loro periodo di appartenenza al collettivo.

Quindi viene accettato in archivio un singolo nuovo enunciato di appartenenza con periodo chiuso t_i-t_j che sostituisce il precedente enunciato di appartenenza con periodo aperto t_i-*t* e, come conseguenza, viene chiuso il periodo di un record esistente che aveva periodo aperto[15] (l'eliminazione del record è solitamente virtuale se l'uscita non è definitiva [16]).

Classe MOD: con le operazioni in questa classe vengono modificate le informazioni relative al possesso di caratteristiche e relazioni da parte degli elementi di un collettivo di tipo popolazione o evento con durata. Comprende tutte le operazioni di registrazione di modifiche alle caratteristiche e relazioni possedute da elementi appartenenti a collettivi di tipo popolazione o evento con durata.

Viene accettato in archivio un nuovo enunciato di possesso di una caratteristica o relazione relativo all'elemento del collettivo, con periodo aperto t_i-*t*. Come conseguenza viene chiuso il periodo di un record esistente che aveva periodo aperto [17] e contestualmente viene accettato in archivio un nuovo record con periodo aperto t_i-*t*.

[15] questa è una descrizione concettuale di ciò che avviene, nella pratica come si è detto può accadere che le informazioni gestite non abbiano esplicitamente associati periodi di validità e sia intesa come uscita di un elemento da un collettivo il semplice verificarsi di un evento di uscita

[16] in concreto a seguito di un'uscita definitiva si dovrebbe avere una contestuale o successiva eliminazione fisica dall'archivio delle informazioni relative all'elemento, con tempi e procedure dipendenti dalle politiche di gestione del singolo archivio, mentre in caso di uscita temporanea tali informazioni saranno in generale conservate in archivio con aggiornamento del periodo di validità se gestito; saranno invece eliminate contestualmente a qualsiasi uscita qualora l'archivio non mantenga uno storico relativo all'appartenenza di elementi al collettivo

[17] se l'archivio mantiene uno storico relativo al possesso di caratteristiche e relazioni l'informazione non più attuale sarà mantenuta, con aggiornamento del relativo periodo di validità se gestito, in caso contrario si ha eliminazione fisica dell'informazione non più attuale

Le operazioni nelle classi REG ed ELIM: ingresso e uscita dai collettivi

Per dettagliare i tipi di operazioni nelle classi *REG* ed *ELIM*, occorre tenere presente che, come illustrato più diffusamente nella PARTE QUARTA, in generale in ogni dato archivio i collettivi di tipo popolazione possono essere collettivi monoingresso, per i quali al primo ingresso di un elemento nel collettivo può seguire solo un'uscita definitiva, oppure collettivi multiingresso, per i quali al primo ingresso di un elemento nel collettivo può succedere una serie di uscite e ingressi temporanei, fino ad un'uscita definitiva, e la stessa distinzione vale per i collettivi di eventi con durata, anche se tra questi i collettivi multiingresso sono più rari.

La classe *REG* comprende prima di tutto la seguente operazione:

- ${}_t\mathrm{I}(E_i^k[t_i])$, registrazione nel momento t di un record relativo ad un evento istantaneo $E_i^k[t_i]$: consiste nella registrazione nel momento t dell'enunciato di appartenenza relativo ad un evento e_i appartenente al collettivo di eventi istantanei E^k che è occorso al momento t_i, ${}_t\mathrm{I}(E^k (e_i, t_i))$, con contestuale registrazione delle caratteristiche e relazioni relative all'evento.

Si osserva immediatamente che l'operazione di registrazione dell'evento istantaneo in archivio ha un suo momento di riferimento t, che può essere uguale o diverso rispetto al momento t_i di occorrenza dell'evento.

Nel caso sia diverso (si può supporre successivo) si configura una specie importante di errore, un *errore di tempestività*. E' importante qui osservare che si tratta di un errore distinto dall'errore nel riferimento t_i, che consiste invece nel registrare l'evento con un riferimento temporale che non è quello reale. In questo capitolo assumeremo per semplicità che l'errore di tempestività non sia presente, e quindi $t = t_i$.

Se l'evento istantaneo registrato è di ingresso o di uscita per altri collettivi si potrà avere come conseguenza l'ingresso in archivio di un record relativo a un'unità di popolazione e/o di un record relativo a un evento con durata, o l'eliminazione temporanea o definitiva di un record relativo a un'unità di popolazione e/o di un record relativo a un evento con durata.

La classe *REG* comprende quindi le seguenti ulteriori operazioni relative alle unità di popolazioni:

- l'operazione ${}_t\mathrm{I}\ (P^k\ (p_i,\ t_i\text{-}t\ ))$ di registrazione nel momento t dell'*unico* ingresso o, se il collettivo è multingresso, del *primo* ingresso al momento t_i di un'unità p_i nel collettivo di tipo popolazione P^k
  - comporta l'operazione ${}_t\mathrm{I}\ (P_i^k[t_i\text{-}t])$ di creazione di un record contenente l'enunciato di appartenenza con periodo aperto $P^k\ (p_i,\ t_i\text{-}t)$ e tutti gli enunciati di possesso di caratteristiche e relazioni riferiti allo stesso periodo
- solo per collettivi multiingresso, l'operazione ${}_t\mathrm{I}_{\mathrm{TEMP}}\ (P^k\ (p_i,\ t_i\text{-}t_j))$, di registrazione nel momento t di un *nuovo* ingresso al momento t_i di un'unità p_i nel collettivo di tipo popolazione P^k, successivo ad un'uscita temporanea (esiste quindi in archivio un record relativo alla stessa unità che risulta chiuso in un momento t_j precedente a t_i)
  - comporta l'operazione ${}_t\mathrm{I}_{\mathrm{TEMP}}\ (P_i^k[t_i\text{-}t])$ di creazione di un record contenente l'enunciato di appartenenza con periodo aperto $P^k\ (p_i,\ t_i\text{-}t)$ e tutti gli enunciati di possesso di caratteristiche e relazioni aperti riferiti allo stesso periodo

e le seguenti analoghe operazioni relative agli eventi con durata:

- l'operazione ${}_t\mathrm{I}\,(D^k\,(d_i,\, t_i\text{-}t))$ di registrazione nel momento t dell'*unico* ingresso o (raramente per gli eventi con durata) del *primo* ingresso al momento t_i di un evento con durata d_i nel collettivo di eventi con durata D^k
    - comporta l'operazione ${}_t\mathrm{I}\,(D_i^k[t_i\text{-}t])$ di creazione di un record contenente l'enunciato di appartenenza con periodo aperto $D^k\,(d_i,\, t_i\text{-}t)$ e tutti gli enunciati di possesso di caratteristiche e relazioni aperti riferiti allo stesso periodo
- solo per collettivi multiingresso (rari per gli eventi con durata), l'operazione ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(D^k\,(d_i,\, t_i\text{-}t))$, di registrazione nel momento t di *nuovo* ingresso al momento t_i di un'unità d_i nel collettivo di eventi con durata D^k, successivo ad un'uscita temporanea (esiste quindi in archivio un record relativo alla stessa unità che risulta chiuso in un momento t_j precedente a t_i)
    - comporta l'operazione ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(D_i^k[t_i\text{-}t])$ di creazione di un record contenente l'enunciato di appartenenza con periodo aperto $D^k\,(d_i,\, t_i\text{-}t)$ e tutti gli enunciati di possesso di caratteristiche e relazioni aperti riferiti allo stesso periodo

La classe *ELIM* comprende le seguenti operazioni relative alle unità di popolazioni:

- l'operazione ${}_t\mathrm{I}\,(P^k\,(p_i,\, t_i\text{-}t_j))$ di registrazione nel momento t di un enunciato di appartenenza con periodo chiuso relativo ad un'unità di popolazione $P^k\,(p_i,\, t_i\text{-}t_j)$: consiste nella registrazione del momento t_j nel quale un'unità p_i cessa *definitivamente* di appartenere ad un collettivo di tipo popolazione P^k
    - comporta la chiusura del periodo di validità per il record $P_i^k[t_i\text{-}t]$ che diventa $P_i^k[t_i\text{-}t_j]$
- solo per collettivi multiingresso, l'operazione ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(P^k\,(p_i,\, t_i\text{-}t_j))$ di registrazione nel momento t di un enunciato di appartenenza con periodo chiuso relativo ad un'unità di popolazione $P^k\,(p_i,\, t_i\text{-}t_j)$: consiste nella registrazione del momento t_j nel quale un'unità p_i cessa *temporaneamente* di appartenere ad un collettivo di tipo popolazione P^k
    - comporta la chiusura del periodo di validità per il record $P_i^k[t_i\text{-}t]$ che diventa $P_i^k[t_i\text{-}t_j]$

e le seguenti analoghe operazioni relative agli eventi con durata:

- l'operazione ${}_t\mathrm{I}\,(D^k\,(d_i,\, t_i\text{-}t_j))$ di registrazione nel momento t di un enunciato di appartenenza con periodo chiuso relativo ad un evento con durata $D^k\,(d_i,\, t_i\text{-}t_j)$: consiste nella registrazione del momento t_j nel quale un evento d_i cessa *definitivamente* di appartenere ad un collettivo di eventi con durata D^k
    - comporta la chiusura del periodo di validità per il record $D_i^k[t_i\text{-}t]$ che diventa $D_i^k[t_i\text{-}t_j]$
- solo per collettivi multiingresso (rari per gli eventi con durata), l'operazione ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(D^k\,(d_i,\, t_i\text{-}t_j))$ di registrazione nel momento t di un enunciato di appartenenza con periodo chiuso relativo ad un evento con durata $D^k\,(d_i,\, t_i\text{-}t_j)$: consiste nella registrazione del momento t_j nel quale un evento con durata d_i cessa *temporaneamente* di appartenere ad un collettivo di eventi con durata D^k
    - comporta la chiusura del periodo di validità per il record $D_i^k[t_i\text{-}t]$ che diventa $D_i^k[t_i\text{-}t_j]$

Tutte le operazioni nella classe *ELIM* comportano quindi la chiusura del periodo di validità di un record P^k o D^k, che passa da $t_i\text{-}t$ a $t_i\text{-}t_j$. Negli enunciati componenti il record che sono relativi al possesso di caratteristiche e relazioni il periodo di validità viene chiuso perché non sono più pertinenti, oppure perché a partire al momento di uscita t_j l'elemento non è più osservato [18].

[18] come sarà spiegato nell'apposito capitolo, cioè è vero solo per i collettivi principali dell'archivio, non per i collettivi sottoinsieme

Le operazioni nella classe MOD: modifiche relative al possesso di caratteristiche e relazioni

La classe MOD comprende la seguente operazione relativa alle unità di popolazioni:

- l'operazione ${}_tI\,(B_h^k\ (p_i, c_i, t_m\text{-}t))$ di registrazione nel momento t di un enunciato con periodo aperto $B_h^k\ (p_i, c_i, t_m\text{-}t)$ di possesso da parte di un'unità di popolazione della modalità o valore c_i , in sostituzione di una modalità o valore precedente, per una caratteristica B_h^k , oppure l'operazione ${}_tI\,(C_l^k\ (p_i, u_j, t_m\text{-}t))$ di registrazione nel momento t di un enunciato con periodo aperto $C_l^k\ (p_i, u_j, t_m\text{-}t)$ di possesso da parte di un'unità di popolazione di un legame con u_j, in sostituzione di un legame precedente con un'altra unità, per una relazione C_l^k
    - entrambe queste operazioni comportano la chiusura temporale del vecchio enunciato relativo a B_h^k o a C_l^k e del record $P_i^k[t_i\text{-}t]$ che lo conteneva che diventa $P_i^k[t_i\text{-}t_j]$, e la contestuale creazione di un nuovo record $P_i^k[t_i\text{-}t]$ che contiene il nuovo enunciato, con $t_i = t_j = t_m$.

e la seguente analoga operazione relativa agli eventi con durata:

- l'operazione ${}_tI\,(B_h^k\ (d_i, c_i, t_m\text{-}t))$ di registrazione nel momento t di un enunciato con periodo aperto $B_h^k\ (d_i, c_i, t_m\text{-}t)$ di possesso da parte di un evento con durata della modalità o valore c_i, in sostituzione di una modalità o valore precedente, per una caratteristica B_h^k, oppure l'operazione ${}_tI\,(C_l^k\ (d_i, u_j, t_m\text{-}t))$ di registrazione nel momento t di un enunciato con periodo aperto $C_l^k\ (d_i, u_j, t_m\text{-}t)$ di possesso da parte di un evento con durata di un legame con u_j, in sostituzione di un legame precedente con un'altra unità, per una relazione C_l^k
    - entrambe queste operazioni comportano la chiusura temporale del vecchio enunciato di possesso relativo a B_h^k o a C_l^k e del record $P_i^k[t_i\text{-}t]$ che lo conteneva che diventa $P_i^k[t_i\text{-}t_j]$, e la contestuale creazione di un nuovo record $P_i^k[t_i\text{-}t]$ che contiene il nuovo enunciato, con $t_i = t_j = t_m$.

Per semplicità abbiamo supposto qui che tutte le caratteristiche e le relazioni siano obbligatorie [19].

Le operazioni nelle classi REG ed ELIM: gli effetti degli eventi istantanei di ingresso e uscita

Nei termini delle operazioni introdotte precedentemente possiamo dire che:

- se E^k è un collettivo di *unico* o *primo ingresso* per P^k, allora ${}_tI(E_i^k[t_i])$ implica ${}_tI\,(P^k\ (p_i, t_i\text{-}\ t))$ e quindi ${}_tI\,(P_i^k[t_i\text{-}t])$
- se E^k è un collettivo di *nuovo ingresso* per P^k, allora ${}_tI(E_i^k[t_i])$ implica ${}_tI_{TEMP}\ (P^k\ (p_i, t_i\text{-}\ t))$ e quindi ${}_tI_{TEMP}\,(P_i^k[t_i\text{-}t])$
- se E^k è un collettivo di *uscita definitiva* per P^k, allora ${}_tI(E_i^k[t_i])$ implica ${}_tI\,(P^k\ (p_i, t_i\text{-}\ t_j))$
- se E^k è un collettivo di *uscita temporanea* per P^k, allora ${}_tI(E_i^k[t_i])$ implica ${}_tI_{TEMP}\,(P^k\ (p_i, t_i\text{-}\ t_j))$

dove in tutti questi casi il record $E_i^k[t_i])$ contiene un enunciato di possesso di relazione $C_l^k\ (e_i, pi, t_i)$ che lega e_i a $p_{i..}$

[19] nella pratica spesso una caratteristica o relazione opzionale può essere considerato come obbligatoria utilizzato una modalità o un elemento legato fittizio nel caso di modalità o elemento legato non noto; considerare caratteristiche e relazioni opzionali non introdurrebbe comunque rilevanti complicazioni.

Analogamente per gli eventi con durata:

- se E^k è un collettivo di *unico* o *primo ingresso* per D^k, allora ${}_t\mathrm{I}(E_i^k[t_i])$ implica ${}_t\mathrm{I}\,(D^k\,(p_i, t_i\text{-}t))$ e quindi ${}_t\mathrm{I}\,(D_i^k[t_i\text{-}t])$
- se E^k è un collettivo di *nuovo ingresso* (raro) per D^k, allora ${}_t\mathrm{I}(E_i^k[t_i])$ implica ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(D^k\,(p_i, t_i\text{-}t))$ e quindi ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(D_i^k[t_i\text{-}t])$
- se E^k è un collettivo di *uscita definitiva* per D^k, allora ${}_t\mathrm{I}(E_i^k[t_i])$ implica ${}_t\mathrm{I}\,(D^k\,(p_i, t_i\text{-}t_j))$
- se E^k è un collettivo di *uscita temporanea* (raro) per D^k, allora ${}_t\mathrm{I}(E_i^k[t_i])$ implica ${}_t\mathrm{I}_{\mathrm{TEMP}}\,(D^k\,(p_i, t_i\text{-}t_j))$

dove in tutti questi casi il record $E_i^k[t_i]$) contiene un enunciato di possesso di relazione $C_l^k\ (e_i, di, t_i)$ che lega e_i a d_i.

Le operazioni nelle classi REG ed ELIM: ESEMPI di successioni di ingressi e uscite per elementi dei collettivi di tipo popolazione o evento con durata

Di seguito si presentano esempi di operazioni di ingresso e uscita, riguardanti elementi di popolazioni o eventi con durata, ed esempi della serie di operazioni possibili nel tempo. Si vedano in proposito i disegni da pagina 50.

Esempi di eventi di ingresso per collettivi di tipo popolazione: l'evento di Immatricolazione (primo ingresso) e l'evento di Iscrizione (nuovo ingresso) per il collettivo Studente, l'evento di Prima accettazione degente (primo ingresso) e di Accettazione degente (nuovo ingresso) per il collettivo Degente, l'evento Primo avvio di un rapporto di lavoro (primo ingresso) e l'evento Avvio dell'unico o principale rapporto di lavoro (nuovo ingresso) per il collettivo Lavoratore.

Esempi di eventi di uscita per collettivi di tipo popolazione: Laurea (uscita temporanea) per il collettivo Studente, Dimissione degente (uscita temporanea) per il collettivo Degente, Dimissione dall'unico rapporto di lavoro (uscita temporanea) o Licenziamento dall'unico rapporto di lavoro (uscita temporanea) per il collettivo Lavoratore.

Si può avere ad esempio la seguente serie di operazioni.

- ${}_t\mathrm{I}\ (Immatricolazione_i\ [1\ novembre\ 2015])$: l'immatricolazione imm_i entra nel collettivo Immatricolazioni, cioè
  - viene registrato un record $Immatricolazione_i$ [1 *novembre 2015*] relativo all'immatricolazione imm_i, contenente data, corso di laurea e altre sue caratteristiche tra cui la relazione $Riguarda\ (imm_i, Rossi, 1\ novembre\ 2015)$
- ${}_t\mathrm{I}\ (Immatricolazione_i\ [1\ novembre\ 2015])$ implica ${}_t\mathrm{I}\ (Studente_{Rossi}[1\ novembre\ 2015\ \text{-}\ t])$, quindi
  - viene registrato un record $Studente_{Rossi}[1\ novembre\ 2015\ \text{-}\ t]$ relativo a Rossi che entra *per la prima volta* nel collettivo Studente, contenente sesso, titolo di studio, e altre caratteristiche e relazioni
- ${}_t\mathrm{I}\,(Laurea_{Rossi}[1\ luglio\ 2018])$: la laurea lau_i entra nel collettivo Lauree, cioè
  - viene registrato un record relativo alla laurea lau_i, contenente data, corso di laurea e altre sue caratteristiche tra cui la relazione $Riguarda\ (lau_i, Rossi, 1\ luglio\ 2018)$
- ${}_t\mathrm{I}\ (Laurea_{Rossi}[1\ luglio\ 2018])$ implica ${}_t\mathrm{I}_{\mathrm{TEMP}}\ (Studente\ (Rossi, 1\ novembre\ 2015\ \text{-}\ 1\ luglio\ 2018)$, quindi

    - Rossi *esce temporaneamente* dal collettivo Studente per cui
    - il record $Studente_{Rossi}$[1 *novembre 1015* - t] viene sostituito dal record $IStudente_{Rossi}$[1 *novembre 1015* - 1 *luglio 2018*] nel quale l'enunciato di appartenenza al collettivo Studente ha periodo chiuso al 1 *luglio 2018*
- ${}_t\mathrm{I}$ ($Iscrizione_{Rossi}$[1 *novembre 2018*]): l'iscrizione *isc*$_i$ entra nel collettivo Iscrizioni, cioè
    - viene registrato un record relativo all'iscrizione *isc*$_i$, contenente data, corso di laurea e altre sue caratteristiche tra cui la relazione $Riguarda$ (*isc*$_i$, *Rossi*, 1 *luglio 2018*)
- ${}_t\mathrm{I}$ ($Iscrizione_{Rossi}$[1 *novembre 2018*]) implica ${}_t\mathrm{I}$ ($Studente_{Rossi}$[1 *novembre 2018* - t]), quindi
    - viene registrato un record relativo a Rossi che entra *di nuovo* nel collettivo Studente, contenente sesso, titolo di studio, e altre caratteristiche e relazioni

Esempi di eventi di ingresso per collettivi di eventi con durata: l'evento di Immatricolazione (unico ingresso) e l'evento di Iscrizione (unico ingresso) per il collettivo Carriera dello studente, l'evento di Prima accettazione degente (unico ingresso) e di Accettazione degente (unico ingresso) per il collettivo Ricovero ospedaliero, l'evento Avvio di un rapporto di lavoro (unico ingresso) per il collettivo Rapporto di lavoro. [20]

Esempi di eventi di uscita per collettivi di eventi con durata: Laurea (uscita definitiva) per il collettivo Carriera dello studente, Dimissione degente (uscita definitiva) per il collettivo Ricovero ospedaliero, Dimissione (uscita definitiva) o Licenziamento (uscita definitiva) per il collettivo Rapporto di lavoro.

Si può avere ad esempio la seguente serie di operazioni.

- ${}_t\mathrm{I}$ ($Avvio\ rapporto\ lavoro_i$ [1 *gennaio 2015*]): l'avvio *avv*$_i$ entra nel collettivo Comunicazioni di avvio, cioè
    - viene registrato un record $Avvio\ rapporto\ lavoro_i$ [1 *gennaio 2015*] relativo all'avvio *avv*, contenente data, modalità di comunicazione ed eventuali altre sue caratteristiche tra cui la relazione $Riguarda\_rapporto$ (*avv*$_i$, *Rapporto*$_i$, 1 *gennaio 2015*),ed anche le relazioni con Lavoratore e Datore di lavoro $Riguarda\_lavoratore$ (*avv*$_i$, *Bianchi*, 1 *gennaio 2015*), $Riguarda\_datore$ (*avv*$_i$, *Fiat*, 1 *gennaio 2015*)
- ${}_t\mathrm{I}$ ($Avvio\ rapporto\ lavoro_i$ [1 *gennaio 2015*]) implica ${}_t\mathrm{I}$ ($Rapporto_i$ [1 *gennaio 2015* - t]), quindi
    - viene registrato un record $Rapporto_i$ [1 *gennaio 2015* - t] relativo a *Rapporto*$_i$ che entra nel collettivo Rapporti di lavoro, contenente contratto, durata, e altre caratteristiche e relazioni tra le quali $Riguarda\_lavoratore$ (*Rapporto*$_i$, *Bianchi*, 1 *gennaio 2015*), $Riguarda\_datore$ (*Rapporto*$_i$, *Fiat*, 1 *gennaio 2015*)
- ${}_t\mathrm{I}$ ($Licenziamento_i$ [1 *gennaio 2018*]): il licenziamento *lic*$_i$ entra nel collettivo Licenziamenti, cioè
    - viene registrato un record $Licenziamento_i$ [1 *gennaio 2018*] relativo al licenziamento *lic*$_i$, contenente data, modalità di comunicazione ed eventuali altre sue caratteristiche tra cui la relazione $Riguarda\_rapporto$ (*lic*$_i$, *Rapporto*$_i$, 1 *gennaio*

[20] In questo esempio vale la pena di notare che il collettivo di eventi Avvio di un rapporto di lavoro, che è di ingresso per l'evento con durata Rapporto di lavoro, ha come sottoinsiemi i collettivi di eventi Primo avvio di un rapporto di lavoro e Nuovo avvio di un rapporto di lavoro (dopo chiusura), che sono di ingresso per il collettivo di tipo popolazione Lavoratore, e considerazioni simili valgono per gli eventi di uscita. Ciò dipende dal fatto che un lavoratore può avere più rapporti di lavoro, per cui tutti gli eventi di avvio danno inizio a un rapporto di lavoro, ma solo alcuni fanno diventare una persona un lavoratore, e analogamente per gli eventi di uscita.

2015),ed anche le relazioni con Lavoratore e Datore di lavoro $Riguarda\_lavoratore$ *(lic$_i$, Bianchi,* 1 *gennaio 2015)*, $Riguarda\_datore$ *(lic$_i$, Fiat,* 1 *gennaio 2015)*

- $_t$I ($Licenziamento_i$ [1 *gennaio 2018*]) implica $_t$I (*Rapporto di lavoro (Rapporto$_i$,* 1 *gennaio 2015* - 1 *gennaio 2018*)), quindi
  - *Rapporto$_i$* esce definitivamente dal collettivo Rapporti di lavoro per cui:
  - il record $Rapporto_i$ [1 *gennaio 2015* - *t*], viene sostituito dal record $Rapporto_i$ [1 *gennaio 2015* - 1 *gennaio 2018* nel quale l'enunciato di appartenenza al collettivo Rapporti di lavoro ha periodo chiuso al 1 *gennaio 2018*
- $_t$I ($Avvio\ rapporto\ lavoro_i$, [1 *gennaio 2019*]): l'avvio *avv$_{i'}$* entra nel collettivo Comunicazioni di avvio, cioè
  - viene registrato un record $Avvio\ rapporto\ lavoro_i$, [1 *gennaio 2019*] relativo all'avvio *avv$_{i'}$*, contenente data, modalità di comunicazione ed eventuali altre sue caratteristiche tra cui la relazione $Riguarda\_rapporto$ *(avv$_{i'}$, Rapporto$_{i'}$,* 1 *gennaio 2019)*,ed anche le relazioni con Lavoratore e Datore di lavoro $Riguarda\_lavoratore$ *(avv$_{i'}$, Bianchi,* 1 *gennaio 2019)*, $Riguarda\_datore$ *(avv$_{i'}$, Fiat,* 1 *gennaio 2019)* (o un diverso datore di lavoro)
- $_t$I (*$Avvio\ rapporto\ lavoro_i$,* [1 *gennaio 2019*]) implica $_t$I ($Rapporto_i$, [1 *gennaio 2019* - *t*]), quindi
  - viene registrato un record $Rapporto_i$, [1 *gennaio 2019* - *t*] relativo a *Rapporto$_{i'}$* che entra nel collettivo Rapporti di lavoro, contenente contratto, durata, e altre caratteristiche e relazioni tra le quali $Riguarda\_lavoratore$ *(Rapporto$_{i'}$, Bianchi,* 1 *gennaio 2019)*, $Riguarda\_datore$ *(Rapporto$_{i'}$, Fiat,* 1 *gennaio 2019)* (o un diverso datore di lavoro).

Come mostrano gli esempi, lo stesso collettivo può essere di ingresso, oppure di uscita, per uno o più collettivi di tipo popolazione o evento con durata.

Le operazioni nella classe MOD: ESEMPI di aggiornamento delle caratteristiche e relazioni per elementi dei collettivi di tipo popolazione o evento con durata

Nel seguito mostriamo a titolo di esempio gli effetti di un cambiamento di residenza e di un cambiamento di tutor, per uno studente. Si vedano anche i disegni da pagina 50.

- $_t$I ($Immatricolazione_i$ [1 *novembre 2015*]): l'immatricolazione *imm$_i$* entra nel collettivo Immatricolazioni, cioè
  - viene registrato un record $Immatricolazione_i$ [1 *novembre 2015*] relativo all'immatricolazione *imm$_i$*, contenente data, corso di laurea e altre sue caratteristiche tra cui la relazione $Riguarda$ *(imm$_i$, Rossi,* 1 *novembre 2015)*
- $_t$I ($Immatricolazione_i$ [1 *novembre 2015*]) implica $_t$I ($Studente_{Rossi}$[1 *novembre 2015* - *t*]), quindi
  - viene registrato un record $Studente_{Rossi}$[1 *novembre 2015* - *t*] relativo a Rossi che entra *per la prima volta* nel collettivo Studente, contenente enunciati relativi a sesso, titolo di studio e altre caratteristiche e relazioni, tra le quali la caratteristica *Residenza*, per cui ($Residenza$ *(Rossi, Milano,* 1 *novembre 2015* - *t)*, e la relazione *Ha tutor*, per cui ($Ha\ tutor$ *(Rossi, Bianchi,* 1 *novembre 2015* - *t)*
- $_t$I ($Residenza$ *(Rossi, Roma, 3 marzo 2016* - *t)*): Rossi cambia la sua residenza, spostandosi a Roma

- il record $Studente_{Rossi}$[1 *novembre 2015 - t*] viene sostituito dal record $Studente_{Rossi}$[1 *novembre 2015 - 3 marzo 2016*], e al suo interno (*Residenza (Rossi, Milano,* 1 *novembre 2015 - t)* viene sostituito da (*Residenza (Rossi, Milano,* 1 *novembre 2015 - 3 marzo 2016)*
- viene registrato un record $Studente_{Rossi}$[*3 marzo 2016 - t*] relativo a Rossi nel quale ci sono gli stessi enunciati di $Studente_{Rossi}$[1 *novembre 2015 - 3 marzo 2016*] salvo che al posto di (*Residenza (Rossi, Milano,* 1 *novembre 2015 - 3 marzo 2016)* si trova (*Residenza (Rossi, Roma, 3 marzo 2016 - t)*

- ${}_{t}$I (*Ha tutor (Rossi, Verdi, 1 gennaio 2017 - t)*): Rossi cambia la sua relazione con il tutor, che ora è Verdi
  - il record $Studente_{Rossi}$[*3 marzo 2016 - t*] viene sostituito dal record $Studente_{Rossi}$[*3 marzo 2016 - 1 gennaio 2017*], e al suo interno (*Ha tutor (Rossi, Bianchi,* 1 *novembre 2015 - t)* viene sostituito da (*Ha tutor (Rossi, Bianchi,* 1 *novembre 2015 - 1 gennaio 2017)*
  - viene registrato un record $Studente_{Rossi}$[*1 gennaio 2017 - t*] relativo a Rossi nel quale ci sono gli stessi enunciati di $Studente_{Rossi}$[*3 marzo 2016 - 1 gennaio 2017*] salvo che al posto di (*Ha tutor (Rossi, Bianchi,* 1 *novembre 2015 - 1 gennaio 2017)* si trova (*Ha tutor (Rossi, Verdi, 1 gennaio 2017 - t)*
  - notare che nel record $Studente_{Rossi}$[*1 gennaio 2017 - t*] ci sono enunciati con momento d'inizio diverso, come: (*Studente (Rossi,* 1 *novembre 2015* - t) (l'enunciato di appartenenza), (*Sesso (Rossi, maschio,* 1 *novembre 2015 - t),* (*Residenza (Rossi, Roma, 3 marzo 2016 – t),* (*Ha tutor (Rossi, Verdi, 1 gennaio 2017 - t)*
- ${}_{t}$I (*Laurea*$_{Rossi}$[1 *luglio 2018*]): la laurea *lau*$_i$ entra nel collettivo Lauree, cioè
  - viene registrato un record relativo alla laurea *lau*$_i$, contenente data, corso di laurea e altre sue caratteristiche tra cui la relazione *Riguarda (lau*$_i$*, Rossi,* 1 *luglio 2018)*
- ${}_{t}$I (*Laurea*$_{Rossi}$[1 *luglio 2018*]) implica ${}_{t}$I$_{TEMP}$ (*Studente (Rossi,* 1 *novembre 2015* - 1 *luglio 2018*)), quindi
  - Rossi *esce temporaneamente* dal collettivo Studente per cui
  - il record $Studente_{Rossi}$[*1 gennaio 2017 - t*] viene sostituito dal record $IStudente_{Rossi}$[*1 gennaio 2017* - 1 *luglio 2018*] nel quale l'enunciato di appartenenza al collettivo Studente ha periodo chiuso al 1 *luglio 2018*
  - anche tutti gli altri enunciati vengono chiusi, perché non sono più pertinenti, come la relazione *Ha tutor* o non sono più osservabili, come la caratteristica *Residenza*[21].

Serie di record presenti in archivio come effetto della dinamica degli aggiornamenti, per ogni elemento di popolazione o evento con durata

Ogni archivio amministrativo ha associato un suo periodo di osservazione T-t_A, che inizia con il momento d'impianto dell'archivio T e termina con il momento attuale t_A, e quindi si allunga con lo scorrere del tempo.

Nel corso del tempo cioè il momento attuale t_A viene a corrispondere a determinazioni temporali successive, in ciascuna delle quali l'archivio compie operazioni di aggiornamento, accettando nuova informazione riferita al momento t_A [22]. Precisamente, per ogni momento t_A l'archivio:

[21] Questo è vero solo perché Studente è un collettivo principale per l'archivio, non è vero per i collettivi sottoinsieme

- accetta una serie di nuovi enunciati di appartenenza ai collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi ponendo in essi $t_i=t_A$, dove t_i è il momento di riferimento, per gli enunciati riferiti a eventi istantanei, oppure il momento d'inizio del periodo aperto di validità dell'enunciato, per gli enunciati riferiti a unità di popolazioni o a eventi con durata
- chiude il periodo di validità, ponendo in esso $t_j=t_A$, per una serie di enunciati di appartenenza ai collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi, riferiti a unità di popolazioni o a eventi con durata.

Consideriamo più in dettaglio il complesso dell'informazione che può essere gestita nell'archivio amministrativo, per ogni possibile elemento di un collettivo.

Per ogni evento istantaneo e_i entrato in archivio in un momento $t_I \geq T$ può esistere al più un unico enunciato, riferito temporalmente al momento $t_i = t_I$.

Invece per ogni unità di popolazione p_i o evento con durata d_i entrato per la prima volta in archivio in un momento $t_I \geq T$ si possono avere in momenti t_A successivi a t_I una serie di aggiornamenti consecutivi relativi alle caratteristiche o relazioni e/o, se l'elemento appartiene a un collettivo multingresso, una serie di uscite ed ingressi temporanei nel collettivo. Questa serie termina con l'uscita definitiva dell'elemento dal collettivo, ad un momento t_U.

Come conseguenza in un generico momento attuale t_A, con $t_A \geq t_I$, per l'elemento p_i o d_i si possono trovare in archivio una serie di record con periodi di validità chiusi consecutivi, ciascuno generato da uno degli aggiornamenti effettuati in momenti precedenti.

Indichiamo con ${}_iC_a$ una serie di aggiornamenti chiusa, con ${}_iA_a$ una serie di aggiornamenti aperta.

Un serie ${}_iC_a$ inizia con un aggiornamento di tipo *REG*, cioè con un ingresso nel collettivo per l'elemento p_i o d_i, può continuare con una serie di aggiornamenti di tipo *MOD*, cioè cambiamenti nelle caratteristiche o nelle relazioni, e termina con un aggiornamento di tipo *ELIM*, cioè con un'uscita dal collettivo dell'elemento p_i o d_i. Se ${}_iN_a$ è il numero di aggiornamenti consecutivi che costituiscono la serie ${}_iA_a$, ci sono in archivio ${}_iN_a - 1$ record chiusi con date d'inizio successive relativi a p_i o d_i.

Come caso particolare, una serie ${}_iC_a$ senza aggiornamenti di tipo *MOD* contiene due soli aggiornamenti di tipo *REG* ed *ELIM* rispettivamente e dà luogo in archivio a un unico record con periodo chiuso riferito a p_i o d_i.

Un serie ${}_iA_a$ inizia con un aggiornamento di tipo *REG*, cioè con un ingresso nel collettivo per l'elemento p_i o d_i, e può continuare con una serie di aggiornamenti di tipo *MOD*, cioè cambiamenti nelle caratteristiche o nelle relazioni. Se ${}_iN_a$ è il numero di aggiornamenti consecutivi che costituiscono la serie ${}_iA_a$, ci sono in archivio ${}_iN_a - 1$ record chiusi con date d'inizio successive relativi a p_i o d_i, e un ultimo record aperto, con la data d'inizio più recente.

Come caso particolare, una serie ${}_iA_a$ senza aggiornamenti di tipo *MOD* contiene il solo aggiornamento di tipo *REG* e dà luogo in archivio a un unico record con periodo aperto riferito a p_i o d_i.

[22] Per effetto dell'errore di tempestività, può accadere che la nuova informazione riferita al momento t_A sia accettata in pratica in un momento t successivo a t_A

Ad ogni generico momento attuale t_A per l'elemento p_i o d_i si saranno succedute K_A serie di aggiornamenti, di cui K_A -1 serie di tipo ${}_iC_a$, ed un'ultima serie che può essere di tipo ${}_iC_a$ oppure ${}_iA_a$.

Si avranno perciò in archivio ${}_iN_1 + \ldots + {}_iN_{K_A} - K_A$ record riferiti a p_i o d_i con periodi chiusi aventi date d'inizio successive, e in più un ultimo record con periodo aperto che ha la data d'inizio più recente, solo se l'ultima serie è aperta.

Nel momento t_A poi, se l'ultima serie è chiusa può intervenire solo un aggiornamento di tipo *REG* che inaugura una nuova serie aperta per l'elemento p_i o d_i, se invece l'ultima serie è aperta può intervenire un aggiornamento di tipo *MOD* che l'allunga oppure un aggiornamento di tipo *ELIM* che la chiude.

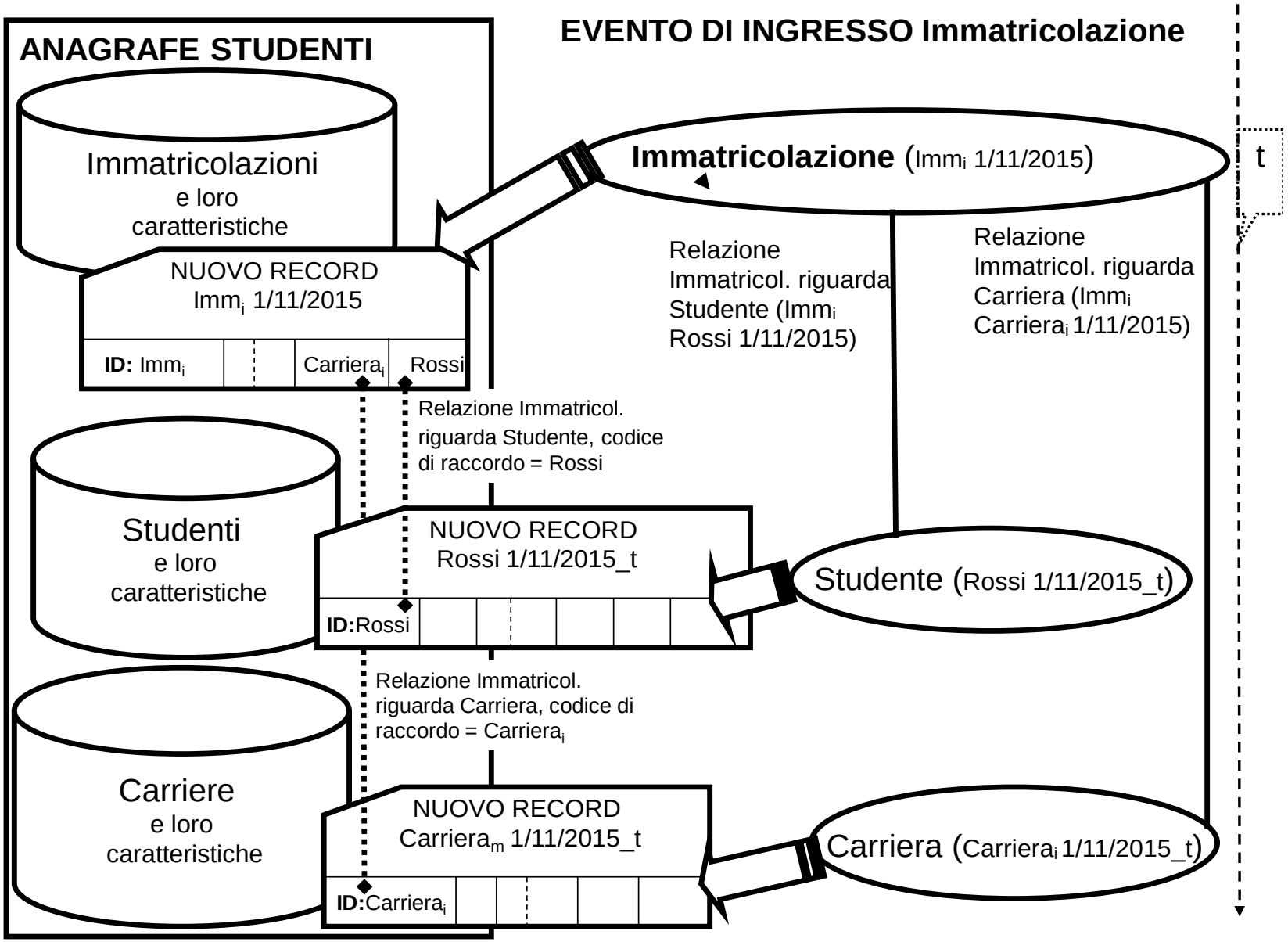
All'uscita definitiva dell'elemento p_i o d_i dal collettivo, per $t_A = t_U$, si saranno succedute K_U serie di aggiornamenti tutte di tipo ${}_iC_a$, si avranno perciò in archivio in via definitiva ${}_iN_1 + \ldots + {}_iN_{K_U} - K_U$ record chiusi con date d'inizio successive relativi a p_i o d_i, senza possibilità di aggiornamento ulteriore.

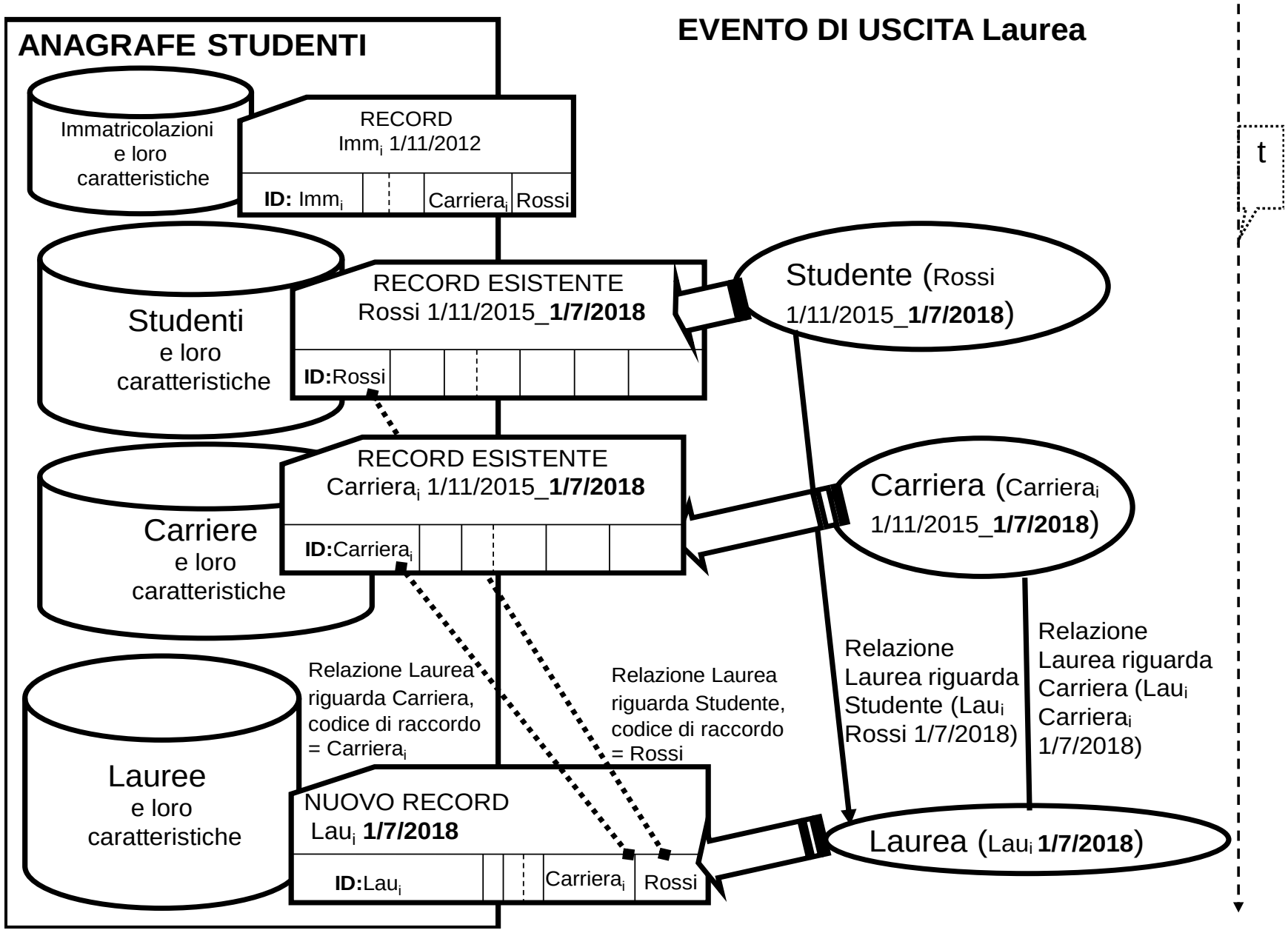
Queste successioni di record rappresentano concettualmente il contenuto dell'archivio, come risultato della dinamica di aggiornamento. Per quanto detto dovrebbe essere evidente che questi record non sono come tali materialmente reperibili in archivio ma sono ricostruibili nella misura in cui l'archivio gestisce uno storico dell'appartenenza ai collettivi e/o delle variazioni relative al possesso di caratteristiche o relazioni.

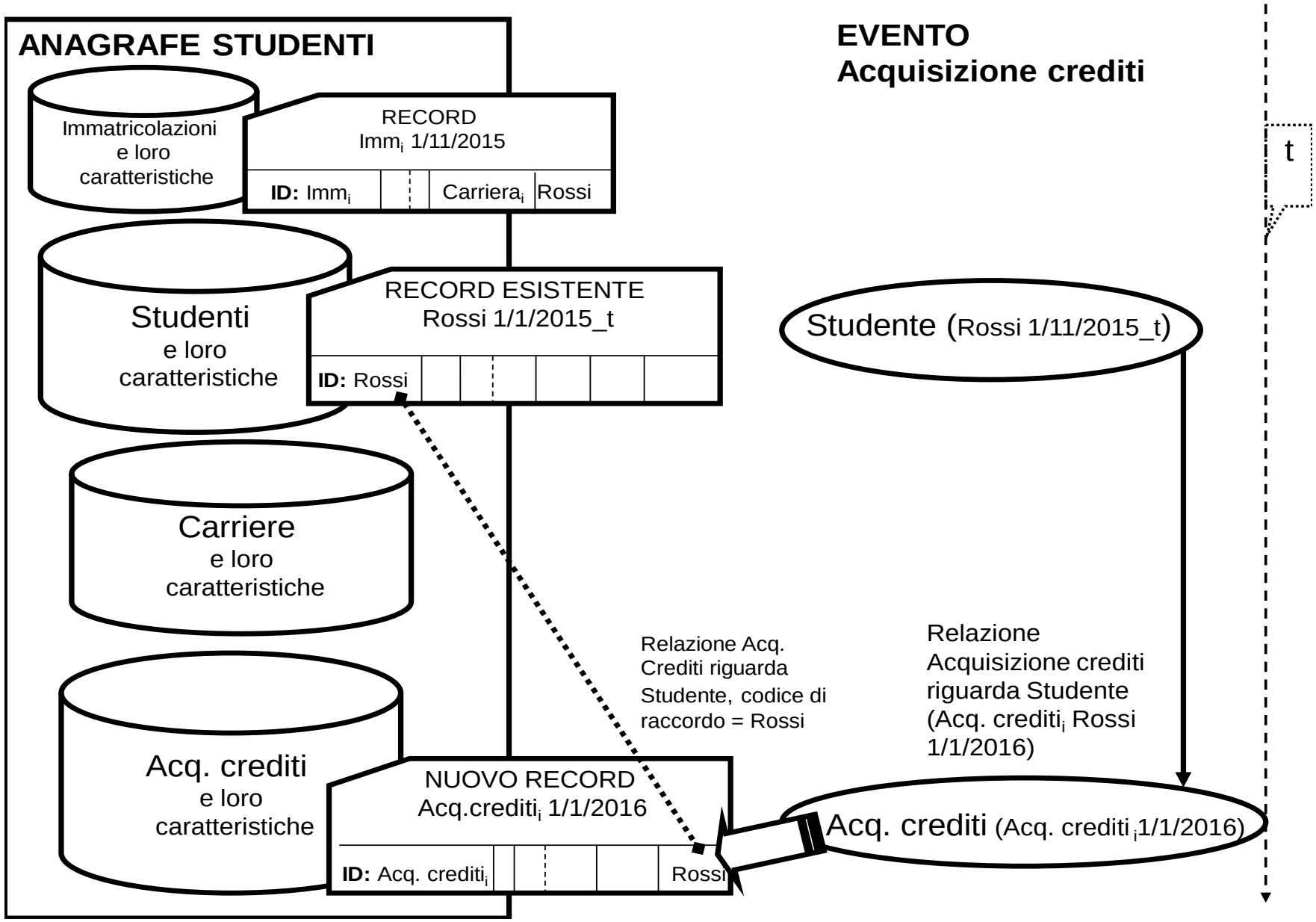
Tutti questi record, sia quelli con periodo chiuso che quelli con periodo aperto, possono essere affetti da errori, quando sono affetti da errori gli enunciati che li compongono. E' appena il caso di notare che in assenza di errori tutti gli enunciati, e di conseguenza tutti i record che li contengono, sono da considerarsi concettualmente veri, compresi gli enunciati con periodo chiuso corrispondenti a situazioni non più attuali ad un momento t.

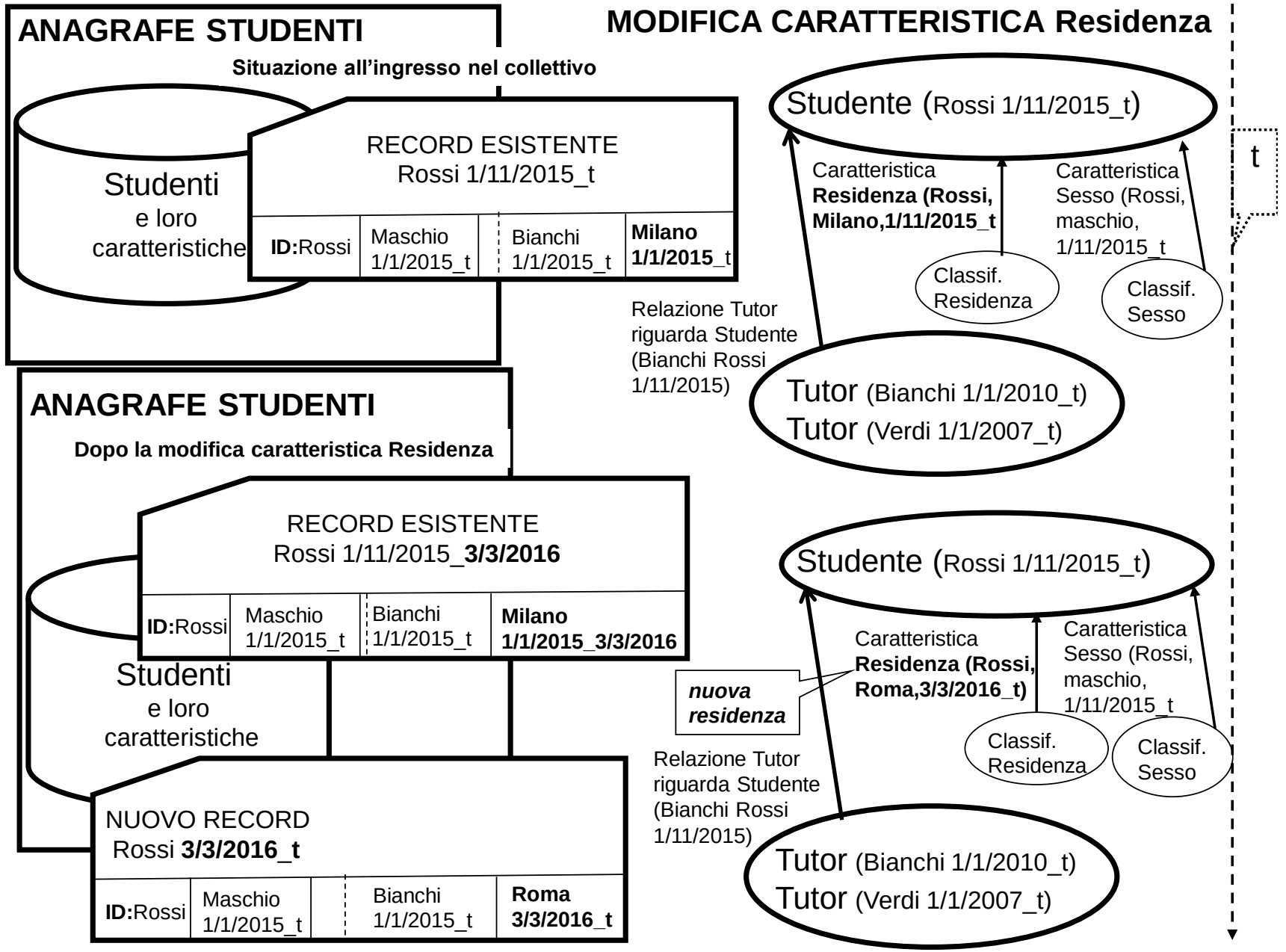
Gli errori si generano a causa di una errata o mancata corrispondenza tra i cambiamenti di qualsiasi tipo relativi agli elementi e_i o p_i o d_i che si verificano nella realtà e gli aggiornamenti che avvengono nell'archivio e, inoltre, a causa di una difformità tra il contenuto degli enunciati accettati in archivio e il contenuto degli enunciati veri, cioè che descrivono la realtà.

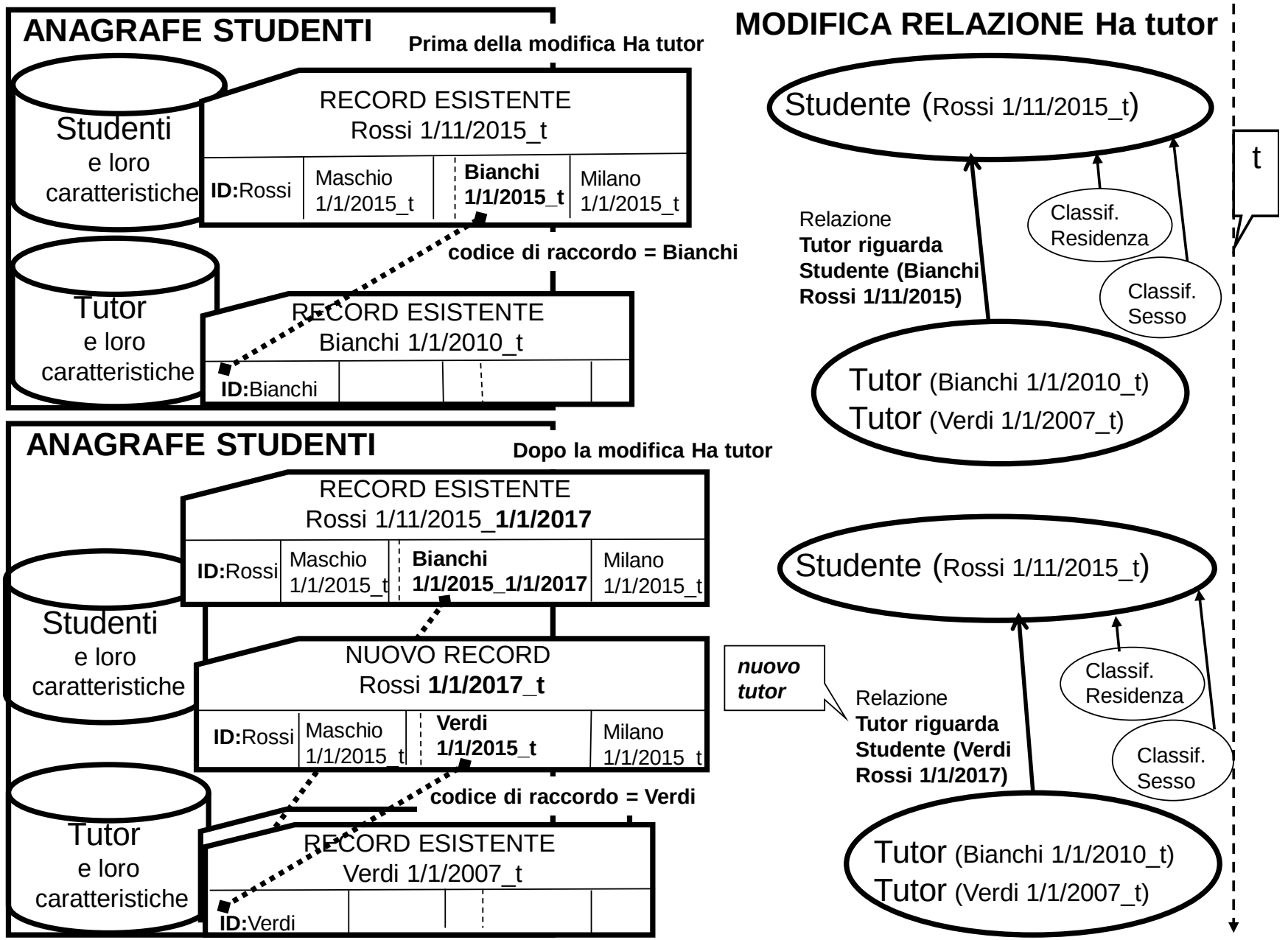
A questo è dedicato il capitolo successivo.


ANAGRAFE STUDENTI
EVENTO DI INGRESSO Immatricolazione
Immatricolazioni
e loro
caratteristiche
Immatricolazione (Imm$_i$ 1/11/2015)
t
NUOVO RECORD
Imm$_i$ 1/11/2015
ID: Imm$_i$
Carriera$_i$
Rossi
Relazione
Immatricol. riguarda
Studente (Imm$_i$
Rossi 1/11/2015)
Relazione
Immatricol. riguarda
Carriera (Imm$_i$
Carriera$_i$ 1/11/2015)
Relazione Immatricol.
riguarda Studente, codice
di raccordo = Rossi
Studenti
e loro
caratteristiche
NUOVO RECORD
Rossi 1/11/2015_t
ID:Rossi
Studente (Rossi 1/11/2015_t)
Relazione Immatricol.
riguarda Carriera, codice di
raccordo = Carriera$_i$
Carriere
e loro
caratteristiche
NUOVO RECORD
Carriera$_m$ 1/11/2015_t
ID:Carriera$_i$
Carriera (Carriera$_i$ 1/11/2015_t)


ANAGRAFE STUDENTI
EVENTO DI USCITA Laurea
t
Immatricolazioni e loro caratteristiche
RECORD Imm$_i$ 1/11/2012
ID: Imm$_i$
Carriera$_i$
Rossi
Studenti e loro caratteristiche
RECORD ESISTENTE Rossi 1/11/2015_1/7/2018
ID:Rossi
Studente (Rossi 1/11/2015_1/7/2018)
Carriere e loro caratteristiche
RECORD ESISTENTE Carriera$_i$ 1/11/2015_1/7/2018
ID:Carriera$_i$
Carriera (Carriera$_i$ 1/11/2015_1/7/2018)
Relazione Laurea riguarda Carriera, codice di raccordo = Carriera$_i$
Relazione Laurea riguarda Studente, codice di raccordo = Rossi
Relazione Laurea riguarda Studente (Lau$_i$ Rossi 1/7/2018)
Relazione Laurea riguarda Carriera (Lau$_i$ Carriera$_i$ 1/7/2018)
Lauree e loro caratteristiche
NUOVO RECORD Lau$_i$ 1/7/2018
ID:Lau$_i$
Carriera$_i$
Rossi
Laurea (Lau$_i$ 1/7/2018)


ANAGRAFE STUDENTI
Immatricolazioni e loro caratteristiche
RECORD
Imm$_i$ 1/11/2015
ID: Imm$_i$
Carriera$_i$
Rossi
Studenti e loro caratteristiche
RECORD ESISTENTE
Rossi 1/1/2015_t
ID: Rossi
Carriere e loro caratteristiche
Acq. crediti e loro caratteristiche
NUOVO RECORD
Acq.crediti$_i$ 1/1/2016
ID: Acq. crediti$_i$
Rossi
Relazione Acq. Crediti riguarda Studente, codice di raccordo = Rossi
EVENTO
Acquisizione crediti
t
Studente (Rossi 1/11/2015_t)
Relazione Acquisizione crediti riguarda Studente (Acq. crediti$_i$ Rossi 1/1/2016)
Acq. crediti (Acq. crediti$_i$ 1/1/2016)


ANAGRAFE STUDENTI
Situazione all'ingresso nel collettivo
Studenti
e loro
caratteristiche
RECORD ESISTENTE
Rossi 1/11/2015_t
ID:Rossi
Maschio
1/1/2015_t
Bianchi
1/1/2015_t
Milano
1/1/2015_t
MODIFICA CARATTERISTICA Residenza
Studente (Rossi 1/11/2015_t)
t
Caratteristica
Residenza (Rossi,
Milano,1/11/2015_t
Caratteristica
Sesso (Rossi,
maschio,
1/11/2015_t
Classif.
Residenza
Classif.
Sesso
Relazione Tutor
riguarda Studente
(Bianchi Rossi
1/11/2015)
Tutor (Bianchi 1/1/2010_t)
Tutor (Verdi 1/1/2007_t)
ANAGRAFE STUDENTI
Dopo la modifica caratteristica Residenza
RECORD ESISTENTE
Rossi 1/11/2015_3/3/2016
ID:Rossi
Maschio
1/1/2015_t
Bianchi
1/1/2015_t
Milano
1/1/2015_3/3/2016
Studenti
e loro
caratteristiche
NUOVO RECORD
Rossi 3/3/2016_t
ID:Rossi
Maschio
1/1/2015_t
Bianchi
1/1/2015_t
Roma
3/3/2016_t
Studente (Rossi 1/11/2015_t)
nuova
residenza
Caratteristica
Residenza (Rossi,
Roma,3/3/2016_t)
Caratteristica
Sesso (Rossi,
maschio,
1/11/2015_t
Classif.
Residenza
Classif.
Sesso
Relazione Tutor
riguarda Studente
(Bianchi Rossi
1/11/2015)
Tutor (Bianchi 1/1/2010_t)
Tutor (Verdi 1/1/2007_t)


ANAGRAFE STUDENTI
Prima della modifica Ha tutor
Studenti
e loro
caratteristiche
RECORD ESISTENTE
Rossi 1/11/2015_t
ID:Rossi
Maschio
1/1/2015_t
Bianchi
1/1/2015_t
Milano
1/1/2015_t
codice di raccordo = Bianchi
Tutor
e loro
caratteristiche
RECORD ESISTENTE
Bianchi 1/1/2010_t
ID:Bianchi
MODIFICA RELAZIONE Ha tutor
Studente (Rossi 1/11/2015_t)
t
Relazione
Tutor riguarda
Studente (Bianchi
Rossi 1/11/2015)
Classif.
Residenza
Classif.
Sesso
Tutor (Bianchi 1/1/2010_t)
Tutor (Verdi 1/1/2007_t)
ANAGRAFE STUDENTI
Dopo la modifica Ha tutor
RECORD ESISTENTE
Rossi 1/11/2015_1/1/2017
ID:Rossi
Maschio
1/1/2015_t
Bianchi
1/1/2015_1/1/2017
Milano
1/1/2015_t
Studenti
e loro
caratteristiche
NUOVO RECORD
Rossi 1/1/2017_t
ID:Rossi
Maschio
1/1/2015_t
Verdi
1/1/2015_t
Milano
1/1/2015_t
codice di raccordo = Verdi
Tutor
e loro
caratteristiche
RECORD ESISTENTE
Verdi 1/1/2007_t
ID:Verdi
Studente (Rossi 1/11/2015_t)
nuovo
tutor
Relazione
Tutor riguarda
Studente (Verdi
Rossi 1/1/2017)
Classif.
Residenza
Classif.
Sesso
Tutor (Bianchi 1/1/2010_t)
Tutor (Verdi 1/1/2007_t)

I DIVERSI TIPI DI ERRORE

I diversi tipi di errore in generale

Come si è visto, si possono distinguere due tipi di errori teorici, *errori relativi agli identificativi contenuti negli enunciati ed errori relativi all'accettazione nella fonte di enunciati falsi e alla mancata accettazione di enunciati veri nonché all'accettazione ripetuta di enunciati (duplicazione)*.

Consideriamo dapprima gli *errori sugli identificativi*.

Infatti dato un enunciato, prima di poterlo giudicare vero o falso occorre che l'enunciato stesso sia riferibile senza errore agli elementi dei quali si vuole enunciare l'appartenenza a collettivi, caratteristiche o relazioni, occorre cioè che sia possibile assegnare un identificativo a tutti gli elementi dei quali si vuole enunciare l'appartenenza, senza ambiguità.

Gli identificativi possono presentare i seguenti errori:

- *errori sintattici*: riguardano la presenza o assenza dell'identificativo, la sua correttezza intesa come appartenenza all'insieme di identificativi previsti dal sistema di identificazione adottato o, se si tratta di un identificativo strutturato, la sua corretta costruzione e la correttezza delle sue componenti
- *errori semantici*: riguardano l'effettiva capacità di identificare, ciò significa che: a) ogni elemento in ogni momento t_A ha un identificativo, b) ogni elemento in ogni momento t_A ha un solo identificativo c) ogni identificativo in ogni momento t_A identifica un solo elemento d) ogni elemento mantiene lo stesso identificativo in tutti i momenti t_A che si succedono nel corso del tempo..

Distinguiamo tra identificativi degli elementi dei collettivi di tipo popolazione o evento, e_i o p_i o d_i (o genericamente u_i, u_j), e identificativi delle modalità o valori c_i.

Per gli identificativi delle modalità, o valori:

- non esiste un errore semantico, in quanto le modalità e i valori appartengono a un insieme prefissato a priori di elementi pre-identificati, appunto la classificazione o il dominio/range utilizzato
- il relativo errore sintattico consiste nell'uso di una modalità o di un valore non previsto ed è comunemente diagnosticato come modalità non prevista per la classificazione utilizzata o valore fuori range.

I diversi tipi di identificativi e gli errori che li riguardano sono discussi nella PARTE QUARTA (in corso di revisione), in generale comunque si può affermare quanto segue.

Gli *errori sintattici* su tutti i tipi di identificativi, e_i o p_i o d_i (o genericamente u_i, u_j), c_i, sono sempre evidenti e direttamente diagnosticabili.

Gli *errori semantici* sugli identificativi riguardano solo gli identificativi degli elementi dei collettivi di tipo popolazione o evento e_i o p_i o d_i (o genericamente u_i, u_j) e sono quelli che compromettono il requisito dell'effettiva capacità di identificare, cioè:

a) nel sistema di identificazione adottato, alcuni elementi non hanno assegnato un proprio identificativo
b) nel sistema di identificazione adottato, ci sono identificativi condivisi tra più elementi
c) nel sistema di identificazione adottato, ci sono identificativi per elementi non esistenti
d) nel sistema di identificazione adottato, ci sono elementi che hanno un doppio identificativo
e) nel sistema di identificazione adottato, ci sono elementi che ricevono identificativi diversi in momenti diversi (analogo al precedente).

Va sottolineato che questi errori riguardano gli identificativi sia quando sono usati per identificare, sia quando sono usati come codici di raccordo, hanno cioè il ruolo u_j, in un enunciato di tipo $C(u_i, u_j, t_i)$, $C(u_i, u_j, t_i - t)$, $C(u_i, u_j, t_i - t_j)$.

Dal punto di vista dell'*accettazione dell'enunciato in archivio*, sono possibili in teoria due tipi di errori: *accettare enunciati falsi o non accettare enunciati veri*.

Di conseguenza in generale gli errori possono essere di tre tipi:

ERRORI DI IDENTIFICAZIONE

ERRONEA INCLUSIONE: accettazione di enunciati falsi

ERRONEA ESCLUSIONE: mancata accettazione di enunciati veri.

In linea di principio gli errori di identificazione, errata inclusione ed errata esclusione riguardano tutti e tre i tipi di enunciati che abbiamo individuato:

- gli enunciati di appartenenza di un elemento u_i al collettivo A di tipo popolazione o evento, istantaneo o con durata: $A(u_i, t_i)$ per i collettivi di tipo evento istantaneo, $A(u_i, t_i - t)$ e $A(u_i, t_i - t_j)$ per i collettivi di tipo popolazione o evento con durata

- gli enunciati relativi al possesso per una caratteristica B di una modalità o di un valore c_i da parte di un elemento u_i, di un collettivo: $B(u_i, c_i, t_i)$ per i collettivi di tipo evento istantaneo, $B(u_i, c_i, t_i - t)$ e $B(u_i, c_i, t_i - t_j)$ per i collettivi di tipo popolazione o evento con durata

- gli enunciati relativi alla relazione C tra l'elemento ei di un collettivo dominio e l'elemento e_j di un collettivo codominio: $C(u_i, u_j, t_i)$ se il collettivo nel ruolo di dominio è di tipo evento istantaneo, $C(u_i, u_j, t_i - t)$ e $C(u_i, u_j, t_i - t_j)$ se il collettivo nel ruolo di dominio è di tipo popolazione o evento con durata.

Inoltre, per tutti gli enunciati una differenza tra i riferimenti temporali effettivi degli enunciati t_i, $t_i - t$, $t_i - t_j$ e i riferimenti temporali per essi registrati in archivio provoca errori di **ERRONEA INCLUSIONE TEMPORANEA** ed **ERRONEA ESCLUSIONE TEMPORANEA.**

Come già osservato gli errori nei riferimenti temporali t_i o t_j sono distinti dagli ***errori di tempestività.***

Questi consistono nella ritardata registrazione in archivio di un enunciato con il suo periodo di riferimento, il quale poi può essere corretto o anch'esso sbagliato. L'errore di tempestività quindi riguarda l'operazione di aggiornamento e non direttamente l'informazione registrata. Ha comunque un suo importante impatto sull'usabilità a fini statistici dell'archivio. I due tipi di errori devono essere ben distinti e valutati separatamente nei casi concreti.

Infine, sia gli enunciati correttamente inclusi che quelli erroneamente inclusi possono essere **DUPLICATI** (si veda l'apposito paragrafo).

Per passare da questa specifica teorica degli errori alla specifica degli errori che si possono incontrare in pratica occorre:

- considerare *come i diversi tipi di errori possano concretamente occorrere in occasione delle diverse operazioni di creazione e modifica di record*, appartenenti alle classi *REG, ELIM, MOD*
- descrivere sistematicamente *tutte le combinazioni possibili tra errate inclusioni, errate esclusioni, duplicazioni e i diversi tipi di errori di identificazione*, distintamente per ciascuno tipo di enunciato, considerando quindi
    - per gli *enunciati che asseriscono l'appartenenza di un elemento a un collettivo*, le possibili combinazioni tra errori di accettazione nella fonte dell'enunciato ed errori sintattici o semantici di identificazione dell'elemento
    - per gli *enunciati che asseriscono l'associazione tra un elemento di un collettivo e una modalità, o valore*, tra quelli assumibili per una particolare caratteristica, le possibili combinazioni tra errori di accettazione nella fonte dell'enunciato ed errori sintattici o semantici di identificazione dell'elemento nonché errori sintattici di identificazione della modalità o valore (modalità non prevista o valore fuori range)
    - per gli *enunciati che asseriscono l'esistenza di una particolare relazione tra due elementi di collettivi* (due diversi collettivi o lo stesso collettivo) le possibili combinazioni tra errori di accettazione nella fonte dell'enunciato ed errori sintattici o semantici di identificazione sia dell'elemento di riferimento del dominio, sia dell'elemento di riferimento del codominio (in questo secondo caso, si tratta dell'errore nel codice identificativo utilizzato come codice di raccordo)

Nei paragrafi seguenti è presentata una prima discussione e classificazione di massima degli errori che possono occorrere in pratica in occasione delle diverse operazioni, anche con riferimento alle tradizionali nozioni di errore di copertura e di accuratezza.

Il lavoro di descrizione sistematica di *tutte le combinazioni possibili tra errate inclusioni, errate esclusioni, duplicazioni e i diversi tipi di errori di identificazione* distintamente per ciascuno tipo di enunciato è condotto nella PARTE QUARTA (in corso di revisione) e nella PARTE QUINTA (in corso di stesura) del presente Framework, dove sono discussi più approfonditamente i diversi tipi di errori.

Gli errori possibili in occasione di operazioni nella classe *REG*

Per le operazioni nella classe *REG*, il primo tipo di errore consiste nell'*accettazione di un falso enunciato di appartenenza di un elemento a un collettivo di tipo popolazione, evento istantaneo o evento con durata, che si traduce nella registrazione di un record falso.*

Si ha cioè **ERRATA INCLUSIONE di un enunciato di appartenenza falso, eventualmente TEMPORANEA,** e quindi di un intero record falso.

Per i collettivi di tipo evento istantaneo, si può avere:

- *errata inclusione* per *registrazione dell'ingresso di un elemento non esistente*, vale a dire accettazione di un enunciato di appartenenza al collettivo relativo a un evento non accaduto
- se il collettivo è sottoinsieme di un collettivo più ampio, *errata inclusione* per *registrazione dell'ingresso di un elemento non appartenente al collettivo*, vale a dire accettazione di un enunciato di appartenenza al collettivo relativo a un evento accaduto ma non appartenente al collettivo
- *molto improbabile dovrebbe essere l'errata inclusione temporanea* per *registrazione di un ingresso con riferimento temporale anticipato*, vale a dire la corretta accettazione di un enunciato di appartenenza al collettivo nel quale però il riferimento temporale è errato in quanto anticipa il riferimento vero.

Per i collettivi di tipo popolazione o evento con durata questo errore dipende da errori negli eventi di ingresso. Si possono avere i seguenti casi:

- *errata inclusione* per *registrazione dell'ingresso di un elemento non esistente*:
    - l'errata inclusione nel collettivo degli eventi di ingresso di un evento legato a un elemento non esistente (per errata inclusione di un evento di ingresso legato a un elemento non esistente) provoca
    - un'errata inclusione di un enunciato con periodo aperto di appartenenza dell'elemento al collettivo di tipo popolazione o evento con durata, cioè un'errata inclusione dell'elemento
- se il collettivo è sottoinsieme di un collettivo più ampio, *errata inclusione* per *registrazione dell'ingresso di un elemento non appartenente al collettivo*:
    - l'errata inclusione nel collettivo degli eventi di ingresso di un evento legato all'elemento (per errata inclusione dell'evento di ingresso e/o per errore nel legame, per cui l'ingresso viene riferito ad un elemento sbagliato) provoca
    - un'errata inclusione di un enunciato con periodo aperto di appartenenza dell'elemento al collettivo di tipo popolazione o evento con durata, cioè un'errata inclusione dell'elemento
- *errata inclusione temporanea* per registrazione di un ingresso temporaneamente non esistente, cioè *registrazione di un ingresso con riferimento temporale anticipato*, quando a un'errata inclusione segue successivamente un ingresso effettivo (improbabile invece un riferimento temporale anticipato nell'evento di ingresso).

Per le operazioni nella classe *REG*, l'altro tipo di errore consiste nella *mancata accettazione di un vero enunciato di appartenenza aperto di un elemento a un collettivo di tipo popolazione, evento istantaneo o evento con durata: questo errore si traduce nella mancata registrazione di un record corrispondente ad un elemento del collettivo.*

Si ha cioè **ERRATA ESCLUSIONE di un enunciato di appartenenza vero, eventualmente TEMPORANEA,** e quindi di un intero record vero.

Per i collettivi di tipo evento istantaneo si possono avere i seguenti casi:

- *errata esclusione* per *mancata registrazione di un ingresso*, vale a dire errata esclusione di un enunciato di appartenenza al collettivo relativo a un evento accaduto e appartenente al collettivo
- *errata esclusione temporanea* per *registrazione di un ingresso con riferimento temporale posticipato*, vale a dire corretta accettazione di un enunciato di appartenenza al collettivo nel quale però il riferimento temporale è errato, in quanto posticipa il riferimento vero.

Per i collettivi di tipo popolazione o evento con durata, questo errore dipende da errori negli eventi di ingresso. Si possono avere i seguenti casi:

- *errata esclusione* per *mancata registrazione di un ingresso*:
    - l'errata esclusione dal collettivo degli eventi di ingresso di un evento legato a un elemento appartenente al collettivo (per errata esclusione dell'evento di ingresso o per errore nel legame [23], per cui l'ingresso viene registrato ma riferito ad un elemento sbagliato) provoca
    - un'errata esclusione di un enunciato con periodo aperto di appartenenza dell'elemento al collettivo di tipo popolazione o evento con durata, cioè un'errata esclusione dell'elemento
- *errata esclusione temporanea* per *registrazione di un ingresso con riferimento temporale posticipato*, dovuta allo stesso errore nel riferimento temporale dell'evento di ingresso.

E' evidente che l'errata inclusione di un elemento dovuta a un errore nel legame con l'evento di ingresso provoca la contemporanea errata esclusione dell'elemento al quale l'ingresso era in realtà legato.

Il record registrato a seguito di un'operazione nella classe *REG* contiene enunciati di possesso di caratteristiche e relazioni.

Evidentemente un'errata esclusione di un elemento e quindi la mancata registrazione del relativo record comporta anche la mancata registrazione degli enunciati di possesso di caratteristiche e relazioni relativi all'elemento: è la situazione tradizionalmente definita ***mancata risposta totale***.

Per i record registrati, *sia che l'enunciato di appartenenza al collettivo, e quindi il record, sia vero, sia che l'enunciato di appartenenza al collettivo, e quindi il record, sia falso*, può accadere che:

- *siano falsi uno o più dei singoli enunciati di possesso di caratteristiche e relazioni che compongono il record*, in quanto legano l'elemento a una modalità o valore di una caratteristica, oppure a un elemento di un collettivo, che non sono quelli reali: si ha in questo caso contemporaneamente **l'ERRATA ESCLUSIONE dell'enunciato vero di possesso della caratteristica o relazione e l'ERRATA INCLUSIONE di un enunciato falso [24]**. Precisamente:

[23] o anche, più improbabile, mancata risposta relativa all'enunciato di possesso della relazione

[24] questo perché assumiamo per semplicità che sia sempre presente nel record un enunciato per ciascuna caratteristica o relazione, ciò significa che per le eventuali caratteristiche o relazioni opzionali, cioè per le quali la modalità o l'elemento legato può non essere osservato, si utilizza una modalità o un legame fittizio.

- Un enunciato con periodo aperto di possesso, da parte dell'elemento di un collettivo di tipo popolazione o evento con durata, della modalità o valore c_i per una caratteristica *B*, in sostituzione di una modalità o valore precedente, è falso quando nella realtà la modalità o valore posseduti dall'elemento sono diversi da c_i
- Un enunciato con periodo aperto di possesso, da parte dell'elemento di un collettivo di tipo popolazione o evento con durata, di un legame con l'elemento u_j per una relazione C, in sostituzione di un legame precedente con un'altra unità, è falso quando nella realtà il legame posseduto dall'elemento è con un elemento diverso da u_j.

Inoltre per i record registrati, *sia che l'enunciato di appartenenza al collettivo, e quindi il record, sia vero, sia che l'enunciato di appartenenza al collettivo, e quindi il record, sia falso*, può accadere che:

- *non siano accettati uno o più singoli enunciati veri di possesso di caratteristiche e relazioni relativi all'elemento*: questi sono i tradizionali errori di ***mancata risposta parziale***, che consistono in un'**ERRATA ESCLUSIONE degli enunciati veri di possesso delle caratteristiche e relazioni**.

Gli errori possibili in occasione di operazioni nella classe *ELIM*

Per le operazioni nella classe ELIM relative ai collettivi di tipo popolazione o evento con durata, un primo tipo di errore consiste nell'accettazione di un falso enunciato di chiusura temporale dell'appartenenza di un elemento a un collettivo, che si traduce nell'errata chiusura temporale di un record.

Si ha cioè **ERRATA INCLUSIONE di un enunciato falso di chiusura** e quindi **ERRATA ESCLUSIONE di un elemento da un collettivo, anche TEMPORANEA,** per la chiusura del relativo record

Questo errore dipende da errori negli eventi di uscita. Si possono avere i seguenti casi:

- *errata esclusione* per *registrazione di una chiusura non effettiva*:
    - l'errata inclusione nel collettivo degli eventi di uscita di un evento legato all'elemento (per errata inclusione dell'evento di uscita e/o per errore nel legame, per cui l'uscita viene riferita ad un elemento sbagliato) provoca
    - un'errata inclusione di un enunciato con periodo chiuso di appartenenza dell'elemento al collettivo e quindi
    - l'errata esclusione dell'elemento dal collettivo di tipo popolazione o evento con durata**,** per la chiusura temporale del relativo record
- *errata esclusione temporanea* per registrazione di un'uscita temporaneamente non esistente, cioè *registrazione di un'uscita con riferimento temporale anticipato*, quando a un'errata esclusione segue successivamente un'uscita effettiva (improbabile invece un riferimento temporale anticipato nell'evento di uscita).

Per le operazioni nella classe ELIM relative ai collettivi di tipo popolazione o evento con durata, l'altro errore consiste nella mancata accettazione di un vero enunciato di chiusura dell'appartenenza di un elemento a un collettivo e si traduce nell'errata mancata chiusura di un record.

Si ha cioè **ERRATA ESCLUSIONE di un enunciato vero di chiusura** e quindi **ERRATA INCLUSIONE di un elemento in un collettivo, anche TEMPORANEA,** per la mancata chiusura del relativo record

Questo errore dipende da errori negli eventi di uscita. Si possono avere i seguenti casi:

- *errata inclusione* per *mancata registrazione di una chiusura*:
  - l'errata esclusione dal collettivo degli eventi di uscita di un evento legato all'elemento (per errata esclusione dell'evento di uscita e/o per errore nel legame, per cui l'uscita viene riferita ad un elemento sbagliato [25]) provoca
  - un'errata esclusione di un enunciato con periodo chiuso di appartenenza dell'elemento al collettivo e quindi
  - l'errata inclusione dell'elemento nel collettivo di tipo popolazione o evento con durata**,** per la mancata chiusura temporale del relativo record
- *errata inclusione temporanea* per *registrazione di un'uscita con riferimento temporale posticipato*, dovuta allo stesso errore nel riferimento temporale dell'evento di uscita.

Anche in questo caso è evidente che l'errata esclusione di un elemento dovuta ad un errore nel legame con l'evento di uscita provoca la contemporanea errata inclusione dell'elemento al quale l'uscita era in realtà legata.

Gli errori che possono praticamente verificarsi in occasione delle operazioni nella classe ELIM dipendono comunque anche da un'ulteriore circostanza.

Per le operazioni nella classe *REG* il record relativo a un elemento del collettivo di tipo popolazione o evento con durata, vero o falso, viene creato contestualmente e a seguito dell'evento di ingresso[26]. Si può assumere quindi che il codice di raccordo dell'evento di ingresso [27], che individua l'elemento legato (giusto o sbagliato), e il codice identificativo di tale elemento nel collettivo di tipo popolazione o evento con durata coincidano per costruzione.

Invece le operazioni nella classe *ELIM* richiedono come pre-requisito che l'evento di uscita sia per prima cosa attribuibile a un elemento del collettivo già presente in archivio con il proprio identificativo.

Viene quindi per prima cosa effettuata un'***operazione di matching*** tra il record relativo all'evento di uscita e il record con periodo aperto relativo all'elemento del collettivo al quale l'evento di uscita è legato, che dev'essere già presente in archivio: ciò significa che si cerca in archivio un record con periodo aperto nel quale il codice identificativo u_i coincida con il codice di raccordo u_j presente nel record relativo all'evento di uscita [28] che individua l'elemento legato.

Il matching è impossibile in caso di precedente errata esclusione dell'elemento del collettivo al quale l'evento di uscita è legato [29], ma anche errori nel codice di raccordo e/o nel codice identificativo dell'elemento da raccordare determinano l'esito del matching, che può essere negativo o positivo e condiziona i successivi effetti, in termini di corretta o errata inclusione, oppure

[25] o anche, più improbabile, mancata risposta relativa all'enunciato di possesso della relazione

[26] ciò vale per i collettivi principali dell'archivio, la situazione è diversa per i sottoinsiemi

[27] si tratta precisamente del codice di raccordo u_j presente nell'apposito enunciato di possesso della relazione contenuto nel record relativo all'evento di ingresso, che lega tale enunciato all'elemento del collettivo cui è riferito

[28] si tratta precisamente del codice di raccordo u_j presente nell'apposito enunciato di possesso della relazione contenuto nel record relativo all'evento di uscita, che lega tale enunciato all'elemento del collettivo cui è riferito

[29] il matching invece è possibile nel caso limite di un evento di chiusura erroneamente incluso legato ad elemento del collettivo che era stato erroneamente incluso

di corretta o errata esclusione [30]. All'analisi di questi effetti è dedicato un apposito capitolo nella PARTE QUINTA.

Gli errori possibili in occasione di operazioni nella classe *MOD*

Per le operazioni nella classe MOD relative ai collettivi di tipo popolazione o evento con durata, un primo tipo di errore consiste nell'errata accettazione di un nuovo enunciato riferito a un elemento e relativo a una caratteristica o relazione, che si traduce nella creazione di un nuovo record con periodo aperto relativo all'elemento che è falso.

Si ha in questo caso contemporaneamente **l'ERRATA ESCLUSIONE dell'enunciato vero di possesso della caratteristica o relazione e l'ERRATA INCLUSIONE di un enunciato falso** [31], anche **TEMPORANEI**, con conseguente **creazione di un nuovo record con periodo aperto relativo all'elemento che è falso** perché contiene un enunciato falso.

Si possono avere i seguenti casi:

- *errata creazione di un nuovo record con periodo aperto falso* per *errata accettazione di un nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione*:
    - l'errata accettazione di un nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione (dove l'accettazione può essere errata perché l'enunciato è falso e/o perché è riferito ad un elemento sbagliato) provoca
    - l'errata chiusura temporale del record esistente relativo all'elemento del collettivo di tipo popolazione o evento con durata e la creazione di un nuovo record con periodo aperto relativo all'elemento che è falso perché contiene un enunciato falso.
- *errata creazione temporanea di un nuovo record aperto falso* per *accettazione di un nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione con riferimento temporale anticipato*, probabilmente raro ma materialmente possibile, se a un'errata modifica segue successivamente un'identica modifica vera (improbabile invece un riferimento temporale anticipato nell'evento di modifica)

ricordando che:

- un enunciato aperto di possesso, da parte dell'elemento di un collettivo di tipo popolazione o evento con durata, della modalità o valore c_i per una caratteristica *B*, in sostituzione di una modalità o valore precedente, è falso quando nella realtà la modalità o valore posseduti dall'elemento sono diversi da c_i
- un enunciato aperto di possesso, da parte dell'elemento di un collettivo di tipo popolazione o evento con durata, di un legame con l'elemento u_j per una relazione C, in sostituzione di un legame precedente con un'altra unità, è falso quando nella realtà il legame posseduto dall'elemento è con un elemento diverso da u_j.

[30] gli effetti degli errori nel codice di raccordo sul matching e sulla correttezza delle inclusione ed esclusioni si combinano in maniera complessa: si può avere un esito positivo del matching che provoca poi errori di inclusione o esclusione, o anche un esito negativo del matching che impedisce errori di inclusione o esclusione.

[31] questo perché assumiamo per semplicità che sia sempre presente nel record un enunciato per ciascuna caratteristica o relazione, ciò significa che per le eventuali caratteristiche o relazioni opzionali, cioè per le quali la modalità o l'elemento legato può non essere osservato, si utilizza una modalità o un legame fittizio.

Un secondo tipo di errore *consiste nella mancata accettazione di un nuovo enunciato vero riferito a un elemento e relativo a una caratteristica o relazione, che si traduce nella mancata creazione di un nuovo record con periodo aperto vero relativo all'elemento.*

Si ha in questo caso un'**ERRATA ESCLUSIONE degli enunciati veri di possesso delle caratteristiche e relazioni, anche TEMPORANEA**

Si possono avere i seguenti casi:

- *mancata creazione di un nuovo record con periodo aperto vero* per *mancata accettazione di un nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione*:
    - la mancata accettazione di un nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione (dove l'accettazione può mancare del tutto, oppure mancare perché l'enunciato è riferito ad altro elemento) provoca
    - la mancata chiusura temporale del record esistente relativo all'elemento del collettivo di tipo popolazione o evento con durata, che è divenuto falso, e la mancata creazione di un nuovo record con periodo aperto vero relativo all'elemento
- *mancata creazione temporanea di un nuovo record con periodo aperto vero* per *accettazione di un nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione con riferimento temporale posticipato*

E' evidente che l'errata accettazione di un nuovo enunciato relativo a una caratteristica o relazione per un errore nel riferimento provoca la contemporanea mancata accettazione dell'enunciato per l'elemento al quale l'enunciato era in realtà riferito.

Come per le operazioni nella classe *ELIM*, anche gli errori che possono praticamente verificarsi in occasione delle operazioni nella classe *MOD* dipendono comunque dall'esito delle operazioni di matching.

Le operazioni nella classe *MOD* richiedono infatti come pre-requisito che il nuovo enunciato riferito all'elemento e relativo a una caratteristica o relazione sia per prima cosa attribuibile ad un elemento del collettivo già presente in archivio con il proprio identificativo.

Viene quindi per prima cosa effettuata un'***operazione di matching*** tra il nuovo enunciato riferito all'elemento u_i e relativo al possesso di una caratteristica o relazione e il record con periodo aperto relativo all'elemento del collettivo al quale il nuovo enunciato è riferito, che dev'essere già presente in archivio: ciò significa che si cerca in archivio un record con periodo aperto nel quale il codice identificativo u_i coincida con il codice identificativo u_i presente nel nuovo enunciato.

Il matching è impossibile in caso di precedente errata esclusione dell'elemento del collettivo u_i al quale l'enunciato di modifica è riferito [32], ma anche errori in uno o entrambi i codici identificativi coinvolti determinano l'esito del matching, che può essere negativo o positivo, e poi i successivi effetti, in termini di corretta o errata creazione di un nuovo record.

Oltre a ciò, quando il *nuovo enunciato riferito all'elemento è relativo a una relazione* viene effettuata un'***ulteriore operazione di matching*** tra il nuovo enunciato riferito all'elemento u_i e il record con periodo aperto relativo all'elemento raccordato, che dev'essere già presente in archivio: ciò significa che si cerca in archivio un record con periodo aperto nel quale il codice identificativo u_i coincida con il codice di raccordo $u_{j.}$ presente nel nuovo enunciato, che individua l'elemento legato.

[32] il matching invece è possibile per elementi dei collettivi che sono stati erroneamente inclusi

Il matching è impossibile in caso di precedente errata esclusione dell'elemento raccordato dal proprio collettivo[33], ma anche errori nel codice di raccordo e/o nel codice identificativo dell'elemento da raccordare determinano l'esito del matching, che può essere negativo o positivo, e condiziona i successivi effetti in termini di corretta o errata creazione di un nuovo record.

All'analisi di questi effetti è dedicato un apposito capitolo nella PARTE QUINTA.

Considerazioni su ERRORI DI COPERTURA, MANCATE RISPOSTE ed ERRORI DI ACCURATEZZA e sui tipi di errori più frequenti in pratica

In questo paragrafo discutiamo brevemente le tradizionali definizioni di errori di copertura, mancate risposte ed errori di accuratezza alla luce dell'analisi fin qui condotta.

Come nella tradizionale definizione di copertura, le errate inclusioni ed esclusioni di elementi da un determinato collettivo determinano la copertura del collettivo da parte dell'archivio. La prevalenza di errori dell'uno o dell'altro tipo, quando è determinabile, si tradurrà in sovracopertura o sottocopertura.

In termini logico-insiemistici, la copertura di un collettivo è corretta quando l'archivio assicura in ogni momento la corretta registrazione dell'estensione del collettivo, cioè la registrazione di tutti e soli gli elementi effettivamente appartenenti al collettivo in quel momento (si veda anche l'Appendice dedicata alla documentazione dell'ontologia degli archivi amministrativi).

La stessa nozione di corretta registrazione dell'estensione si applica alle caratteristiche e alle relazioni osservate dall'archivio: in ogni momento l'archivio deve assicurare la registrazione di tutti e soli gli enunciati effettivamente appartenenti in quel momento all'estensione di una caratteristica o di una relazione, vista come un insieme di legami tra elementi.

Si potrebbe quindi parlare di una corretta copertura dell'estensione delle caratteristiche e delle relazioni osservate da parte dell'archivio.

Nel nostro approccio in effetti il concetto di copertura è quello che si riferisce nel modo più generale agli effetti degli errori sull'osservazione della realtà condotta dalla specifica fonte.

Tuttavia la precedente analisi ha evidenziato che gli errori di inclusione ed esclusione relativi al possesso di caratteristiche e relazioni hanno un loro proprio meccanismo di azione connesso alla corrispondenza con la realtà delle singole modalità, dei singoli valori, dei singoli elementi in relazione che vengono legati agli elementi dei collettivi sia all'ingresso dell'elemento nel collettivo, sia in occasione delle successive operazioni di aggiornamento.

Si può perciò mantenere la distinzione tradizionale tra *ERRORI DI COPERTURA, che riguardano la corretta registrazione dell'appartenenza di specifici elementi ad un collettivo*, ed *ERRORI DI ACCURATEZZA, che sono relativi alle caratteristiche e relazioni possedute dagli elementi di un collettivo.*

L'approccio adottato in BLUE-ETS di attribuire simmetricamente ai collettivi e alle variabili tanto errori di copertura quanto errori di accuratezza può trovare una spiegazione concettuale proprio nel fatto che tutti gli errori sono di copertura in senso lato in quanto riguardano sempre estensioni di

[33] Il matching invece è possibile nel caso di un nuovo enunciato riferito all'elemento u_i e a un elemento erroneamente raccordato u_j per il quale esiste in archivio un record erroneamente incluso

insiemi o di relazioni tra insiemi, e quindi tutti gli errori potrebbero essere interpretati come errori di copertura dando un senso largo a questo termine, e d'altra parte l'errore di identificazione degli elementi dei collettivi potrebbe essere interpretato in qualche modo come un errore di accuratezza.

Tuttavia, avendo chiarito quali possono essere le cause concettuali di questa esigenza, sembra preferibile mantenere la terminologia tradizionale.

Rispetto all'approccio tradizionale, l'approccio adottato nel presente Framework permette di meglio descrivere la dinamica degli aggiornamenti e gli effetti degli errori che possono occorrere in occasione di aggiornamenti tanto sulla copertura dei collettivi quanto sull'accuratezza di registrazione delle modalità o valori, per le caratteristiche, e degli elementi legati, per le relazioni.

La situazione più simile a quella delle indagini è non a caso quella che si determina in occasione delle operazioni nella classe *REG*, nelle quali sono contestualmente registrati l'appartenenza di un elemento a un collettivo e tutte le modalità e valori assunti per le caratteristiche nonché i legami per le relazioni, proprio come avviene in un'indagine. *A questa situazione infatti sono appropriati concetti come mancate risposte totali e parziali*.

Tuttavia *nella realtà delle altre fonti, e in particolare degli archivi amministrativi, hanno altrettanta importanza gli errori che possono generarsi in occasione di aggiornamenti*.

Inoltre assumono un'importanza particolare le relazioni che legano tra loro i collettivi, in particolare quelle che legano i collettivi di eventi di ingresso e di uscita ai collettivi di tipo popolazione o evento con durata, e di conseguenza assumono importanza tutte le possibilità di errore sulle relazioni e sui codici di raccordo nonché gli effetti del matching, aspetti generalmente molto poco presi in considerazione nello studio della qualità delle indagini.

Nella tabellina di seguito è presentata una sintesi degli errori possibili.

Infine, sul piano pratico, è importante osservare che gli errori descritti in precedenza non hanno tutti la stessa frequenza e importanza.

L'errata inclusione di elementi in un collettivo di tipo popolazione o evento con durata dovuta alla registrazione di elementi che non vi appartengono presuppone una qualche forma di volontarietà o un particolare malfunzionamento dell'archivio, perciò è lecito presumere che non sia frequente, mentre *è certamente molto presente l'errata inclusione in un collettivo di tipo popolazione o evento con durata dovuta all'errata esclusione di eventi di uscita*, circostanza molto influenzata da errori sulle relazioni e da errori di matching.

Viceversa, è relativamente *più frequente l'errata esclusione di elementi da un collettivo di tipo popolazione o evento con durata al momento del loro ingresso, che dà luogo a mancata risposta totale*, rispetto a un'errata esclusione dovuta a errata inclusione di un evento di chiusura, che anch'essa sembra presupporre una qualche forma di volontarietà o un particolare malfunzionamento dell'archivio.

Molto frequenti sono invece gli *errori sul possesso di caratteristiche e relazioni che si verificano in occasione di aggiornamenti, soprattutto i mancati aggiornamenti*, errori che sono anch'essi influenzati da errori sul riferimento e da errori di matching.

Si tratta anche di *errori meno facilmente diagnosticabili che nel caso delle indagini, perché il mancato aggiornamento non si traduce in mancata risposta parziale ma nel permanere di una precedente modalità, di un precedente valore, di un precedente legame che non sono più veri*.

TIPO OGGETTO	TIPO ENUNCIATO	**ERRORI DI ERRONEA INCLUSIONE**	**ERRORI DI ERRONEA ESCLUSIONE**	**ERRONEA INCLUSIONE ed ERRONEA ESCLUSIONE**
Collettivo di tipo popolazione	*Appartenenza al collettivo* **A (u_i, t_i – t),** esempi: Studente (id studente, t_i – t), Lavoratore (id lavoratore, t_i – t), Datore di lavoro (id datore, t_i – t), Degente (id degente, t_i – t)	**I_A** Erronea inclusione di elementi non appartenenti al collettivo **Se prevalente: SOVRACOPERTURA del collettivo** Si registra in archivio un record per un elemento che al momento t_i non esiste OPPURE si registra in archivio un record per un elemento che al momento t_i esiste ma non appartiene in realtà al collettivo	**E_A** Erronea esclusione di elementi appartenenti al collettivo **Se prevalente: SOTTOCOPERTURA del collettivo** NON si registra in archivio un record per un elemento che esiste e che a partire dal momento t_i appartiene al collettivo	
Collettivo di tipo evento	*Appartenenza al collettivo* **Istantaneo A (u_i, t_i)**, esempi: Avvio rapporto di lavoro (id avvio rapporto di lavoro, t_i), Ricovero ospedaliero(id ricovero ospedaliero, t_i); Dimissione degente (id Dimissione degente, t_i), Dimissione lavoratore (id dimissione lavoratore, $\mathbf{t_j}$), Licenziamento (id licenziamento, $\mathbf{t_j}$), **Con durata A (u_i, t_i – t),** esempi: Rapporto di lavoro (id rapporto di lavoro, t_i – t), Degenza (id degenza, t_i – t)	**I_A** Erronea inclusione di elementi non appartenenti al collettivo **Se prevalente: SOVRACOPERTURA del collettivo** Si registra in archivio un record per un evento che nel momento t_i non si è davvero verificato	**E_A** Erronea esclusione di elementi appartenenti al collettivo **Se prevalente: SOTTOCOPERTURA del collettivo** NON si registra in archivio un record per un evento che si è verificato al momento t_i	

TIPO OGGETTO	TIPO ENUNCIATO	**ERRORI DI ERRONEA INCLUSIONE**	**ERRORI DI ERRONEA ESCLUSIONE**	**ERRONEA INCLUSIONE ed ERRONEA ESCLUSIONE**
Caratteristiche degli elementi di un collettivo di tipo popolazione o evento	*Possesso della caratteristica* **B (u_i, c_i, t_i)** (per caratteristiche degli eventi istantanei) **B (u_i, c_i, t_i – t)** (per caratteristiche delle popolazioni o eventi con durata) Sesso(id studente, femmina, t_i – t), Residenza (id degente, Roma, t_i – t), Valore della produzione(id impresa, 100000 euro, t_i – t), Motivo ricovero(id ricovero, incidente, t_i), Spesa per degenza (id degenza, 550 euro,t_i – t)		**E_B** Erronea esclusione di un enunciato relativo al possesso di una caratteristica per un elemento che appartiene al collettivo **MANCATA RISPOSTA su caratteristiche** NON si registra in archivio il possesso di una di una modalità o valore per la caratteristica	**M_B** Errata inclusione di un enunciato relativo al possesso di una caratteristica per un elemento che appartiene al collettivo = viene registrato un enunciato B (u_i, c_j, t_i) o B (u_i, c_j, t_i – t) non vero INOLTRE errata esclusione di un enunciato relativo al possesso di una caratteristica per l'elemento = non viene registrato un enunciato B (u_i, c_i, t_i) o B (u_i, c_i, t_i – t) vero **ERRORE DI ACCURATEZZA su caratteristiche** Si registra in archivio il possesso di una modalità o valore per la caratteristica, ma in realtà l'elemento al momento t_i possiede per tale caratteristica una modalità o valore diverso, in questo caso è stato accettato un enunciato falso al posto di uno vero, esempio: viene registrata per una persona una residenza diversa da quella reale, si ha quindi contemporaneamente un errore di errata inclusione ed errata esclusione

Partecipazione in una relazione per un elemento del collettivo dominio, che viene legato a un elemento del collettivo codominio	*Possesso della relazione* **C (u_i, e_j, t_i)** (per relazioni il cui dominio è un collettivo di tipo evento istantaneo) **C (u_i, e_j, t_i – t)** (per relazioni il cui dominio è un collettivo di tipo popolazione o evento con durata Unità locale appartiene Impresa(id unità locale, id impresa, t_i – t), Rapporto di lavoro riguarda Lavoratore (id lavoratore, id rapporto di lavoro, t_i – t), Rapporto di lavoro riguarda Datore di lavoro (id datore di lavoro, id rapporto di lavoro, t_i – t), Avvio rapporto di lavoro inizia Rapporto di lavoro (id avvio rapporto di lavoro, id rapporto di lavoro, t_i)		**E_C** Erronea esclusione di un enunciato relativo alla partecipazione in una relazione per un elemento che appartiene al collettivo dominio **MANCATA RISPOSTA su relazioni** NON si registra in archivio la partecipazione in una relazione che lega un elemento esistente e appartenente al collettivo dominio ad un elemento del codominio t_i	**M_C** Errata inclusione di un enunciato relativo alla partecipazione in una relazione per un elemento che appartiene al collettivo dominio = viene registrato un enunciato C (u_i, u_j, t_i) o C (u_i, u_j, t_i – t) non vero INOLTRE errata esclusione di un enunciato relativo alla partecipazione in una relazione per l'elemento = non viene registrato un enunciato C (u_i, u_j, t_i) o C (u_i, u_j, t_i – t) vero **ERRORE DI ACCURATEZZA su relazioni** Si registra in archivio la partecipazione in una relazione che lega un elemento esistente e appartenente al collettivo dominio ad un elemento del codominio, ma in realtà al momento t_i l'elemento è legato per tale relazione a un elemento del codominio diverso, in questo caso è stato accettato un enunciato falso al posto di uno vero, esempio: viene registrato per un rapporto di lavoro un datore di lavoro diverso da quello reale, si ha quindi contemporaneamente un errore di errata inclusione ed errata esclusione

Errori di duplicazione

Gli errori di duplicazione possono essere visti come un caso particolare di errata inclusione. E' comunque utile descriverli separatamente per poterne poi puntualizzare meglio le proprietà e i possibili metodi di diagnosi, anche perché sono oggetto di una vasta letteratura.

Consideriamo di seguito la duplicazione di enunciati veri di appartenenza a collettivi e di enunciati veri di possesso di caratteristiche o relazioni da parte di elementi effettivamente appartenenti al collettivo. Ovviamente comunque si può avere duplicazione anche di enunciati falsi, erroneamente inclusi.

Duplicazione di enunciati veri di appartenenza a collettivi

D_A Erronea inclusione dovuta a duplicazione: si decide di creare in archivio un enunciato di appartenenza per un elemento che appartiene effettivamente al collettivo e per il quale esiste già in archivio un enunciato di appartenenza aperto, e quindi un record corretto aperto.

Ci sono tre possibilità:

- nell'enunciato il codice identificativo è corretto e quindi identico a quello dell'enunciato già accettato e inoltre tutti gli enunciati di possesso di relazioni e di caratteristiche contestualmente creati sono identici, vale a dire viene registrato un record identico a uno già esistente: in questo caso l'errore è evidente e individuabile con un apposito controllo preliminare;

- nell'enunciato il codice identificativo è corretto e quindi identico a quello dell'enunciato già accettato ma non tutti gli enunciati di possesso di relazioni e di caratteristiche contestualmente creati sono identici, in questo caso l'errore non è immediatamente evidente perché può essere confuso con l'errore semantico sull'identificativo consistente nella condivisione dello stesso identificativo tra elementi diversi; per discriminare i due casi è necessario un controllo specifico sull'identità degli elementi, utilizzando caratteristiche con una buona capacità di identificazione, ad esempio Nome, Cognome, Indirizzo, Luogo di nascita, o comunque combinazioni di caratteristiche;

- nell'enunciato il codice identificativo è diverso rispetto a quello dell'enunciato già accettato, questo errore è necessariamente provocato da un errore semantico sull'identificativo consistente nell'esistenza di un doppio codice identificativo per uno stesso elemento, anche questo caso può essere diagnosticato con un controllo specifico sull'identità degli elementi, utilizzando caratteristiche con una buona capacità di identificazione, ad esempio Nome, Cognome, Indirizzo, Luogo di nascita, o comunque combinazioni di caratteristiche.

Duplicazione di enunciati veri di possesso di caratteristiche o relazioni da parte di elementi effettivamente appartenenti al collettivo

D_B Erronea inclusione dovuta a duplicazione: si decide di creare in archivio un enunciato di possesso di una caratteristica corretto ma per il quale esiste già in archivio un enunciato aperto di possesso di una caratteristica relativo allo stesso elemento e alla stessa modalità o valore.

D_C Erronea inclusione dovuta a duplicazione: si decide di creare in archivio un enunciato di partecipazione in una relazione corretto ma per il quale esiste già in archivio un enunciato aperto di partecipazione in una relazione relativo allo stesso elemento del collettivo dominio e allo stesso elemento legato del collettivo codominio.

Questi errori si possono verificare all'ingresso dell'elemento in archivio, e quindi nel contesto delle operazioni di tipo *REG*, oppure successivamente, in occasione di operazioni di aggiornamento delle caratteristica o della relazione, e quindi nel contesto di operazioni di tipo *MOD*.

Nel primo caso questi errori sulle caratteristiche e relazioni si verificano in pratica solo in concomitanza di errori di tipo D_A, a causa del fatto che gli enunciati di possesso di caratteristiche o relazioni sono creati contestualmente a quelli di appartenenza e devono essere unici: in pratica sono duplicabili interi record, che possono risultare identici o parzialmente diversi per il codice identificativo o per altre caratteristiche, ma non possono presentare duplicazioni di campi, in base ai vincoli di unicità che sono generalmente automaticamente soddisfatti o comunque direttamente controllabili.

Nel contesto di operazioni di tipo *MOD* invece si può avere una doppia inclusione di un enunciato di possesso di una caratteristica o relazione. In questo caso come si è detto il nuovo enunciato di possesso di una caratteristica o relazione deve trovare il riferimento ad un elemento che risulti già appartenente al collettivo, per il quale cioè esiste già un record aperto in archivio.

Ipotizzando che tale record esista e in esso l'elemento sia correttamente identificato, consideriamo separatamente i nuovi enunciati di possesso di caratteristiche e i nuovi enunciati di possesso di relazioni.

Per i nuovi enunciati di possesso di caratteristiche si avranno in questo caso due possibilità:

- nell'enunciato il codice identificativo dell'elemento del collettivo al quale è riferita la nuova modalità o valore è corretto e quindi identico a quello del record presente in archivio: in questo caso la duplicazione è evidente e individuabile con un apposito controllo preliminare

- nell'enunciato il codice identificativo dell'elemento del collettivo al quale è riferita la nuova modalità o valore è diverso rispetto a quello del record presente in archivio, questo errore è necessariamente provocato da un errore semantico sull'identificativo consistente nell'esistenza di un doppio codice identificativo per uno stesso elemento: in questo caso nell'ipotesi di correttezza del record già presente in archivio si hanno due possibilità:

    - l'identificativo scorretto non coincide con un codice di altro elemento, in questo caso il nuovo enunciato duplicato non trova il riferimento ed è facilmente individuato come errato
    - l'identificativo scorretto coincide con un codice di altro elemento, in questo caso il nuovo enunciato duplicato viene scorrettamente riferito ad un altro elemento

Per i nuovi enunciati di possesso di relazioni si avranno in questo caso due possibilità:

- nell'enunciato sia il codice identificativo dell'elemento del dominio sia il codice di raccordo con l'elemento del codominio sono corretti e quindi identici a quello del record presente in archivio: in questo caso la duplicazione è evidente e individuabile con un apposito controllo preliminare

- nell'enunciato o il codice identificativo dell'elemento del dominio o il codice di raccordo con l'elemento del codominio è diverso rispetto a quello del record già presente in archivio, questo errore è necessariamente provocato da un errore semantico sull'identificativo consistente nell'esistenza di un doppio codice identificativo per uno stesso elemento, che riguarderà rispettivamente il dominio o il codominio o entrambi: in questo caso nell'ipotesi di correttezza dei record già presenti in archivio

- nessuno dei due codici identificativi eventualmente scorretti coincide con un codice di altro elemento, in questo caso il nuovo enunciato duplicato non trova il riferimento ed è facilmente individuato come errato
- uno o entrambi i codici identificativi coinvolti coincidono anche con un codice di altro elemento, in questo caso il nuovo enunciato duplicato viene scorrettamente riferito ad un altro elemento del dominio e/o del codominio.

Se invece nel record presente in archivio l'elemento non è correttamente identificato, sono possibili diverse situazioni, analogamente se anche il record relativo all'elemento è duplicato.

Infine anche i record falsi e gli enunciati di possesso di caratteristiche e relazioni falsi possono essere duplicati.

Queste situazioni sono analizzate nella PARTE QUARTA caso per caso, occorre in generale tenere presente che si può comunque avere in tutti questi casi il matching tra il nuovo enunciato di possesso di una caratteristica o relazione e un record presente in archivio, per cui la presenza di errori può rimanere non evidenziata.

LE POSSIBILITA' DI DIAGNOSI PER I DIVERSI TIPI DI ERRORE

Le tipologie di controlli applicabili

Occorre ricordare che, come anticipato nel primo paragrafo e come viene spiegato in seguito, in base agli strumenti di cui disponiamo per individuare gli errori, i diversi tipi di errore non sono sempre discriminabili.

Si possono anzitutto enucleare i controlli strutturali e, all'interno di questi, i controlli iniziali di errori evidenti, che comprendono anche i controlli sintattici sugli identificativi.

Controlli strutturali

Controlli iniziali di errori evidenti:

1. controlli di errori sintattici sugli identificativi degli elementi dei collettivi
2. controlli di errori sintattici sugli identificativi delle modalità o valori delle caratteristiche (fuori dominio, fuori range)
3. ricerca di interi record duplicati
4. in occasione di operazioni di tipo *REG*, ricerca di record con valori plurimi per caratteristiche o relazioni (se questo errore è tecnicamente possibile)

Altri controlli strutturali

5. ricerca di duplicazioni di codici identificativi
6. ricerca di elementi che sono identici anche se non si presentano come record duplicati perché differiscono in alcune caratteristiche o relazioni
7. mancato matching in occasione di operazioni di tipo *ELIM*: il record relativo all'evento di uscita non trova il raccordo con un record aperto relativo all'elemento del collettivo al quale l'evento di uscita è legato
8. mancato matching in occasione di operazioni di tipo *MOD*: il nuovo enunciato di possesso di una caratteristica o relazione non trova il raccordo con un record aperto relativo all'elemento del collettivo al quale è riferito
9. mancato matching in occasione di operazioni di tipo *MOD*: il nuovo enunciato di possesso di una relazione non trova il raccordo con un record aperto relativo all'elemento raccordato

I controlli 1-4 individuano direttamente record con errori dovuti ad una causa specifica, che possono essere eliminati o sottoposti a correzione prima di cercare eventuali altri errori.

I controlli 5 -8 individuano record con errori che, anche quando sono immediatamente evidenti come nel caso della duplicazione di codici, possono essere dovuti a cause diverse e devono essere correttamente interpretati prima di intraprendere azioni di correzione.

Di seguito si dettagliano questi tipi di controlli.

Controlli iniziali di errori evidenti

Alcuni controlli preliminari permettono di diagnosticare immediatamente un record con uno specifico tipo di errore, che può essere eliminato o corretto prima di diagnosticarne altri, precisamente:

1. Controlli per *errori sintattici sugli identificativi degli elementi dei collettivi*: per gli identificativi che hanno una struttura, si controlla la correttezza della loro composizione (che tutti gli elementi della struttura siano presenti ed eventualmente coerenti con le caratteristiche e relazioni possedute)

2. Controlli per *errori sintattici sugli identificativi delle modalità o valori delle caratteristiche*: si controlla la loro appartenenza all'insieme delle modalità previste per la classificazione utilizzata o al range di valori predefiniti
3. Presenza di *record interamente duplicati*: immediatamente rilevabile, indica duplicazione di un enunciato di appartenenza e si può correggere immediatamente eliminando il record duplicato
4. In occasione di operazioni di tipo *REG*, nel caso in cui ciò sia tecnicamente possibile, presenza di valori plurimi per caratteristiche o relazioni: immediatamente rilevabile, indica duplicazione dei relativi enunciati e permette di diagnosticare direttamente un record che contiene questo tipo di errori, che può essere eliminato o sottoposto a correzione.

Altri metodi di controllo strutturali sono sempre preliminari ma non permettono l'identificazione puntuale di un singolo errore e indicano piuttosto la presenza possibile di errori di tipo diverso, eventualmente concomitanti.

5. Presenza di *codici identificativi duplicati, cioè identificativo identico in due record diversi*: immediatamente rilevabile, può indicare duplicazione di un enunciato di appartenenza oppure dipendere dalla condivisione dello stesso identificativo tra elementi diversi, la distinzione richiede il controllo dell'identità degli elementi utilizzando, se disponibile, un eventuale insieme di caratteristiche e relazioni con un potenziale di identificazione (esempio Nome, Cognome e Indirizzo, Luogo di nascita) o metodologie più sofisticate
6. *Controllo dell'identità degli elementi* per individuare la duplicazione di un enunciato di appartenenza: *utilizza, se disponibile, un eventuale insieme di caratteristiche e relazioni con un potenziale di identificazione* (esempio Nome, Cognome e Indirizzo Luogo di nascita) o metodologie più sofisticate e consente di *diagnosticare la duplicazione dello stesso elemento, registrato con due identificativi differenti*
7. *Mancato matching* in occasione di operazioni di tipo *ELIM*, cioè presenza di record relativi a eventi di uscita che non trovano il raccordo con un record aperto relativo all'elemento del collettivo al quale l'evento di uscita è legato: è immediatamente rilevabile, ma può dipendere da combinazioni di errori dei diversi tipi sui record che devono essere raccordati
8. *Mancato matching* in occasione di operazioni di tipo *MOD*, cioè presenza di enunciati di possesso di caratteristiche o relazioni che *non trovano il raccordo con un record aperto relativo all'elemento del collettivo al quale sono riferiti*: è immediatamente rilevabile, ma può dipendere da combinazioni di errori dei diversi tipi sull'enunciato e sul record che deve essere raccordato
10. *Mancato matching* in occasione di operazioni di tipo *MOD*, cioè presenza di enunciati di possesso di relazioni con codici di raccordo che *non trovano il raccordo con un record aperto relativo all'elemento raccordato*: è immediatamente rilevabile, ma può dipendere da combinazioni di errori dei diversi tipi sull'enunciato e sul record che deve essere raccordato.

Esistono poi ulteriori possibili metodi di controllo che non sono di immediata applicazione e vanno utilizzati congiuntamente per diagnosticare errori che possono poi essere dovuti a cause diverse:

Altri controlli

9. Raccolta di *informazioni a priori*, ad esempio mediante le istruttorie sugli archivi: questo metodo è abbastanza peculiare degli archivi amministrativi e serve oltre che ad una prima diagnosi degli errori anche a cercare di ricostruire l'origine dell'errore, utilizzato in combinazione con gli altri metodi
10. *Confronto con le informazioni gestite in altri archiv*i mediante operazioni di linkage esatto
11. Uso di *vincoli statici di obbligatorietà e di incompatibilità*, che limitano le combinazioni accettabili di modalità o valori e/o di elementi legati, nonché gli insiemi di modalità o valori

e/o di elementi legati compatibili con l'appartenenza al collettivo, in ogni momento specifico t

12. Per i collettivi di tipo popolazione o evento con durata, uso di *vincoli dinamici di obbligatorietà e di incompatibilità*, che limitano le combinazioni di modalità o valori e/o di elementi legati che possono succedersi nel tempo

Due possibili ulteriori controlli possono essere utilizzati per la *diagnosi degli errori nel riferimento temporale* e per la *diagnosi degli errori di tempestività*, rispettivamente.

Questi errori riguardano tutti gli enunciati e quindi tutti i record presenti in archivio, ma saranno usualmente cercati sui nuovi record riferiti a eventi istantanei e sui nuovi enunciati di possesso di caratteristiche e relazioni riferiti a unità di popolazioni o eventi con durata, poiché come si è visto tutti gli altri enunciati presenti in archivio sono creati o modificati di conseguenza.

13. Per la *diagnosi degli errori nel riferimento temporale*: rilevazione, laddove possibile, della *distanza media tra il momento di riferimento effettivo di un enunciato e il momento di riferimento t_i registrato in archivio*, dove t_i è il momento di occorrenza, per i record riferiti a eventi istantanei, oppure il momento d'inizio del periodo aperto di validità, per i nuovi enunciati di possesso di caratteristiche e relazioni riferiti a unità di popolazioni o a eventi con durata.
14. Per la *diagnosi degli errori di tempestività*: rilevazione della *distanza media tra il momento di riferimento t_i registrato in archivio per un enunciato e il momento dell'immissione in archivio della relativa informazione*, dove t_i è il momento di occorrenza, per i record riferiti a eventi istantanei, oppure il momento d'inizio del periodo aperto di validità, per i nuovi enunciati di possesso di caratteristiche e relazioni riferiti a unità di popolazioni o a eventi con durata.

Nel successivo paragrafo "Tabelle di associazione tra tipologie di controlli e tipologie di errore" si presentano alcuni prospetti che pongono in relazione i diversi tipi di controllo con i diversi tipi di errore.

Come più volte osservato, gli identificativi possono essere affetti da propri errori che influenzano tanto la copertura dell'archivio quanto le possibilità di effettuare correttamente il linkage con altri archivi e, quindi, anche i risultati della corrispondenza con altri archivi, influenzando quindi la possibilità di controllare l'errata inclusione e l'errata esclusione. Per questo motivo *i controlli sulla copertura, l'identificazione e la linkabilità non dovrebbero essere condotti indipendentemente*.

Inoltre, il tipo di identificativo utilizzato in un archivio influenza le possibilità di controllo.

Tutti questi aspetti relativi al ruolo degli identificativi sono discussi nel paragrafo sottostante.

Le tipologie di controlli applicabili e gli identificativi

Riguardo in particolare i controlli sugli identificativi, si ricorda che per quanto detto gli identificativi possono avere due possibili ruoli in un enunciato, che comportano controlli diversi:

- *ruolo di identificazione*: in tutti gli enunciati di qualsiasi tipo serve un identificativo per designare l'elemento del collettivo cui è riferito l'enunciato
- *ruolo di raccordo*: negli enunciati che servono a dichiarare le relazioni tra due collettivi un apposito identificativo serve a designare qual è l'elemento dell'altro collettivo raccordato.

Inoltre, nel loro ruolo di identificazione gli identificativi possono essere utilizzati anche per un ulteriore uso:

- il *linkage esatto (cioè non probabilistico) con gli elementi dello stesso collettivo eventualmente gestiti in altri archivi, che utilizzino lo stesso codice*

Nel seguito si suppone che nell'archivio esista sempre almeno un sistema di identificazione degli elementi per ogni collettivo, una condizione in teoria non sempre verificata ma che comunque dovrebbe essere assicurata in archivi da utilizzare a scopo statistico.

Come specificato nella precedente descrizione dei controlli 5 e 6, l'esistenza ulteriore per un collettivo di un insieme di caratteristiche e relazioni con una certa capacità di identificare gli elementi (ad esempio Nome, Cognome, Indirizzo, Luogo di nascita) può essere sfruttata per il controllo sull'identità degli elementi interno all'archivio e, quindi, per discriminare tra gli errori sugli identificativi e quelli di duplicazione o di errata inclusione.

Può anche essere sfruttata congiuntamente all'identificativo nelle operazioni di linkage esatto con gli altri archivi, in controlli di tipo 10.

In tutti questi casi si possono utilizzare anche tecniche più sofisticate che coinvolgono tutte le caratteristiche e le relazioni per calcolare la probabilità che due elementi, sia nello stesso archivio che in archivi diversi, siano identici. Non illustriamo qui queste tecniche, studiate sia nell'ambito informatico che in quello statistico, nel quale definiscono le metodologie di linkage probabilistico sulle quali esiste una letteratura specifica.

I controlli che coinvolgono gli identificativi si applicano e interpretano poi in modo diverso a seconda del tipo di identificativo.

In generale si può dire che i controlli di tipo 5 e 6 e di tipo 10, in particolare in quest'ultimo caso quando è possibile il controllo della corrispondenza con una lista certificata di identificativi esterna all'archivio (ad esempio un elenco di codici fiscali), consentono di cogliere contestualmente la presenza di un errore sull'identificativo e/o di un errore di duplicazione e/o di una errata inclusione, ma la maggiore o minore possibilità di discriminare tra questi errori dipende da:

- il tipo di identificativo
- la disponibilità di un insieme di caratteristiche e relazioni con una certa capacità di identificare gli elementi (ad esempio Nome, Cognome, Indirizzo, Luogo di nascita).

Distinguiamo i seguenti tipi di identificativi:

1. *Identificativo semplice*, senza una struttura, ad esempio costituito da un numero d'ordine
2. *Identificativo con una struttura, ad esempio costituito da uno o più codici di corrispondenza con altri collettivi e da un numero d'ordine*, ad esempio identificativo del rapporto di lavoro = identificativo del lavoratore + = identificativo del datore di lavoro + numero d'ordine
3. *Identificativo con una struttura che tiene conto di altre caratteristiche e relazioni con capacità di identificare gli elementi* (ad esempio Nome, Cognome, Indirizzo, Luogo di nascita), come il codice fiscale

Per ciò che riguarda infine l'utilizzabilità dell'identificativo per il linkage esatto con gli elementi dello stesso collettivo eventualmente gestiti in altri archivi è importante osservare che:

- l'utilizzabilità dell'identificativo per il linkage esatto con gli elementi dello stesso collettivo eventualmente gestiti in altri archivi, che utilizzino lo stesso codice, è naturalmente in primo

luogo dipendente dalla correttezza dell'identificativo nel suo ruolo di identificazione interna all'archivio: se l'identificativo è scorretto in questo ruolo anche la corrispondenza sarà impossibile o scorretta

- tuttavia si può verificare che un identificativo corretto internamente all'archivio non consenta il linkage con altri archivi o trovi un linkage scorretto, con un elemento che non è lo stesso, semplicemente perché gli identificativi utilizzati nei due archivi sono diversi per alcuni elementi (con la possibilità, più grave, che nei due archivi sia attribuito uno stesso identificativo ad elementi diversi, determinando un linkage scorretto): ciò a stretto rigore non configura un errore interno all'archivio
- tuttavia occorre considerare che, anche se in effetti nel caso di corrispondenze con archivi qualsiasi la precedente possibilità può essere vista solo come una circostanza che inficia la possibilità di controlli, se questo si verifica rispetto ad una lista certificata di identificativi esterna all'archivio (ad esempio un elenco di codici fiscali) si configura una limitazione effettiva all'usabilità statistica dell'archivio per la sua diminuita integrabilità.

Tabelle di associazione tra tipologie di controlli e tipologie di errore

Per interpretare correttamente le tabelle seguenti occorre tenere presenti le seguenti avvertenze.

A) *Per collettivo di provenienza di un elemento si intende un collettivo più ampio*, di cui il collettivo considerato è una parte.

Nel caso di collettivi di tipo evento il collettivo di provenienza va semplicemente inteso come un collettivo più generale, in senso insiemistico.

Ad esempio il collettivo di tipo evento istantaneo *Costituzione di un'impresa* è un sottoinsieme del più ampio collettivo di tipo evento istantaneo *Costituzione di un'unità che svolge attività economica*, che si può considerare suo collettivo di provenienza.

Anche nel caso di collettivi di tipo popolazione il collettivo di provenienza può semplicemente essere inteso come un collettivo più generale, in senso insiemistico, ma può anche essere inteso come collettivo di provenienza in senso temporale, nel senso che un elemento proveniente da tale collettivo può entrare nel collettivo considerato in un certo momento oppure, se è un collettivo multiingresso, anche più volte.

Ad esempio il collettivo di tipo popolazione *Impresa* è un sottoinsieme del più ampio collettivo di tipo popolazione *Unità che svolge attività economica*, che si può considerare suo collettivo di provenienza.

Oppure, il collettivo *Degente* ha come collettivo di provenienza *Persona* nel senso che un elemento del collettivo *Persona* può entrare nel collettivo *Degente* e poi uscirne più volte, ogni volta che si verificano le circostanze previste, dipendenti dalla definizione di *Degente*.

Lo stesso vale per i collettivi di tipo evento con durata, anche se per essi la seconda eventualità è meno frequente.

Ad esempio il collettivo di tipo evento con durata *Rapporto di lavoro dipendente* è un sottoinsieme del più ampio collettivo dello stesso tipo *Rapporto di lavoro*, che si può considerare suo collettivo di provenienza.

Oppure, il collettivo di tipo evento con durata *Rapporto di lavoro part-time* ha come collettivo di provenienza *Rapporto di lavoro* nel senso che un elemento del collettivo *Rapporto di lavoro* può

entrare nel collettivo *Rapporto di lavoro part-time* e poi uscirne più volte, ogni volta che si verificano le circostanze previste, dipendenti dalla definizione di *Rapporto di lavoro part-time*.

B) *I vincoli di obbligatorietà e incompatibilità sono generalmente espressi in termini di combinazioni di modalità e valori e di elementi legati, che un elemento del collettivo può possedere per specifiche caratteristiche o relazioni pertinenti al collettivo. Tuttavia questi vincoli possono anche coinvolgere modalità, valori o legami derivati da quelli direttamente posseduti con una vasta gamma di operazioni specifiche, tra le quali anche la quantificazione sulle quelle relazioni 1-n nelle quali il collettivo è codominio*. Per ciò che riguarda la creazione di nuove caratteristiche sfruttando le relazioni esistenti tra collettivi si veda l'Appendice 3.

Ad esempio possono esistere vincoli di incompatibilità che legano la modalità posseduta da un elemento del collettivo *Rapporto di lavoro*, per la sua caratteristica *Qualifica professionale* [34], alla modalità della caratteristica *Livello di istruzione* posseduta dall'elemento del collettivo *Lavoratore* al quale il rapporto di lavoro è legato.

Oppure può esistere un vincolo di obbligatorietà che lega l'esistenza di eventi di *Acquisizione crediti* legati ad un elemento del collettivo *Studenti* all'esistenza di un evento *Iscrizione annuale a corso di laurea* legato allo stesso studente, vale a dire: il possesso di almeno un evento di *Acquisizione crediti* legato ad uno studente (un legame ricavato per quantificazione dalla relazione tra Studente e Acquisizione crediti, dove Studente è il codominio) implica obbligatoriamente l'esistenza di almeno un evento di *Iscrizione annuale a corso di laurea* per lo stesso studente.

C) *Per ciò che riguarda le corrispondenze con altri archivi, il risultato dell'operazione di linkage esatto dipende evidentemente dalla qualità relativa degli archivi tra i quali si effettua*: in generale gli errori potrebbero dipendere dall'archivio che si sta controllando così come dall'archivio utilizzato per il controllo. Una buona organizzazione dei controlli richiederebbe la disponibilità preliminare di valutazioni della qualità relativa degli archivi tra i quali si può effettuare la corrispondenza.

[34] Nelle Comunicazioni obbligatorie la caratteristica Qualifica professionale è pertinente al collettivo di eventi con durata Rapporto di lavoro

TIPI DI CONTROLLO: **CONTROLLI STRUTTURALI**	ENUNCIATI DI APPARTENENZA Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di identificazione	ENUNCIATI DI POSSESSO CARATTERISTICHE Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di raccordo	ENUNCIATI DI POSSESSO RELAZIONI Errata inclusione, Errata esclusione, Duplicazione
Controllo sintattico della **correttezza degli identificativi** degli elementi dei collettivi: si controlla la correttezza della loro struttura o la corrispondenza con una lista certificata di identificativi		**X** (per identificativi degli elementi)		**X**	
Controllo sintattico della **correttezza degli identificativi** delle modalità o valori per le caratteristiche: si controlla la loro appartenenza all'insieme delle modalità previste per la classificazione utilizzata, o al range di valori predefiniti		**X** (per identificativi delle modalità o valori)			
Controllo della presenza di record interamente duplicati	**X**				
In occasioni di operazioni di tipo REG, nel caso in cui ciò sia tecnicamente possibile, presenza di valori plurimi per caratteristiche o relazioni, che indica duplicazione dei relativi enunciati			**X**		**X**
Controllo della presenza di codici identificativi duplicati: può indicare duplicazione di un enunciato di appartenenza oppure dipendere dalla condivisione dello stesso identificativo tra elementi diversi, la distinzione richiede il	**X**	X			

controllo dell'identità degli elementi utilizzando particolari caratteristiche e relazioni					
Controllo dell'identità dell'elemento utilizzando particolari caratteristiche e relazioni, per individuare la duplicazione di un enunciato di appartenenza (se l'identificativo è identico nei due record, serve per distinguere la duplicazione dall'errore di condivisione di codici tra elementi differenti, se invece l'identificativo è diverso serve per diagnosticare la duplicazione del record, in presenza di errore di doppio codice per l'elemento)	**X**	**X**			
In occasione di operazioni di tipo ELIM, presenza di record relativi a eventi di uscita che non trovano il raccordo con un record aperto relativo all'elemento al quale l'evento di uscita è legato	**X**	**X**		**X**	**X**
In occasione di operazioni di tipo MOD presenza di enunciati di possesso di caratteristiche o relazioni che non trovano il raccordo con un record aperto relativo all'elemento	**X**	**X**	**X**		**X**
In occasione di operazioni di tipo MOD presenza di enunciati di possesso di relazioni con codici di raccordo che non trovano il raccordo con un record aperto relativo all'elemento raccordato	**X**	**X**		**X**	**X**

TIPI DI CONTROLLO: **USO DI INFORMAZIONI A PRIORI**	ENUNCIATI DI APPARTENENZA Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di identificazione	ENUNCIATI DI POSSESSO CARATTERISTICHE Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di raccordo	ENUNCIATI DI POSSESSO RELAZIONI Errata inclusione, Errata esclusione, Duplicazione
informazioni a priori sull'esistenza di cause che provocano errori nella procedura di scelta appartenenza/non appartenenza	**X**				
informazioni a priori su cattiva progettazione o malfunzionamento procedura di identificazione		**X**			
informazioni a priori su cattiva progettazione o malfunzionamento procedure di formulazione enunciati su caratteristiche			**X**		
informazioni a priori su cattiva progettazione o malfunzionamento procedura di attribuzione codici di raccordo				**X**	
informazioni a priori su cattiva progettazione o malfunzionamento procedure di formulazione enunciati su relazioni					**X**

TIPI DI CONTROLLO: **LINKAGE CON ALTRI ARCHIVI**	ENUNCIATI DI APPARTENENZA Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di identificazione	ENUNCIATI DI POSSESSO CARATTERISTICHE Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di raccordo	ENUNCIATI DI POSSESSO RELAZIONI Errata inclusione, Errata esclusione, Duplicazione
linkage con anagrafe relativa al *collettivo di provenienza* con identificativi certificati (esempio Anagrafe fiscale), usando l'identificativo dell'elemento ed eventualmente anche insieme di caratteristiche e relazioni con potenzialità di identificazione (esempio Nome, Cognome) o comunque metodi di controllo dell'identità degli elementi posti in corrispondenza	**X**	**X**			
linkage con archivio esterno relativo al *collettivo di provenienza* con identificativi *non* certificati, usando l'identificativo dell'elemento ed eventualmente anche insieme di caratteristiche e relazioni con potenzialità di identificazione (esempio Nome, Cognome) *MA possibilità di errore nell'archivio esterno*	**X**	**X**			
linkage con archivio esterno relativo al *collettivo considerato*, usando l'identificativo dell'elemento ed eventualmente anche insieme di caratteristiche e relazioni con potenzialità di identificazione (esempio Nome, Cognome)	**X**	**X**			

MA possibilità di errore nell'archivio esterno					
raramente: linkage con archivio esterno relativo al *collettivo complementare al collettivo considerato* (cioè contenente quegli elementi del collettivo di provenienza che non appartengono al collettivo considerato), usando l'identificativo ed eventualmente anche insieme di caratteristiche e relazioni con potenzialità di identificazione (esempio Nome, Cognome) *MA possibilità di errore nell'archivio esterno*	**X**	**X**			
linkage con archivio esterno relativo al *collettivo considerato* che contenga la *caratteristica da controllare* *MA possibilità di errore nell'archivio esterno*			**X**		
linkage con archivio esterno relativo al *collettivo considerato* che contenga la *relazione da controllare* *MA possibilità di errore nell'archivio esterno*					**X**
linkage con anagrafe relativa al *collettivo raccordato* con identificativi certificati (esempio Anagrafe fiscale), usando l'identificativo dell'elemento legato				**X**	**X**
linkage con archivio esterno relativo al *collettivo raccordato*, con identificativi *non* certificati usando l'identificativo dell'elemento legato *MA possibilità di errore nell'archivio esterno*				**X**	**X**

TIPI DI CONTROLLO: **USO DI VINCOLI STATICI O DINAMICI DI OBBLIGATORIETA' O INCOMPATIBILITA'**	ENUNCIATI DI APPARTENENZA Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di identificazione	ENUNCIATI DI POSSESSO CARATTERISTICHE Errata inclusione, Errata esclusione, Duplicazione	IDENTIFICATIVI nel ruolo di raccordo	ENUNCIATI DI POSSESSO RELAZIONI Errata inclusione, Errata esclusione, Duplicazione
Controllo sulla *presenza in un record* di *modalità o valori di caratteristiche* che comportano obbligatoriamente *l'appartenenza al collettivo*, oppure che sono incompatibili con l'appartenenza al collettivo	**X**		**X**		
Controllo sulla *presenza in un record* di *legami con elementi* che comportano obbligatoriamente *l'appartenenza al collettivo* oppure sono incompatibili con l'appartenenza al collettivo	**X**			**X**	**X**
Controllo basato su relazioni di conseguenza o incompatibilità tra *modalità o valori di caratteristiche presenti in un record*			**X**		
Controllo basato su relazioni di conseguenza o incompatibilità *tra elementi legati, e tra modalità o valori, per relazioni e caratteristiche presenti in un record*			**X**	**X**	**X**
Controllo basato su relazioni dinamiche di conseguenza o incompatibilità tra *modalità o valori di caratteristiche presenti in record*			**X**		

successivi relativi allo stesso elemento					
Controllo basato su relazioni di conseguenza o incompatibilità dinamiche tra elementi legati, e tra modalità o valori, per *relazioni e caratteristiche presenti in record successivi* relativi allo stesso elemento			**X**	**X**	**X**

TIPI DI CONTROLLO: **CALCOLO DELLA DISTANZA MEDIA tra il momento di riferimento effettivo degli enunciati e il momento di riferimento t_i registrato in archivi**	ENUNCIATI DI APPARTENENZA Errata inclusione e Errata esclusione TEMPORANEE	IDENTIFICATIVI nel ruolo di identificazione	ENUNCIATI DI POSSESSO CARATTERISTICHE Errata inclusione e Errata esclusione TEMPORANEE	IDENTIFICATIVI nel ruolo di raccordo	ENUNCIATI DI POSSESSO RELAZIONI Errata inclusione e Errata esclusione TEMPORANEE
Calcolo della distanza media tra il momento di riferimento effettivo degli *enunciati di appartenenza relativi a un collettivo di eventi istantanei* e il momento di riferimento t_i registrato in archivio (e di conseguenza delle distanze medie tra i momenti effettivi di *inizio* e *chiusura* e i momenti t_i e t_j registrati in archivio, per *enunciati di appartenenza a collettivi di tipo popolazione o evento con durata*)	**X**				
Per gli elementi di un collettivo di tipo popolazione o evento con durata, calcolo della distanza media tra il momento di inizio effettivo dei *nuovi enunciati di possesso di una caratteristica* e il momento di inizio t_i registrato in archivio (e di conseguenza della distanza media tra i momenti effettivi di *inizio* e *chiusura* e i momenti t_i e t_j registrati in archivio, per tutti gli *enunciati di possesso di una caratteristica*)			**X**		

Per gli elementi di un collettivo di tipo popolazione o evento con durata, calcolo della distanza media tra il momento di inizio effettivo dei *nuovi enunciati di possesso di una relazione* e il momento di inizio t_i registrato in archivio (e di conseguenza della distanza media tra i momenti effettivi di *inizio* e *chiusura* e i momenti t_i e t_j registrati in archivio, per tutti gli *enunciati di possesso di una relazione*)					**X**

TIPI DI CONTROLLO: **CALCOLO DELLA DISTANZA MEDIA tra il momento di riferimento t_i registrato in archivio per un enunciato e il momento t dell'immissione in archivio della relativa informazione**	ENUNCIATI DI APPARTENENZA	IDENTIFICATIVI nel ruolo di identificazione	ENUNCIATI DI POSSESSO CARATTERISTICHE	IDENTIFICATIVI nel ruolo di raccordo	ENUNCIATI DI POSSESSO RELAZIONI
Calcolo della distanza media tra il momento di riferimento t_i registrato in archivio per gli *enunciati di appartenenza relativi a un collettivo di eventi istantanei* e il momento t dell'immissione in archivio della relativa informazione (e di conseguenza delle distanze medie tra i momenti t_i e t_j di *inizio* e *chiusura* registrati in archivio e il momento t dell'immissione in archivio della relativa informazione, per *enunciati di appartenenza a collettivi di tipo popolazione o evento con durata*)	**X**				

Per gli elementi di un collettivo di tipo popolazione o evento con durata, calcolo della distanza media tra il momento di inizio t_i registrato in archivio per i *nuovi enunciati di possesso di una caratteristica* e il momento t dell'immissione in archivio della relativa informazione (e di conseguenza della distanza media tra i momenti t_i e t_j di *inizio* e *chiusura* registrati in archivio e il momento t dell'immissione in archivio della relativa informazione, per tutti gli *enunciati di possesso di una caratteristica*)			**X**		
Per gli elementi di un collettivo di tipo popolazione o evento con durata, calcolo della distanza media tra il momento di inizio t_i registrato in archivio per i *nuovi enunciati di possesso di una relazione* e il momento t dell'immissione in archivio della relativa informazione (e di conseguenza della distanza media tra i momenti t_i e t_j di *inizio* e *chiusura* registrati in archivio e il momento t dell'immissione in archivio della relativa informazione, per tutti gli *enunciati di possesso di una relazione*)					**X**

APPENDICE 1 Statistical Network on Administrative Data: methodologies for an integrated use of administrative data in the statistical process

Preliminary technical report on a generic quality assessment framework

Background

Administrative data sources are data holdings containing information which is not primarily collected for statistical purposes (UNECE 2011). Under this definition, administrative data is not restricted to data collected and maintained for the purpose of implementing administrative regulations and can include commercial data for example held by the private sector.

The Statistical Network responsible for developing methodologies for an integrated use of administrative data in the statistical process (SN-MIAD) is chaired by the Italian National Institute (Istat) and comprises representatives from the Australian Bureau of Statistics (ABS), Statistics Canada (StatCan), Statistics New Zealand (SNZ) and Statistics Sweden (SCB). The development of the project charter of the SN-AD started in October 2012 and led to its approval in June 2013.

The project charter states that a framework to assess the quality of administrative data and its statistical usability will be developed. This framework will be applied prior to the statistical usage of the administrative data, to determine if an administrative data source can be used for statistical purpose and how. This report constitutes the first deliverable under this objective.

It is understood that preceding and complementing this work will be a mapping of how administrative data can be used into a statistical process (e.g., applying the Generic Statistical Business Process Model (GSBPM) as the foundation). Once this mapping is done and the framework to assess the quality of administrative data and determine the statistical usability is completed, the SN-MIAD will be in a position to move to a second stage of a project and develop a general framework for the integration of administrative and survey data for each one of the GSBPM phases and for the reporting of the quality of the resulting statistical output.

Objective of this report

This report constitutes deliverable B.1 of the SN-MIAD project charter. In progressing towards the development of a generic quality assessment framework that could be used across national statistical offices (NSO), this report provides the foundation of this work by including a set of elements that need to be considered when developing this framework taken into account that the context may be different from country to county, an inventory of existing approaches with a focus on those that stem as relevant and good practices and the resulting elements that are proposed to be included in the generic quality assessment framework.

Approval of this preliminary report from the Steering Committee will be sought before finalizing the second deliverable, which is the final technical report detailing the generic quality assessment framework of administrative data.

Elements to be considered

Assessment is likely to start with no data in hands

Accessing and using administrative data for official statistics is bounded by the legislation in place in a given country. Legal frameworks vary widely from country to country, and frequently reflect historical and cultural differences around issues of privacy and security. In some countries, central

population registers, personal identification numbers and large-scale data matching to produce official statistics are seen as sensible approaches that reduce respondent burden and costs, while in other countries such approaches meet with strong opposition from privacy advocates and the general public.

For example, Section 13 of the *Statistics Act* gives StatCan the authority to access virtually any administrative record to fulfil the purposes of the Act. By comparison, the United Kingdom must obtain Parliament's authority to access administrative data on a case-by-case basis according to the dispositions of the *Statistics and Registration Act 2007*. Dispositions in legislation can also ease the access such as in the *Statistics Act* of Ireland where the Act specifies that administrative records held by public authorities must be provided free of charge, that the right of access applies notwithstanding other enactments, other than those pertaining to the protection of public order or the security of the state. Such dispositions are not part of the Canadian *Statistics Act* and this absence often slows down the acquisition of administrative data.

All this to say that to meet the needs of a variety of NSOs, the generic quality assessment framework for administrative data must include an initial step of assessment where the data is assessed by the NSO with no data in hands, relying mostly on consultation with the data provider. A subsequent step of assessment can be planned once the data is received by the NSO.

Assessment is likely to be conducted with no, one or many statistical usages in mind

References

United Nations Economic Commission for Europe (UNECE). 2011. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.* United Nations Publication.

Usage of Administrative Data Sources for Statistical Purposes

Register vs Survey based Systems

A review of literature shows a clear relationship between the usage of ADS and the degree to which the NSO has adopted a register-based statistical system.

Countries which have a largely register based system (e.g. Nordic countries) typically make far greater use of ADS, usually in direct tabulation. They exist in an environment of both ready availability and high quality of ADS. A fully-register based system typically involves a suite of registers built from ADS and a unique ID that enables linking of records between registers. Surveys in register based systems are often used to supplement/assist the ADS with the ADS considered the primary reliable source.

At other end of spectrum are countries that primarily have a survey-based statistical system (e.g. Australia). Due to legislative environment and other factors ADS are less available, often not at the statistical unit level of interest, and of variable quality. ADS in these situations can be used for direct tabulation for some statistics where quality is deemed sufficient but often without ability to form links between different sources. Traditionally ADS are used to supplement/assist surveys through creation of sampling frames and/or as auxiliary variables used in estimation, with ADS usage built around carefully designed surveys. At macro levels survey data can be considered the 'truth' and be used to adjust the ADS.

Various NSOs exist between these ends of the register/survey spectrum. Some have registers but no unique ID to link them together (e.g. Germany), and often within the NSO itself there may be differing degrees to which ADS are relied on e.g. economic versus social statistics depending on availability and quality of ADS.

With increasing levels of ADS becoming available and advances in probabilistic linking methods/tools, ADS are expected to play more prominent role, replacing surveys or components thereof and/or augmenting existing survey data. The integration of data sources could be considered more complex in as survey based system as compared to a clean register based system.

The register/survey dichotomy may help frame the context for the inventory of ADS usage below and assist in establishing quality dimensions required. It will also guide the next phase of this project in summarising the different 'informative contexts' for ADS use, in particular the roles ADS can play in inference when integrated with survey data.

Uses for admin data

Following is a draft inventory of ways in which administrative data sources (ADS) are currently used by NSOs in the production of official statistics. The context assumes descriptive statistical outputs (e.g. counts, totals, percentiles) are the primary objective as opposed to analytical studies, and that outputs meet NSO's quality parameters as deemed fit for relevant 'official statistics' status.

We list usages of ADS as described in Lavallée's paper "Administrative Data Usage in the Framework of Social Statistics: Current and Future Picture" (2007), with one dimension "survey evaluation" replaced by "Data validation/confrontation". We found this categorization as good as any other we could come up with.

The usages described in Lavallée's paper are a little survey centric, the one exception of 'direct tabulation' where an ADS has full coverage of population of interest. Cases where partial coverage ADS are utilised in combination with surveys and/or other ADS data integration exercises should be noted in the inventory, though it is considered they can still fit within the broad 7 usages described.

1. Creation and maintenance of registers and survey frames
    - Registers – definition, requirements
    - Sampling frames
    - Statistical units
    - Auxiliary data (e.g. size and classificatory items)

2. Construction of sampling designs
    - measures of variability for design variables
    - size measures to improve efficency and/or targetting sub-populations of interest (stratification, pps etc)
    - facilitation of specialised selection mechanisms e.g. screening
    - ADS to facilitate sample design vs sample design to facilitate ADS register based system

3. Substitution for direct collection
    - Whole and/or partial substitution for directly collected survey variables for sub-populations and/or variables of interest
    - Augmentation of directly collected survey variables

- *Data ‘fusion’; integrating multiple sources data representing the same object to produce synthetic data that is more informative than original

4. Editing and imputation
   - construction of edit rules
   - auxiliary data to construct imputation models
   - for surveys and/or other ADS

5. Indirect estimation and weighting
   - Creation of population benchmarks (independent of frame)
   - Improve efficiency of estimation through a model assisted or model based framework
   - Address quality issues e.g. non-response
   - ‘harmonising’ – creating new figures that agree between sources
   - Prediction – using a early available ADS to predict later more reliable estimates

6. Direct tabulation
   - For full coverage ADS, simple computation of totals, means, percentiles etc.
   - Without a system of registers or unique ID, can only directly tabulate based on a single data set., the more integrated the admin datasets are, the more varied the tabled statistics can be

7. Data validation/confrontation
   - validation of survey estimates and/or other ADS
   - micro or macro level
   - assess the quality of other potential ADS

APPENDICE 2: Indicatori Blue–Ets

Introduzione

Nell'approccio BLUE - ETS si propone di adottare quattro dimensioni descrittive della qualità statistica dei dati amministrativi.

Nella **Tabella 1** le dimensioni sono riportate specificando anche la necessità di definire due livelli di lettura: uno per le variabili e uno per gli oggetti (in genere unità ed eventi). Tale distinzione è strategica per i dati amministrativi poiché le unità amministrative non hanno necessariamente una connotazione statistica e la valutazione della qualità della fonte deve comprendere quest'ottica bivalente.

Nelle **Tabelle 2 – 5** si riportano gli indicatori considerati maggiormente rilevanti nell'ambito della presente attività.

Si sottolinea infatti che gli obiettivi del WP4 Blue Ets mirano a valutare la qualità dei dati amministrativi per il loro utilizzo statistico all'interno dell'Istituto di statistica. L'ottica alla base del presente Framework è invece di interagire con l'ente produrre e definire in modo condiviso gli elementi descrittivi da considerare. La proposta presentata riadatta, quindi, in questa prospettiva la ricerca condotta nell'ambito del progetto Blue Ets.

Il framework Blue Ets prevede una flessibilità necessaria a comprendere le specificità delle fonti amministrative: se le dimensioni e gli indicatori sono fissati, i metodi di misura sono adattabili al caso. Nel progetto Blue Ets sono stati proposti complessivamente circa 50 metodi di misura.

Per quanto riguarda la definizione dei metodi di misura dei singoli indicatori, occorre riflettere sugli interessi specifici che devono essere pertinenti alla fonte e condivisi.

Principali riferimenti

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Cerroni, F., Di Bella, G., Laitila, T., Wallgren, A., Wallgren, B. (2011) Report on methods preferred for the quality indicators of administrative data sources. Deliverable 4.2 Blue Ets, September 28.

Di Bella G., Galiè L., Bonardo D., Cerroni F., Talucci V. (2012) Methodological case study for testing and evaluating WP4 input data quality indicators, Istat Report for WP8 Blue Ets, December 31 (draft).

http://www.blue-ets.istat.it

Tabella 1. Dimensioni descrittive della qualità statistica dei dati amministrativi

DIMENSIONI	LIVELLO	SCOPO DEGLI INDICATORI
1.Integrabilita'	Oggetti/Variabili	Verificare la capacità della fonte di essere integrata nei processi di produzione statistica
2. Accuratezza	Oggetti/Variabili	Misurare la precisione dei dati, della loro affidabilità dal punto di vista statistico
3. Completezza	Oggetti/Variabili	Verificare la capacità dei dati di comprendere le informazioni in modo esaustivo rispetto all'insieme degli oggetti di riferimento della popolazione target statistica
4. Dimensione connessa agli aspetti temporali	Oggetti/Variabili	Verificare l'utilizzabilità dei dati in termini di tempestività, misurare la dinamicità degli oggetti e la stabilità delle variabili

Tabella 2. Indicatori di Integrabilità

Dimensione	Livello	Indicatori	Descrizione
1. Integrabilità	Oggetti	1.1. Comparabilità degli oggetti	Corrispondenza tra gli oggetti della fonte e gli oggetti statistici
	Variabili	1.2. Comparabilità delle variabili	Corrispondenza delle variabili della fonte con quelle statistiche
	Variabili	1.3. Variabili di linkage	Utilizzabilità delle variabili di linkage presenti nella fonte

Tabella 3. Indicatori di Accuratezza

Dimensione	Livello	Indicatori	Descrizione
2. Accuratezza	Oggetti	2.1.Autenticità	Legittimità degli oggetti della fonte
	Oggetti	2.2. Oggetti inconsistenti	Grado di coerenza nelle relazioni tra diversi tipi di unità

	Oggetti	2.3. Oggetti dubbi	Presenza di relazioni ambigue tra diversi tipi di oggetti contenuti nella fonte
	Variabili	2.4. Errori di misura	Presenza di errori di misura sui valori delle variabili
	Variabili	2.5. Valori inconsistenti	Presenza di valori (o di combinazioni di valori) non corretti statisticamente per una o più variabili
	Variabili	2.6 Valori dubbi	Presenza di valori (o di combinazioni di valori) ambigui statisticamente per una o più variabili

Tabella 4. Indicatori di Completezza

Dimensione	Livello	Indicatori	Descrizione
3. Completezza	Oggetti	3.1 Sottocopertura	Oggetti target mancanti nell'archivio
	Oggetti	3.2 Sovracopertura	Presenza di oggetti non-target nell'archivio
	Oggetti	3.3 Selectivity	Copertura per sottopopolazioni statistiche
	Oggetti	3.4 Ridondanza	Presenza di registrazioni multiple degli oggetti
	Variabili	3.5 Valori mancanti	Assenza di valori per oggetti di interesse
	Variabili	3.6 Valori imputati	Presenza di valori derivanti da procedure di imputazione effettuate dal produttore di dati

Tabella 5. Indicatori della dimensione temporale

Dimensione	Livello	Indicatori	Descrizione
4. Dimensione temporale	Archivio	4.1 Tempestività	Tempo totale trascorso tra la fine del periodo di riferimento dei dati nella

			fonte e il momento di disponibilità
	Archivio	4.2 Ritardo	Ritardi di registrazione nell'archivio dell'evento
	Oggetti	4.3 Dinamicità degli oggetti	Variazioni della popolazione di oggetti nel tempo
	Variabili	4.4. Stabilità delle variabili	Variazioni delle variabili o dei valori nel tempo

INTEGRABILITA’		
Indicatori	**Livello**	**Metodi di misura**
Comparabilità degli oggetti Corrispondenza tra gli oggetti della fonte e gli oggetti statistici (oggetti identici, corrispondenti, comparabili)	Oggetti	% di oggetti dell’archivio identici a quelli presenti in una popolazione statistica. (Analogamente per gli oggetti corrispondenti e incomparabili)
		% di oggetti dell’archivio non corrispondenti, a livello aggregato, a quelli presenti in una popolazione statistica
		% di oggetti in una popolazione statistica non allineati, a livello aggregato, a quelli presenti nell’archivio
Variabili di linkage Utilizzabilità delle variabili di linkage presenti nella fonte	Variabili	% di oggetti dell’archivio con valore mancante nella variabile di linkage
		% di oggetti dell’archivio con valore della variabile di linkare diverso da quello utilizzato in Istat (errori sintattici)
		% di oggetti dell’archivio per cui il valore della variabile di linkare risulta convertibile in quello utilizzato in Istat
Comparabilità delle variabili Corrispondenza tra le variabili della fonte e quelle presenti in altre fonti	Variabili	Metodi grafici o analitici per confrontare a livello aggregato i valori che una variabile presenta nell’archivio in esame con quelli assunti dalla stessa variabile in un altro archivio (Grafici a barre, scatter plot, MAPE, V di Cramer)
		% di oggetti dell’archivio aventi esattamente lo stesso valore della variabile in esame in un altro archivio (comparabili)

ACCURATEZZA		
Indicatori	**Livello**	**Metodi di misura**
Autenticità Legittimità degli oggetti	Oggetti	% di record/oggetti che hanno un valore dell'identificativo errato (incompatibile con la definizione teorica della sintassi dell'identificativo o incoerente con quello presente in una lista di riferimento)
		% di record non autentici dichiarata dal fornitore della fonte di dati
Oggetti inconsistenti Grado di coerenza nelle relazioni tra diversi tipi di unità	Oggetti	% di oggetti coinvolti in relazioni non logiche con altri oggetti o eventi
Oggetti dubbi Presenza di relazioni ambigue tra diversi tipi di oggetti contenuti nella fonte	Oggetti	% di oggetti coinvolti in relazioni ambigue ma non necessariamente errate con altri oggetti
Errori di misura Presenza di errori di misura sui valori delle variabili	Variabili	% di valori per ciascuna variabile segnalati come errati dal fornitore
		Informazioni richieste al fornitore della fonte sulla gestione della qualità nella fase di raccolta dei dati Esiste un processo di controllo durante la fase di acquisizione dei dati, quali regole vengono utilizzate? Esiste un processo di controllo sui dati? Quali regole vengono utilizzate? I dati incoerenti sono segnalati/corretti? Come? Esiste un piano per la gestione dei valori mancanti in fase

		di acquisizione e in fase di trattamento?
Valori inconsistenti Presenza di valori (o di combinazioni di valori) non corretti per una o più variabili	Variabili	% di oggetti con valori (o combinazioni di valori) su una o più variabili coinvolti in relazioni non logiche
Valori dubbi Presenza di valori (o di combinazioni di valori) ambigui per una o più variabili	Variabili	% di oggetti con valori (o combinazioni di valori) su una o più variabili coinvolti in relazioni ambigue ma non necessariamente errate

COMPLETEZZA		
Indicatori	**Livello**	**Metodi di misura**
Sottocopertura Oggetti target mancanti nell'archivio	Oggetti	% di oggetti della lista di riferimento mancanti nell'archivio
Sovracopertura Presenza di oggetti non-target nell'archivio	Oggetti	% di oggetti dell'archivio non inclusi nella popolazione di riferimento
		% di oggetti dell'archivio non appartenenti alla popolazione target
Selectivity Copertura per sottopopolazioni statistiche	Oggetti	Metodi statistici (distanze) per confrontare le distribuzioni degli oggetti nell'archivio e nella popolazione di riferimento rispetto ad una o più variabili di stratificazione
		Metodi grafici (istogrammi, tableplots)
Ridondanza Presenza di registrazioni multiple degli oggetti	Oggetti	% di oggetti duplicati nell'archivio (con lo stesso codice identificativo)
		% di oggetti duplicati nell'archivio con gli stessi valori per un insieme di variabili
		% di oggetti duplicati nell'archivio con gli stessi valori per tutte le variabili
Valori mancanti Assenza di valori per variabili di interesse	Variabili	% di oggetti con valore mancante per una particolare variabile
		% di oggetti con tutti valori mancanti per un insieme (limitato) di variabili
		Metodi grafici per il controllo di valori mancanti sulle

		variabili
Valori imputati Presenza di valori derivanti da procedure di imputazione effettuate dal fornitore dei dati	Variabili	% di oggetti con valori imputati per ciascuna variabile nell'archivio
		% di valori imputati dichiarata dal fornitore dei dati per ciascuna variabile

DIMENSIONE TEMPORALE		
Indicatori	**Livello**	**Metodi di misura**
Tempestività Tempo totale trascorso tra la fine del periodo di riferimento dei dati nella fonte e il momento di disponibilità	Archivio	Differenza totale tra la data finale del periodo di riferimento dei dati e la data di disponibilità dell'archivio
Ritardo Ritardi di registrazione nell'archivio dell'evento	Archivio	Ritardi nelle registrazioni comunicati dal fornitore dei dati
		Differenza tra la data di registrazione degli eventi nella fonte da parte del fornitore e la data di accadimento dell'evento nella popolazione
Dinamicità degli oggetti Variazioni della popolazione di oggetti nel tempo	Oggetti	% di oggetti presenti al tempo t ma non al tempo t-1 (nuovi oggetti) rispetto al totale di oggetti al tempo t
		% di oggetti presenti al tempo t-1 ma non al tempo t (vecchi oggetti) rispetto al totale di oggetti al tempo t (o al totale di oggetti al tempo t-1)
Stabilità delle variabili Variazioni delle variabili o dei valori nel tempo	Variabili	Metodi grafici (grafici a barre, scatter plot) per il confronto dei valori di specifiche variabili assunti dagli oggetti persistenti in differenti forniture della fonte
		% di oggetti con valore cambiato da t-1 a t (di una variabile a valori non missing) rispetto al totale di oggetti persistenti
		Indici di associazione (V di Cramer), per variabili categoriali, o indici di correlazione, per variabili numeriche, tra i valori di una stessa variabile al tempo t e al tempo t-1

APPENDICE 3: Documentare l'Ontologia di un Archivio Amministrativo

Perché documentare l'ontologia di un archivio amministrativo

Caratterizzare gli archivi amministrativi come strumenti di raccolta di informazioni di potenziale interesse per lo statistico comporta documentarne il contenuto informativo in funzione di questo scopo.

Il criterio di partenza per documentare il contenuto di un archivio amministrativo, e in generale di una collezione di dati ottenuti dall'osservazione di specifici aspetti del mondo reale, in funzione del suo potenziale utilizzo statistico è descrivere sistematicamente tutto ciò che in esso può essere utilizzato per definire statistiche, indipendentemente da un eventuale uso statistico attuale o pianificato.

Empiricamente, una statistica è definibile come un indicatore numerico calcolabile a partire dai dati, ad esempio una frequenza o il totale di una variabile numerica relativo a un collettivo, eventualmente partizionato utilizzando variabili di classificazione.

Si tratta quindi di *documentare l'ontologia dell'archivio amministrativo*, e in generale di una collezione di dati, basandosi però *su un modello concettuale che finalizzi la documentazione dell'ontologia al potenziale uso dell'archivio per la produzione di statistiche*.

Ciò comporta enucleare concettualmente tutti gli aspetti del mondo reale ai quali sono riferite le informazioni gestite nell'archivio, con una specifica attenzione però al loro possibile uso per definire statistiche.

All'ontologia dell'archivio così definita può essere poi ancorata la definizione dei diversi tipi di errori.

Sulla base di questi presupposti, l'ontologia di un archivio amministrativo (come del resto di un'indagine sufficientemente complessa) risulta spesso specificata come una rete di diversi collettivi di tipo popolazione o evento connessi da relazioni.

Questo tipo di specifica del contenuto informativo dell'archivio si discosta solo in apparenza dai presupposti comunemente esposti nei testi accademici di statistica.

In essi si assume che lo statistico sia interessato allo studio della distribuzione congiunta di un insieme di variabili su un unico collettivo di riferimento (indicato anche come popolazione di riferimento, o come unità d'analisi).

A rigore una statistica è definita come parametro di sintesi della distribuzione congiunta di un insieme di variabili osservate su un collettivo, o come parametro derivato da parametri di sintesi della distribuzione congiunta. Questo punto di vista è proprio dello statistico interessato allo studio approfondito di uno specifico fenomeno, in funzione del quale definisce il collettivo e le variabili di proprio interesse.

D'altra parte, in presenza di collezioni di dati disponibili, in particolare di archivi amministrativi, lo statistico dovrà preliminarmente valutare l'opportunità di "estrarre" da queste collezioni di dati l'informazione relativa al singolo collettivo e alle variabili di proprio interesse.

Questa operazione si colloca a valle della descrizione delle ontologie di tali collezioni di dati, descrizione che d'altra parte è necessaria a compiere questa operazione, e deve essere specificata proprio in funzione di essa.

Quanto più gli statistici ricorreranno per lo studio dei fenomeni all'utilizzo di collezioni di dati già esistenti, di qualsiasi origine, ed in particolare quindi di archivi amministrativi, quanto più diventerà importante descrivere l'ontologia di tali collezioni di dati in modo standard e comprensibile a tutti i potenziali utilizzatori, e al tempo stesso indipendente da ogni successiva scelta di "estrazione" che può essere operata dal singolo utilizzatore.

Inoltre a differenza dello studioso, il quale può seguire un approccio più esplorativo, la statistica ufficiale è chiamata a produrre informazioni a supporto delle decisioni dei soggetti della vita collettiva, in una forma perciò necessariamente sintetica e ponendo spesso in relazione informazioni relative a diversi fenomeni, diffondendo quindi direttamente un insieme di statistiche relative a numerosi collettivi pianificate sulla base del concetto empirico di indicatore numerico di un fenomeno.

Per considerazioni di costi infine, l'uso potenziale di una stessa collezione di dati per ricavare numerosi indicatori di sintesi, anche relativi a collettivi diversi, è un valore aggiunto per la statistica ufficiale.

Tutte queste considerazioni motivano la descrizione completa e sistematica delle ontologie delle collezioni di dati disponibili.

Nel nostro approccio, l'ontologia di una fonte d'informazione è descritta mediante una rete di *collettivi* i cui elementi posseggono *caratteristiche (variabili)*, le quali assumono modalità o valori in *classificazioni* o *domini* rispettivamente, e sono connessi da *relazioni* (si veda il disegno a pagina 113 per una prima presentazione sintetica delle metarelazioni tra questi concetti).

Documentare l'ontologia di un archivio amministrativo: i collettivi

Primo passo nella definizione dell'ontologia di un archivio amministrativo è l'individuazione di tutti i *collettivi* che lo caratterizzano. Questa attività è guidata dalle seguenti definizioni.

Definizione di collettivo: un collettivo d'interesse statistico è un'entità del mondo reale costituita da un insieme di uno o più elementi, detti istanze del collettivo, che presentano una o più proprietà (caratteristiche e relazioni con altri collettivi), entità che può essere oggetto di operazioni di rilevazione o misura che diano luogo a statistiche non ricavabili altrimenti.

Questa definizione offre un criterio per discriminare che cosa conviene considerare come un collettivo a sé stante, e non semplicemente come una proprietà di un altro collettivo: ogni insieme di elementi osservabili su cui può avere *potenzialmente* interesse effettuare un'operazione *indipendente* di misura.

Definizione di collettivo di tipo popolazione: un collettivo d'interesse statistico costituito da istanze che hanno una loro esistenza indipendente e durata e che è sottoinsieme di uno dei due grandi collettivi di base delle persone e degli organismi che svolgono attività economica, o di un collettivo ad essi connessi.

Esempi di collettivi che sono direttamente sottoinsieme dei due grandi collettivi di base: Studenti, Docenti, Degenti, Lavoratori, Richiedenti asilo, Contribuenti persone fisiche, che sono sottoinsiemi della popolazione di base Persone, oppure Aziende agricole, Ospedali, Scuole, Atenei, Istituti di previdenza, Datori di lavoro, Contribuenti persone giuridiche, che sono sottoinsiemi della popolazione di base Organismi che svolgono attività economica.

Esempi di collettivi che coincidono o sono sottoinsiemi di altri grandi collettivi che sono connessi alle persone o agli organismi che svolgono attività economica da relazioni di raggruppamento o composizione: Famiglia, Nucleo familiare, Gruppo di imprese, Unità territoriale dell'impresa, Istituto comprensivo (che riunisce più scuole), Circolo didattico, oppure da relazioni organizzative o funzionali: Corso di laurea, Dipartimento, Facoltà, Unità funzionale di un'impresa o di un'istituzione.

Tenendo conto di quanto detto nel primo paragrafo del Framework sul ciclo di vita dell'informazione amministrativa, in primo luogo possono essere descritti come collettivi principali di tipo popolazione tutti i soggetti coinvolti nell'attività amministrativa a supporto della quale è costituito l'archivio, quindi da una parte l'insieme delle istituzioni che esercitano l'attività (che sono sottoinsieme della popolazione di base Organismi che svolgono attività economica), dall'altra quegli specifici sottoinsiemi delle due popolazioni di base delle persone e degli organismi che svolgono attività economica, o di popolazioni ad esse connesse, sui quali si esercita l'attività.

Dato che il fine è descrivere il contenuto effettivo della versione di archivio, vanno descritti come collettivi di tipo popolazione tutti e soli quegli insiemi di elementi per i quali l'archivio registra effettivamente delle caratteristiche o relazioni con altri collettivi, o che è comunque utile introdurre per descrivere compiutamente il contenuto dell'archivio.

Ad esempio nell'Anagrafe studenti universitari si può non introdurre esplicitamente il collettivo Atenei, anche se gli atenei sono soggetti dell'attività amministrativa che l'archivio supporta e infatti come tali forniscono i dati, perché l'Anagrafe non registra esplicitamente informazioni su di essi, mentre si può introdurre il collettivo Corso di laurea perché ad esso sono legate le carriere degli studenti. Nello stesso ordine di idee nella descrizione di archivi nei quali il soggetto principale è ad esempio il richiedente asilo, o il contribuente, si può introdurre rispettivamente il collettivo Figli del richiedente asilo, o il collettivo Familiari a carico del contribuente, se si vuole dare conto del fatto che l'archivio registra informazioni specifiche su questi collettivi.

Per i collettivi principali di tipo popolazione è in genere estraibile dalla documentazione una *definizione* articolata. Una definizione è una serie di *condizioni necessarie* di appartenenza al collettivo, che costituisce nel suo insieme una *condizione sufficiente* di appartenenza.

Ricordando che i collettivi principali di tipo popolazione sono sottoinsiemi delle due popolazioni delle persone e degli organismi che svolgono attività economica, o di popolazioni ad esse connesse, ci sono due possibilità relativamente al contenuto di tali definizioni.

Nel caso più semplice, la definizione enumera una serie di condizioni che congiuntamente vanno a determinare, al loro verificarsi, l'appartenenza di un elemento delle popolazioni più generali allo specifico collettivo.

Tali condizioni comprendono quindi una condizione di appartenenza ad una delle popolazioni più generali (ad esempio, per l'appartenenza al collettivo Studente una condizione è l'appartenenza al collettivo Persona) oppure ad un collettivo più ampio a sua volta contenuto in una delle popolazioni più generali, e una serie di condizioni di possesso di determinate modalità per alcune caratteristiche,

e/o una serie di condizioni che coinvolgono le relazioni possedute dall'elemento, come codominio o come dominio (si veda più avanti), in quest'ultimo caso con l'uso di quantificatori logici.

Spesso negli archivi amministrativi a differenza che nelle indagini sono principalmente le relazioni con eventi di ingresso nel collettivo che concorrono a definire i collettivi di tipo popolazione: ad esempio uno studente è una persona che ha una relazione con un evento appartenente al collettivo Immatricolazione.

Nel caso più complesso, il collettivo di tipo popolazione è definito come unione di più collettivi, ognuno dei quali ha una sua definizione, che può essere espressa come una serie di condizioni oppure essere espressa anch'essa a sua volta come unione di più collettivi. Ad esempio il collettivo Contribuente persona fisica è ottenuto come unione di più collettivi che sono comunque sottoinsieme del collettivo Persona, ciascuno dei quali è definito da condizioni specifiche o è a sua volta unione di altri collettivi. In questi casi andrebbero individuati e descritti esplicitamente i singoli collettivi che riuniti formano il collettivo principale.

Infine alcuni collettivi, in particolare quelli più generali e composti di istanze costituite da elementi naturali, come il collettivo Persona, possono non avere associata una vera e propria definizione, ma solo una o più condizioni necessarie di appartenenza, ad esempio l'esistenza di una relazione con un evento di nascita.

Talvolta anche per collettivi la cui definizione è molto complessa, ad esempio perché fa riferimento a una pluralità di norme, non è possibile in pratica arrivare a specificare un insieme di condizioni necessarie e nel loro complesso anche sufficienti, e ci si deve limitare a enumerare una serie di condizioni necessarie.

Definizione di collettivo di tipo evento: un collettivo d'interesse statistico costituito da istanze corrispondenti ad azioni, fatti o accadimenti che riguardano, o coinvolgono, elementi dei collettivi di tipo popolazione.

Nella maggior parte dei casi gli elementi dei collettivi di tipo evento non hanno esistenza indipendente, in quanto per natura dipendono dagli elementi dei collettivi di tipo popolazione cui sono legati, ma possono esistere eventi indipendenti che vengono registrati perché coinvolgono elementi dei collettivi di tipo popolazione, e sono in genere direttamente riferiti a un territorio.

Esempi di eventi dipendenti da elementi dei collettivi di tipo popolazione sono Immatricolazione, Iscrizione, Acquisizione crediti, Assunzione, Ricovero ospedaliero, Avvio rapporto di lavoro, Laurea, Licenziamento, Dimissione ospedaliera, Carriera dello studente, Rapporto di lavoro, Dichiarazione dei redditi.

Esempi di eventi indipendenti che coinvolgono elementi dei collettivi di tipo popolazione possono essere Incidente, Epidemia.

Si distinguono poi gli *eventi semplici*, legati a un solo collettivo di tipo popolazione (esempio Immatricolazione, che è legato al collettivo Studenti) dagli *eventi di tipo associativo*, legati a più collettivi di tipo popolazione (esempio Avvio rapporto di lavoro, Rapporto di lavoro, che sono entrambi legati ai collettivi Datore di lavoro e Lavoratore).

Gli eventi possono essere *istantanei* o *con durata* (ad esempio l'evento Immatricolazione è istantaneo, l'evento Avvio rapporto di lavoro è istantaneo, l'evento Rapporto di lavoro è con durata). Questa è una particolare qualità dell'evento, non però una qualità intrinseca, ma una qualità che viene definita in funzione delle finalità dell' archivio.

Assume particolare importanza ai fini della dinamica delle informazioni gestite dall'archivio, perché gli eventi che hanno una durata sono affini alle popolazioni in quanto, avendo una durata, possono cambiare nel tempo le loro caratteristiche e le relazioni che li coinvolgono e inoltre, come le popolazioni, hanno relazioni con eventi istantanei che determinano l'ingresso e l'uscita di un elemento nel collettivo.

Tenendo conto di quanto detto nel primo paragrafo sul ciclo di vita dell'informazione amministrativa, in primo luogo possono essere descritti come collettivi di tipo evento tutti gli specifici eventi, fatti, accadimenti che sono oggetto dell'attività amministrativa a supporto della quale è costituito l'archivio (ad esempio nascita, immatricolazione all'università, ricovero ospedaliero, rapporto di lavoro), inclusa l'erogazione di servizi ad essi diretti (ad esempio l'erogazione di una pensione) o l'adempimento di obblighi da essi dovuti (ad esempio il pagamento di una tassa, e la relativa dichiarazione), in secondo luogo potranno essere descritti come collettivi principali di tipo evento tutti gli altri eventi, fatti, accadimenti relativi ai soggetti dell'attività amministrativa che l'archivio comunque registra. Tra questi possono esservi eventi relativi a collettivi di tipo popolazione più ampi che includono come sottoinsieme il collettivo dei soggetti dell'attività amministrativa a supporto della quale è costituito l'archivio, è il caso dell'evento decesso per lo studente.

Ci si può chiedere in quali casi individuare un collettivo di tipo evento anziché limitarsi a definire il fatto, l'azione, o l'accadimento come una semplice caratteristica degli elementi del collettivo di tipo popolazione, o una semplice relazione che li lega ad altri elementi.

In questo, e in altri casi di dubbio, occorre tenere presente la definizione generale dei collettivi d'interesse statistico presentata all'inizio, che offre un criterio per discriminare che cosa conviene considerare come un collettivo a sé stante, e non semplicemente come una proprietà di un altro collettivo: ogni insieme di elementi osservabili e distintamente enumerabili su cui può avere potenzialmente interesse effettuare un'operazione *indipendente* di misura.

In base a questa definizione si adotterà di massima il seguente criterio guida: definire un collettivo di tipo evento, anziché una semplice caratteristica della popolazione o una relazione tra popolazioni, in tutti quei casi in cui ha senso un'operazione indipendente di misura.

Ciò avviene sempre quando l'evento ha caratteristiche sue proprie oppure si verifica che per ogni elemento di popolazione vi sono più eventi associati. Inoltre se una relazione tra collettivi è m-n, nel senso che per ogni elemento di ciascuna popolazione vi sono più elementi dell'altra popolazione associati, essa va descritta come un evento di tipo associativo, in quanto può essere oggetto di un'operazione indipendente di misura, rispetto alla misura dei collettivi legati. Analoghe considerazioni valgono per le relazioni tra più di due collettivi.

Per questo motivo l'ontologia di un archivio amministrativo si presenta come una rete di relazioni tra collettivi nella quale le relazioni sono solo di tipo 1-1 o 1-n.

Anche i collettivi di tipo evento hanno associata una *definizione*: questa in genere, a differenza della definizione dei collettivi di tipo popolazione, enuncia una condizione di riconoscibilità dell'evento espressa in termini di altri eventi, fatti, accadimenti che concorrono a determinarlo (si pensi alla definizione degli eventi nascita, decesso, negli archivi demografici), ma può accadere che, come per le popolazioni, il collettivo di tipo evento osservato da un archivio sia un sottoinsieme di un più ampio collettivo di tipo evento, o un'unione di eventi sottoinsieme di un più ampio collettivo di tipo evento.

Documentare l'ontologia di un archivio amministrativo: le caratteristiche

Analogamente alle indagini, l'archivio amministrativo raccoglie e gestisce informazioni relative a *caratteristiche osservabili, qualitative e quantitative, per ciascun elemento di ogni collettivo, di tipo popolazione o evento* (esempi: Sesso, Fatturato, Durata della degenza).

Da un punto di vista statistico, una caratteristica degli elementi di un collettivo costituisce una variabile osservabile per il collettivo. Come illustrato in seguito peraltro, ulteriori variabili possono essere costruire combinando le relazioni e le caratteristiche osservate dall'archivio.

Le caratteristiche sono di diverso *tipo*. Una prima distinzione può essere effettuata tra *variabile di classificazione*, *variabile numerica*, *data* e *altro*. Queste tipologie sono evidentemente mirate a fornire una prima indicazione sul tipo di usabilità statistica della caratteristica.

In particolare, è evidente che le caratteristiche di tipo variabile numerica sono quelle che possono essere direttamente oggetto di operazioni di aggregazione, ad esempio Numero crediti, Reddito, Spesa, mentre le caratteristiche di tipo variabile di classificazione sono quelle utilizzabili per partizionare il collettivo in classi, ad esempio Sesso, Residenza, Forma giuridica, Motivo del licenziamento.

Questa distinzione non coincide esattamente con la distinzione tra caratteristiche quantitative e qualitative: tra le caratteristiche di tipo variabile di classificazione si possono trovare caratteristiche che per loro natura sarebbero quantitative ma sono registrate suddivise in classi. Ad esempio Classe di età è di tipo variabile di classificazione, mentre Età in anni compiuti può essere considerata di tipo variabile numerica. Sono considerate di tipo variabile di classificazione anche le caratteristiche associate alla speciale classificazione che ha come modalità le due modalità sì, no, o a classificazioni simili (ad esempio con modalità sì, no, non registrato).

Le variabili numeriche assumono *valori* in uno specifico *dominio numerico*, per esse può essere definita un'unità di misura.

Le variabili di classificazione assumono *modalità* in una specifica *classificazione* ad esse associata. Variabili diverse possono usare la stessa classificazione, e una stessa variabile può essere associata a classificazioni diverse in archivi diversi o in uno stesso archivio nel corso del tempo. Ad esempio un archivio può utilizzare una stessa classificazione di stati esteri per rilevare la residenza e la cittadinanza, così come può in tempi diversi adottare classificazioni diverse per queste variabili.

Le caratteristiche di tipo data sono ovviamente elaborabili solo in modi particolari, infine possono essere catalogate come caratteristiche di tipo altro le caratteristiche del tipo Indirizzo, Cap, Numero di telefono, e altre caratteristiche simili non direttamente elaborabili.

Le caratteristiche possono avere una loro *definizione* e in teoria possono essere *obbligatorie* od *opzionali*, a seconda che ogni elemento del collettivo debba necessariamente avere associato almeno un valore numerico o una modalità per la caratteristica (è il caso più frequente), o possa non avere associato alcun valore o modalità per la caratteristica. Spesso le caratteristiche opzionali vengono in pratica trattate come obbligatorie adattando opportunamente la classificazione ad esse associata con l'aggiunta di opportune modalità corrispondenti all'assenza o alla non osservabilità della caratteristica.

Documentare l'ontologia di un archivio amministrativo: gli identificativi

Particolari caratteristiche associate a ciascun elemento del collettivo sono gli identificativi: un *identificativo* deve permettere di individuare univocamente l'elemento appartenente a un collettivo. Un identificativo può essere *semplice* o *strutturato*, quando è ottenuto a partire da diverse caratteristiche o relazioni possedute dall'elemento.

Gli identificativi possono essere utilizzati come *codici di raccordo*, per rappresentare praticamente una relazione legando ogni elemento del dominio al corrispondente elemento del codominio (si veda di seguito).

Con il nome di identificativo si intende, per gli elementi dei collettivi, quella particolare proprietà, o combinazione di proprietà, che viene espressamente attribuita ad un elemento con il preciso scopo di identificarlo distinguendolo dagli altri elementi.

Ne discende tra l'altro che un identificativo può essere direttamente utilizzato per il linkage esatto con altri archivi.

Un identificativo strutturato può essere costruito in diversi modi.

In primo luogo può esser costruito associando diverse caratteristiche dell'elemento, con l'aggiunta se necessario di numero sequenziale. In particolare, a definire il codice identificativo dell'elemento possono concorrere uno o più dei codici di raccordo che lo legano ad elementi di altri collettivi. Questo è possibile solo quando l'elemento identificato ha una relazione di dipendenza con l'elemento raccordato (si veda più avanti), cosa che generalmente avviene per gli eventi. Spesso perciò un evento contiene nel proprio identificativo il codice di raccordo con l'elemento di una popolazione, o se è un evento di tipo associativo, i codici di raccordo con gli elementi legati. Ad esempio un evento di Acquisizione crediti conterrà nell'identificativo il codice fiscale dello studente, e un evento Rapporto di lavoro conterrà nell'identificativo i codici fiscali del lavoratore e del datore di lavoro. L'identificativo di un evento di Immatricolazione, che è unico per un dato studente, coincide con l'identificativo dello studente.

In secondo luogo un identificativo strutturato può anche essere internamente costruito in modo complesso tenendo conto di una combinazione di caratteristiche atta in una certa misura a identificare un elemento (ad esempio Nome, Cognome, Indirizzo, Luogo di nascita): è il caso ad esempio del codice fiscale.

Tale combinazione di caratteristiche atta in una certa misura a identificare un elemento, pur non costituendo da sola un identificativo, può anche essere utilizzata per il linkage con altri archivi, in particolare per il linkage probabilistico, oppure a supporto del linkage esatto mediante codice identificativo. Da questo punto di vista può essere interessante misurare il potenziale di identificazione e di linkage con altri archivi di queste combinazioni di caratteristiche. Completamente diverso è il caso dell'utilizzo, in particolari contesti statistici, di una combinazione di caratteristiche per provare a ricostruire a posteriori collettivi non gestiti nell'archivio, in relazione a specifici utilizzi statistici.

Infine una combinazione di caratteristiche atta in una certa misura a identificare un elemento è utile per discriminare tra errori di tipo diverso, come illustrato nel Framework.

Un requisito di base per l'utilizzabilità statistica dell'archivio, controllabile mediante indicatori nella iperdimensione dei Metadati, è che esista per ogni collettivo un sistema di identificazione degli elementi mirante in linea di principio ad attribuire codici identificativi unici, non duplicati e stabili nel tempo. Il soddisfacimento di tale requisito non impedisce naturalmente che si generino

errori sugli identificativi nell'aggiornamento dell'archivio, non sempre peraltro individuabili in pratica, e comunque da controllare direttamente sui dati.

Documentare l'ontologia di un archivio amministrativo: le relazioni

Infine un archivio amministrativo raccoglie informazioni sulle *relazioni che legano gli elementi di diversi collettivi*. Le relazioni che legano gli elementi di un collettivo ad elementi degli altri collettivi costituiscono a tutti gli effetti una proprietà degli elementi, così come le caratteristiche osservate, e sono importanti da un punto di vista statistico perché utilizzabili per costruire ulteriori caratteristiche, e quindi variabili, indirettamente attribuibili agli elementi del collettivo.

Le relazioni possono essere *semplici* o di *dipendenza*: sono di dipendenza quando un elemento dipende per la sua esistenza dalla relazione che lo lega all'altro elemento (o anche entrambi dipendono uno dall'altro per la loro esistenza, se la relazione è 1-1), e quindi rimane legato per tutta la sua durata a tale altro elemento, è semplice se per entrambi gli elementi legati la loro esistenza è indipendente dal fatto che siano legati.

In particolare sono spesso di dipendenza le relazioni dei collettivi di tipo evento da collettivi di tipo popolazione. Infatti un collettivo di tipo evento, istantaneo o con durata, proprio perché costituito da fatti che accadono a elementi di uno o più collettivi di tipo popolazione, ha nella maggior parte dei casi una relazione speciale di dipendenza con tali collettivi, precisamente dipende da un collettivo di tipo popolazione se è semplice, da più collettivi di tipo popolazione se è di tipo associativo (esempi: Immatricolazione dipende da Studente, Acquisizione crediti dipende da Studente, Rapporto di lavoro dipende da Lavoratore e da Datore di lavoro).

Importanti relazioni sono quelle che legano un collettivo di tipo popolazione o evento con durata ai collettivi di tipo evento istantaneo che determinano gli ingressi e le uscite dei suoi elementi (esempi: Nascita inizia Persona, Avvio rapporto di lavoro inizia Rapporto di lavoro). Altre relazioni, semplici o di dipendenza, possono legare tra loro collettivi di tipo popolazione o evento.

Esempi di relazioni tra collettivi sono presentati nelle pagine 114-118.

Per come viene definita l'ontologia di un archivio, le relazioni sono *sempre di tipo funzionale, vale a dire sono relazioni 1-n*, per le quali si verifica sempre che per ciascuno degli elementi di uno dei due collettivi coinvolto nella relazione esiste un solo elemento legato dell'altro collettivo**,** *oppure sono relazioni 1-1*.

Ad esempio c'è un unico conduttore per ogni azienda agricola, un'unica impresa di appartenenza per ogni unità locale, un unico studente per ogni acquisizione crediti, un unico lavoratore e un unico datore di lavoro per ogni evento di avvio di rapporto di lavoro e anche per ogni rapporto di lavoro.

Infatti *una relazione m-n, non funzionale, tra elementi di due o più collettivi costituisce sempre un ulteriore collettivo*, precisamente *un evento di tipo associativo* legato funzionalmente a due o più collettivi, come nel caso degli eventi Avvio rapporto di lavoro, Rapporto di lavoro. Gli elementi di una relazione m-n possono infatti essere in numero qualsiasi, indipendentemente dalla numerosità dei collettivi legati, sono quindi potenzialmente oggetto di operazioni di osservazione e misura indipendenti dalle operazioni di misura dei collettivi legati.

Esempi di eventi di tipo associativo e dei loro legami con altri collettivi sono presentati alle pagine 117 e 118.

In particolare le relazioni in una versione di archivio possono essere_anche 1-1, specialmente per alcuni tipi di eventi di ingresso, come nel caso della relazione tra immatricolazione e studente o della relazione tra avvio rapporto di lavoro e rapporto di lavoro.

Quando una relazione è 1-n si può distinguere un *collettivo dominio* da un *collettivo codominio*. In particolare in tutte le relazioni di dipendenza che legano l'evento di tipo associativo ai collettivi di riferimento l'evento di tipo associativo ha il ruolo di dominio.

Ad esempio: nella relazione tra conduttori e aziende agricole il collettivo delle aziende agricole è il dominio e il collettivo dei conduttori il codominio, nella relazione tra imprese e unità territoriali il collettivo delle unità territoriali è il dominio e il collettivo delle imprese il codominio, nella relazione tra eventi di acquisizione crediti e studenti il collettivo degli eventi di acquisizione crediti è il dominio e il collettivo degli studenti il codominio, nella relazione tra il collettivo degli eventi di avvio rapporto di lavoro e il collettivo dei lavoratori il collettivo degli eventi di avvio rapporto di lavoro è il dominio e il collettivo dei lavoratori il codominio, nella relazione tra lo stesso collettivo degli eventi di avvio rapporto di lavoro e il collettivo dei datori di lavoro il collettivo degli eventi di avvio rapporto di lavoro è il dominio e il collettivo dei datori di lavoro il codominio, analogamente, nella relazione tra il collettivo degli eventi (con durata) rapporto di lavoro e il collettivo dei lavoratori, il collettivo degli eventi rapporto di lavoro è il dominio e il collettivo dei lavoratori il codominio, nella relazione tra lo stesso collettivo degli eventi rapporto di lavoro e il collettivo dei datori di lavoro il collettivo degli eventi rapporto di lavoro è il dominio e il collettivo dei datori di lavoro il codominio.

Per rappresentare in pratica il legame esistente tra un elemento del collettivo dominio e un elemento del collettivo codominio si utilizza l'identificativo di quest'ultimo come *codice di raccordo*.

Documentare l'ontologia di un archivio amministrativo: costruire nuove caratteristiche

Poiché le relazioni sono 1-n oppure 1-1, agli elementi di un collettivo possono essere attribuite nuove caratteristiche costruite sfruttando i suoi legami con altri collettivi. Si veda in proposito il disegno a pagina 119. Ciò spiega perché è interessante documentare le relazioni tra collettivi.

Nel caso di *relazioni 1-n* la costruzione di nuove caratteristiche attribuibili agli elementi di un collettivo coinvolto avviene in modo diverso a seconda che il collettivo ricopra il ruolo di dominio o di codominio nella relazione. Precisamente, per ogni collettivo:

- data una relazione che lo coinvolge come dominio, agli elementi del collettivo possono essere indirettamente attribuite le caratteristiche dell'elemento legato, come *caratteristiche indirettamente possedute*; esempi: alle unità territoriali si può attribuire la forma giuridica dell'impresa di riferimento, agli eventi di avvio rapporto di lavoro si può attribuire la qualifica del corrispondente lavoratore e il tipo del corrispondente datore di lavoro, analogamente per gli eventi rapporto di lavoro;
- data una relazione che lo coinvolge come codominio, per gli elementi del collettivo possono essere costruite nuove caratteristiche che tengano conto della pluralità degli elementi legati, cioè *caratteristiche ottenute per quantificazione sulle relazioni*, ad esempio esistenza di almeno un elemento legato, o di almeno un elemento legato con una certa caratteristica, numero degli elementi legati, eccetera; esempi: allo studente può essere attribuita la caratteristica Possesso di almeno un evento di acquisizione crediti, al lavoratore la

caratteristica Numero di rapporti di lavoro a tempo determinato, al datore di lavoro la caratteristica Numero di rapporti di lavoro.

Nel caso di *relazioni 1-1* agli elementi dei collettivi sono sempre attribuibili le caratteristiche dell'elemento legato come *caratteristiche indirettamente possedute*. Esempi: allo studente può essere attribuito il tipo di immatricolazione, al rapporto di lavoro può essere attribuita la data di avvio, all'avvio del rapporto di lavoro il tipo del rapporto di lavoro avviato.

Si noti che la costruzione di nuove caratteristiche sfruttando tutta la rete delle relazioni può risultare anche molto complessa: infatti agli elementi del collettivo possono essere indirettamente attribuite, come nuove caratteristiche, non sole le caratteristiche che appartengono direttamente all'elemento legato, ma anche le caratteristiche che appartengono all'elemento legato in quanto indirettamente possedute od ottenute per quantificazione.

Più in generale, *le catene di relazioni tra collettivi definiscono percorsi lungo i quali possono essere costruite e propagate nuove caratteristiche*, utilizzando se necessario specifici operatori per la quantificazione. Si veda in proposito il disegno a pagina 120.

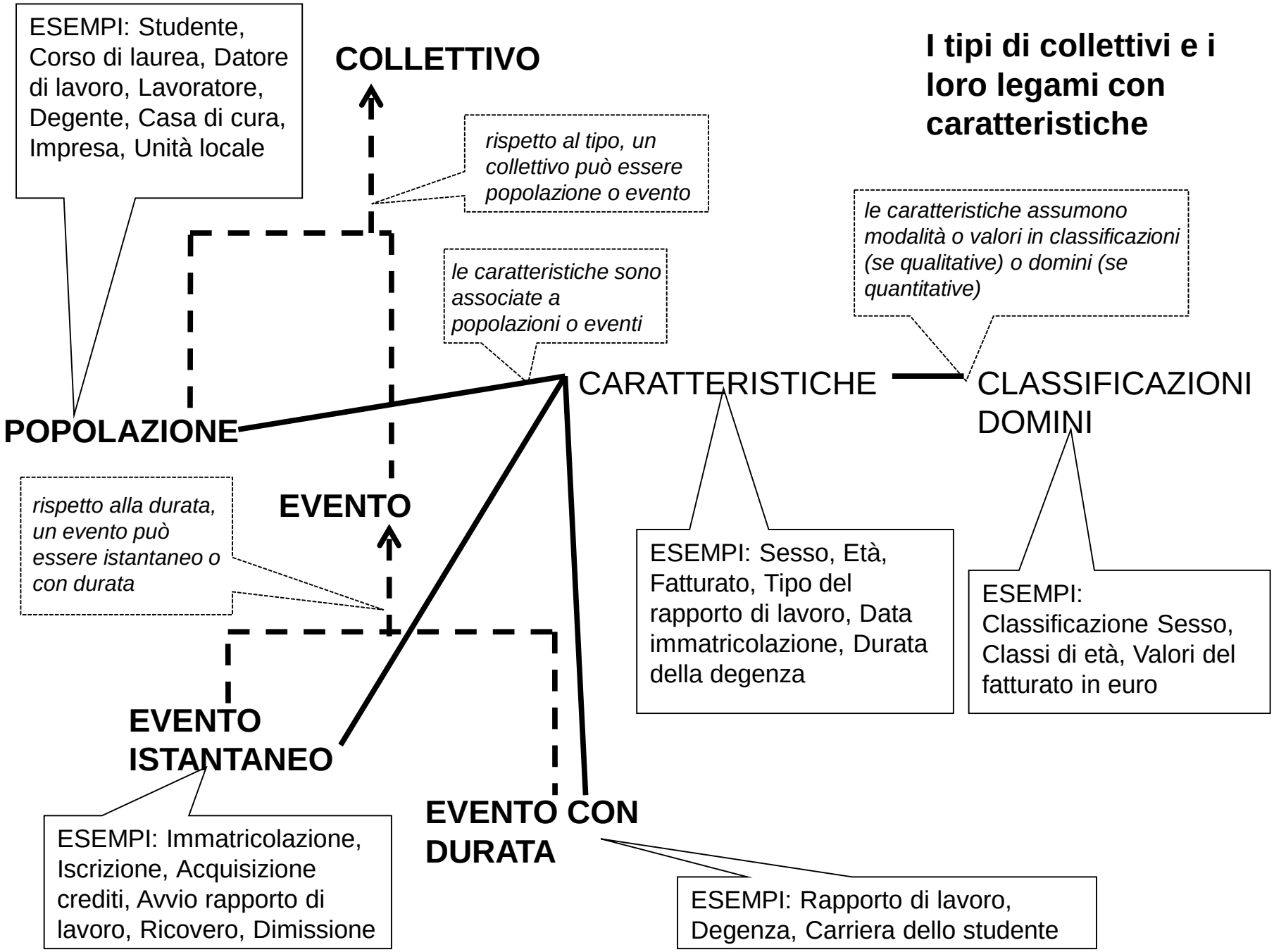
Questa possibilità di costruzione di nuove caratteristiche attribuibili agli elementi dei collettivi è particolarmente importante per gli archivi amministrativi nei quali si può supporre che, rispetto alle indagini, esista un maggior numero di relazioni tra i collettivi per via del maggior numero e della maggiore importanza dei collettivi di tipo evento, connessi da molteplici relazioni ai collettivi di tipo popolazione e anche tra loro. Le nuove caratteristiche attribuite agli elementi di un collettivo possono essere oggetto di interesse statistico diretto, o essere utili ai fini del controllo di qualità.

E' comune per esempio negli archivi amministrativi l'esigenza del controllo dei vincoli di obbligatorietà o incompatibilità che possono esistere tra eventi riguardanti uno stesso elemento di un collettivo di tipo popolazione: ad esempio l'esistenza di un evento del tipo acquisizione crediti legato ad uno studente implica obbligatoriamente l'esistenza di almeno un evento di iscrizione a corso di laurea per lo stesso studente.

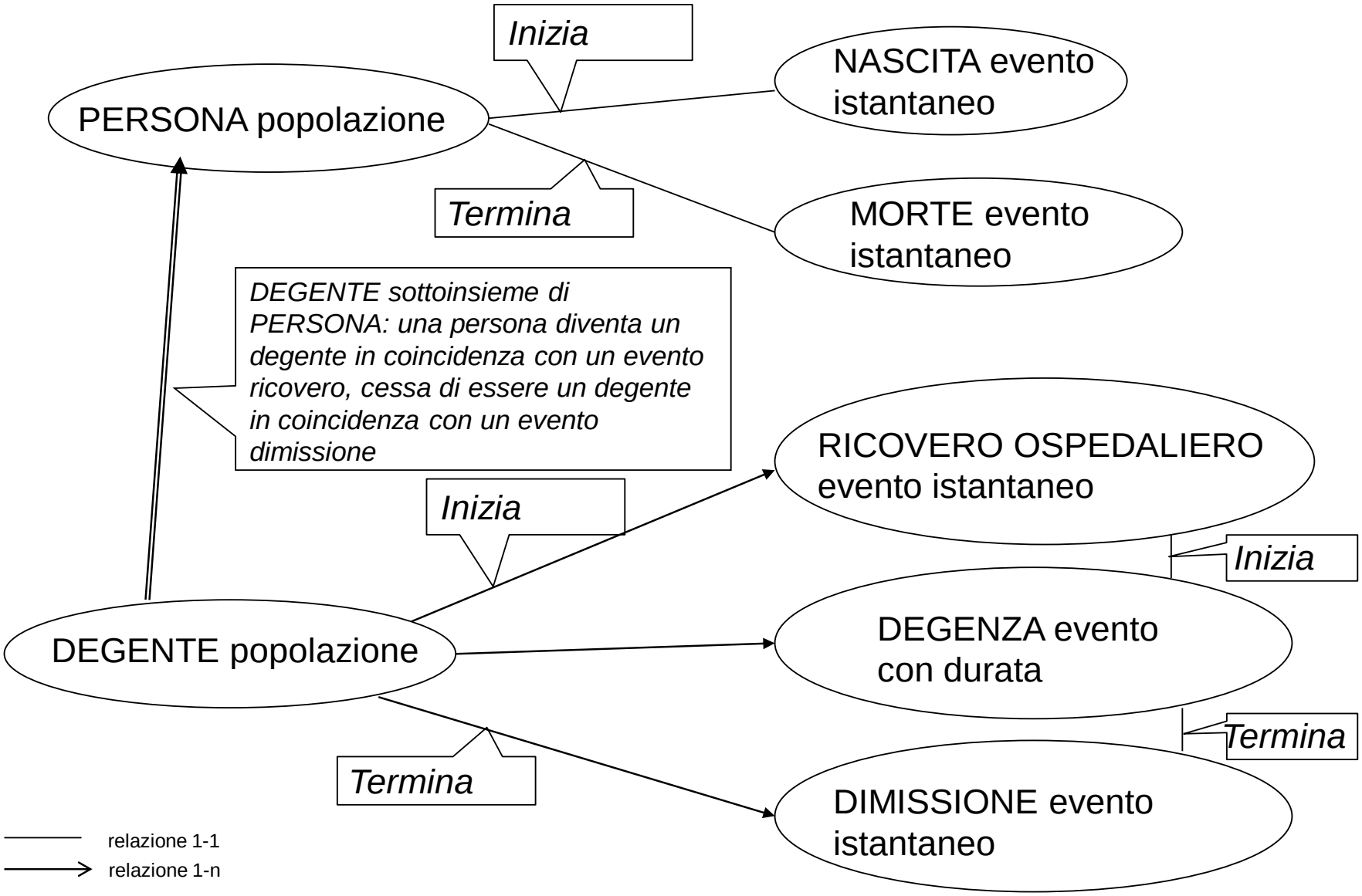
Questi vincoli sono formalizzabili come vincoli tra un evento riferito a un elemento e la caratteristica dello stesso elemento ottenuta considerando l'esistenza di almeno una relazione con l'altro evento. Con riferimento all'esempio, c'è un vincolo di obbligatorietà che lega un evento di acquisizione crediti riferito ad uno studente alla caratteristica dello stesso studente consistente nell'esistenza per esso di almeno un evento di iscrizione a un corso di laurea.

Le relazioni possono avere una propria *definizione*. Inoltre per ciascuno dei collettivi coinvolti possono essere *obbligatorie* od *opzionali*, a seconda che ogni elemento del collettivo debba necessariamente avere associata almeno una relazione o meno. Per ciascuno dei collettivi coinvolti possono essere inoltre definiti un numero minimo o massimo di elementi associabili a ciascun elemento del collettivo. Tutte queste peculiarità della relazione possono essere specificate nella sua descrizione, se ritenute importanti.

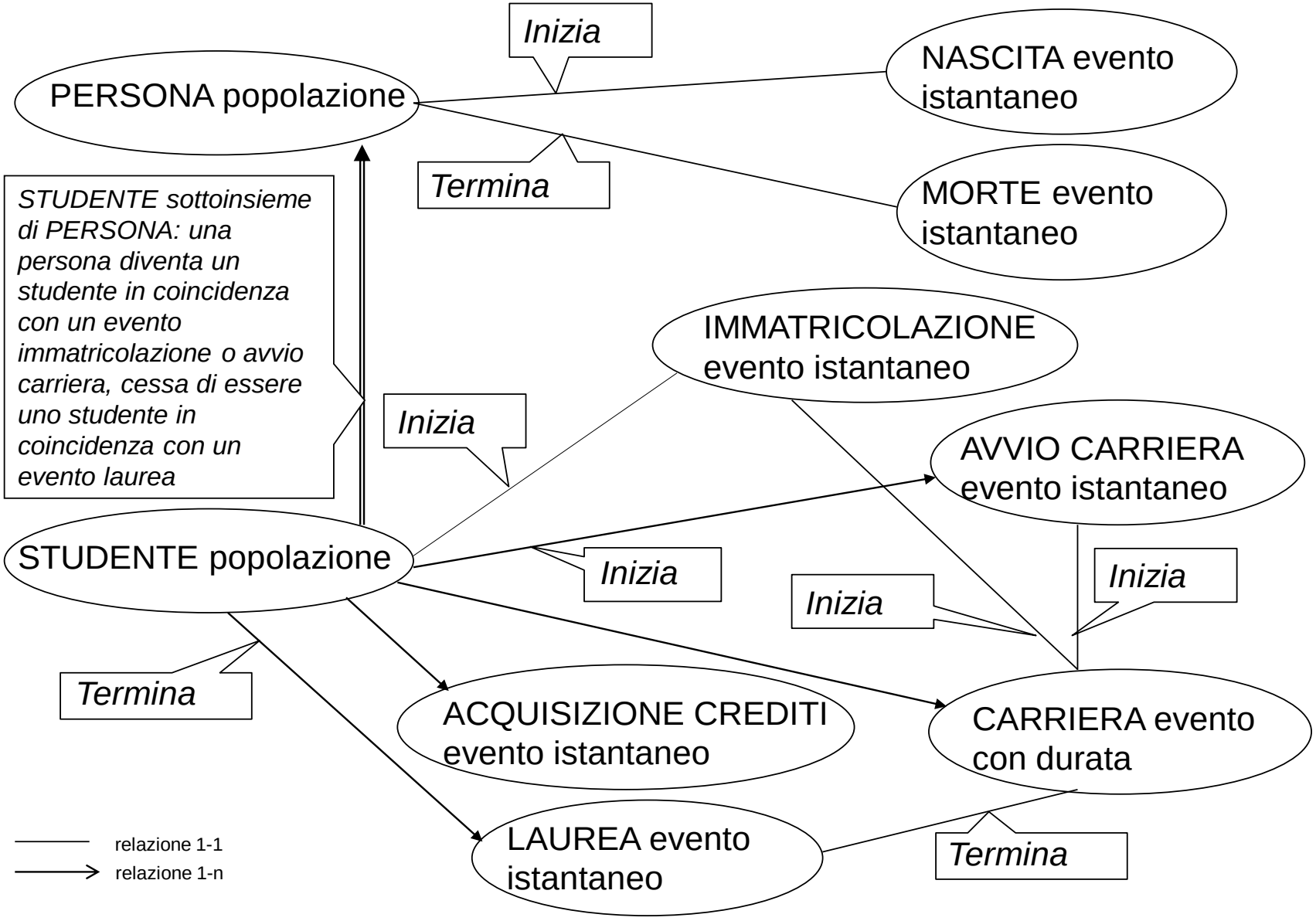
I tipi di collettivi e i loro legami con caratteristiche



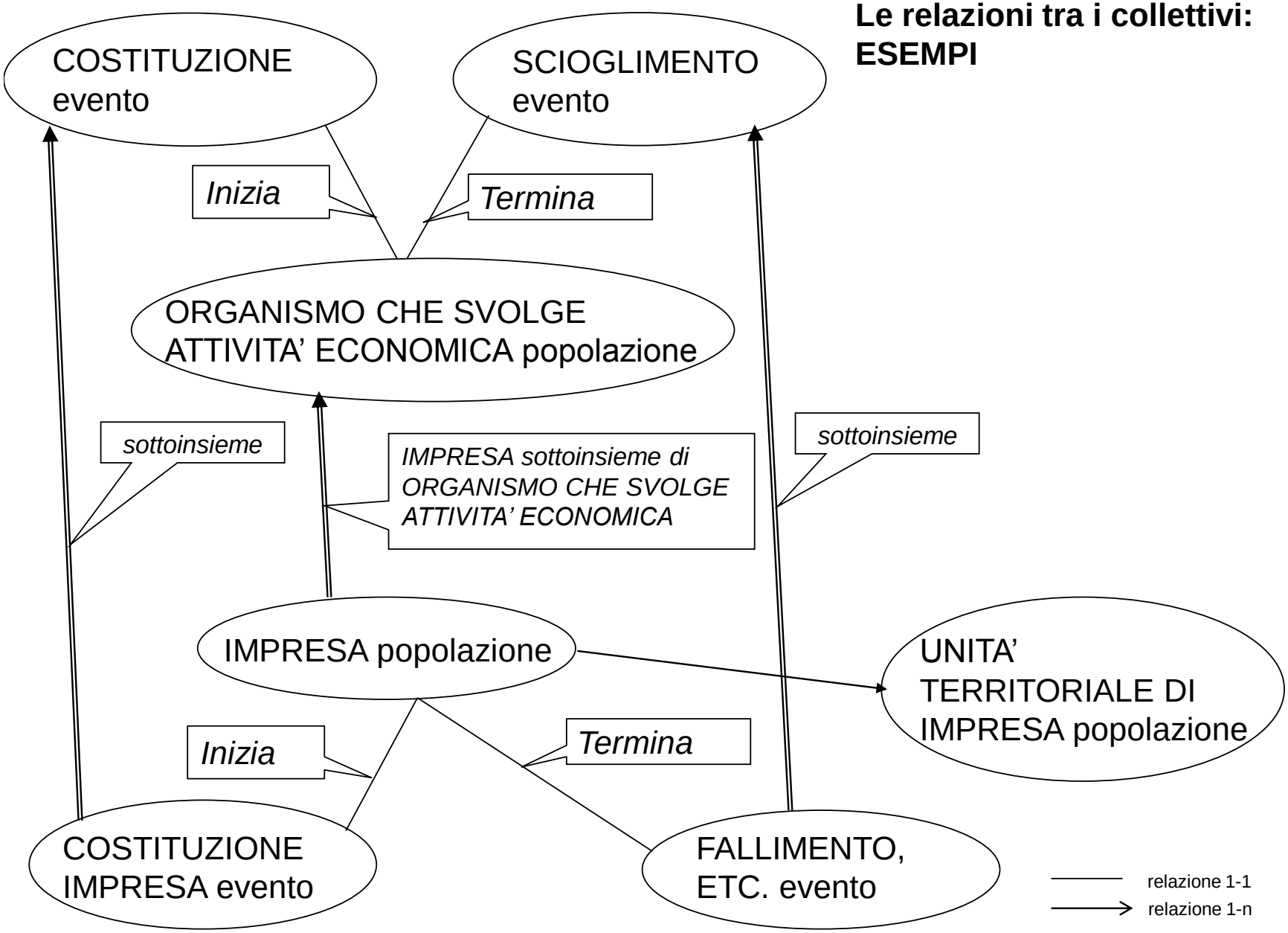
Le relazioni tra i collettivi: ESEMPI



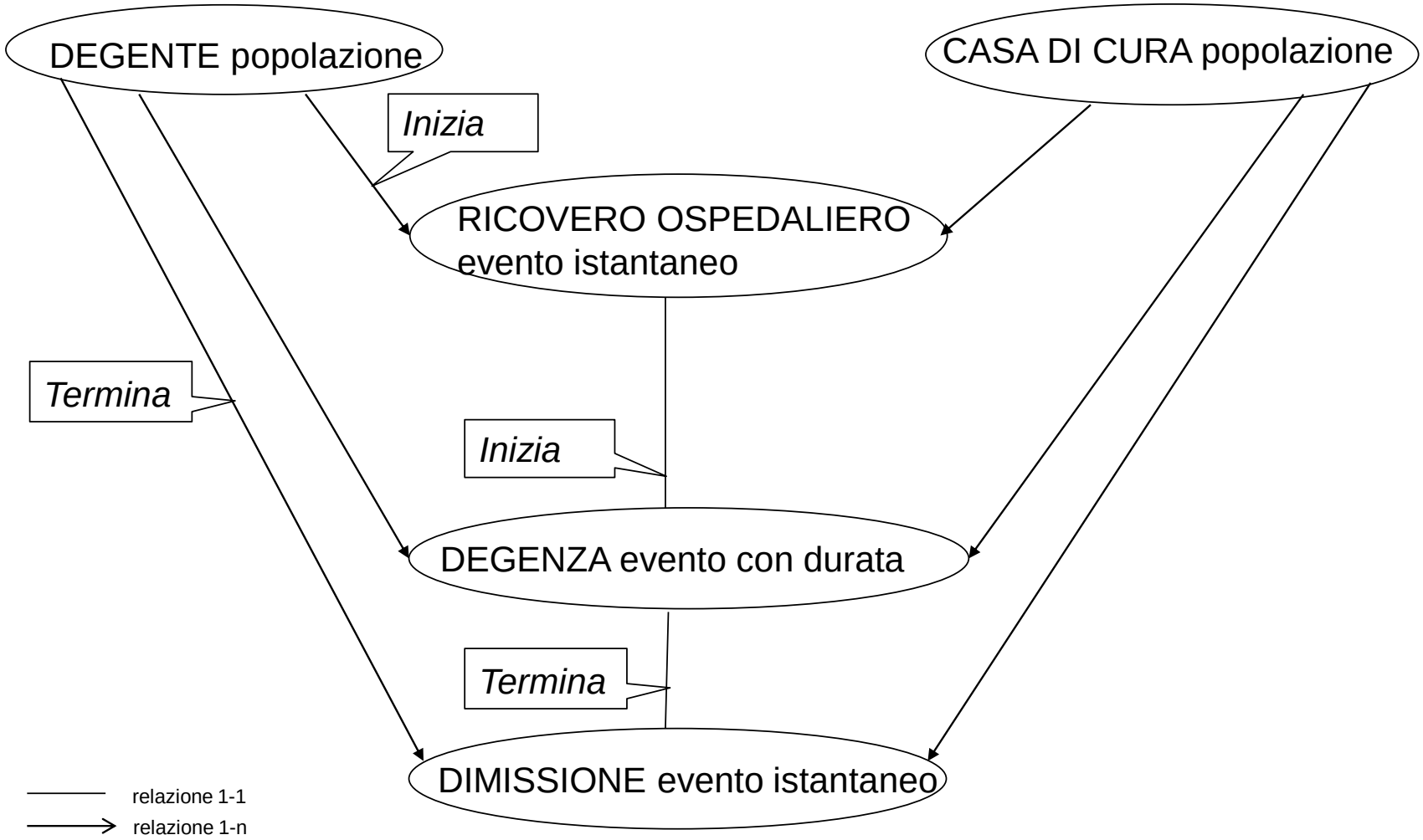
Le relazioni tra i collettivi: ESEMPI



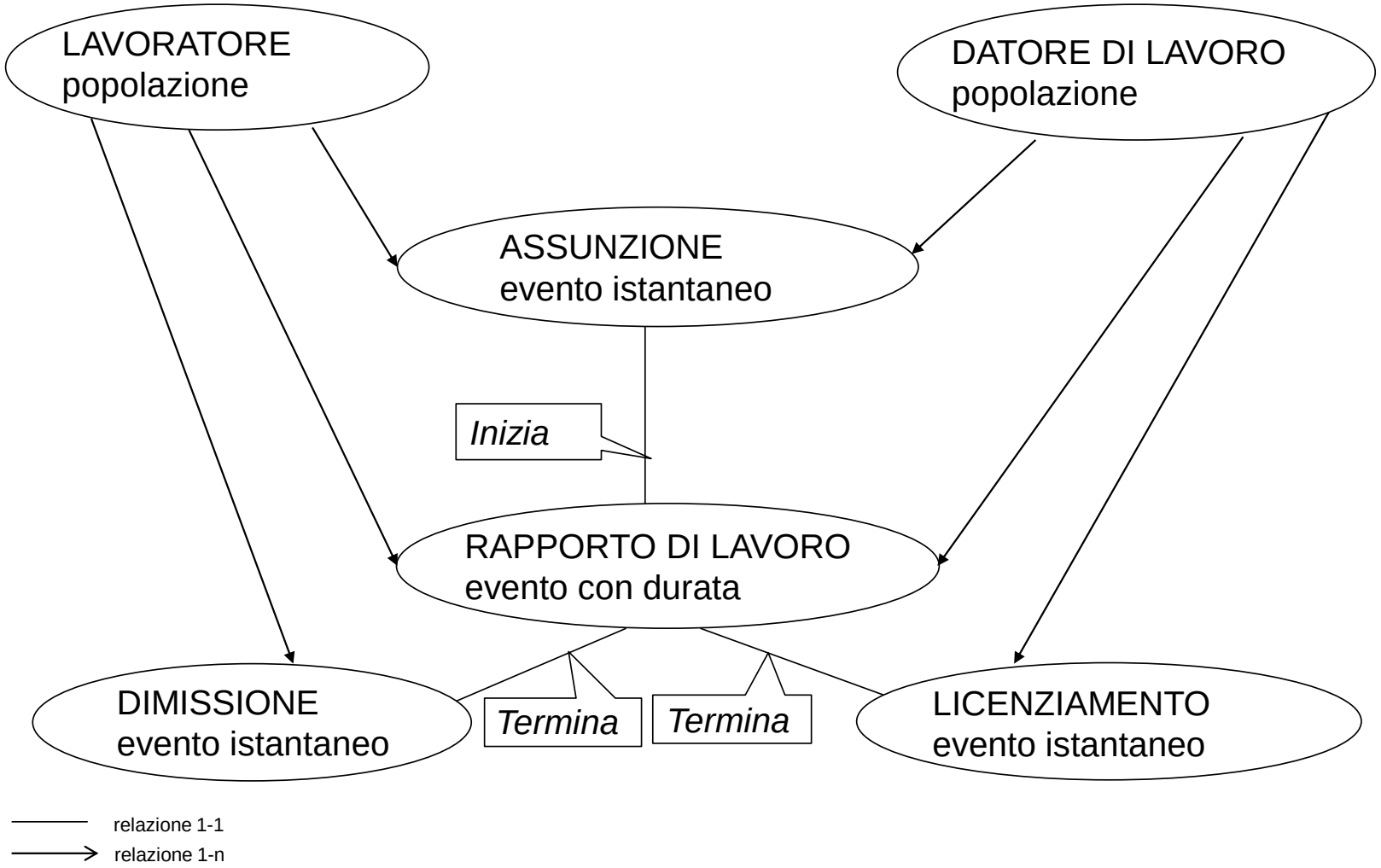
Le relazioni tra i collettivi: ESEMPI



Le relazioni tra i collettivi: ESEMPI di eventi di tipo associativo, cioè legati a più popolazioni *sono collettivi corrispondenti a relazioni m-n*

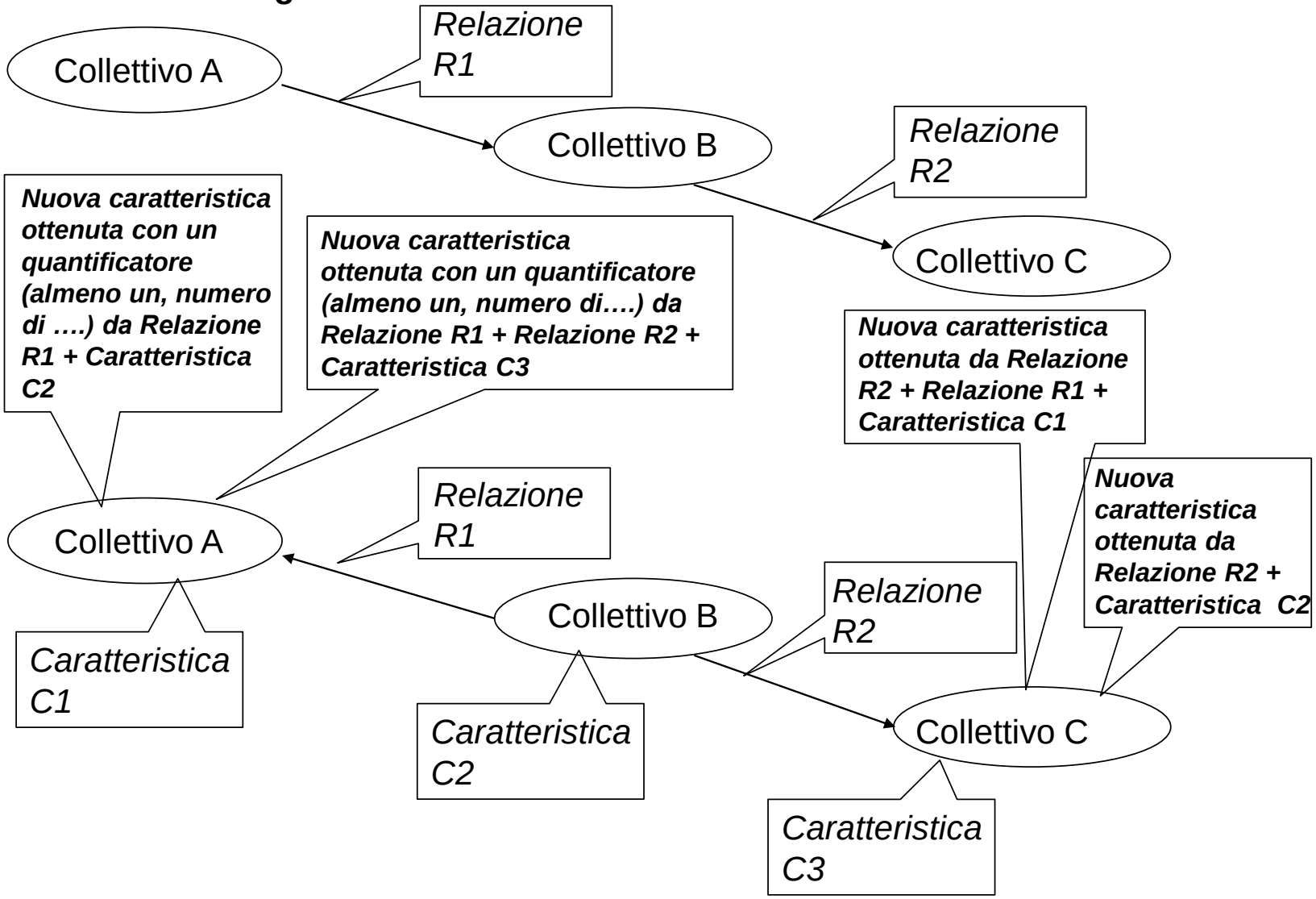


Le relazioni tra i collettivi: ESEMPI di eventi di tipo associativo, cioè legati a più popolazioni *sono collettivi corrispondenti a relazioni m-n*



Nuove caratteristiche dei collettivi possono essere costruite utilizzando le relazioni che li legano

Nuove caratteristiche dei collettivi possono essere costruite utilizzando le relazioni che li legano



Documentare l'ontologia di un archivio amministrativo: l'archivio come rete di relazioni tra collettivi

In sintesi quindi gli oggetti che descrivono il contenuto informativo di un archivio amministrativo ai fini del suo utilizzo statistico possono essere dei seguenti tipi:

Collettivi di tipo popolazione (esempi Studente, Corso di laurea, Datore di lavoro, Lavoratore, Degente, Casa di cura, Impresa, Unità locale) o di tipo evento, istantaneo (esempi Immatricolazione, Iscrizione, Acquisizione crediti, Avvio rapporto di lavoro, Ricovero, Dimissione), o con durata (esempi Rapporto di lavoro, Degenza).

In base alle definizioni precedenti, nel descrivere l'ontologia di un archivio amministrativo si adotterà di massima il seguente criterio guida: definire un collettivo di tipo evento, anziché una semplice caratteristica della popolazione, in tutti quei casi in cui ha senso un'operazione indipendente di misura, perché l'evento ha caratteristiche sue proprie oppure si verifica che per ogni elemento di popolazione vi sono più eventi associati, definire inoltre un evento di tipo associativo per tutte le relazioni m-n tra due collettivi, o che legano più di due collettivi.

Caratteristiche dei collettivi di tipo popolazione o evento (esempi Sesso, Età, Fatturato, Tipo del rapporto di lavoro, Data immatricolazione, Durata della degenza), possedute dagli elementi del collettivo, tra le quali assumono ruoli speciali gli identificativi. Ulteriori caratteristiche possono essere derivate utilizzando relazioni con altri collettivi, queste possono essere possedute indirettamente oppure ottenute per quantificazione..

Relazioni tra collettivi di tipo popolazione o evento (esempi Persona conduce Azienda agricola, Unità funzionale appartiene Impresa, Nascita inizia Persona, Immatricolazione inizia Studente, Acquisizione crediti riguarda Studente, Avvio rapporto di lavoro riguarda Lavoratore, Avvio rapporto di lavoro riguarda Datore di lavoro, Avvio rapporto di lavoro inizia Rapporto di lavoro, Rapporto di lavoro riguarda Lavoratore, Rapporto di lavoro riguarda Datore di lavoro), rilevanti da un punto di vista statistico perché consentono di costruire e attribuire agli elementi dei collettivi caratteristiche possedute indirettamente o per quantificazione.

Nel caso più generale l'ontologia di un archivio si presenta come *una rete di collettivi principali, di tipo popolazione o evento, connessi da relazioni 1-1 o 1-n, che possono essere semplici o di dipendenza, nella quale ciascun collettivo ha la propria definizione e le proprie caratteristiche di pertinenza, anch'esse con la propria definizione e, se qualitative, le proprie classificazioni associate, con le relative modalità*.

Documentare l'ontologia di un archivio amministrativo: i sottoinsiemi dei collettivi principali

Oltre a ciò, *all'interno di ogni collettivo principale possono essere enucleati dei sottoinsiemi*, vale a dire altri collettivi che comprendono solo alcuni elementi del collettivo dato e sono quindi insiemisticamente contenuti all'interno di esso.

Per via di questa relazione di contenimento gli elementi di un collettivo sottoinsieme hanno quindi automaticamente associate tutte le caratteristiche e le relazioni tipiche degli elementi del collettivo nel quale sono contenuti: si dice che ereditano le proprietà del collettivo in cui sono contenuti. In più possono poi avere associate caratteristiche e relazioni che sono loro peculiari.

Ad esempio il collettivo principale dell'Anagrafe degli studenti universitari, Studente, ha come proprietà Sesso, Luogo di residenza, la relazione con l'evento Nascita ed eventualmente la relazione con l'evento Decesso, e inoltre la relazione con l'evento Immatricolazione e la caratteristica Tipo impegno; il collettivo Studente part-time che è contenuto nel collettivo Studente ne eredita tutte le proprietà, in più può avere associata la caratteristica Data inizio part-time.

Vale la pena di osservare che dato che perlopiù i collettivi principali di un archivio sono anch'essi nella maggior parte dei casi per definizione contenuti in collettivi più ampi alcune delle loro proprietà sono ereditate: ad esempio Studente è contenuto nel collettivo Persona e da esso eredita Sesso, Luogo di residenza, la relazione con l'evento Nascita e l'eventuale relazione con l'evento Decesso.

Nella definizione dell'ontologia di un archivio questa situazione si documenta creando un collettivo corrispondente al collettivo sottoinsieme e legandolo con una *relazione di sottoinsieme* al collettivo che lo contiene, si dice che il collettivo sottoinsieme ha il ruolo di figlio nella relazione di sottoinsieme, mentre il collettivo che lo contiene ha il ruolo di padre.

A partire da un collettivo si possono avere catene di relazioni di sottoinsieme tra collettivi l'uno contenuto nell'altro. Ad esempio a partire dal collettivo Studente posso definire una catena di relazioni di sottoinsieme che lega in successione i collettivi Studente part-time, Studente part time maschio, e un'altra catena che lega in successione i collettivi Studente con carriera aperta, Studente con carriera aperta maschio. Ogni collettivo nella catena ha un padre diretto e una serie di padri indiretti.

I sottoinsiemi dei collettivi principali dell'archivio hanno associate proprie *definizioni*, ottenute da quella del collettivo che è loro padre diretto mediante l'aggiunta di ulteriori condizioni di appartenenza, sempre definite in termini di caratteristiche o relazioni, anche con eventi, possedute dall'elemento del collettivo, condizioni che congiuntamente vanno a determinare, al loro verificarsi, l'appartenenza di un elemento del collettivo principale dell'archivio al collettivo sottoinsieme.

Un gruppo di collettivi sottoinsieme che ricoprono completamente il collettivo che li contiene senza intersecarsi, in modo che ogni elemento del collettivo che li contiene appartiene a uno e uno solo di essi, definisce una *partizione* del collettivo che li contiene.

Nella definizione dell'ontologia di un archivio questa situazione si documenta creando un collettivo corrispondente a ciascuno dei collettivi sottoinsieme che formano la partizione e legando simultaneamente tutti questi collettivi con una *relazione di partizione* al collettivo che li contiene, si dice che i collettivi sottoinsieme hanno il ruolo di figlio nella relazione di partizione, mentre il collettivo che li contiene ha il ruolo di padre.

A partire da un collettivo principale dell'archivio si possono definire catene di relazioni di partizione: ad esempio a partire dal collettivo Studente posso definire una partizione per sesso che lega come figli i collettivi Studente maschio e Studente femmina, e poi per il collettivo Studente maschio posso definire una partizione per tipo impegno che lega come figli i collettivi Studente maschio a tempo pieno e Studente maschio part-time.

In questo esempio è interessante notare che, qualora abbia interesse definire la stessa partizione per tipo impegno sia per gli studenti maschi che per gli studenti femmine, dal punto di vista della chiarezza documentativa può convenire scegliere di definire un'unica partizione per sesso e per tipo impegno, che lega il collettivo Studente nel ruolo di padre ai collettivi Studente maschio a tempo pieno, Studente maschio part-time, Studente femmina a tempo pieno, Studente femmina part-time, nel ruolo di figli.

In generale a partire da un collettivo principale dell'archivio, di tipo popolazione o evento, posso definire diverse catene di relazioni di sottoinsieme e/o di partizione, con la possibilità, nei casi più complessi, di fare scelte documentative diverse che dovranno essere guidate da considerazioni di chiarezza per l'utente finale della documentazione. Come risultato *si potrà avere, a partire da un dato collettivo principale dell'archivio, un albero di relazioni di sottoinsieme e/o partizione più o meno complesso che lega tra loro i diversi collettivi contenuti nel collettivo principale***.**

E' possibile definire un collettivo che sia sottoinsieme contemporaneamente di più collettivi. Ad esempio data la catena che lega in successione i collettivi Studente, Studente part-time, e Studente part time maschio, e l'altra catena che lega in successione i collettivi Studente, Studente con carriera aperta, Studente con carriera aperta maschio, è possibile introdurre un ulteriore collettivo Studente part-time, con carriera aperta, e maschio che è simultaneamente sottoinsieme di Studente part time maschio e Studente con carriera aperta maschio.

Ci sono però dei vincoli. Il primo è che ovviamente un collettivo non può risultare figlio diretto o indiretto di due collettivi che sono a loro volta figli in una relazione di partizione, dato che questi sono per definizione mutuamente esclusivi: in pratica ad esempio nessun collettivo può essere figlio simultaneamente dei due collettivi Studente maschio e Studente femmina. Il secondo è che tutti i collettivi devono risultare comunque figli diretti o indiretti di un unico collettivo padre: in pratica ad esempio nessun collettivo può essere simultaneamente figlio dei due collettivi Lavoratore e Datore di lavoro.

Ci si può chiedere in generale cosa determina, o suggerisce, la necessità di enucleare collettivi sottoinsieme e specificare relazioni di sottoinsieme o partizione.

Dato un qualsiasi collettivo di partenza, *è necessario enucleare un collettivo sottoinsieme in esso contenuto ogni volta che esistono caratteristiche o relazioni che vengono osservate per una parte soltanto degli elementi del collettivo di partenza*, individuata in base ad una condizione.

Ad esempio se la caratteristica Data inizio part-time è osservata solo per lo studente che ha ottenuto il part-time, verrà creato il collettivo Studente part-time in relazione nel ruolo di figlio con il collettivo di partenza Studente.

Se una o più caratteristiche o relazioni vengono osservate da più collettivi, può essere consigliabile creare un collettivo unione al quale attribuire tali caratteristiche o relazioni condivise e al quale legare i collettivi che le condividono mediante relazioni di sottoinsieme o partizione (se sono mutuamente esclusivi) nel ruolo di figli.

Ad esempio se la caratteristica Data del matrimonio è osservata sia per le persone coniugate che per quelle vedove, si potrà creare il collettivo Persona coniugata o vedova al quale attribuire tale caratteristica, legato come padre da una relazione di partizione ai due collettivi Persona coniugata e Persona vedova, che continuano ad aver associata la caratteristica per ereditarietà.

Le precedenti sono le circostanze principali nelle quali vengono introdotti nuovi collettivi legati a quelli esistenti da relazioni di sottoinsieme o partizione. Oltre a ciò si hanno altre due circostanze che possono suggerire la specifica di nuove relazioni di sottoinsieme o partizione.

Dati due o più collettivi, l'analisi accurata delle loro definizioni può suggerire che tra di loro esistono relazioni di sottoinsieme o partizione: come conseguenza di ciò, il collettivo o i collettivi nella condizione di figlio si troveranno esplicitamente associate per ereditarietà le caratteristiche e relazioni associate al collettivo padre.

Ad esempio un'analisi più accurata delle definizioni può suggerire che i collettivi Persona coniugata e Persona vedova siano entrambi sottoinsiemi del collettivo Persona che ha contratto almeno un matrimonio.

Infine si è visto che un collettivo può essere definito come unione di più collettivi, ognuno dei quali ha una sua definizione, che può essere espressa come una serie di condizioni oppure essere espressa anch'essa a sua volta come unione di più collettivi. *Quando un collettivo ha una tale definizione complessa, si può scegliere di enucleare i sottoinsiemi da cui è costituito, strutturando la definizione come un albero di relazioni di partizione*.

Ad esempio il collettivo Contribuente persona fisica è ottenuto come unione di più collettivi che sono comunque sottoinsieme del collettivo Persona, ciascuno dei quali è definito da condizioni specifiche o è a sua volta unione di altri collettivi. Per l'utente finale della documentazione in questo caso è utile trovare esplicitato il contenuto della definizione come unione di sottoinsiemi, piuttosto che in forma testuale.

A questo scopo si analizzerà la definizione per estrarre tutti i sottoinsiemi del collettivo Contribuente in essa impliciti, cercando per maggior chiarezza di definire sottoinsiemi mutuamente esclusivi, in modo da associare al collettivo Contribuente una relazione di partizione che lo leghi nel ruolo di padre ai suoi collettivi componenti nel ruolo di figli. La stessa analisi può essere effettuata su tali collettivi componenti, e se possibile proseguita ulteriormente sui collettivi loro componenti, cercando di ottenere un albero di relazioni di partizione il più possibile ramificato, che evidenzi al meglio la struttura della definizione.

Gli oggetti nell'ontologia dell'archivio e i relativi tipi di enunciati

Come discusso nella parte introduttiva del Framework, ad ogni oggetto di un archivio amministrativo può essere associato un tipo di enunciato di appartenenza.

In sintesi:

I collettivi di tipo popolazione o di tipo evento sono concettualizzati come insiemi di elementi, e a ciascuno di loro corrisponde un tipo di enunciato di appartenenza riferibile ad un solo elemento, ad esempio Studente (x), Ricovero (x), Degenza (x).

Le relazioni tra collettivi sono concettualizzate come insiemi di coppie. Quindi anche ad ogni relazione tra collettivi di tipo popolazione o evento corrisponde un tipo di enunciato di appartenenza, che è però riferito non ad un singolo elemento, ma ad una coppia di elementi.

Ad esempio Azienda agricola è condotta da Conduttore (x, y), Unità locale appartiene Impresa (x, y), Immatricolazione inizia Studente (x, y), Acquisizione crediti riguarda Studente (x, y), Avvio rapporto di lavoro riguarda Lavoratore (x, y), Avvio rapporto di lavoro riguarda Datore di lavoro (x, y), Avvio rapporto di lavoro inizia Rapporto di lavoro (x, y), Rapporto di lavoro riguarda Lavoratore (x, y), Rapporto di lavoro riguarda Datore di lavoro (x, y).

Per ciò che riguarda *le caratteristiche dei collettivi di tipo popolazione o evento*, si assume anzitutto che siano definite *collezioni predefinite di elementi che costituiscono insieme di stati assumibili per tali caratteristiche*. Queste collezioni di elementi corrispondono alle *classificazioni per le caratteristiche qualitative*, e ai *domini per le caratteristiche quantitative*.

Le classificazioni, o i domini, sono concettualizzati come insiemi di elementi, ai quali corrisponde un tipo di enunciato di appartenenza riferibile ad un solo elemento, ad esempio Classificazione del sesso (m), Gruppi ATECO 2007 (m), Classi di età quinquennali (m), Ammontare della spesa per degenza (v). A differenza dei collettivi di tipo popolazione o evento, le classificazioni e i domini sono costituiti di elementi predefiniti, non oggetto di osservazione.

Quindi ad esempio in Classificazione del sesso (m), m può assumere le modalità prefissate {maschio, femmina}, in Gruppi ATECO 2007 (m), m può assumere le modalità prefissate {Coltivazione di colture agricoli non permanenti, ….Siderurgia}, in Classi di età quinquennali (m), m può assumere le modalità prefissate {0-5, 5-10, …..}, in. Ammontare della spesa per degenza (v), v può assumere i valori prefissati {0, 10, 20, …..}.

Una caratteristica degli elementi di un collettivo di tipo popolazione o evento è allora concettualizzata come una relazione tra il collettivo di tipo popolazione o evento e una classificazione (se qualitativa) o un dominio (se quantitativa), e ad essa corrisponde un tipo di enunciato di appartenenza riferibile ad una coppia di elementi, ad esempio Sesso (x ,m), Attività economica principale (x ,m), Classi di età (x, m), Spesa per degenza (x, v).

Può essere utile osservare che, da un punto di vista strettamente *formale*, l'ontologia dell'archivio è specificabile come una collezione di relazioni a un posto (i collettivi e le classificazioni) o a due posti (le caratteristiche, viste come relazioni tra elementi e modalità, oppure valori, e le relazioni in senso proprio); queste relazioni a uno o a due posti sono rispettivamente interpretabili come insiemi e relazioni tra elementi di insiemi. Da questo punto di vista sono necessarie poi apposite *regole logico-formali* per completare l'ontologia dell'archivio stabilendo le definizioni dei collettivi, i collettivi di appartenenza delle caratteristiche, le classificazioni o i domini associati ad ogni caratteristica, i collettivi legati dalle relazioni, l'obbligatorietà o meno di caratteristiche e relazioni [35].

Inoltre come è noto è generalmente pre-definita una serie di *vincoli di incompatibilità/obbligatorietà*, anch'essi formalizzabili come regole, che possono coinvolgere modalità o valori assumibili dalle caratteristiche, legami stabiliti dalle relazioni, appartenenza ai collettivi, e che sono utilizzabili per i controlli di qualità.

Gli enunciati precedentemente introdotti corrispondenti ai collettivi, alle caratteristiche, alle relazioni sono formalmente enunciati aperti, che danno luogo a enunciati chiusi, i quali possono essere veri o falsi, ponendo al posto della x e d eventualmente della y gli identificativi di elementi osservabili e al posto della m o della v una modalità di una classificazione o un valore numerico.

Le diverse *tipologie di informazione* che nel corso del tempo vengono raccolte da una fonte, in particolare da un archivio amministrativo, sono proprio tali enunciati chiusi, cioè riferiti a singoli elementi osservabili (eventi o unità di popolazione), ad esempio Studente (Rossi), Residenza (Rossi, Roma).

Per questo possiamo dire che una fonte d'informazione accetta e gestisce *diverse tipologie di enunciati riferiti ai collettivi, alle caratteristiche, alle relazioni componenti l'ontologia della fonte e riferiti a specifici elementi osservati*, precisamente:

[35] Un'ontologia può essere specificata mediante appositi linguaggi basati sulla logica e dotati di una semantica formale.

A(u_i), enunciati che asseriscono l'appartenenza di elementi singoli ai collettivi di tipo popolazione o evento, ad esempio Studente (Rossi), Ricovero (ricovero$_i$), Degenza (degenza$_i$);

B(u_i, c_i), enunciati che asseriscono il possesso da parte di un elemento u_i, di un collettivo di tipo popolazione o evento, istantaneo o con durata, di una specifica modalità o di uno specifico valore c_i per una caratteristica, vista come relazione tra collettivo e classificazione oppure tra collettivo e dominio, ad esempio Classe di età (Rossi, 20-25), Spesa per degenza (degenza$_i$, 550 euro);

C(u_i, u_j) enunciati che asseriscono l'esistenza di una particolare relazione (per quanto detto, sempre di tipo funzionale) tra due elementi u_i e u_j appartenenti a due collettivi di tipo popolazione o evento (o anche allo stesso collettivo), dei quali uno è il dominio della relazione e l'altro il codominio, ad esempio Azienda agricola è condotta da Persona (azienda$_i$, Verdi), Immatricolazione inizia Studente (immatricolazione$_i$, Rossi), Acquisizione crediti riguarda Studente (acquisizione crediti$_i$, Rossi).

Ad ogni collettivo di tipo popolazione o evento, ad ogni caratteristica, ad ogni relazione corrisponde quindi uno specifico tipo di enunciato di appartenenza, della forma rispettivamente *A(u_i)*, *B(u_i, c_i)* , o *C(u_i, u_j)*.

Tuttavia, come precisato nell'Introduzione, un archivio amministrativo è un processo di raccolta di informazione continuo nel tempo.

Un archivio amministrativo raccoglie e gestisce con continuità informazioni relative all'appartenenza di singoli elementi osservabili ai collettivi di tipo popolazione o evento, istantaneo o con durata, che lo caratterizzano, all'appartenenza di coppie elemento –modalità o elemento-valore alle caratteristiche che lo caratterizzano, all'appartenenza di coppie di elementi osservabili alle relazioni che lo caratterizzano.

Una specifica degli enunciati che caratterizzano gli archivi amministrativi che non tenga conto della evoluzione temporale della validità degli enunciati è quindi troppo generica per fornire un'adeguata base concettuale alla specifica dei diversi tipi di errore: *è necessario descrivere la dinamica delle informazioni raccolte dall'archivio e la conseguente dinamica della validità temporale degli enunciati*.

Gli oggetti nell'ontologia dell'archivio e i relativi tipi di enunciati, riformulati tenendo conto dei riferimenti temporali

La caratterizzazione di un archivio amministrativo come processo che attua la registrazione di informazioni con continuità nel tempo, con i meccanismi precedentemente descritti, obbliga a introdurre un riferimento temporale negli enunciati di tipo *A, B, C* elencati nel paragrafo precedente.

Definire un linguaggio per la specifica *formale* delle ontologie degli archivi amministrativi che tenga conto del parametro tempo richiederebbe un'attività di ricerca dedicata.

Introduciamo comunque di seguito informalmente una specifica riferita al tempo degli enunciati registrati in un archivio amministrativo, con l'unico scopo di rendere più chiaro il discorso successivo sulle diverse tipologie di errori e quindi più rigorosa la classificazione degli errori alla quale si tende.

Introduciamo quindi in tutti i tipi di enunciati accettati e gestiti in un archivio amministrativo un *riferimento temporale*, precisamente:

- un istante t_i per gli enunciati riferiti a elementi dei collettivi di eventi istantanei
- una durata $t_i - t$ oppure $t_i - t_j$, per gli enunciati riferiti a elementi dei collettivi di tipo popolazione o agli eventi con durata, enunciati che possono essere riferiti ad un periodo aperto $t_i - t$ o chiuso $t_i - t_j$.

Di conseguenza i tipi di enunciati d'interesse statistico gestiti in un archivio amministrativo sono:

A) enunciati di appartenenza di un elemento u_i ad un collettivo di tipo popolazione o evento (istantaneo o con durata) A, riferiti ad un momento o un periodo, precisamente:

> $A(u_i, t_i)$ per i collettivi di tipo evento istantaneo, dove e u_i è l'identificativo dell'evento istantaneo appartenente al collettivo *A*, e $t_{i\,i}$ identifica il momento di occorrenza
>
> $A(u_i, t_i - t)$ per i collettivi di tipo popolazione o evento con durata, dove u_i è l'identificativo dell'unità o evento con durata appartenente al collettivo *A*, t_i identifica il momento di ingresso dell'unità nel collettivo, *t* identifica un momento di uscita dell'unità dal collettivo ancora indefinito
>
> $A(u_i, t_i - t_j)$ per i collettivi di tipo popolazione o evento con durata, dove u_i è l'identificativo dell'unità o evento con durata appartenente al collettivo *A*, t_i identifica il momento di ingresso dell'unità nel collettivo, t_j identifica il momento di uscita dell'unità dal collettivo.

Esempi (nei quali l'identificativo di un evento è indicato come evento_i): Studente (Verdi, $t_i - t$), Avvio rapporto di lavoro (avvio rapporto di lavoro_i, t_i), Ricovero (ricovero_i, t_i), Rapporto di lavoro (rapporto di lavoro_i, $t_i - t$), Degenza (degenza_i, $t_i - t$);

B) enunciati riguardanti il possesso da parte di un'unità u_i di un collettivo di tipo popolazione o evento, istantaneo o con durata, di una specifica modalità c_i di una classificazione o di uno specifico valore c_i di un dominio numerico per una certa caratteristica B, corrispondente all'appartenenza della coppia (u_i, c_i) alla caratteristica B, ad uno specifico momento o per un periodo, precisamente:

> $B(u_i, c_i, t_i)$ per i collettivi di tipo evento istantaneo, dove u_i, è l'identificativo dell'evento istantaneo, c_i è l'identificativo di una modalità appartenente alla classificazione utilizzata per la caratteristica o di un valore appartenente al dominio numerico della caratteristica, t_i identifica il momento di occorrenza dell'evento istantaneo
>
> $B(u_i, c_i, t_i - t)$ per i collettivi di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo, c_i è l'identificativo di una modalità appartenente alla classificazione utilizzata per la caratteristica o di un valore appartenente al dominio numerico della caratteristica, t_i identifica il momento di acquisizione della modalità c_i, *t* identifica un momento nel quale la modalità o valore c_i sarà sostituito da un'altra modalità o valore, momento che è ancora indefinito
>
> $B(u_i, c_i, t_i - t_j)$ per i collettivi di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo, c_i è l'identificativo di una modalità appartenente alla classificazione utilizzata per la caratteristica o di un valore appartenente al dominio numerico della caratteristica, t_i identifica il momento di

acquisizione della modalità c_i, mentre t_j identifica il momento nel quale la modalità o valore c_i è stato sostituito da un'altra modalità o valore.

Esempi (nei quali l'identificativo di un evento è indicato come $evento_i$): Classe di età (Verdi, 20-25, $t_i - t$), Sesso(Verdi, femmina, $t_i - t$), Residenza (Rossi, Roma, $t_i - t$), Valore della produzione (Fiat, 100000 euro, $t_i - t$), Motivo ricovero ($ricovero_i$, incidente, t_i), Spesa per degenza ($degenza_i$, 550 euro, $t_i - t$);

C) enunciati riguardanti l'esistenza di una particolare relazione tra due elementi u_i, u_j appartenenti a due collettivi di tipo popolazione o evento (o anche allo stesso collettivo) corrispondente all'appartenenza della coppia *(u_i, u_j)* alla relazione C ad uno specifico momento o per un periodo, precisamente:

C(u_i, u_j, t_i) se il collettivo nel ruolo di dominio è di tipo evento istantaneo, dove u_i, è l'identificativo dell'unità appartenente al collettivo dominio, u_j è l'identificativo dell'unità legata appartenente al collettivo codominio, e t_i identifica il momento di occorrenza dell'evento istantaneo

C(u_i, u_j, $t_i - t$) se il collettivo nel ruolo di dominio è di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo dominio, u_j, è l'identificativo dell'unità legata appartenente al collettivo codominio, t_i identifica il momento di acquisizione del legame con l'elemento u_j,, t identifica un momento nel quale l'elemento legato u_j, sarà sostituito da un altro elemento legato, momento che è ancora indefinito

C(u_i, u_j, $t_i - t_j$) se il collettivo nel ruolo di dominio è di tipo popolazione o evento con durata, dove u_i, è l'identificativo dell'unità o evento con durata appartenente al collettivo dominio, u_j, è l'identificativo dell'unità legata appartenente al collettivo codominio, t_i identifica il momento di acquisizione del legame con l'elemento u_j,, t_j identifica il momento nel quale l'elemento legato u_j, è stato sostituito da un altro elemento.

Esempi (nei quali l'identificativo di un evento è indicato come $evento_i$): Azienda agricola è condotta da Conduttore (Le querce, Bianchi, $t_i - t$), Unità locale appartiene Impresa (Sede Torino, Fiat, $t_i - t$), Immatricolazione inizia Studente ($immatricolazione_i$, Verdi, t_i), Acquisizione crediti riguarda Studente ($esame_i$, Verdi, t_i), Avvio rapporto di lavoro riguarda Lavoratore (avvio rapporto di $lavoro_i$, Bianchi, t_i), Avvio rapporto di lavoro riguarda Datore di lavoro (avvio rapporto di $lavoro_i$, Neri, t_i), Avvio rapporto di lavoro inizia Rapporto di lavoro (avvio rapporto di $lavoro_i$, rapporto di $lavoro_i$, t_i), Rapporto di lavoro riguarda Lavoratore (rapporto di $lavoro_i$, Bianchi, $t_i - t$), Rapporto di lavoro riguarda Datore di lavoro (rapporto di $lavoro_i$, Neri, $t_i - t$).

Nella pratica, come descritto in dettaglio nel Framework, conviene considerare questi enunciati strutturati in *record*. Un record riunisce concettualmente tutti gli enunciati di appartenenza e di possesso di caratteristiche e relazioni pertinenti ad ogni elemento di uno specifico collettivo.

L'archivio amministrativo come strumento di raccolta di enunciati di appartenenza a collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi

Si è detto che nel corso del tempo una fonte d'informazione, in particolare un archivio amministrativo, accetta e gestisce *diverse tipologie di enunciati riferiti ai collettivi, alle*

caratteristiche, alle relazioni componenti l'ontologia della fonte e relativi a specifici elementi osservati.

L'informazione gestita in un archivio amministrativo è concettualmente costituita da raccolte di tali enunciati, organizzati in *record*.

Tenuto conto di ciò possiamo cominciare a introdurre il concetto di *estensione* di un collettivo, di una caratteristica, o di una relazione tra collettivi, che ci permetterà di arrivare a definire la qualità dell'archivio come corretta osservazione dell'estensione dei collettivi, delle caratteristiche e delle relazioni osservate dall'archivio.

L'estensione di un collettivo, di una caratteristica, o di una relazione tra collettivi è l'insieme degli enunciati di appartenenza al collettivo, di possesso della caratteristica, di esistenza della relazione tra elementi che sono veri.

Nel successivo paragrafo questo concetto di estensione viene precisato tenendo conto del carattere continuo nel tempo che caratterizza la raccolta di informazioni attuata dagli archivi amministrativi, in generale dalle fonti non statistiche.

Nel complesso un archivio amministrativo mira a gestire informazioni sullo stato degli elementi dei collettivi osservati, di tipo popolazione o evento, in ciascun momento d'osservazione, registrando quindi le caratteristiche e relazioni possedute da ogni elemento in ogni momento.

E' importante però osservare che nella maggior parte dei casi un archivio amministrativo, e in generale una fonte non statistica, a differenza di un'indagine non osserva direttamente gli elementi dei collettivi di tipo popolazione.

Raccoglie anzitutto informazioni relative agli eventi istantanei che sono oggetto dell'attività amministrativa a supporto della quale è costituito l'archivio (ad esempio immatricolazione, ricovero ospedaliero, avvio di un rapporto di lavoro, erogazione di una pensione, dichiarazione dei redditi).

Quando questi eventi istantanei sono di ingresso o di uscita per gli elementi dei collettivi di tipo popolazione o evento con durata, aggiorna di conseguenza le informazioni relative all'appartenenza di elementi a tali collettivi.

Tipicamente poi l'archivio aggiorna le informazioni relative alle unità delle popolazioni e agli eventi con durata per tutto il tempo della loro appartenenza allo specifico collettivo caratteristico dell'archivio, registrando tutti i cambiamenti nelle caratteristiche e relazioni possedute (ad esempio il cambio di residenza di una persona, il cambio di tipologia di un rapporto di lavoro). Per le caratteristiche e relazioni coinvolte nella definizione delle condizioni di appartenenza al collettivo, tali cambiamenti possono anche comportare l'entrata o l'uscita dell'elemento dal collettivo.

Gli aggiornamenti delle caratteristiche o relazioni per le unità delle popolazioni e gli eventi con durata si possono considerare come particolari specie di eventi istantanei che non hanno caratteristiche proprie e in generale non rivestono interesse di per sé, se non per studi longitudinali, ma concorrono a determinare la dinamica delle informazioni nell'archivio. Si può dire quindi che in un archivio amministrativo di solito tutte le informazioni relative ai collettivi di tipo popolazione e anche ai collettivi di tipo evento con durata vengono aggiornate a seguito di eventi istantanei.

L'archivio amministrativo come strumento di osservazione dell'estensione dei collettivi, delle caratteristiche e delle relazioni

Si è detto che in generale *l'estensione di un collettivo, di una caratteristica, o di una relazione tra collettivi è l'insieme degli enunciati di appartenenza al collettivo, di possesso della caratteristica , di esistenza della relazione tra elementi che sono veri*, e che la qualità dell'archivio dipende dalla corretta registrazione di tali enunciati.

Questa definizione dev'essere approfondita tenendo conto della natura continua nel tempo della registrazione attuata mediante gli archivi amministrativi, che ha obbligato a introdurre riferimenti temporali negli enunciati.

Una rigorosa definizione formale del concetto di *estensione* di un collettivo, di una caratteristica, o di una relazione tra collettivi in un contesto di continuità temporale è un compito difficile che non si può certo affrontare in questo documento.

Nel seguito ci si limita a definizioni pratiche il cui unico scopo è poter continuare a considerare la qualità dell'archivio come corretta osservazione dell'estensione dei collettivi, delle caratteristiche e delle relazioni osservate, fornendo così alla nozione di qualità una base concettuale comunque più rigorosa anche se non ineccepibile dal punto di vista logico-formale.

A questo scopo definiamo l'estensione di un collettivo, di una caratteristica, o di una relazione tra collettivi con riferimento ad un periodo d'osservazione finito, che ha un momento d'inizio convenzionale T e si estende fino al momento attuale t_A. Al nostro scopo si può assumere che T coincida con il momento d'impianto del processo d'osservazione considerato, in particolare dell'archivio amministrativo.

L'estensione di un collettivo, di una caratteristica, o di una relazione tra collettivi nel periodo d'osservazione T- t_A è l'insieme degli enunciati veri di appartenenza al collettivo, di possesso della caratteristica, di esistenza della relazione tra elementi che hanno riferimenti temporali compresi nel periodo considerato.

Questa definizione di estensione è per così dire longitudinale, comprende cioè per ogni elemento che può appartenere ad un collettivo l'intera "storia" della sua appartenenza al collettivo a partire dal momento T, e analogamente per le coppie elemento-modalità, oppure elemento-valore, che possono appartenere alle caratteristiche, e per le coppie di elementi che possono appartenere alle relazioni.

Appartengono all'estensione così intesa di un collettivo, di una caratteristica, o di una relazione tra collettivi nel periodo d'osservazione T- t_A tutti gli enunciati veri riferiti ad un momento t_i con $T \leq t_i \leq t_A$, dove:

- t_i per gli enunciati riferiti a eventi istantanei è il momento di riferimento
- t_i per gli enunciati riferiti a unità di popolazioni e a eventi con durata è il momento d'inizio del periodo di validità che può essere aperto o chiuso, nel qual caso si ha di conseguenza $T \leq t_j \leq t_A$.

Per le unità di popolazioni e gli eventi con durata in particolare l'estensione così intesa può comprendere al momento t_A una serie di enunciati con periodo chiuso ed un eventuale enunciato con periodo aperto e lo stesso vale per l'estensione delle caratteristiche e delle relazioni ad essi relative.

Da questa definizione per così dire longitudinale si ricava la definizione istantanea di estensione, riferita ad un momento *t'* con $T \leq t' \leq t_A$

L'estensione di un collettivo, di una caratteristica, o di una relazione tra collettivi in un momento t' con $T \leq t' \leq t_A$ *è l'insieme degli enunciati veri di appartenenza al collettivo, di possesso della caratteristica o di esistenza della relazione tra elementi i cui riferimenti temporali includono il momento t'.*

Questa definizione istantanea di estensione corrisponde alla nozione intuitiva dell'insieme degli elementi che sono effettivamente appartenenti ad un collettivo in un momento considerato *t'*, e dell'insieme di coppie che rappresentano il possesso di una caratteristica o l'esistenza di una relazione tra elementi nel momento considerato *t'* .

Si interpreta diversamente per gli enunciati riferiti a eventi istantanei, da una parte, o piuttosto ad unità di popolazione oppure eventi con durata, dall'altra, a seconda cioè che gli enunciati siano riferiti ad un momento t_i o viceversa ad un periodo aperto $t_i - t$, oppure chiuso $t_i - t_j$.

Appartengono all'estensione così intesa di un collettivo di eventi istantanei tutti gli enunciati di appartenenza al collettivo veri riferiti ad un a un momento t_i con $t_i = t'$.

Analogo criterio vale per determinare gli enunciati che rappresentano il possesso nel momento *t'* di una caratteristica da parte di un evento istantaneo, oppure l'esistenza nel momento *t'* di una relazione tra un evento istantaneo e un altro elemento.

L'estensione di un collettivo di eventi istantanei ad un dato momento è data quindi dall'insieme degli eventi istantanei che occorrono in quel momento, e lo stesso vale per le relative caratteristiche e relazioni.

Appartengono all'estensione così intesa di un collettivo di tipo popolazione o evento con durata tutti gli enunciati di appartenenza al collettivo veri riferiti ad un periodo $t_i - t$ oppure $t_i - t_j$ nei quali $t_i \leq t'$ ed eventualmente $t' < t_j$. Se $t' = t_A$ (t_A è il momento attuale), appartengono all'estensione così intesa tutti e soli gli enunciati con periodo aperto.

Analogo criterio vale per determinare gli enunciati che rappresentano il possesso nel momento *t'* di una caratteristica da parte di un'unità di popolazione o di un evento con durata, oppure l'esistenza nel momento *t'* di una relazione tra un'unità di popolazione o un evento con durata e un altro elemento.

L'estensione di un collettivo di tipo popolazione o evento con durata ad un dato momento è data quindi dall'insieme delle unità di popolazione o degli eventi con durata che entrano nel collettivo in quel momento, oppure continuano ad appartenervi, e lo stesso vale per le relative caratteristiche e relazioni.

Su questa base si può anche definire l'estensione di un collettivo, una caratteristica, o una relazione riferita ad un periodo *t'-t''*, con $T \leq t' \leq t_A$, $T \leq t'' \leq t_A$, separatamente per gli enunciati riferiti a eventi istantanei, da una parte, o piuttosto ad unità di popolazione oppure eventi con durata, dall'altra.

L'estensione di un collettivo di tipo evento istantaneo in un periodo t'-t'', con $T \leq t' \leq t_A$, $T \leq t'' \leq t_A$ *è l'insieme degli enunciati veri di appartenenza al collettivo riferiti ad un momento* t_i *con* $t' < t_i < t''$, *analogo criterio vale per determinare gli enunciati appartenenti all'estensione di una caratteristica o relazione relativa a eventi istantanei.*

L'estensione di un collettivo di eventi istantanei in un dato periodo è data quindi dall'insieme degli eventi istantanei che sono occorsi in quel periodo, e lo stesso vale per le relative caratteristiche e relazioni.

Per i collettivi di tipo popolazione o evento con durata è necessario specificare la definizione precisando le condizioni di intersezione tra il periodo di validità dell'enunciato e il periodo considerato *t'-t''*, in generale si può affermare quanto segue.

L'estensione di un collettivo di tipo popolazione o evento con durata in un periodo t'-t'', con $T \leq t' \leq t_A$, $T < \leq t'' \leq t_A$ *è l'insieme degli enunciati veri di appartenenza al collettivo riferiti ad un periodo* $t_i - t$, *oppure* $t_i - t_j$ *che interseca t'-t'', potendo in particolare comprendere, essere compreso o coincidere con t'-t'', analogo criterio vale per determinare gli enunciati appartenenti all'estensione di una caratteristica o relazione relativa a unità di popolazione o eventi con durata.*

L'estensione di un collettivo di tipo popolazione o evento con durata ad un dato periodo è data quindi dall'insieme delle unità di popolazione o degli eventi con durata la cui durata interseca il periodo considerato in un modo da precisare, e lo stesso vale per le relative caratteristiche e relazioni.

In base alla più generale definizione di estensione di un collettivo, di una caratteristica, o di una relazione tra collettivi che è stata formulata si può affermare che *un archivio amministrativo osserva con continuità le estensioni degli oggetti che caratterizzano la sua ontologia*, vale a dire:

un archivio amministrativo raccoglie e gestisce con continuità, nell'intero periodo di osservazione T- t_A *che lo caratterizza, enunciati di appartenenza ai collettivi, enunciati di possesso di caratteristiche, enunciati di esistenza di relazioni tra elementi rispettivamente riferiti ai collettivi, alle caratteristiche, alle relazioni che caratterizzano la sua ontologia*.

Nei disegni alle pagine 134-138 sono mostrati esempi di enunciati di appartenenza ai collettivi di tipo popolazione o evento, di possesso delle caratteristiche, di esistenza di relazioni tra elementi.

Nel disegno a pagina 139 è presentato un esempio di successione temporale di enunciati di appartenenza a collettivi, che evidenzia quanto affermato nel precedente paragrafo sul ruolo prioritario degli eventi istantanei.

A pagina 140 è visualizzato il rapporto tra enunciati di appartenenza veri e relativo aggiornamento dell'archivio.

In funzione dei propri interessi di ricerca, lo statistico sceglie uno o più collettivi d'interesse con relative caratteristiche (ad essi direttamente connesse o ad essi attribuibili tenendo conto delle relazioni tra collettivi), e definisce l'ambito temporale di osservazione, tipicamente un momento *t'* per i collettivi di tipo popolazione, e un periodo *t'-t''* per i collettivi di tipo evento istantaneo, mentre per i collettivi di tipo evento con durata potrà riferire l'osservazione ad un momento *t'* o a un periodo *t'-t''* a seconda di quanto sia rilevante la durata dell'evento per i suoi scopi.

Per ricavare le distribuzioni d'interesse, *lo statistico lavora poi, classificando e misurando, proprio sulle estensioni di questi collettivi al momento t' o nel periodo t'-t'' prefissati*.

Per quanto detto, precisamente queste sono le informazioni alle quali lo statistico è tipicamente interessato (a parte casi specifici come particolari studi longitudinali):

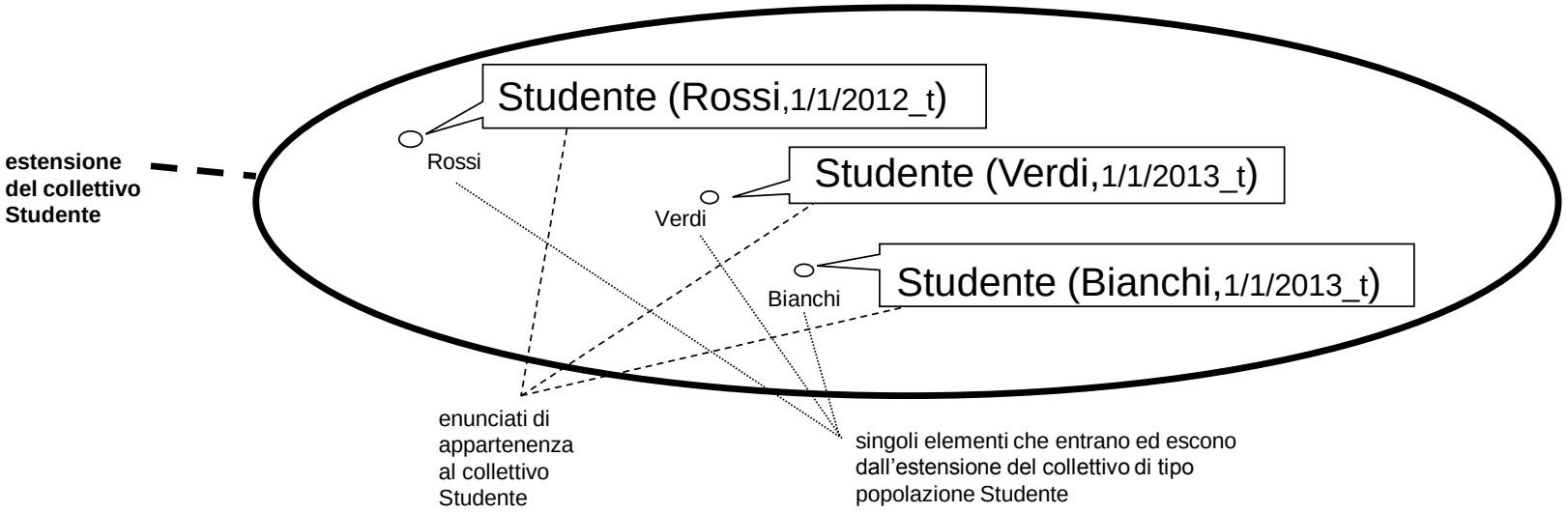
- l'insieme degli enunciati veri di appartenenza ai collettivi d'interesse di tipo evento istantaneo $A(u_i, t_i)$ per i quali $t' < t_i < t''$
- l'insieme degli enunciati veri di possesso di caratteristiche $B(u_i, c_i, t_i)$ e di esistenza di relazioni con altri elementi $C(u_i, u_j, t_i)$ riferiti agli elementi di questi collettivi
- l'insieme degli enunciati veri di appartenenza ai collettivi d'interesse di tipo popolazione o evento con durata $A(u_i, t_i - t)$, $A(u_i, t_i - t_j)$ per i quali $t_i \leq t'$ ed eventualmente $t' < t_j$; se $t' = t_A$ (t_A è il momento attuale) questo insieme comprende tutti e soli gli enunciati con periodo aperto $A(u_i, t_i - t)$
- l'insieme degli enunciati veri di possesso di caratteristiche $B(u_i, c_i, t_i - t)$, $B(u_i, c_i, t_i - t_j)$ e di esistenza di relazioni con altri elementi $C(u_i, u_j, t_i - t)$, $C(u_i, u_j, t_i - t_j)$ riferiti agli elementi di questi collettivi, nei quali sia $t_i \leq t'$ ed eventualmente $t' < t_j$; se $t' = t_A$ (t_A è il momento attuale) questo insieme comprende tutti e soli gli enunciati con periodo aperto $B(u_i, c_i, t_i - t)$ e $C(u_i, u_j, t_i - t)$.

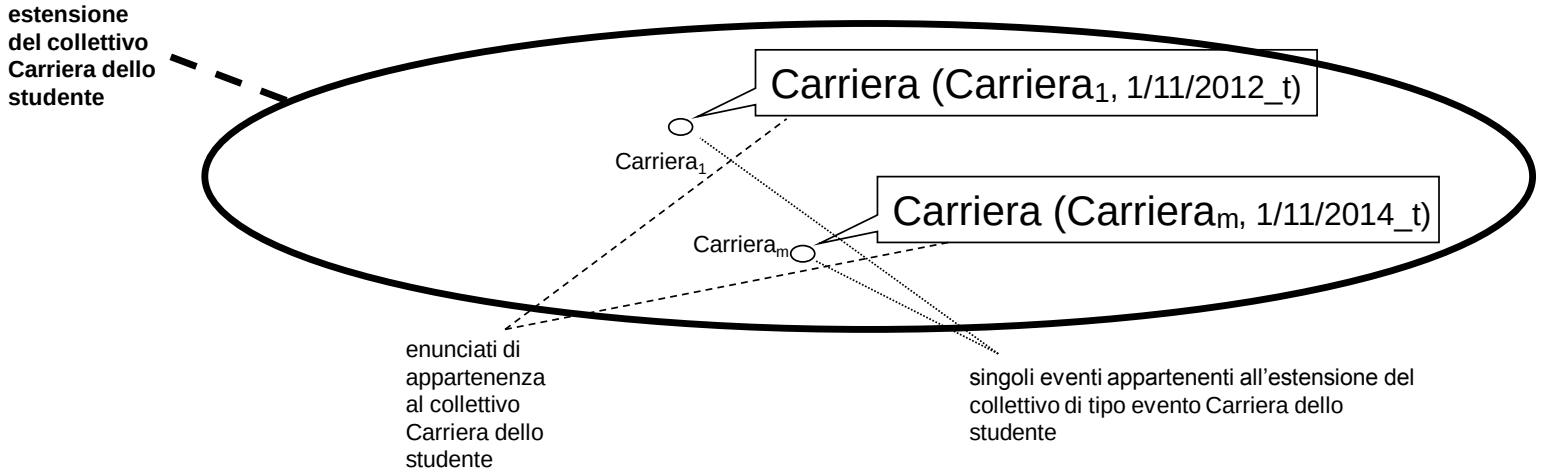
Nella pratica, come descritto in dettaglio nel Framework, conviene considerare questi enunciati strutturati in *record*. Un record riunisce concettualmente tutti gli enunciati di appartenenza e di possesso di caratteristiche e relazioni pertinenti ad ogni elemento di uno specifico collettivo.

.

I collettivi di tipo popolazione o evento con durata e le loro estensioni

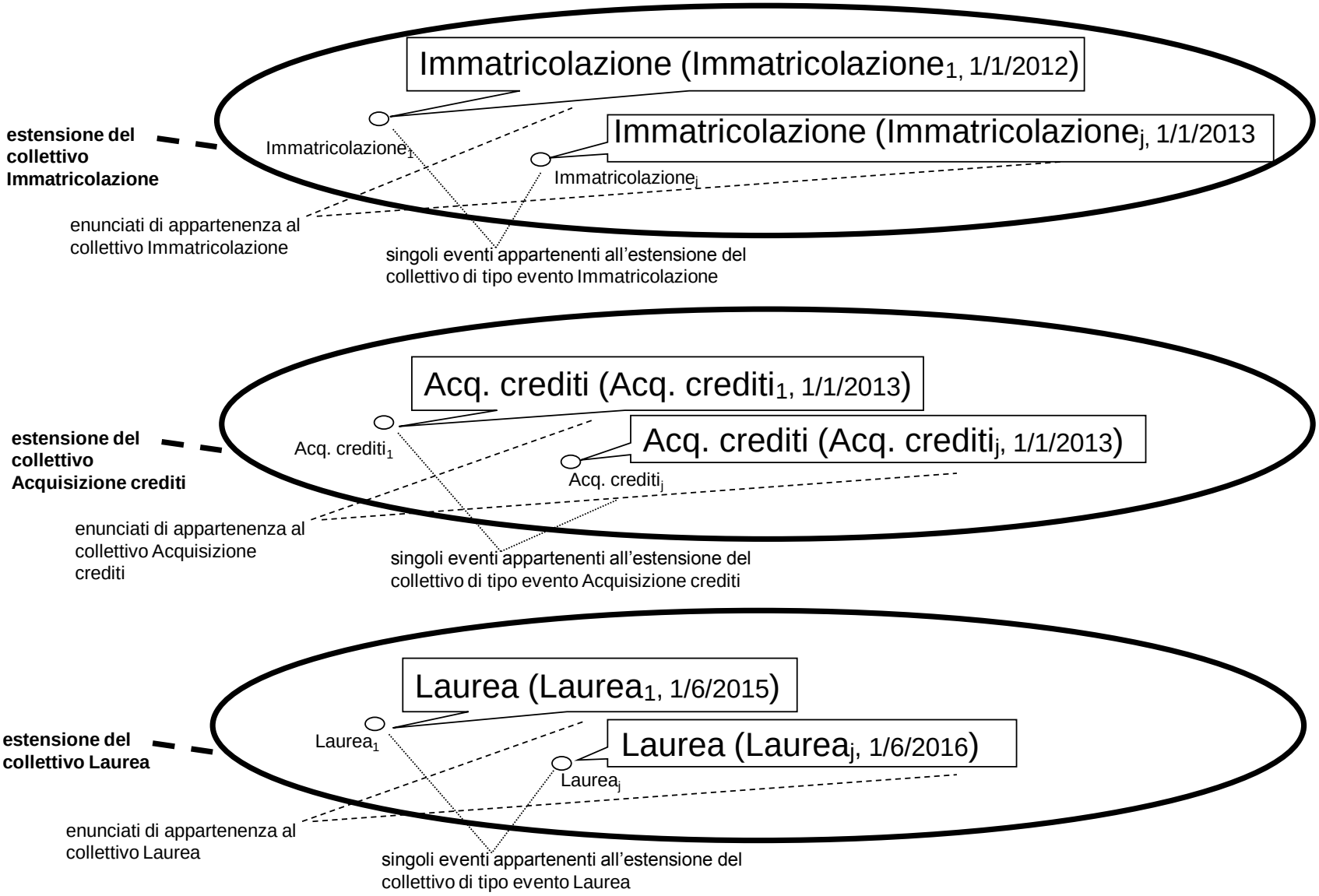
Estensione del collettivo = insieme degli enunciati di appartenenza al collettivo





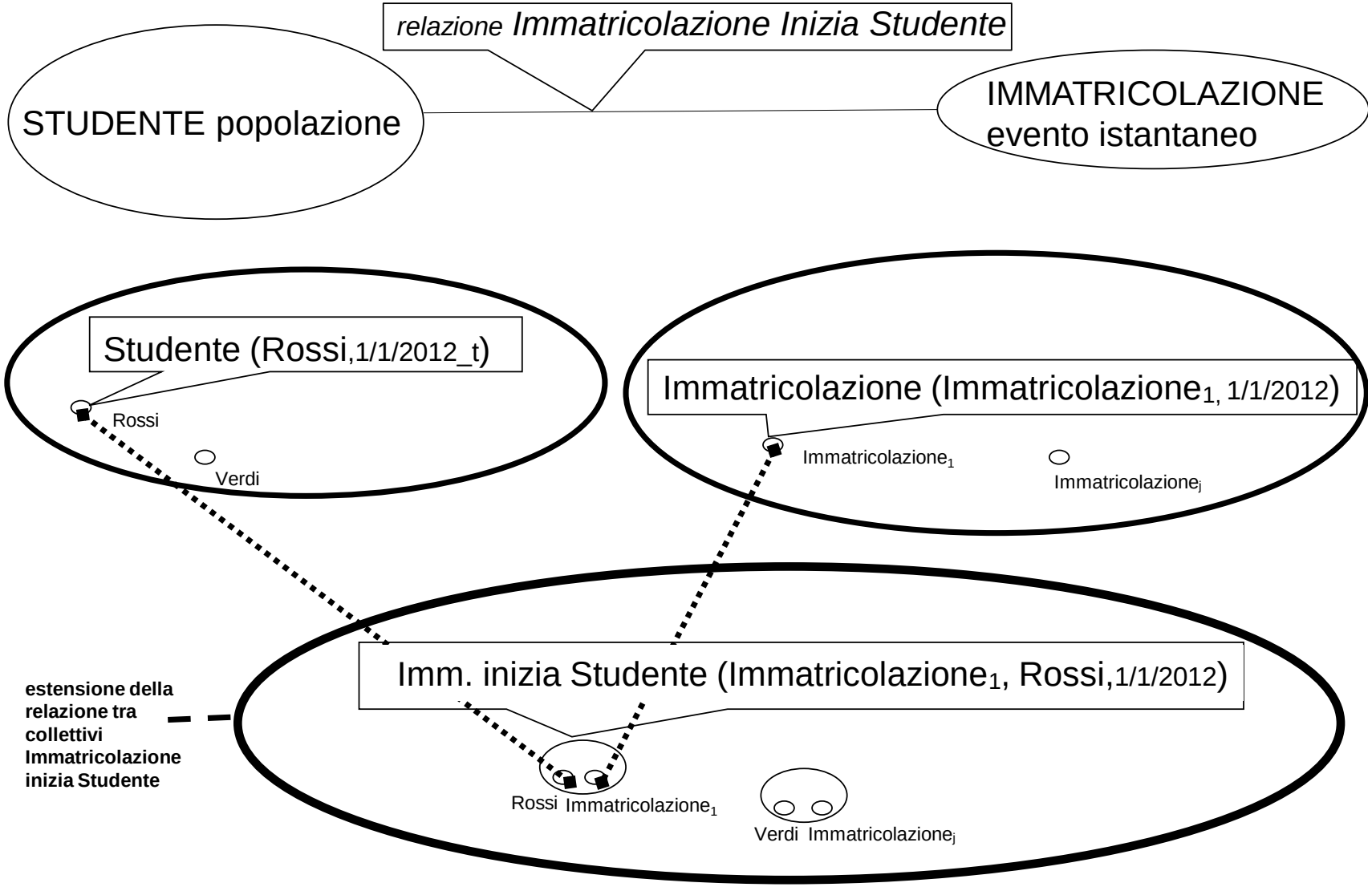
I collettivi di tipo evento istantaneo e le loro estensioni

Estensione del collettivo = insieme degli enunciati di appartenenza al collettivo



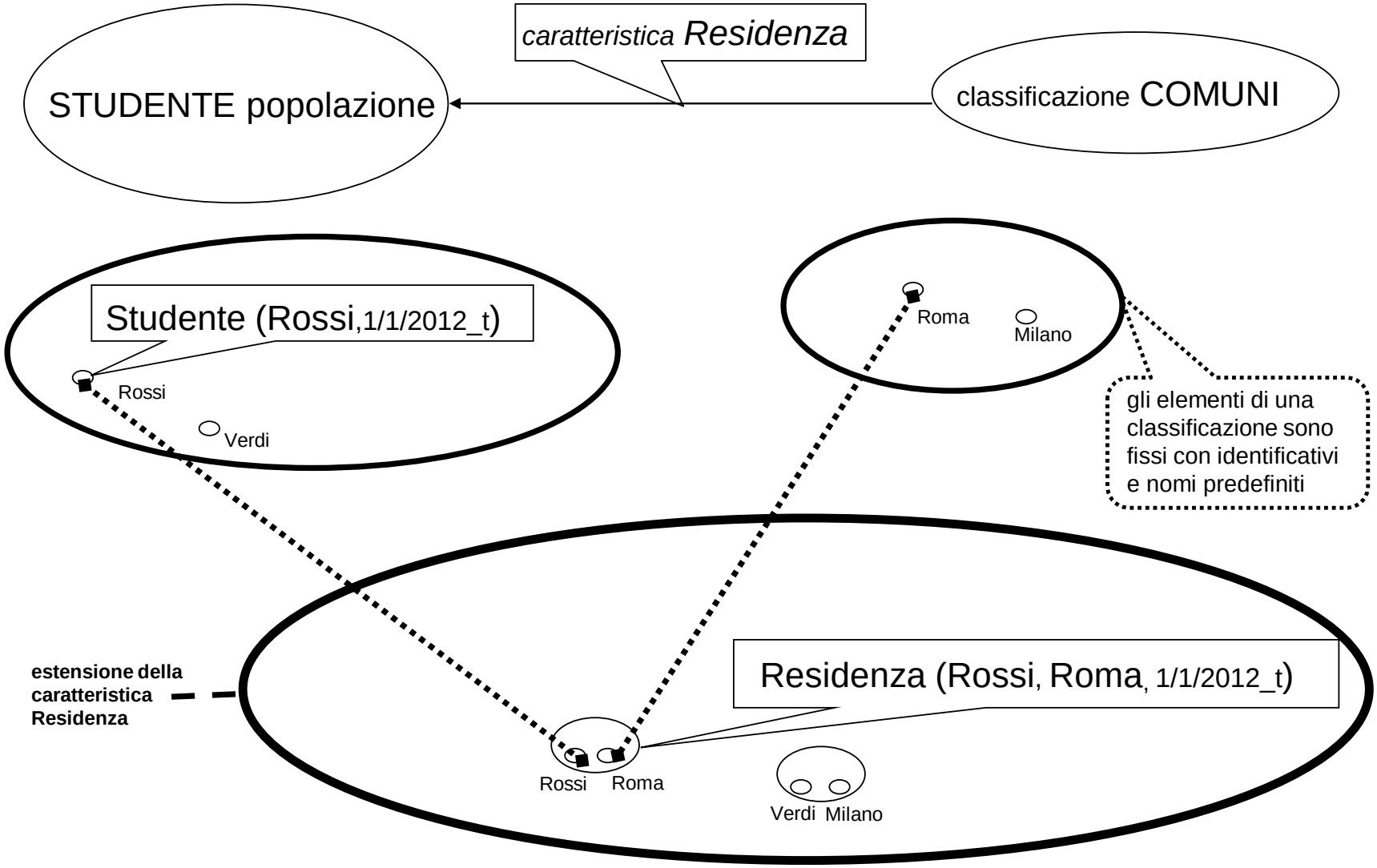
Una relazione è un insieme di coppie, con una sua estensione

Estensione della relazione= insieme degli enunciati di appartenenza alla relazione, riferiti a coppie di elementi dei collettivi



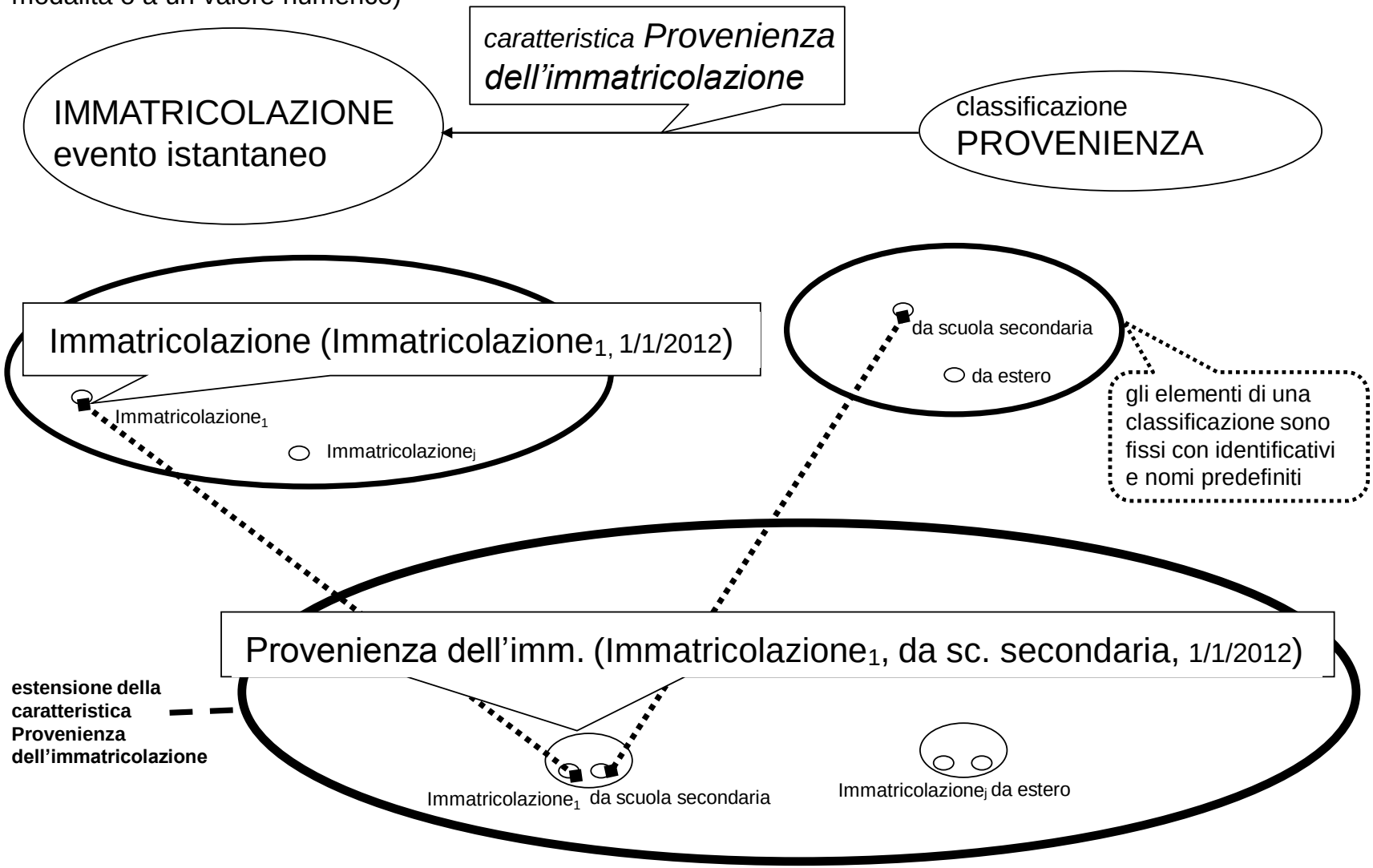
Una caratteristica (variabile) è un insieme di coppie, con una sua estensione

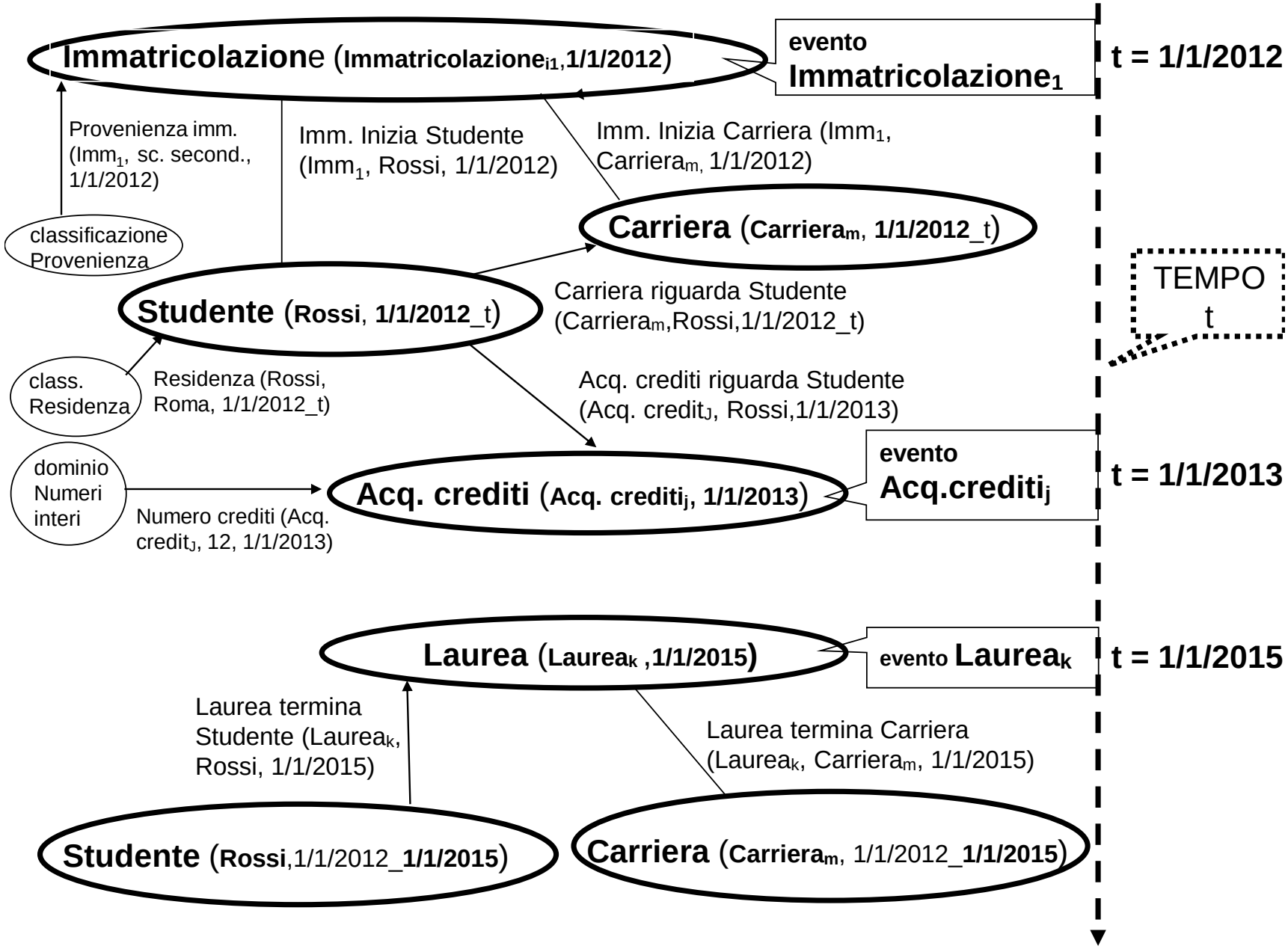
Estensione della caratteristica= insieme degli enunciati di appartenenza alla caratteristica, riferiti a coppie che associano un elemento del collettivo a un elemento di una classificazione o di un dominio (cioe' a una modalità o a un valore numerico)

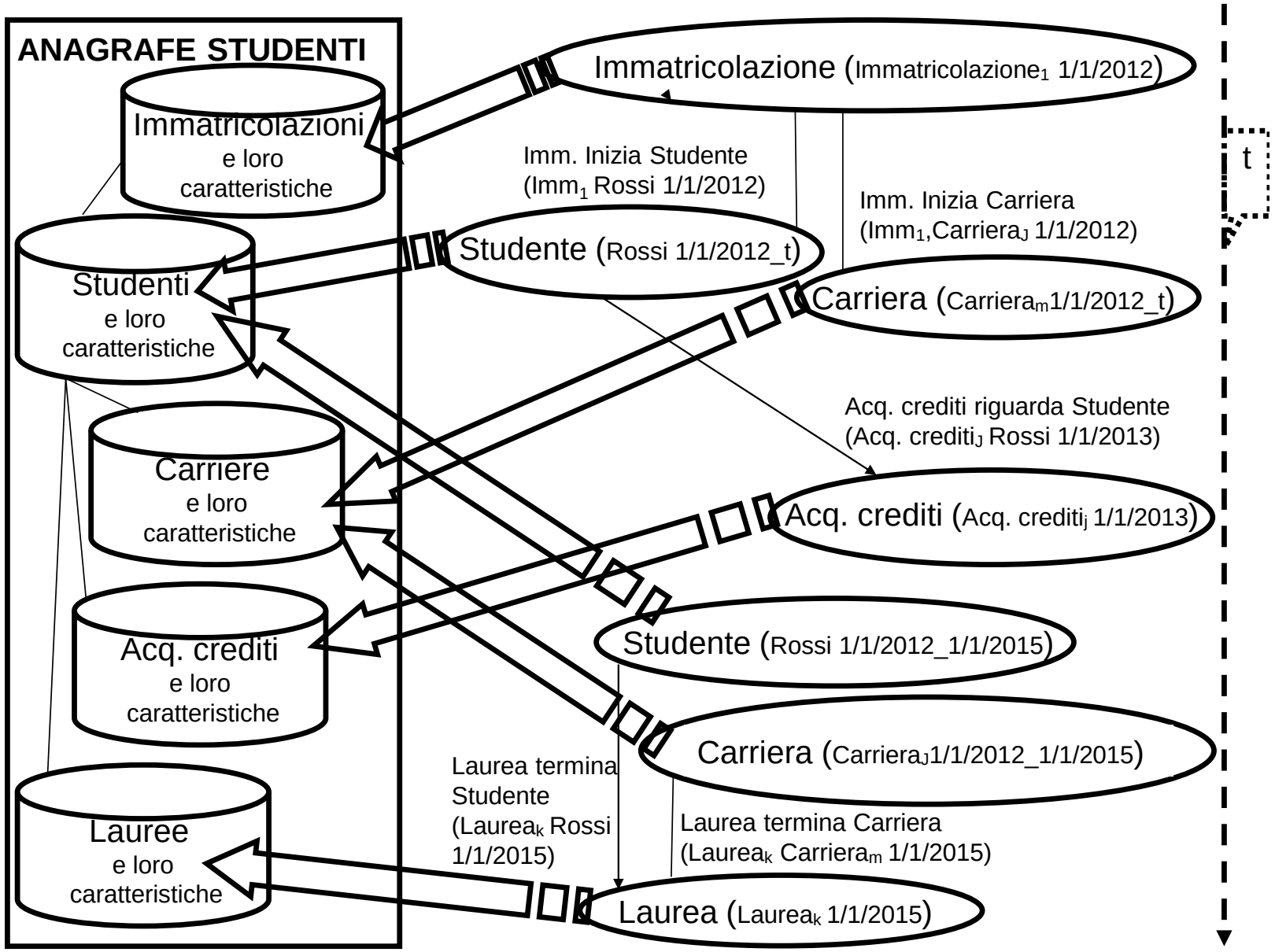


Una caratteristica (variabile) è un insieme di coppie, con una sua estensione

Estensione della caratteristica= insieme degli enunciati di appartenenza alla caratteristica, riferiti a coppie che associano un elemento del collettivo a un elemento di una classificazione o di un dominio (cioe' a una modalità o a un valore numerico)




Immatricolazione (Immatricolazione$_{i1}$,1/1/2012)
evento Immatricolazione$_1$
t = 1/1/2012
Provenienza imm. (Imm$_1$, sc. second., 1/1/2012)
Imm. Inizia Studente (Imm$_1$, Rossi, 1/1/2012)
Imm. Inizia Carriera (Imm$_1$, Carriera$_m$, 1/1/2012)
classificazione Provenienza
Carriera (Carriera$_m$, 1/1/2012_t)
TEMPO t
Studente (Rossi, 1/1/2012_t)
Carriera riguarda Studente (Carriera$_m$,Rossi,1/1/2012_t)
class. Residenza
Residenza (Rossi, Roma, 1/1/2012_t)
Acq. crediti riguarda Studente (Acq. credit$_J$, Rossi,1/1/2013)
dominio Numeri interi
Acq. crediti (Acq. crediti$_j$, 1/1/2013)
evento Acq.crediti$_j$
t = 1/1/2013
Numero crediti (Acq. credit$_J$, 12, 1/1/2013)
Laurea (Laurea$_k$,1/1/2015)
evento Laurea$_k$
t = 1/1/2015
Laurea termina Studente (Laurea$_k$, Rossi, 1/1/2015)
Laurea termina Carriera (Laurea$_k$, Carriera$_m$, 1/1/2015)
Studente (Rossi,1/1/2012_1/1/2015)
Carriera (Carriera$_m$, 1/1/2012_1/1/2015)


ANAGRAFE STUDENTI
Immatricolazioni
e loro
caratteristiche
Studenti
e loro
caratteristiche
Carriere
e loro
caratteristiche
Acq. crediti
e loro
caratteristiche
Lauree
e loro
caratteristiche
Immatricolazione (Immatricolazione1 1/1/2012)
Imm. Inizia Studente
(Imm1 Rossi 1/1/2012)
Imm. Inizia Carriera
(Imm1,CarrieraJ 1/1/2012)
Studente (Rossi 1/1/2012_t)
Carriera (Carrieram1/1/2012_t)
Acq. crediti riguarda Studente
(Acq. creditiJ Rossi 1/1/2013)
Acq. crediti (Acq. creditij 1/1/2013)
Studente (Rossi 1/1/2012_1/1/2015)
Carriera (CarrieraJ1/1/2012_1/1/2015)
Laurea termina
Studente
(Laureak Rossi
1/1/2015)
Laurea termina Carriera
(Laureak Carrieram 1/1/2015)
Laurea (Laureak 1/1/2015)
t

La qualità dell'archivio come corretta osservazione dell'estensione dei collettivi, delle caratteristiche e delle relazioni

In quanto gestisce informazioni relative alle estensioni dei collettivi di tipo popolazione o evento, delle caratteristiche, delle relazioni che lo caratterizzano, *un archivio amministrativo è uno strumento di osservazione, in analogia ad un'indagine, e come strumento di osservazione può raccogliere e gestire informazioni affette da errore.*

Come anticipato nel paragrafo precedente, un archivio amministrativo raccoglie e gestisce con continuità, nell'intero periodo di osservazione T- t_A che lo caratterizza, enunciati di appartenenza ai collettivi, enunciati di possesso di caratteristiche, enunciati di esistenza di relazioni tra elementi rispettivamente riferiti ai collettivi, alle caratteristiche, alle relazioni che caratterizzano la sua ontologia.

A partire da T, il periodo di osservazione dell'archivio si allunga con lo scorrere del tempo. Nel corso del tempo cioè il momento attuale t_A viene a corrispondere a determinazioni temporali successive per ciascuna delle quali l'archivio compie operazioni di aggiornamento, accettando nuova informazione riferita al momento t_A [36] . Precisamente per ogni momento t_A l'archivio:

- accetta una serie di nuovi enunciati di appartenenza ai collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi ponendo in essi $t_i = t_A$, dove t_i è il momento di riferimento, per gli enunciati riferiti a eventi istantanei, oppure il momento d'inizio del periodo aperto di validità dell'enunciato, per gli enunciati riferiti a unità di popolazioni e a eventi con durata
- chiude il periodo di validità, ponendo in esso $t_j = t_A$, per una serie di enunciati di appartenenza ai collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi, riferiti a unità di popolazioni e a eventi con durata.

Nel Framework sono specificate in dettaglio queste operazioni di aggiornamento, precisando inoltre i loro effetti sui *record* oltre che sui singoli enunciati.

La qualità dell'archivio è ottimale quando l'archivio osserva correttamente le estensioni degli oggetti che caratterizzano la sua ontologia, vale a dire*:*

un archivio amministrativo ha qualità ottimale quando raccoglie e gestisce <u>tutti</u> e <u>soli</u> gli enunciati di appartenenza <u>veri</u> ai collettivi, gli enunciati <u>veri</u> di possesso di caratteristiche, gli enunciati <u>veri</u> di esistenza di relazioni tra elementi rispettivamente riferiti ai collettivi, alle caratteristiche, alle relazioni che caratterizzano la sua ontologia.

Per quanto detto dovrebbe essere ovvio che per corretta osservazione dell'estensione si intende qui non semplicemente la corretta numerosità dei collettivi, ma proprio la corretta elencazione degli elementi che realmente ne fanno parte, e la corretta determinazione delle loro caratteristiche e relazioni.

Nel corso del tempo, ciò comporta che ad ogni momento t_A l'archivio:

- accetta <u>tutti</u> e <u>soli</u> gli enunciati <u>veri</u> di appartenenza ai collettivi, gli enunciati <u>veri</u> di possesso di caratteristiche, gli enunciati <u>veri</u> di esistenza di relazioni tra elementi per i quali

[36] Per effetto dell' errore di tempestività, può accadere che la nuova informazione riferita al momento t_A sia accettata in pratica in un momento t successivo a t_A

effettivamente t_i=t_A, dove t_i è il momento di riferimento, per gli enunciati riferiti a eventi istantanei, oppure il momento d'inizio del periodo aperto di validità dell'enunciato, per gli enunciati riferiti a unità di popolazioni e a eventi con durata

- chiude il periodo di validità, ponendo in esso t_j=t_A, per tutti e soli quegli enunciati di appartenenza ai collettivi, di possesso di caratteristiche, di esistenza di relazioni tra elementi, riferiti a unità di popolazioni e a eventi con durata, per i quali effettivamente t_j= t_A.

Queste condizioni si possono anche esprimere facendo ricorso al concetto di *copertura*: un archivio ha qualità ottimale quando garantisce nel corso del tempo la corretta copertura dell'estensione dei collettivi, delle caratteristiche e delle relazioni osservate.

Gli errori si generano a causa di una mancata corrispondenza tra gli enunciati di appartenenza veri ai collettivi, gli enunciati veri di possesso di caratteristiche, gli enunciati veri di esistenza di relazioni tra elementi e gli enunciati accettati e gestiti in archivio e, inoltre, a causa di una difformità tra il contenuto degli enunciati accettati in archivio e il contenuto degli enunciati veri, dovuta ad errori nei codici identificativi degli elementi coinvolti o nei riferimenti temporali.

Anzitutto ad ogni momento t_A di registrazione si possono distinguere enunciati veri o falsi. D'altra parte dato un enunciato qualsiasi, prima di poterlo giudicare vero o falso occorre che l'enunciato stesso sia riferibile senza errore agli elementi dei quali si vuole enunciare qualcosa, vale a dire, per formulare enunciati occorre che sia possibile assegnare un identificativo a tutti gli elementi dei quali si vuole enunciare qualcosa, senza ambiguità.

Ciò implica che il dispositivo di assegnazione degli identificativi di cui è dotato l'archivio garantisca i seguenti requisiti:

a) ogni elemento in ogni momento t_A ha un identificativo, b) ogni elemento in ogni momento t_A ha un solo identificativo c) ogni identificativo in ogni momento t_A identifica un solo elemento d) ogni elemento mantiene lo stesso identificativo in tutti i momenti t_A che si succedono nel corso del tempo.

Per gli enunciati correttamente formulati si può poi valutare la verità o falsità. In linea teorica occorre quindi:

- diagnosticare la presenza di errori sugli identificativi per i diversi tipi di enunciati formulabili
- diagnosticare la verità o falsità degli enunciati. Sono possibili due tipi di errori: accettare enunciati falsi o non accettare enunciati veri. Si veda il disegno a pagina 144.

Di conseguenza in generale gli errori possono essere di tre tipi fondamentali, che si presentano e articolano differentemente per i diversi tipi di enunciati:

ERRORI DI IDENTIFICAZIONE

ERRONEA INCLUSIONE

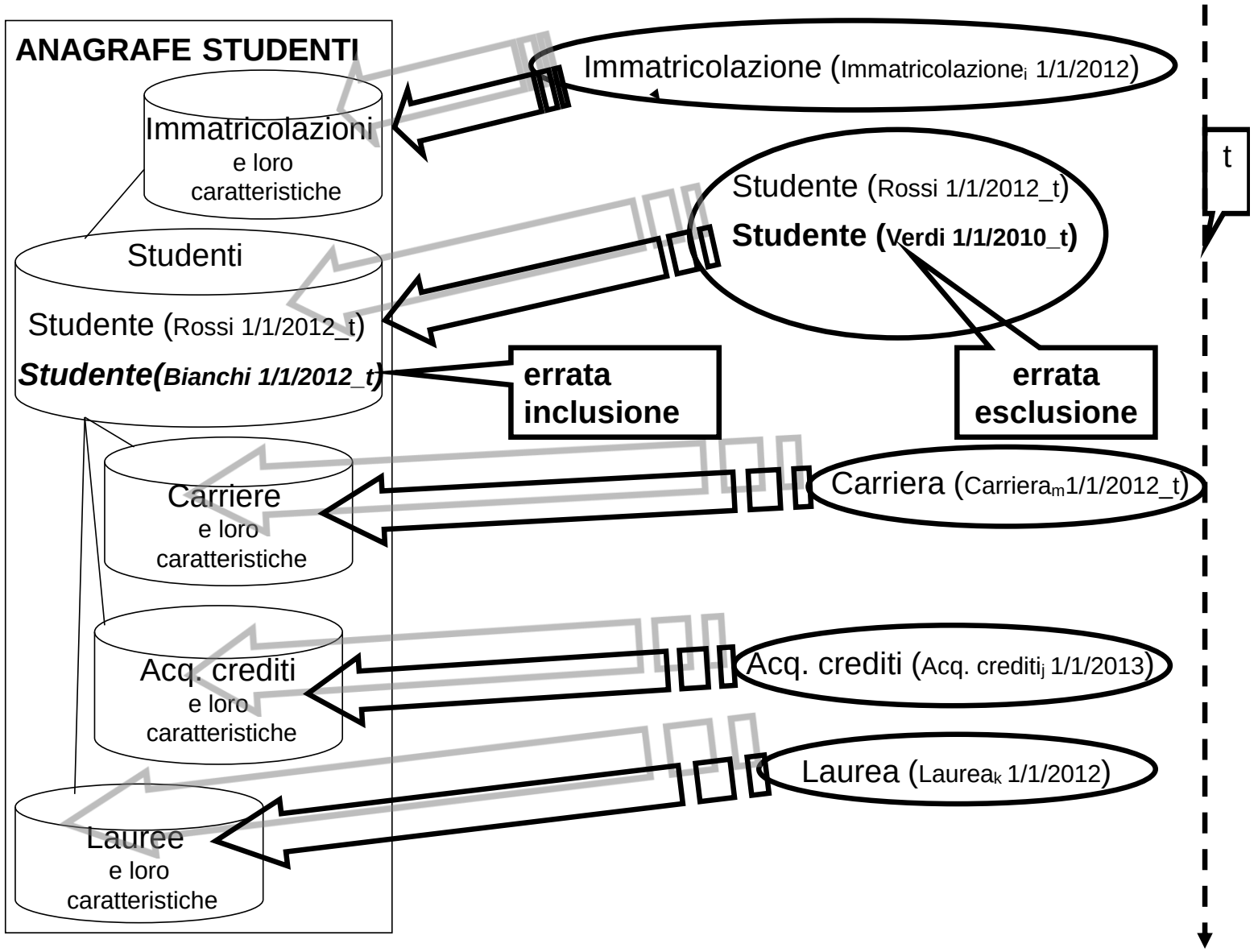
ERRONEA ESCLUSIONE.

Gli errori possono avere diverse origini: problemi definitori, distorsione volontaria da parte di chi fornisce l'informazione o da parte dell'organismo che la raccoglie, cattiva progettazione o cattivo funzionamento di qualche aspetto del dispositivo di assegnazione degli identificativi e/o della procedura operativa di registrazione e accettazione degli enunciati dei diversi tipi.

E' bene ricordare sempre che, come evidenziato nel Framework, in base agli strumenti di cui disponiamo per individuare gli errori i diversi tipi di errore non sono sempre discriminabili.

Per passare da questa specifica teorica degli errori alla disamina dei diversi tipi di errori concretamente possibili, nella PARTE TERZA del Framework si descrive come è attuata in concreto la procedura di formulazione degli enunciati e la loro accettazione nell'archivio, attraverso la creazione o la modifica dei record gestiti in archivio..

Nelle parti QUARTA e QUINTA del Framework son poi dettagliati gli errori dei diversi tipi, appartenenti, come discusso nel Framework, alle due grandi classi degli ERRORI DI COPERTURA e DI ACCURATEZZA.


ANAGRAFE STUDENTI
Immatricolazioni
e loro
caratteristiche
Studenti
Studente (Rossi 1/1/2012_t)
Studente(Bianchi 1/1/2012_t)
Carriere
e loro
caratteristiche
Acq. crediti
e loro
caratteristiche
Lauree
e loro
caratteristiche
Immatricolazione (Immatricolazione$_i$ 1/1/2012)
Studente (Rossi 1/1/2012_t)
Studente (Verdi 1/1/2010_t)
errata
inclusione
errata
esclusione
Carriera (Carriera$_m$1/1/2012_t)
Acq. crediti (Acq. crediti$_j$ 1/1/2013)
Laurea (Laurea$_k$ 1/1/2012)
t