

Progetto *Carriere Studenti* (modellazione dimensionale)



Paolo Atzeni, Luigi Bellomarini



Obiettivi

- Realizzare alcuni data mart per la autovalutazione di Ateneo:
 - **carriere degli studenti**
 - ne abbiamo studiato i requisiti
 - **progettazione dimensionale** (parte I, ne vediamo alcune soluzioni)
 - **implementazione data mart in Business Analytics** (parte II)
 - **implementazione ETL in Pentaho Data Integration** (parte II)
 - **offerta didattica**
 - lo analizzeremo, ma non lo realizzeremo
- **Un esercizio, ma anche un progetto DWH reale**



Homework II

- **Modalità di lavoro**
 - piccoli gruppi (1, 2 o 3 persone)
 - il lavoro può essere svolto in collaborazione
 - gli esperimenti devono essere svolti, almeno in parte, in maniera individuale
 - la consegna è individuale
- **Realizzazione di un progetto DWH**
 - implementazione dello schema dimensionale (OLAP) e dei flussi di ETL
 - **discussioni (31 marzo, 5 aprile)**

Homework

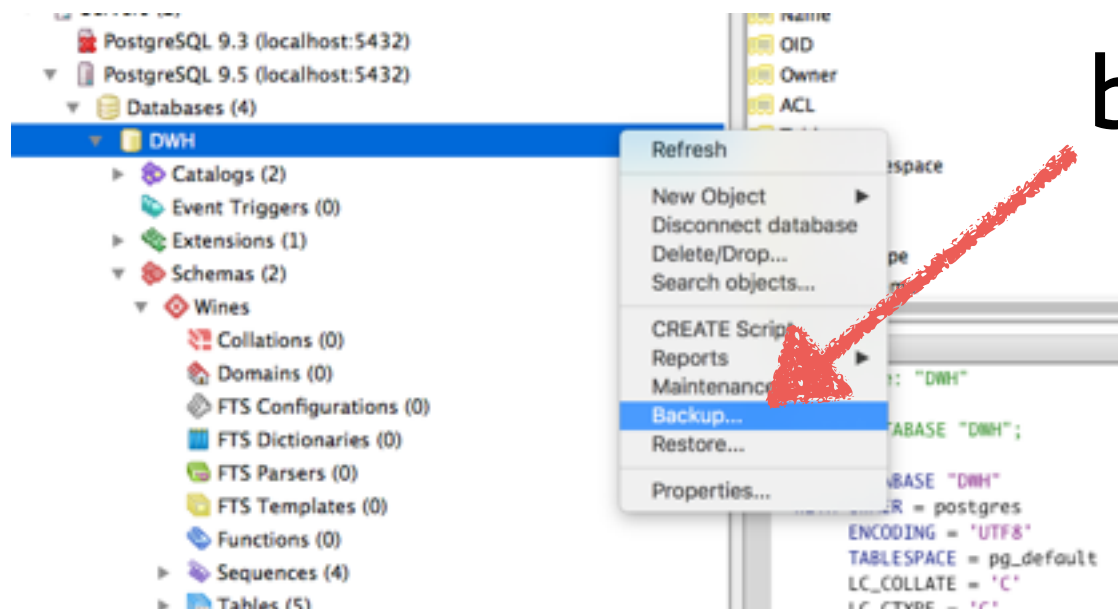
(gestione codice)

- ogni **gruppo** crea un repository (pubblico) su GitHub (<https://github.com/>)
- un repository avrà quindi URL:
 - <https://github.com/<nome gruppo>/<nome repo>>
 - es: <https://github.com/myGroup/myRepo>
- nel repository si creano 2 **directory**:
 - **olap/**
 - **etl/**
 - e un file README.md (da salvare nella root del repository, fuori da tutte le directory)
 - contenente **nome e cognome di tutti i membri del gruppo** (senza virgole, su righe diverse in ordine alfabetico)

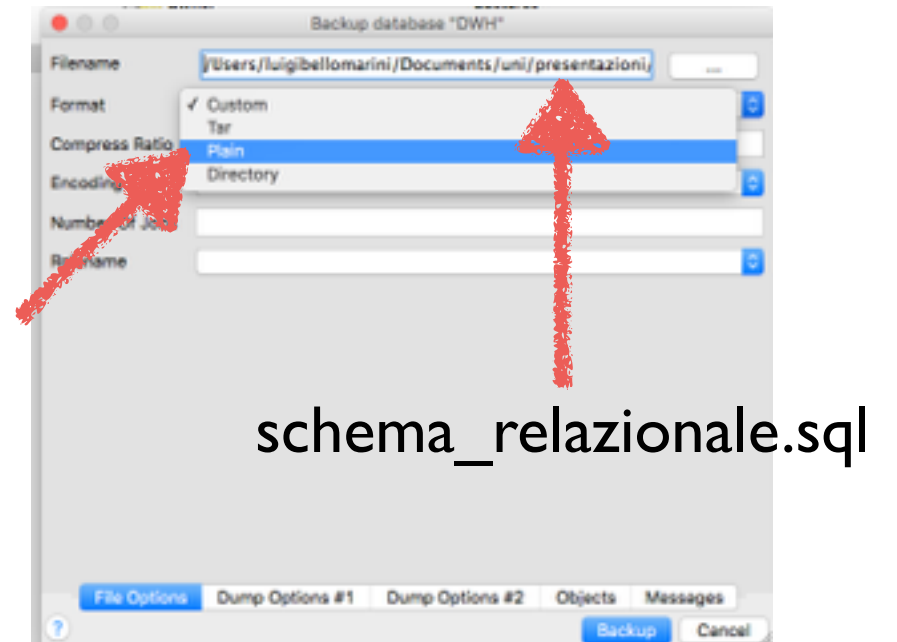
Gestione del codice olap/

- La directory olap/ deve contenere
 - un file **schema_relazionale.sql** contenente gli script DDL necessari alla creazione dello schema creato in Postgres
 - i file dei metadati dello schema dimensionale creato in Pentaho.

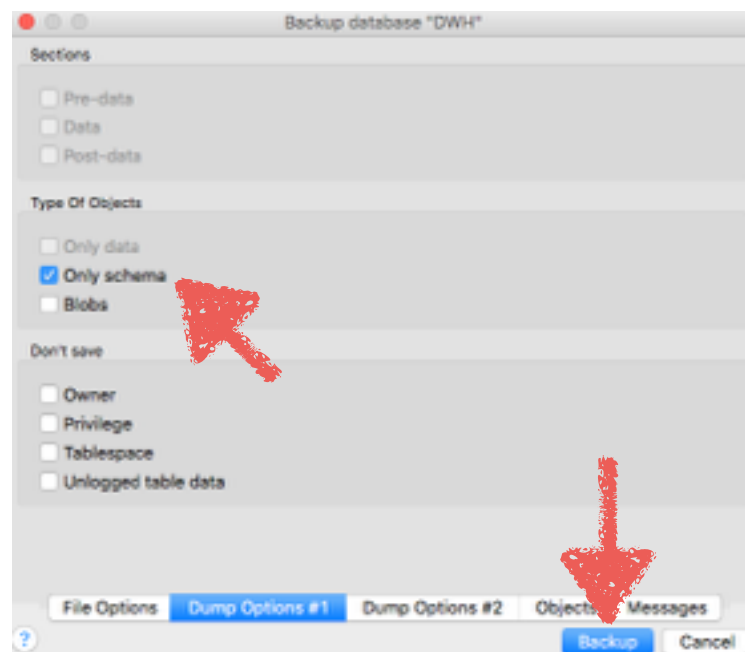
/olap schema relazionale



backup



schema_relazionale.sql

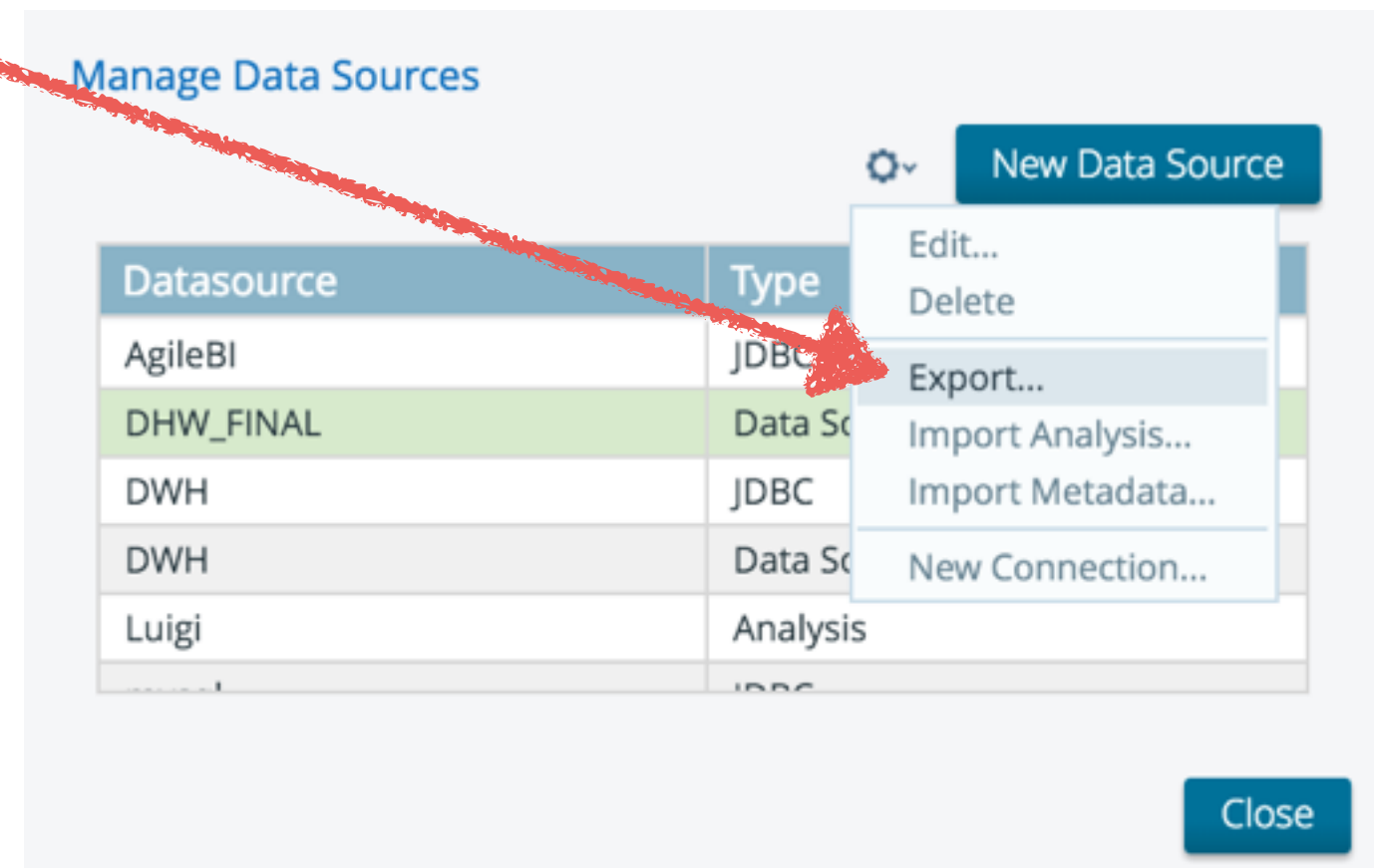


- Si effettua un **backup** del database creato in Postgres
- Formato **Plain**
- File nominato **schema_relazionale.sql**
- Only schema (non esportare i dati)

/olap schema dimensionale

- creare un data source con nome <nome_schema> per ogni fact table e costruire ogni star schema come abbiamo visto
- esportare tutti i data source (cioè gli schemi) creati

- per ogni data source esportato, decomprimere il file ZIP creato
- copiare i due file contenuti:
<nome_schema>.mondrian.xml e
<nome_schema>.xmi
- nella directory /olap
- Alla fine, si dovranno consegnare quindi due file per ogni data source (cioè per ogni fact table)



Gestione del codice /etl

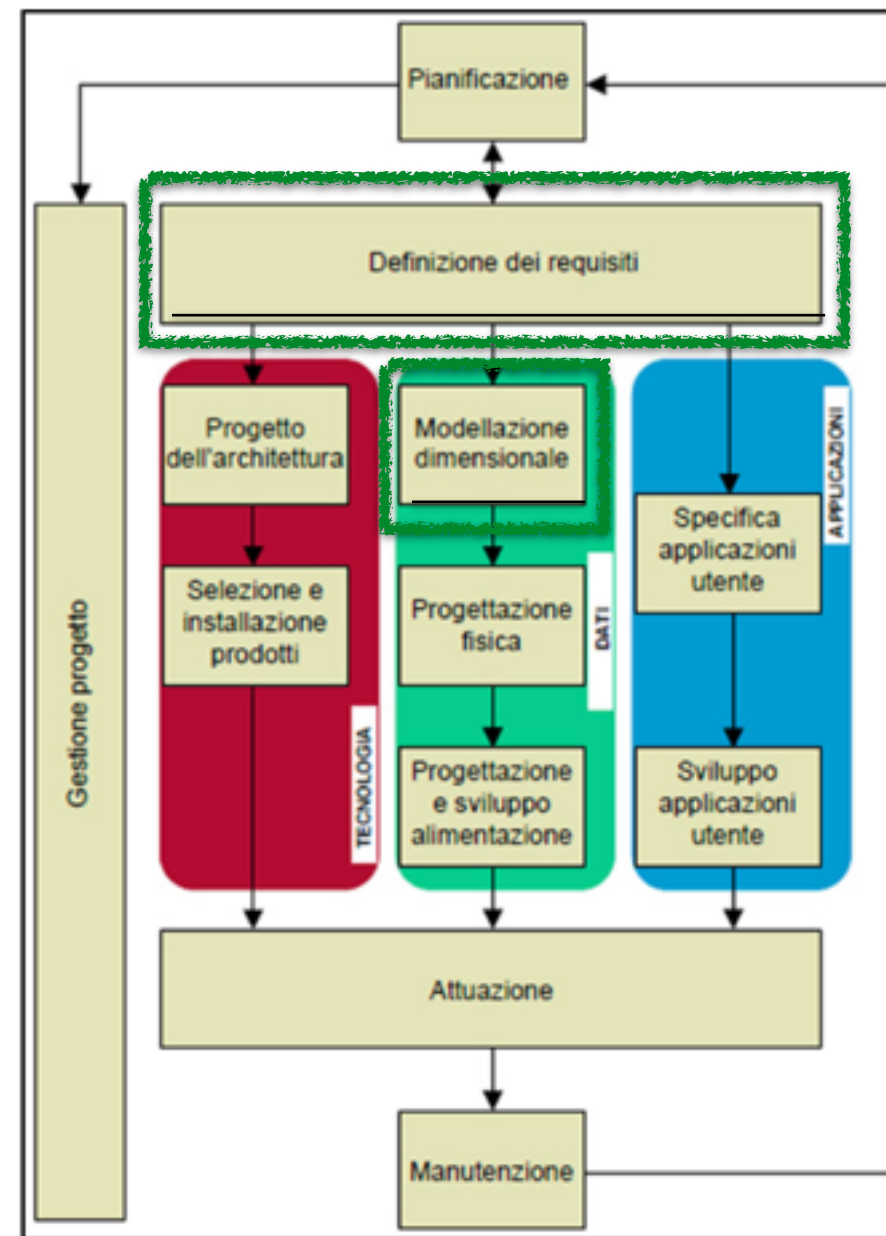
- La directory **etl/** dovrà contenere:
 - un file **<nome_dimensione>.ktr** per ogni dimensione del dwh progettato.
 - un file **<nome_fact_table>.ktr** per ogni fact table progettata.
 - un file **dimensions.kjb** con un job che coordina tutte le trasformazioni delle dimensioni.
 - un file **fact_tables.kjb** con un job che coordina tutte le trasformazioni di tutte le fact table.
- I file **ktr** e **kjb** si salvano direttamente da **Pentaho Data Integration**.

Homework II

(modalità di consegna)

- ogni studente
 - indica su **Moodle**:
 - “Consegna homework: <URL del repository del proprio gruppo>”
 - es: consegna homework: - <https://github.com/bellomarini/myDWproject>
 - quindi gli studenti di uno stesso gruppo indicheranno in Moodle tutti la stessa URL e i loro nomi saranno tutti anche in README.md
- entro la data di consegna, ogni gruppo
 - effettua il PUSH su **GitHub** del codice relativo

Business Dimensional Lifecycle (Kimball, 1998)

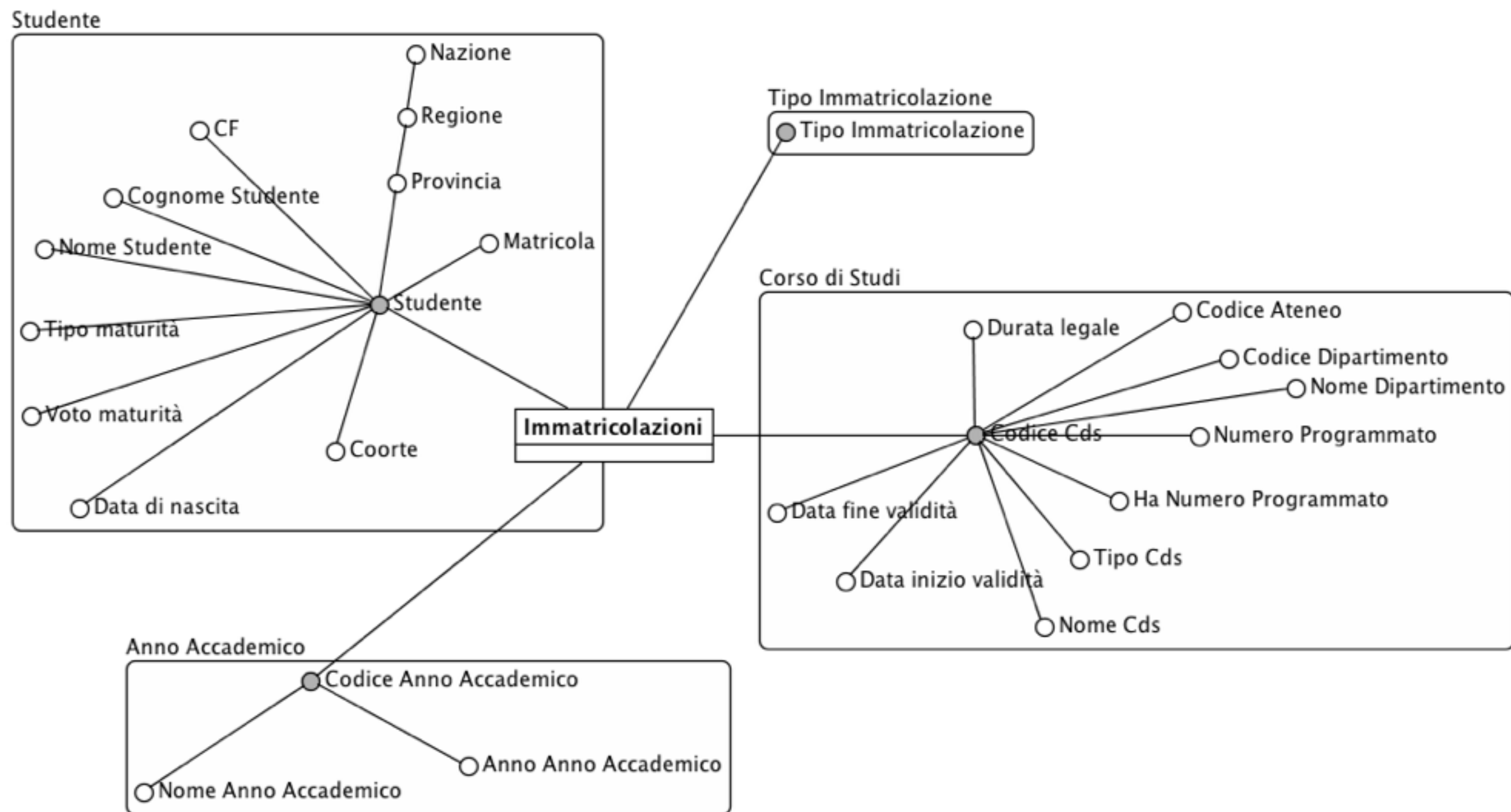


Operational Data Store

Matrice dell'architettura a bus del DW

	Tempo	Anno Accademico	Studente	Docente	Corso	Corso di Studi
Preimmatricolazioni		X	X			X
Immatricolazioni		X	X			X
Test ammissione		X	X			X
Iscrizioni	X	X	X			X
Trasferimenti	X	X	X			X
Rinunce	X	X	X			X
Esami	X	X	X		X	X
Lauree	X	X	X			X
Docenze		X		X	X	X

Immatricolazioni



Immatricolazioni (Fact table)

- Rappresentiamo i fatti delle immatricolazioni.
 - Uno Studente si è immatricolato in un certo Anno Accademico in un certo Corso di Studi.
- La fact table descrive le transazioni, cioè i fatti di immatricolazione.
- Factless fact table.
 - Interessa l'evento "immatricolazione".

Immatricolazioni

(Dimensione Studente)

- La dimensione Studente e tutti i suoi dettagli.
- Gerarchia geografica e data di nascita.
- Non è conveniente normalizzare queste gerarchie o creare degli *outrigger*.
- Non si vogliono fare analisi particolari sulla data di nascita.
- Conviene evitare di referenziare una dimensione temporale esterna.

Immatricolazioni

(Dimensione Studente)

- Se lo Studente si immatricola per una prima volta ad un Corsi di Studi universitario, allora rientra in una coorte.
- Indicazione sintetica dell'Anno Accademico di immatricolazione.
- Altrimenti lo Studente è già collocato in una coorte.
- Una volta acquisita, la coorte a cui appartiene non cambia.
- In un certo senso SCD di tipo I.
- Attributo della dimensione.
- Per utilizzarlo poi come dimensione di analisi.

Immatricolazioni

(Dimensione Anno Accademico)

- Un Anno Accademico e i suoi attributi.
- Per le analisi sulle Immatricolazioni, non interessa invece la data solare in cui l'Immatricolazione è stata chiesta o ottenuta.

Immatricolazioni

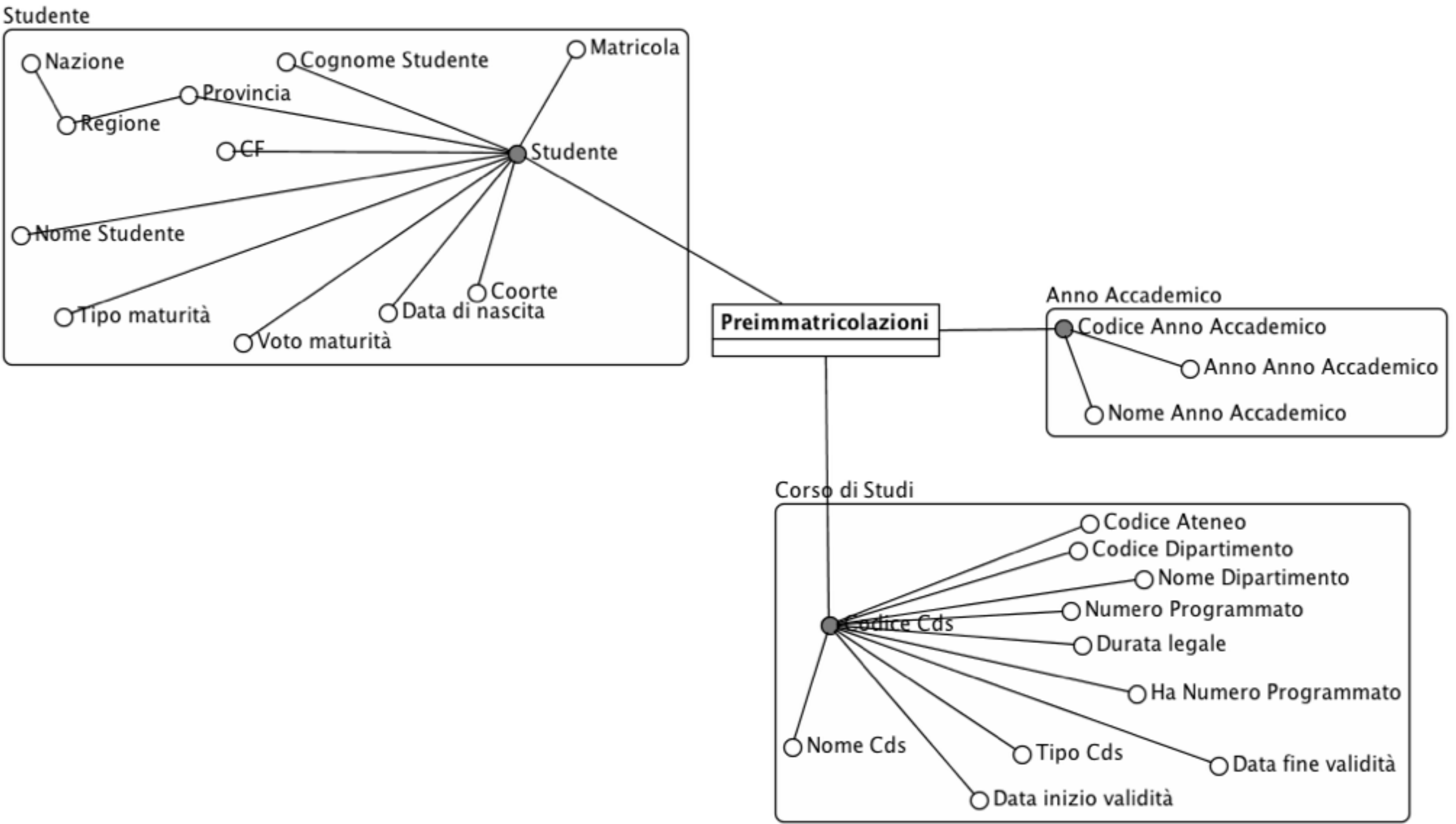
(Dimensione Tipo Immatricolazione)

- Qualifica il tipo di immatricolazione.
- Prima volta nel sistema universitario italiano (immatricolazione MIUR).
 - Ne consegue la collocazione in una coorte.
- Già iscritti in passato in altro ateneo.
- Già iscritti in passato in questo ateneo.

Immatricolazioni (Corso di Studi)

- Un Corso di Studi con le sue proprietà.
- Include le proprietà del dipartimento di riferimento e l'ateneo.
- Un Corso di Studi può avere o meno numero programmato.
- Questa proprietà può variare negli anni.
- Nelle analisi si deve tener conto del fatto che in un determinato istante di tempo il Cds è a numero programmato e qual è il numero massimo di studenti.
- Si vogliono anche fare analisi sul Cds indipendentemente dal fatto che negli anni è stato a numero programmato o meno.
- SCD tipo 2.
 - Data inizio e fine validità.
 - Cambia chiave surrogata, ma chiave business resta uguale.

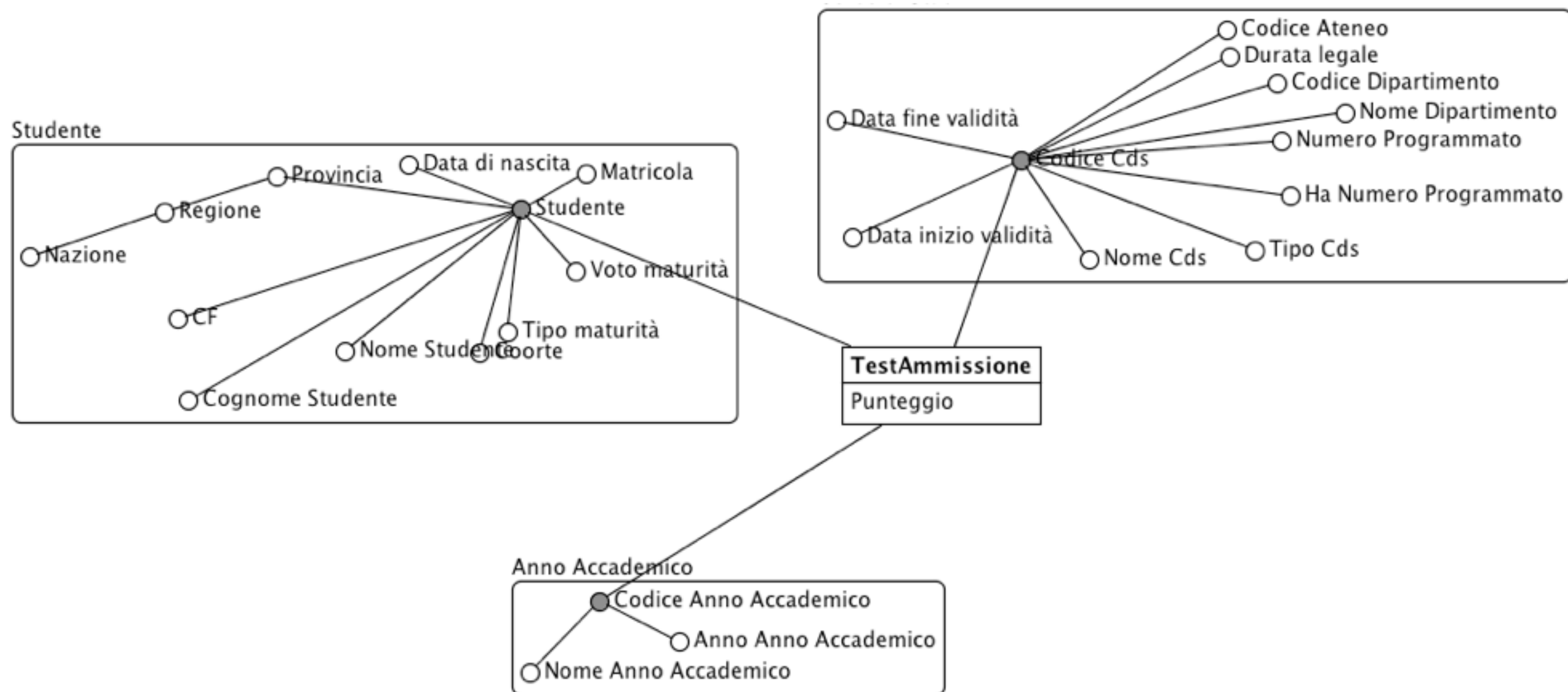
Preimmatricolazioni



Preimmatricolazioni

- La fact table rappresenta il fatto che un determinato Studente si è preimmatricolato in un determinato Anno Accademico in un Corso di Studi.
- Fact table degli eventi “preimmatricolazione”, descrive le transazioni.
- Al momento della preimmatricolazione, uno Studente appartiene ad una coorte solo se si è già precedentemente immatricolato (ad altro corso etc.).
- Si possono fare analisi congiunte del tipo “Preimmatricolati di cui immatricolati”
 - Con un join tra Immatricolazioni e Preimmatricolazioni su Studente, Anno Accademico, Corso di Studi.
 - Fact table derivata.

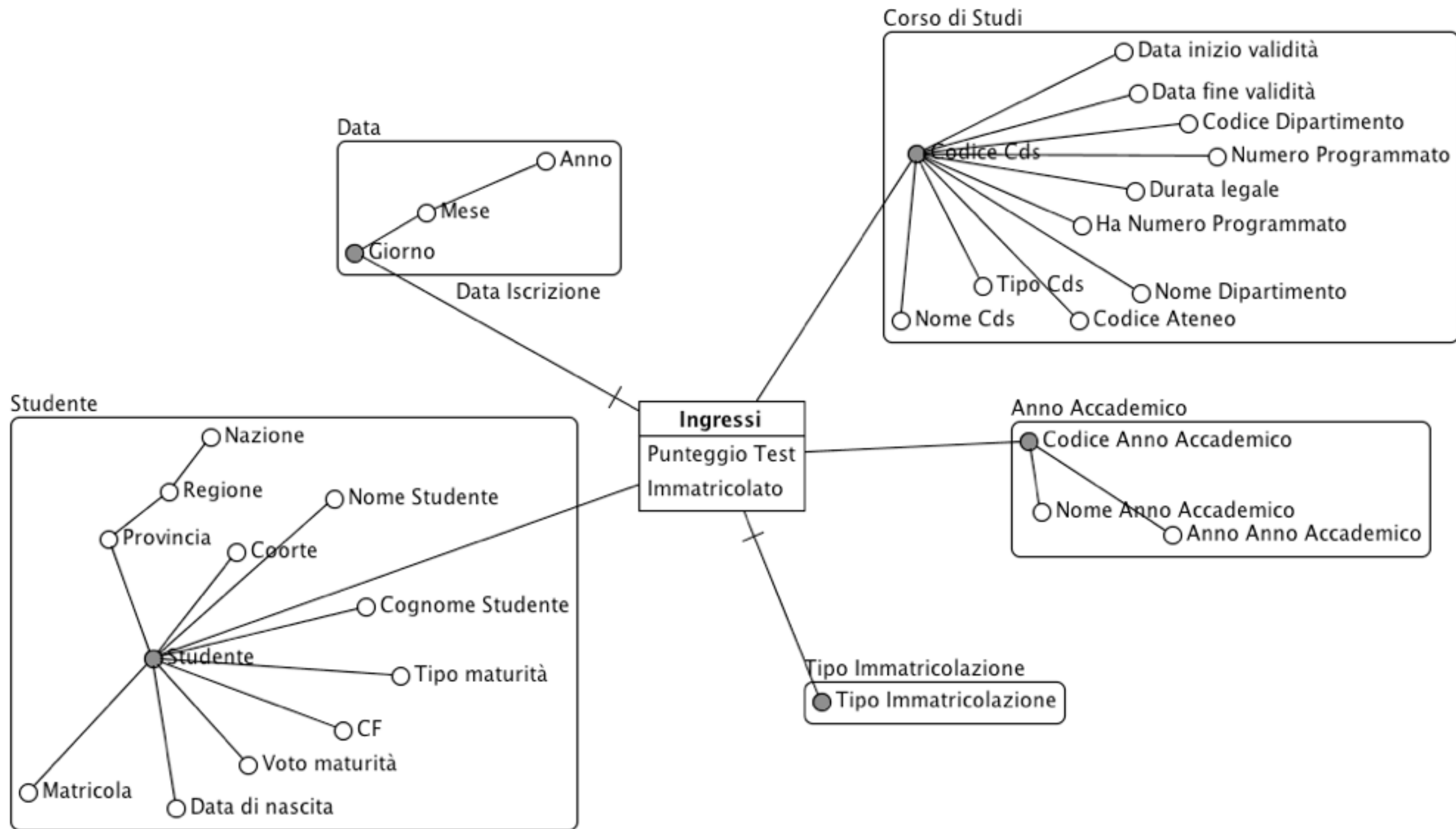
Test ammissione



Test ammissione

- La fact table rappresenta il fatto che uno Studente in un certo Anno Accademico ha sostenuto il test di ammissione per un certo Corso di Studi.
- Interessa misurare il punteggio conseguito al test.
- La media è l'aggregazione di default.
- Il test può essere sostenuto solo per corsi a numero programmato alla data del test.
- Non interessano comunque indagini sulla data solare in cui è stato svolto il test.
- Si possono fare analisi congiunte del tipo “Preimmatricolati di cui hanno superato il test”
- Con un join con Preimmatricolazioni su Studente, Anno Accademico, Corso di Studi.
- Fact table derivata.

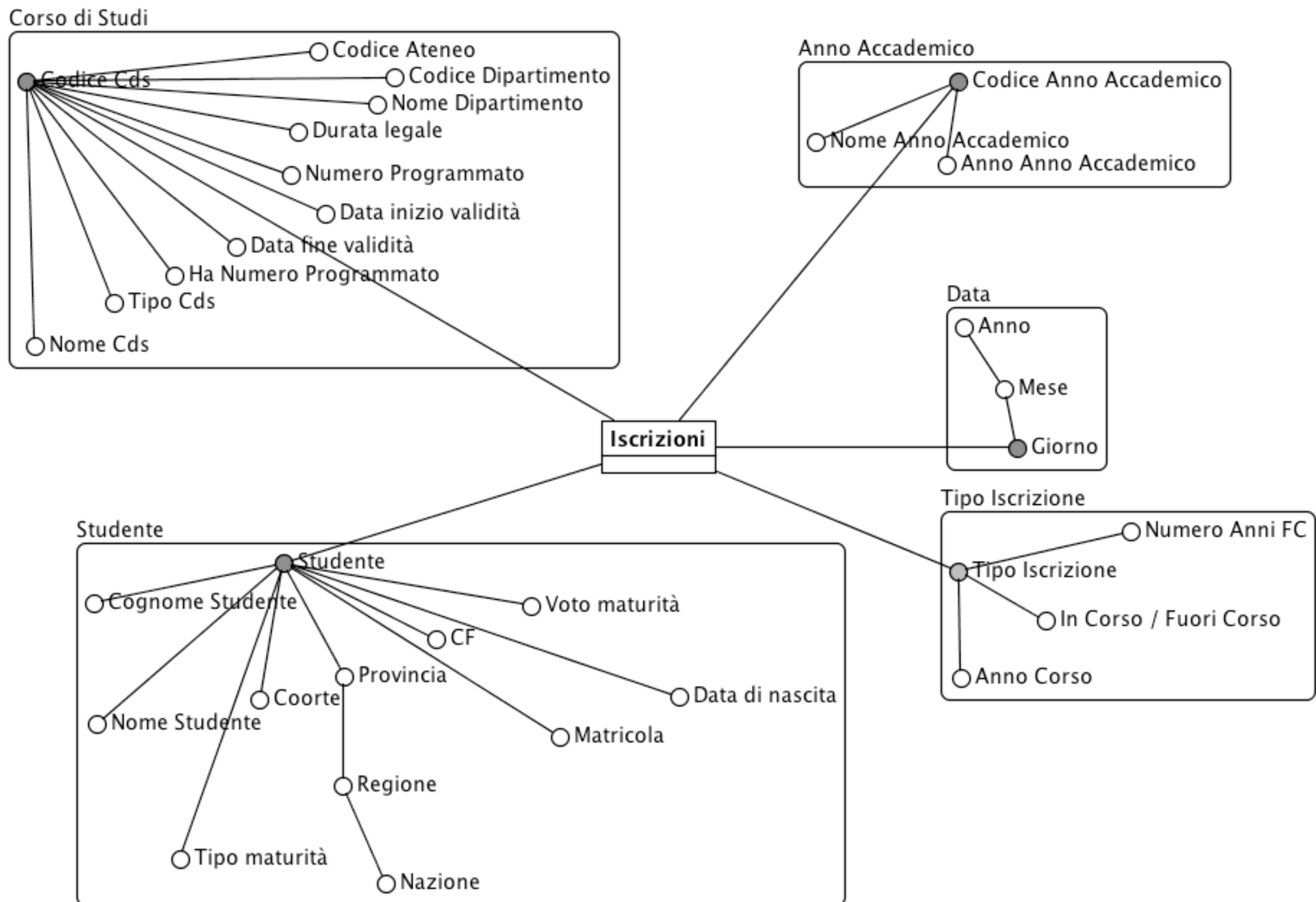
Preimmatricolazioni, Test e Immatricolazioni



Ingressi

- Può essere utile costruire una fact table derivata dalle tre precedenti.
- Memorizzando per ciascuno Studente che si è preimmatricolato:
 - Se ha sostenuto il test e con quale punteggio.
 - Se si è immatricolato (misura Immatricolato $\{0,1\}$)
 - Quale tipo di immatricolazione.
 - Eventualmente la data solare di iscrizione.
- Tipo Immatricolazione e Data Iscrizione sono, naturalmente, opzionali.
- Anche altre modellazioni:
 - Ad esempio senza misura “Immatricolato”, ma con un tipo di immatricolazione “Non immatricolato”.
- Fact table di tipo *transaction* che però riassume più transazioni (alla fine dell’ultima).
- Si poteva modellare come *accumulating snapshot*, ma il processo preimmatricolazione-iscrizione è molto breve e le analisi si assumono essere a posteriori.
- In effetti, è la Fact table delle transazioni delle iscrizioni.

Iscrizioni



Iscrizioni (Fact table)

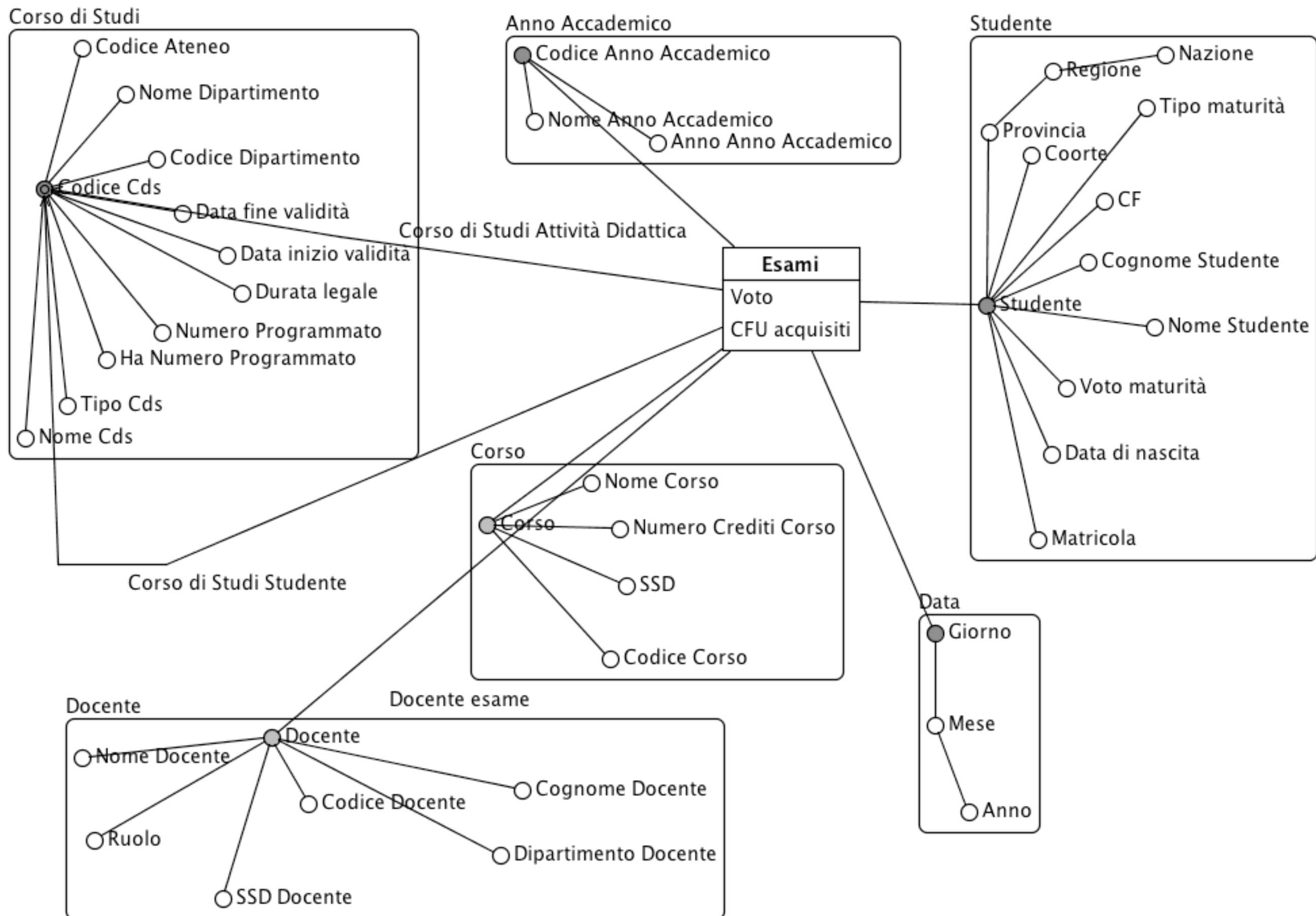
- Riporta la situazione delle Iscrizioni nell'ateneo in un determinato istante di tempo.
- La dimensione Data rappresenta l'istante di tempo a cui si riferiscono i fatti.
- Conteggi non additivi su Data.
- Per una certa Data, Per un certo Anno Accademico, uno Studente risulta iscritto ad un Corso di Studi.
- Fact table di tipo *periodic snapshot*
 - Non si rappresentano le transazioni di iscrizione, ma varie fotografie (alla granularità dello Studente) in vari istanti di tempo.
 - Complementata dalla fact table Ingressi che, invece, mantiene gli eventi veri e propri di iscrizione.

Iscrizioni

(Tipo iscrizione)

- Uno studente è iscritto ad un particolare anno di corso (primo, secondo, etc.)
- Uno studente può essere in corso o fuori corso.
 - Se fuori corso, allora lo è di un certo numero di anni, altrimenti 0.
- La dimensione raccoglie tutte le combinazioni (ragionevolmente possibili) e permette di qualificare un'iscrizione ad un certo istante di tempo (*junk dimension*).
- (I anno, IC, 0), (I anno, FC, 1), (I anno, FC, 2), (II anno, IC, 0).

Esami



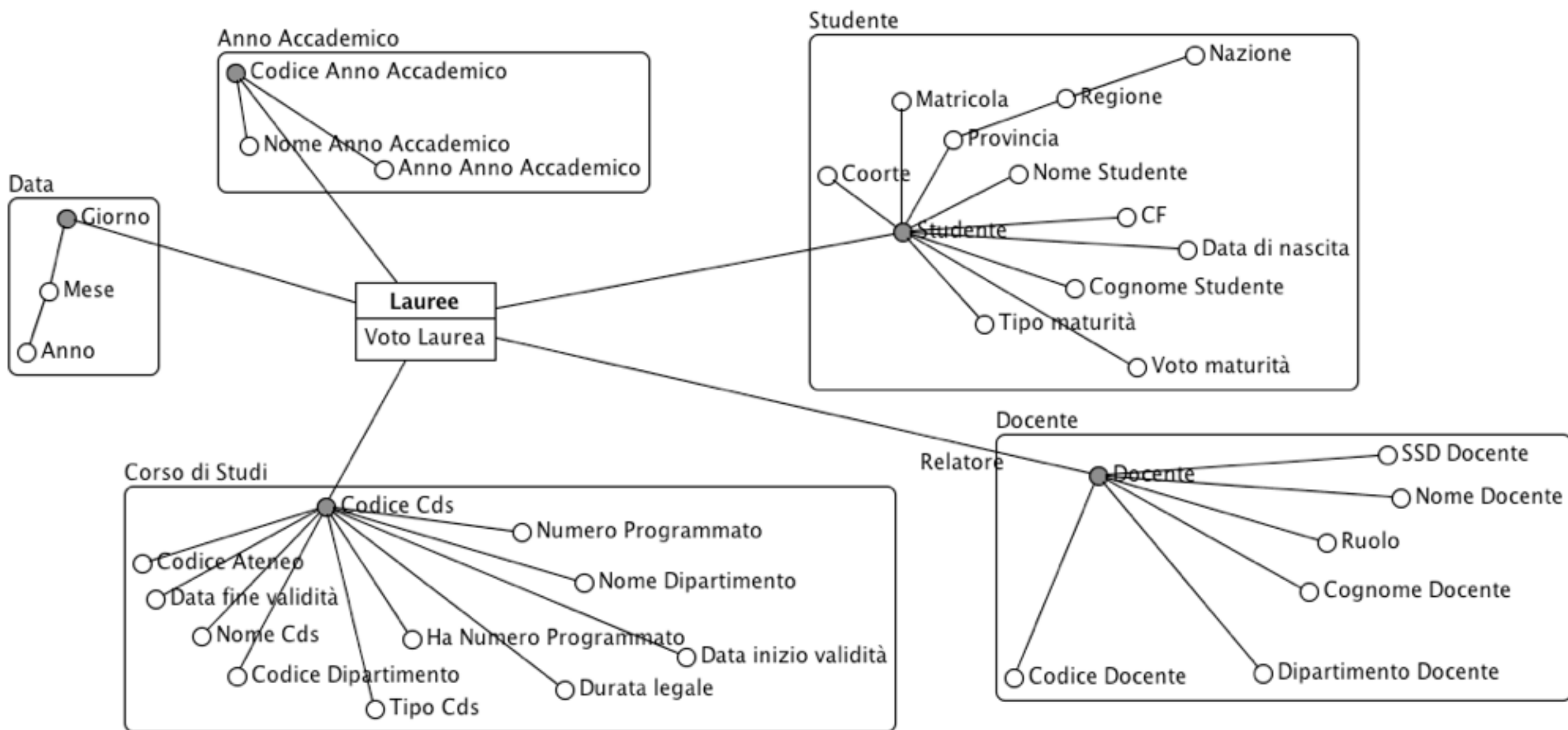
Esami (fact table)

- Rappresenta il fatto che uno **Studiante** ha sostenuto un **Esame** di un determinato **Corso** in una certa **Data** per un **Anno Accademico**.
- Interessano anche i **Cds** e il **Docente** che presiedeva.

Esami (Cds)

- Un esame è relativo a un Corso, che può però appartenere a più Cds.
- Un'edizione di un corso, cioè in un determinato Anno Accademico, è associata ad una Attività Didattica.
- L'Attività Didattica è invece erogata nell'ambito di un unico Cds.
- Interessa questo Cds (Cds Attività Didattica) e il Cds a cui lo Studente è iscritto al momento dell'esame.
- Quest'ultimo si poteva anche ricavare da Iscrizioni.

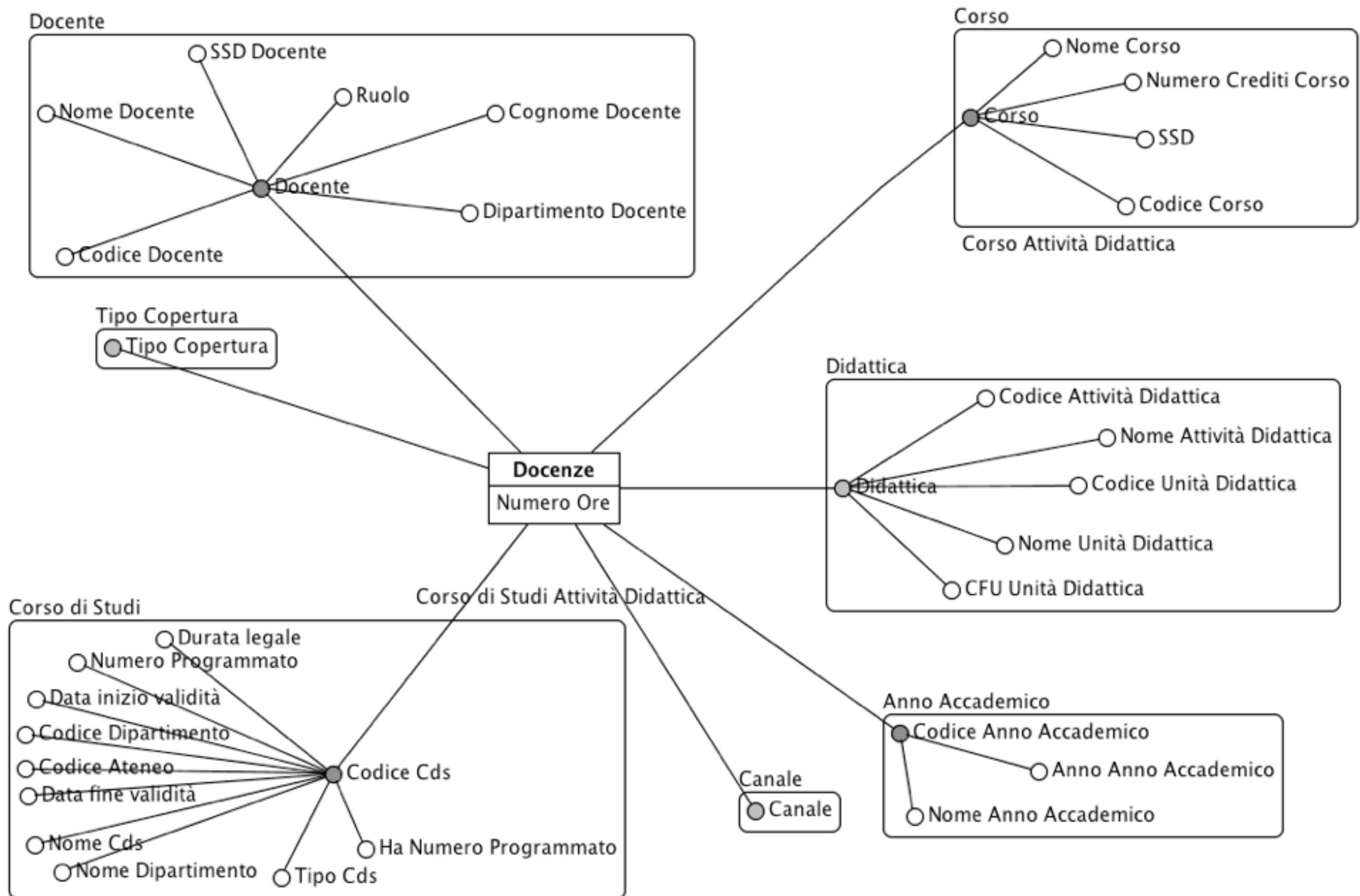
Lauree



Lauree

- Fact table di tipo *transaction* che rappresenta il fatto che uno Studente, ad una Data, in un Anno Accademico ... etc, si è laureato.
- Interessa il voto di laurea.
- Naturalmente non additivo.

Docenze



Docenze

- Fact table di tipo *transaction*.
- Rappresenta il fatto che un Docente ha effettuato un insegnamento.
- Sono di interesse il numero di ore spese nella docenza.

Docenze (Cds)

- Un corso, in una sua edizione (quindi in un Anno Accademico), corrisponde ad una Attività Didattica.
- L'Attività Didattica può essere partizionata in una o più Unità Didattiche (ad es: modulo I, modulo II).
- Corsi diversi corrispondono, nelle rispettive edizioni, ad Attività Didattiche diverse.
- Queste Attività Didattiche possono condividere Unità Didattiche.
- Corsi diversi possono appartenere a Cds diversi (e, in generale, un Corso può appartenere a più Cds).
- Qui è di interesse nell'ambito di quale Unità Didattica è effettivamente avvenuta la docenza.
- Anche se condivisa tra più AD, una UD è erogata una sola volta. Qui è di interesse la AD che ha effettivamente erogato la UD.
- La Docenza è effettiva ed è relativa ad una specifica edizione di un Corso, che fa riferimento ad una sola AD. Questa AD è erogata nell'ambito di un solo Cds.

Docenze (Didattica)

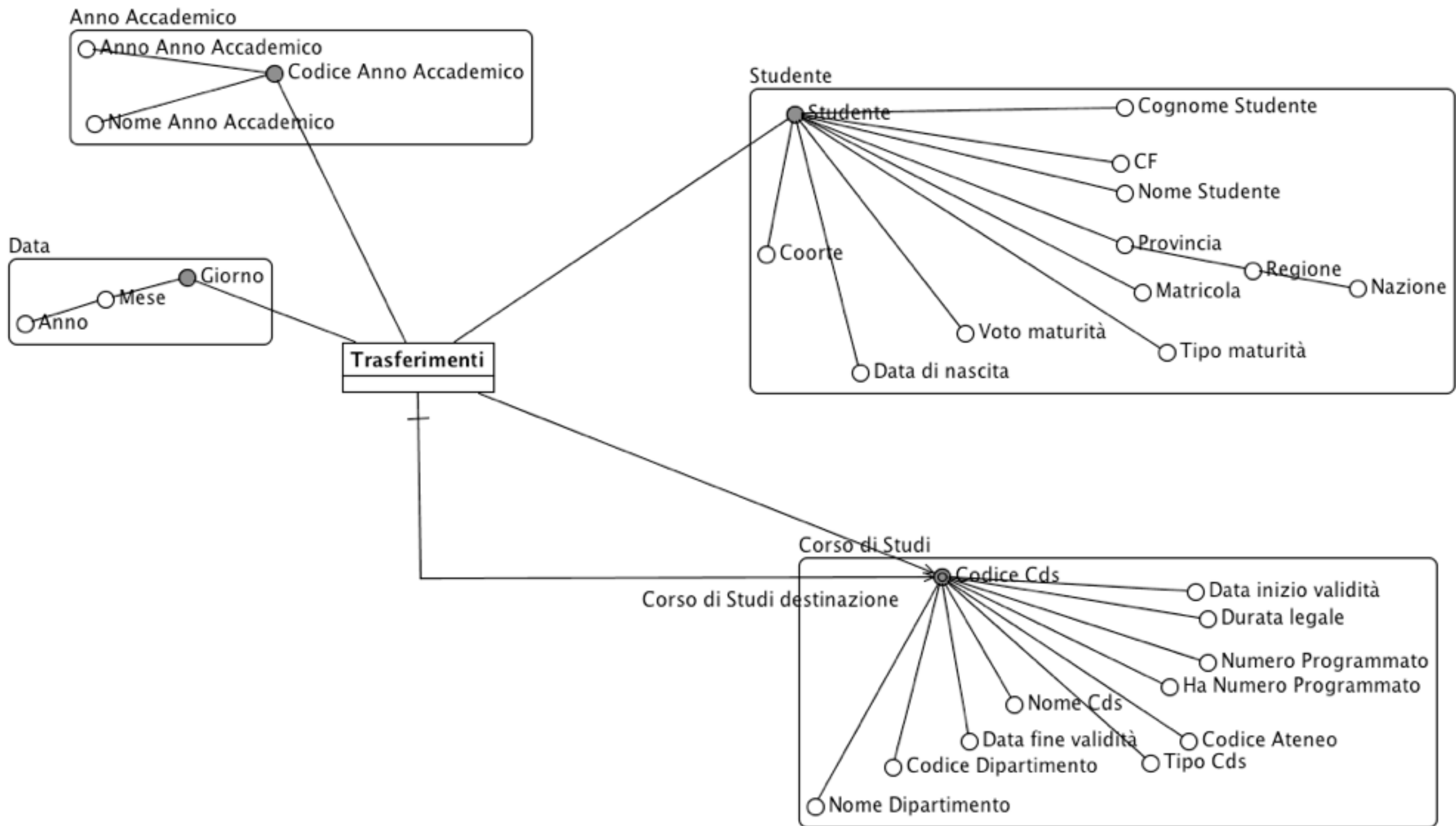
- Dimensione che tiene conto di:
 - Attività Didattiche
 - I CFU sono quelli dell'unico corso relativo
 - Unità Didattiche
 - CFU delle Unità Didattiche
- Non esiste una vera e propria gerarchia UD — AD.
 - Fissato un Anno Accademico e un Corso (quindi un'edizione di corso), allora una UD appartiene ad una sola AD.
- Si memorizzano quindi le possibili combinazioni UD — AD e dati Anno Accademico e Corso, si può scegliere univocamente.
- Scelta l'AD, si può determinare anche il Cds del Corso.

Docenze

(altre dimensioni)

- Tipo Copertura
 - “Compito didattico”, “Docenza esterna”, etc.
- Canale
 - Eventuale suddivisione in canali.
 - “Unico”.

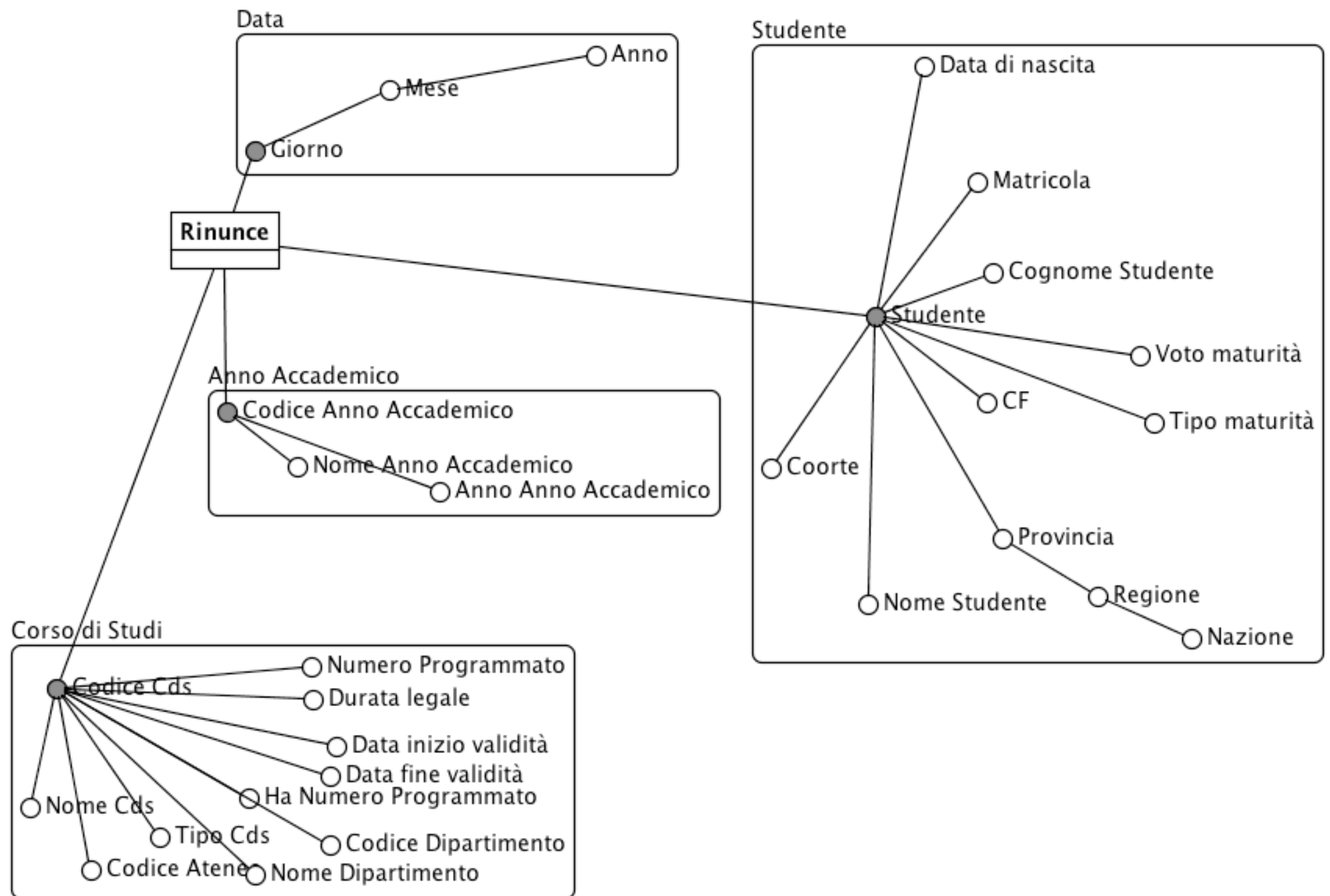
Trasferimenti



Trasferimenti

- Fact table di tipo *transaction*
- Rappresenta il fatto che uno Studente ha richiesto (e ottenuto) in una certa data il trasferimento ad altro Cds.
- Interessa il Cds di provenienza (di questo ateneo).
- Il Cds di destinazione è facoltativo.
 - Si rappresenta solo se di questo ateneo.
- Interessa l'Anno Accademico di riferimento.

Rinunce

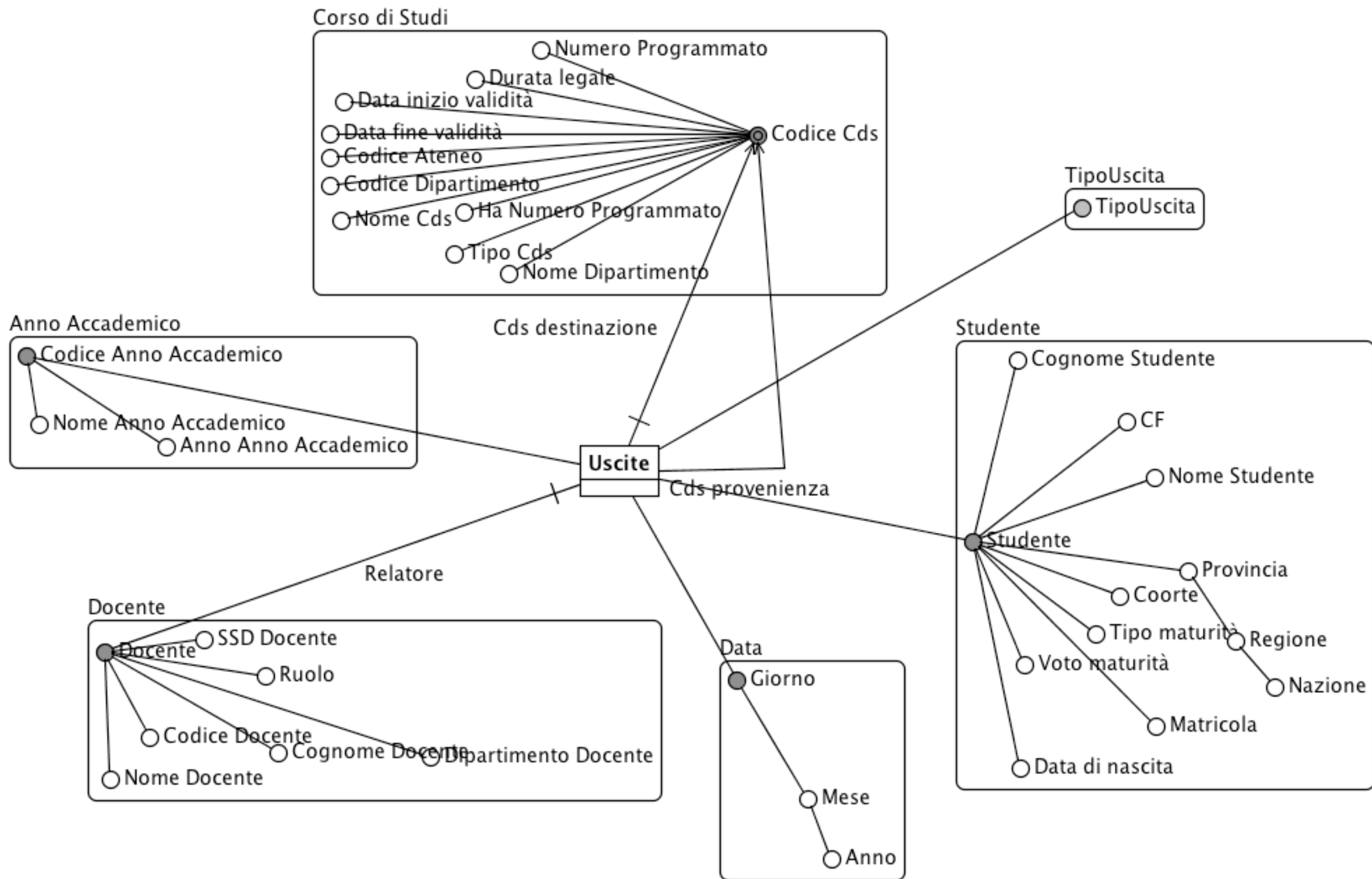


Rinunce

- Fact table di tipo *transaction*.
- Rappresenta il fatto che uno Studente ha richiesto (e ottenuto) in una certa data la rinuncia agli studi.
- Interessano il Cds di provenienza e l'Anno Accademico di riferimento.

Uscite

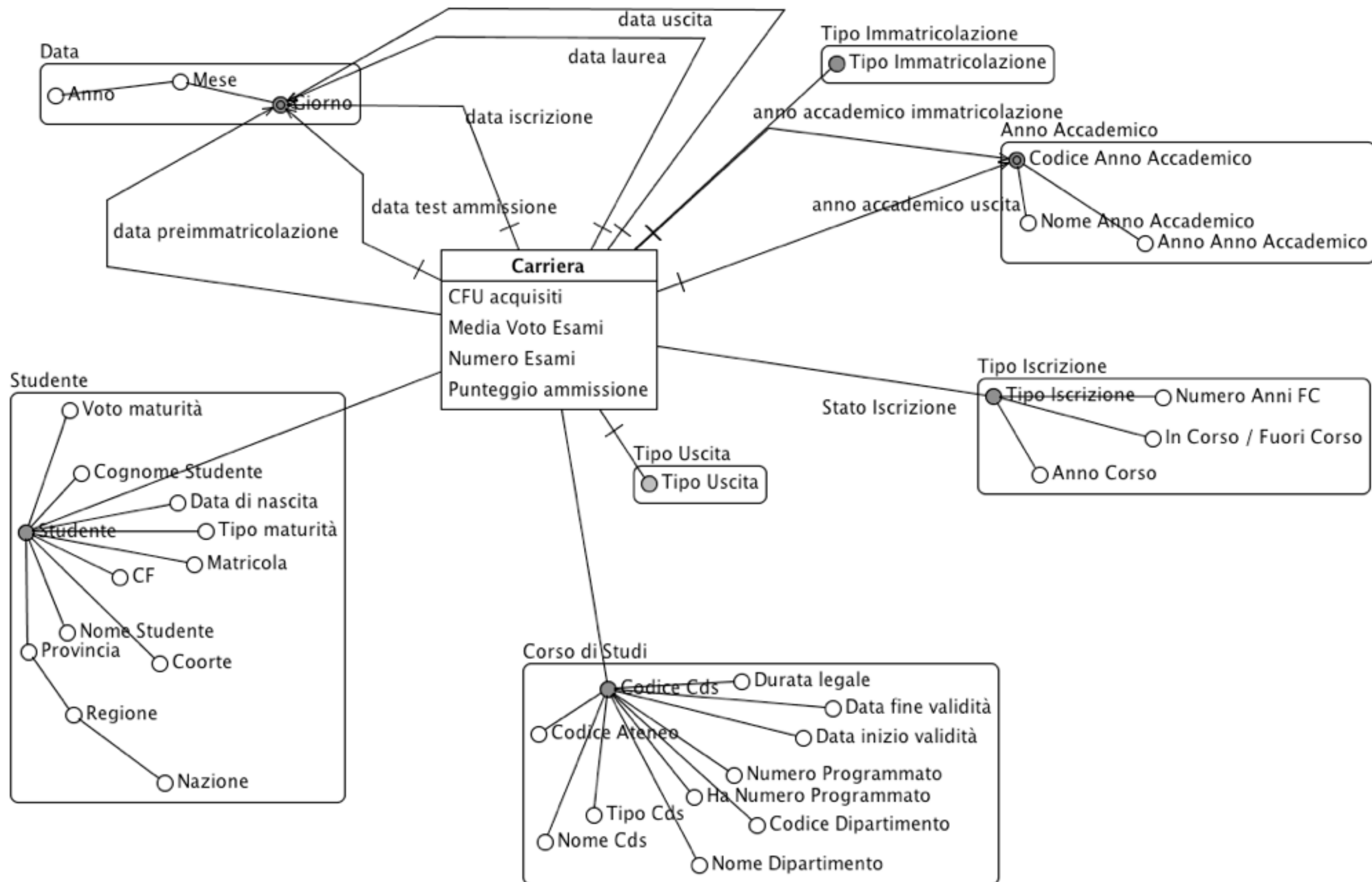
- Conviene progettare una fact table che metta insieme Trasferimenti, Rinunce e Lauree per modellare le uscite?
- Forse, visti i report.
- Attenzione però a creare fact table troppo generiche.
 - Troppo sparse, misure nulle, etc.
- In questo caso però:
 - Factless.



Uscite (Tipo di uscita)

- Dimensione che caratterizza il motivo di uscita da un Cds.
- Trasferimento altro Cds stesso ateneo
- Trasferimento altro Ateneo
 - In questo caso ha senso Cds destinazione.
- Laurea
 - In questo caso ha senso Relatore.

Carriera



Carriera

- Fact table di tipo *accumulating snapshot*.
- Ha lo scopo di monitorare nel continuo l'andamento complessivo della carriera degli studenti.
- Al momento dell'ingresso viene inserito un record che è man mano aggiornato nel tempo.
- Le foreign key verso le dimensioni vengono aggiornate quando significativo nel processo.
- I dettagli (sulle lauree, sugli esami, trasferimenti, etc.) non sono riportati, ma si rimanda agli schemi dimensionali specifici.

Carriera

(alcune dimensioni)

- Tipo Uscita
- Tipo Immatricolazione
- Stato Iscrizione
 - Può aver senso anche un codice per “non rinnovata”.
- Forte uso delle dimensioni condivise.

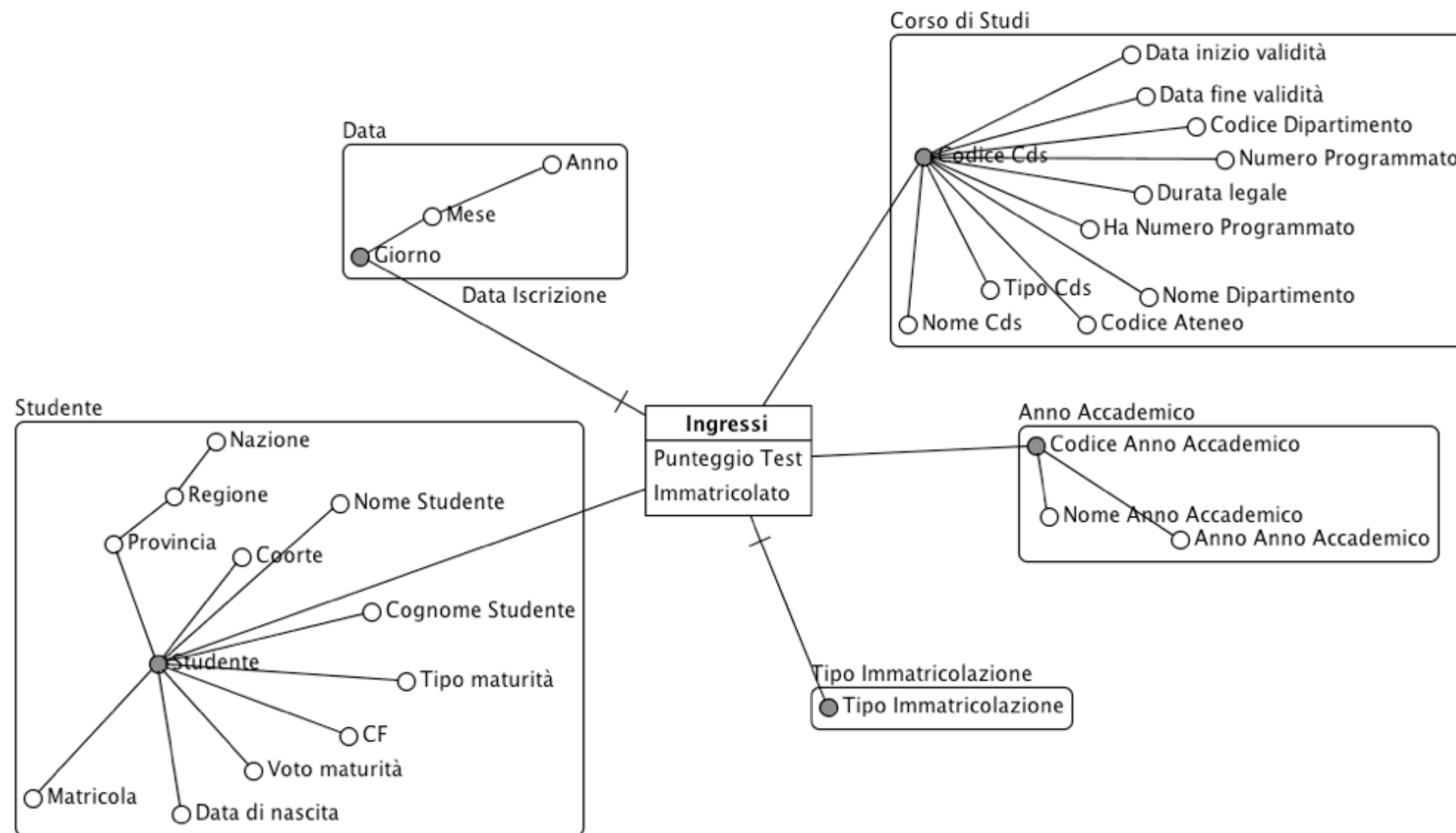
Carriera (misure)

- CFU acquisiti:
 - additiva ma poco significativa
 - aggregabile, ad esempio, per media.
- Media voto esami:
 - non additiva
 - non aggregabile direttamente
 - si può combinare con il numero di esami e aggregare per media.
- Punteggio ammissione:
 - aggregabile, ad esempio, per media.
- Numero esami:
 - additiva ma poco significativa
 - aggregabile, ad esempio, per media.

Test del carico di lavoro

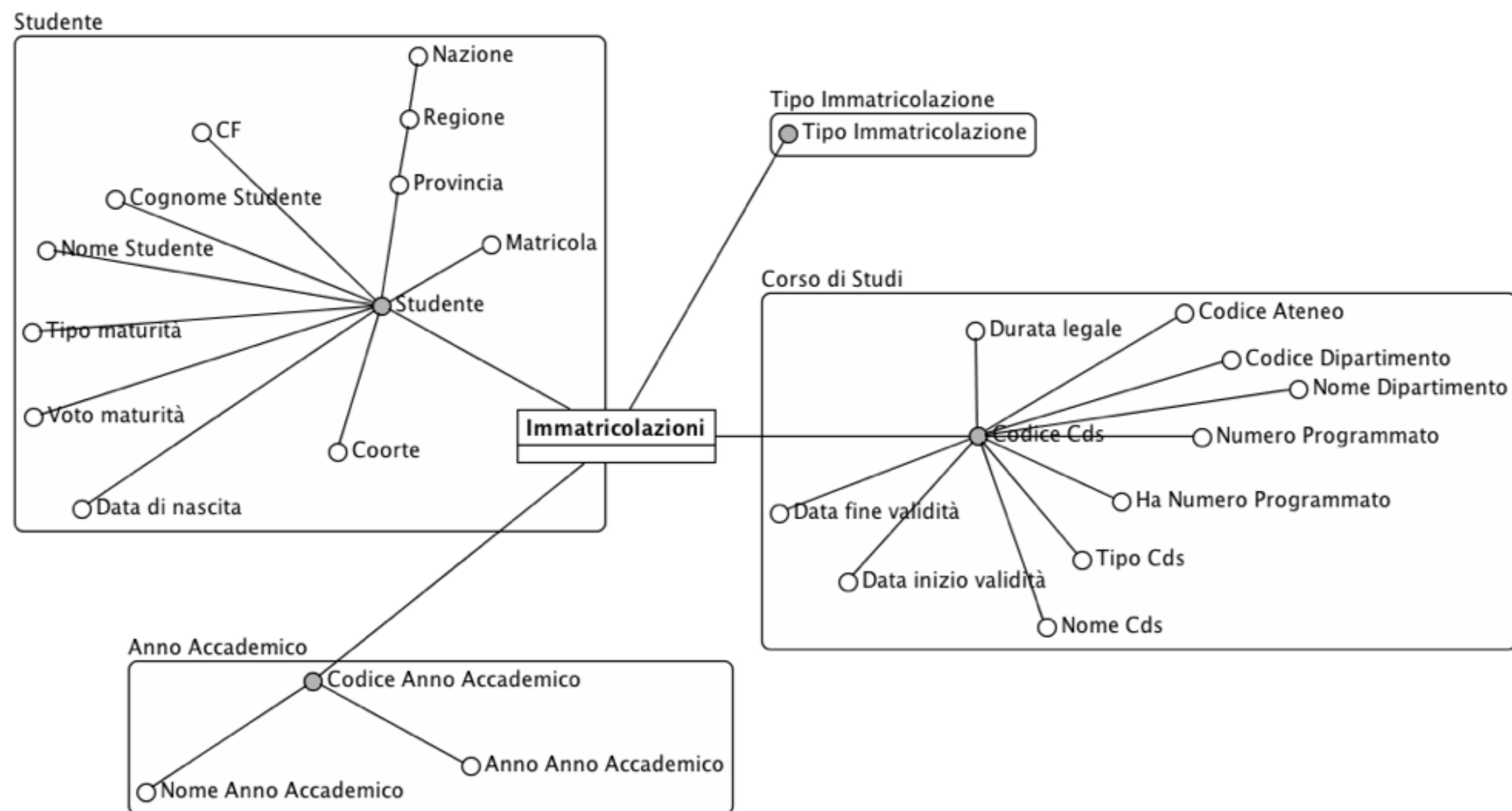
- AA COORTE
- POSTI O UTENZA SOST
- DOMANDE PREIMM
- DI CUI PARTECIPANTI TEST
- DI CUI IMMATRICOLATI
- MEDIA PUNTEGGIO
- MEDIA PUNTEGGIO IMM

RC.I



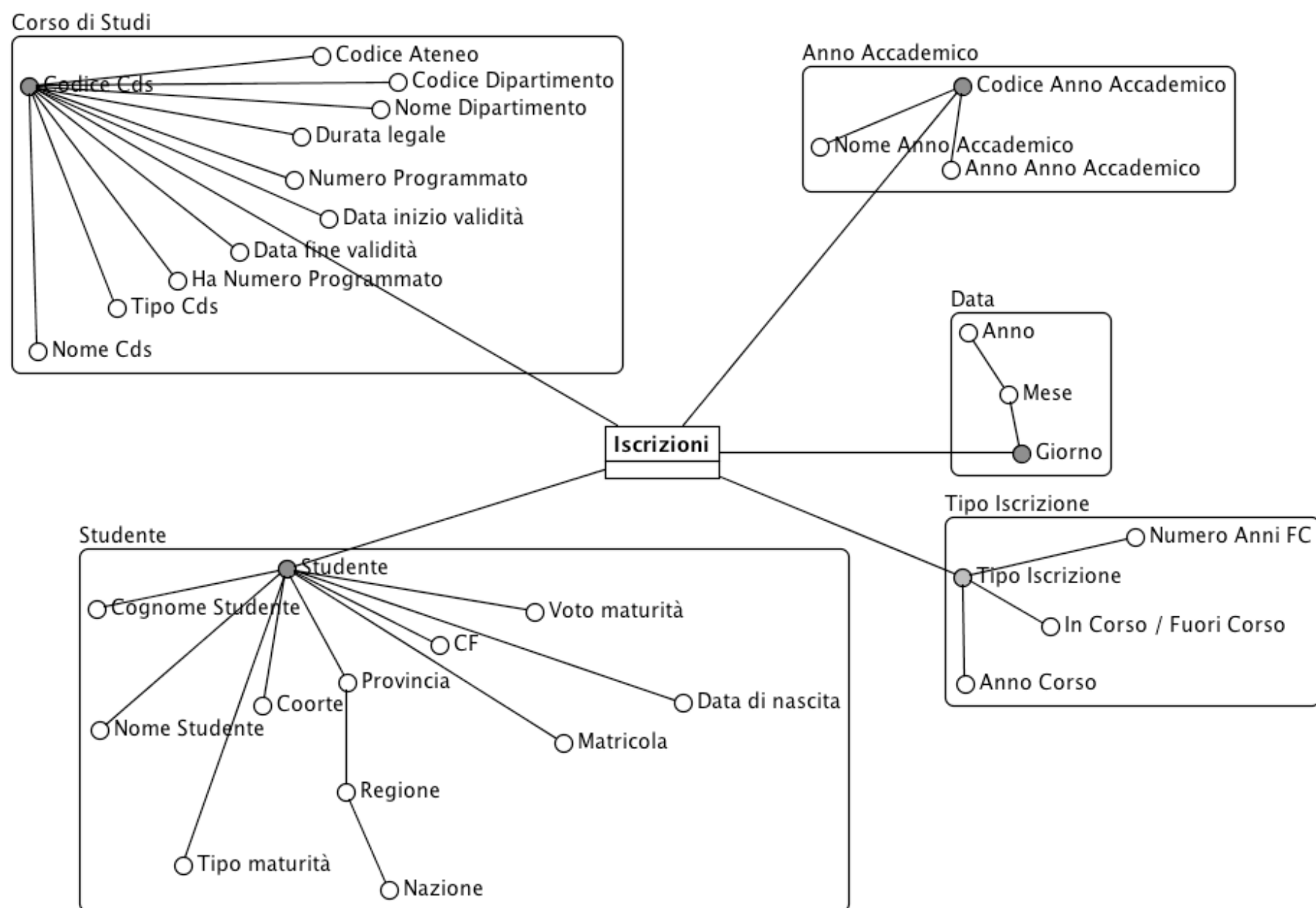
- AA COORTE
- IMM TOT
- PERC TRIENNALE ALTRO ATENEO
- ISCR II ANNO INCORSO
- ISCR I ANNO RIP
- TASSO ABBANDONO

RC.2



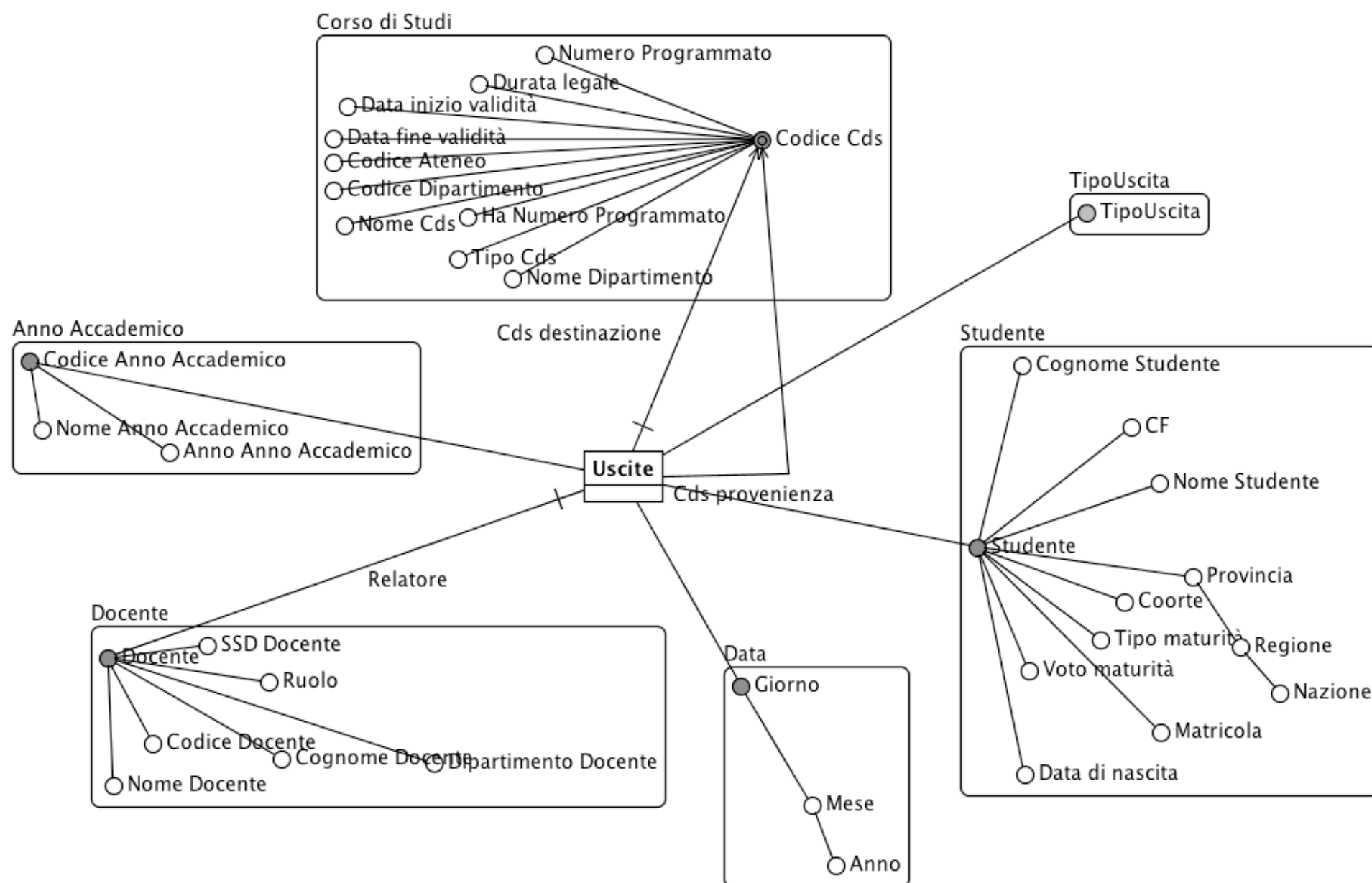
- AA COORTE
- IMM TOT
- PERC TRIENNALE ALTRO ATENEO
- ISCR II ANNO INCORSO
- ISCR I ANNO RIP
- TASSO ABBANDONO

RC.2



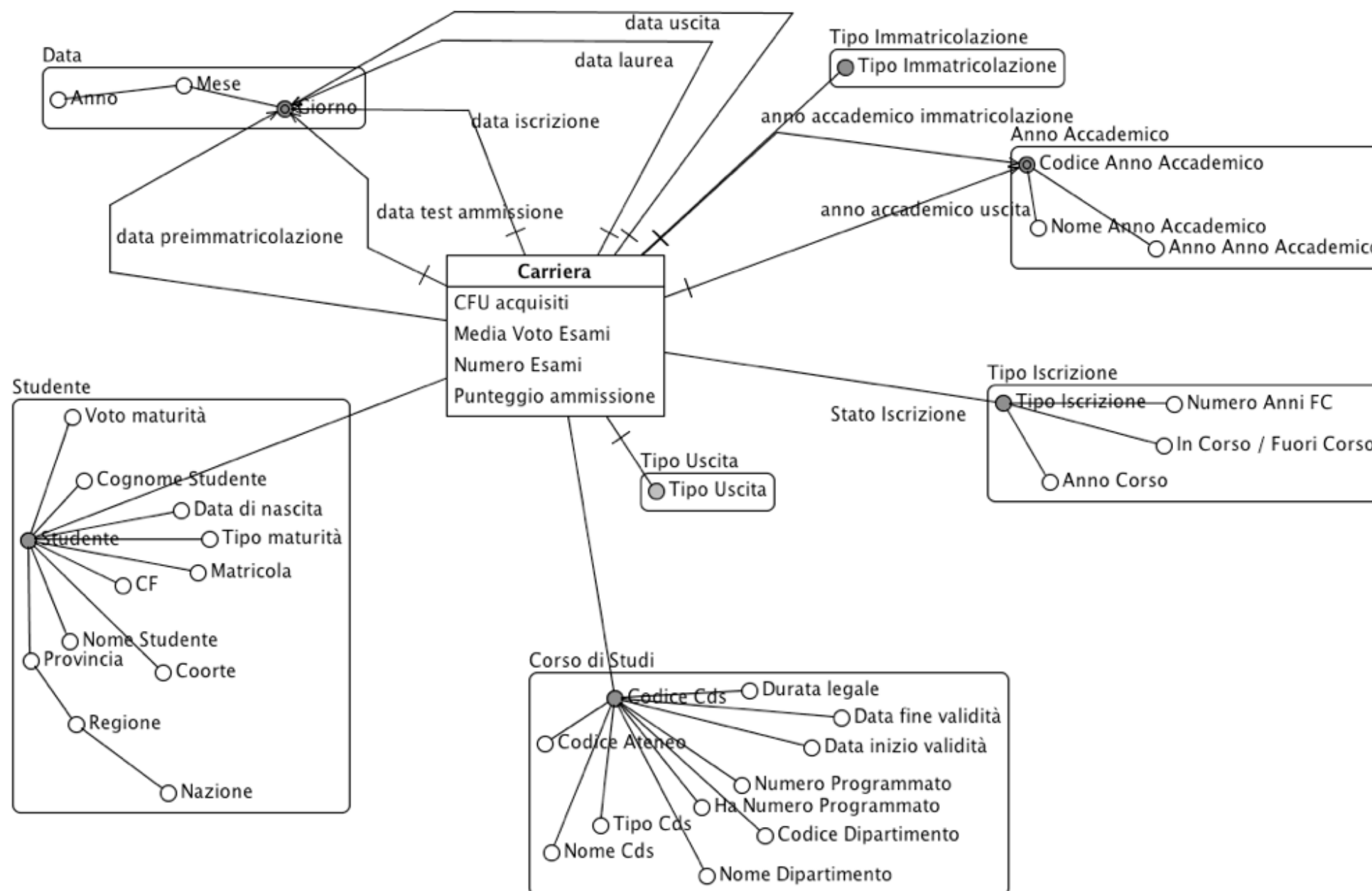
- AA COORTE
- IMM TOT
- PERC TRIENNALE ALTRO ATENEO
- ISCR II ANNO INCORSO
- ISCR I ANNO RIP
- TASSO ABBANDONO

RC.2



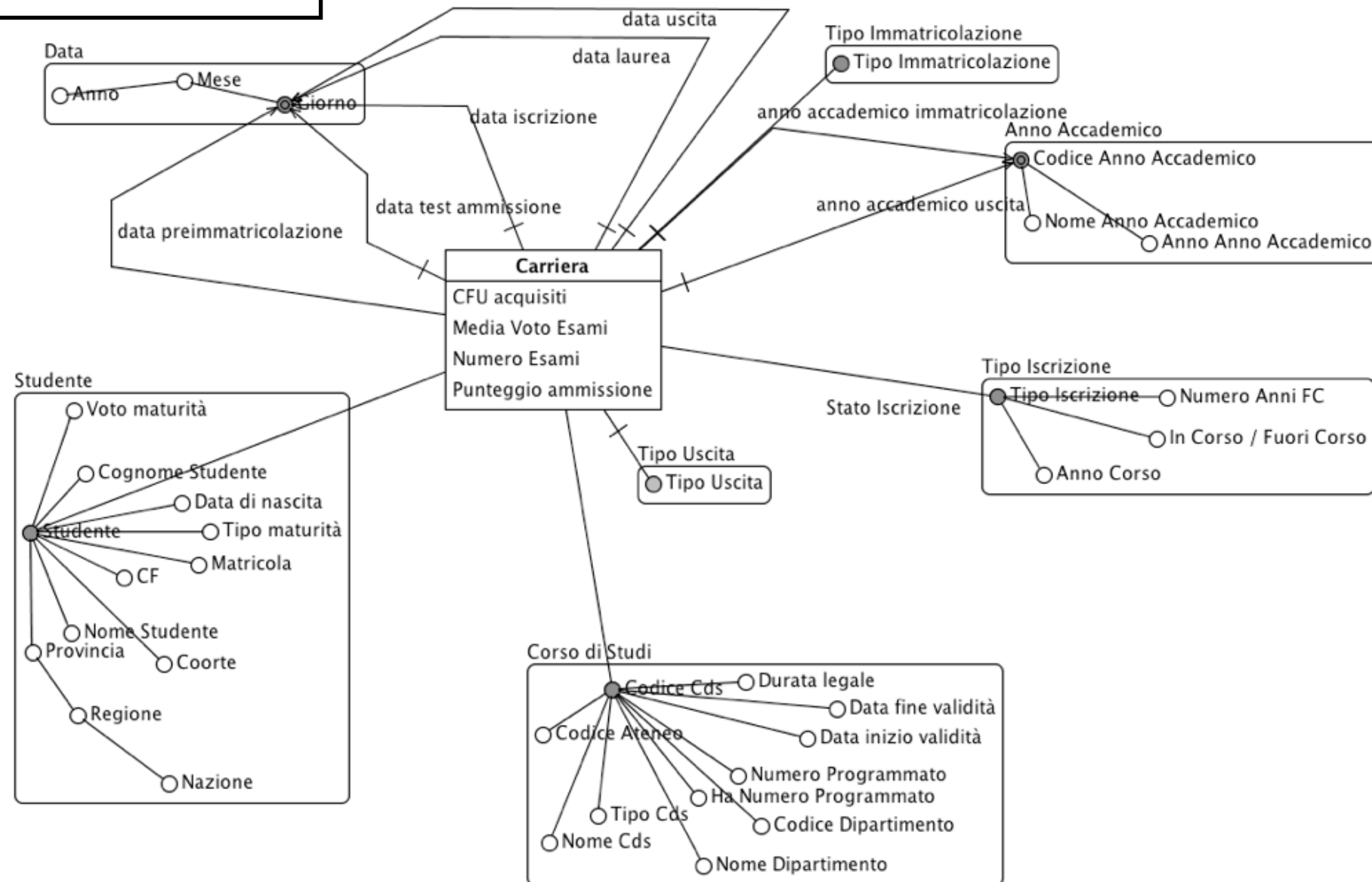
- AA COORTE
- IMM TOT
- PERC TRIENNALE ALTRO ATENEO
- ISCR II ANNO INCORSO
- ISCR I ANNO RIP
- TASSO ABBANDONO

RC.2



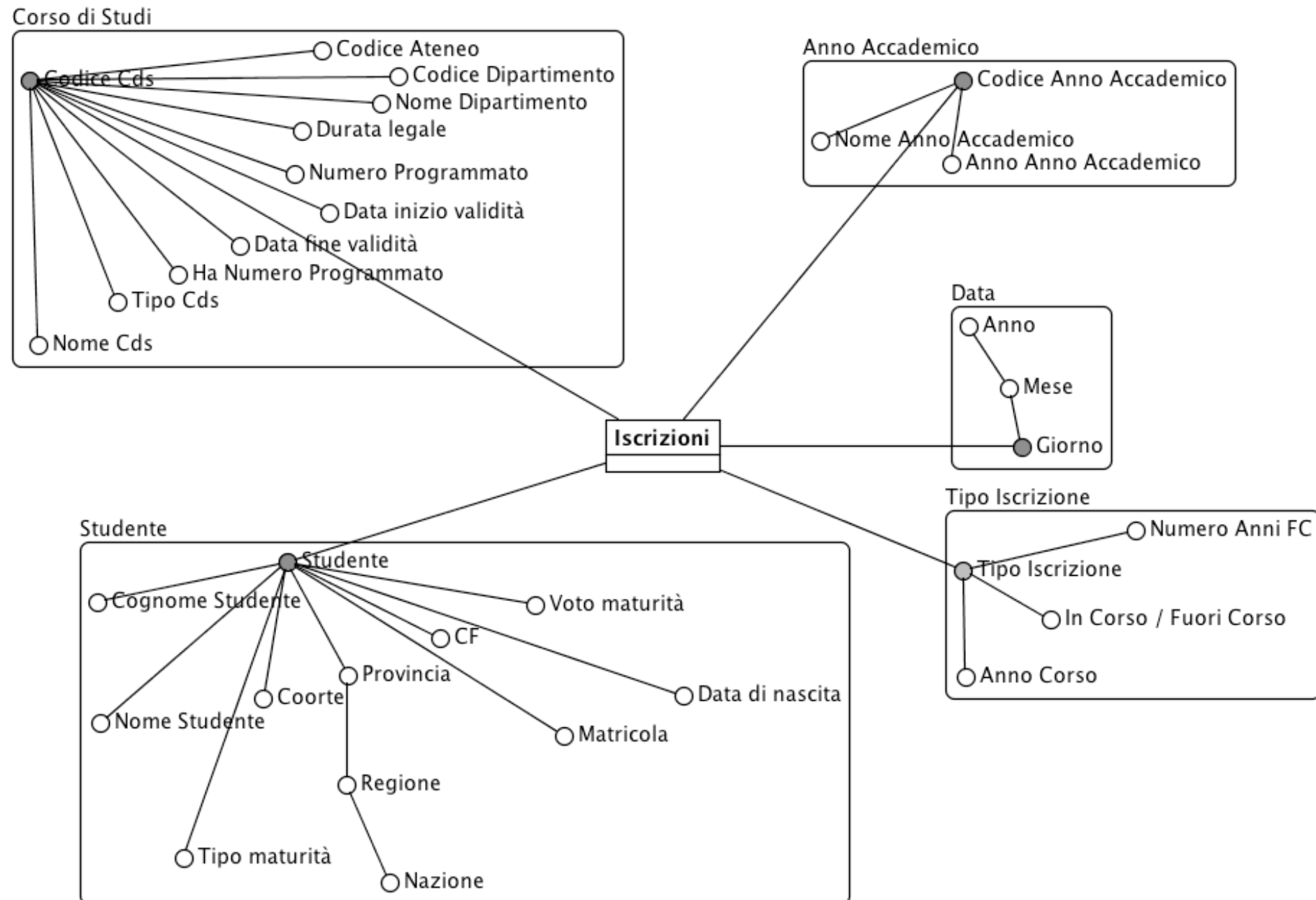
- AA COORTE
- IMM TOT
- PCT PASS INT
- PCT PASS EST
- PCT RIN
- PCT TRASF
- PCT MANC RINN
- PCT LAU

RC.3



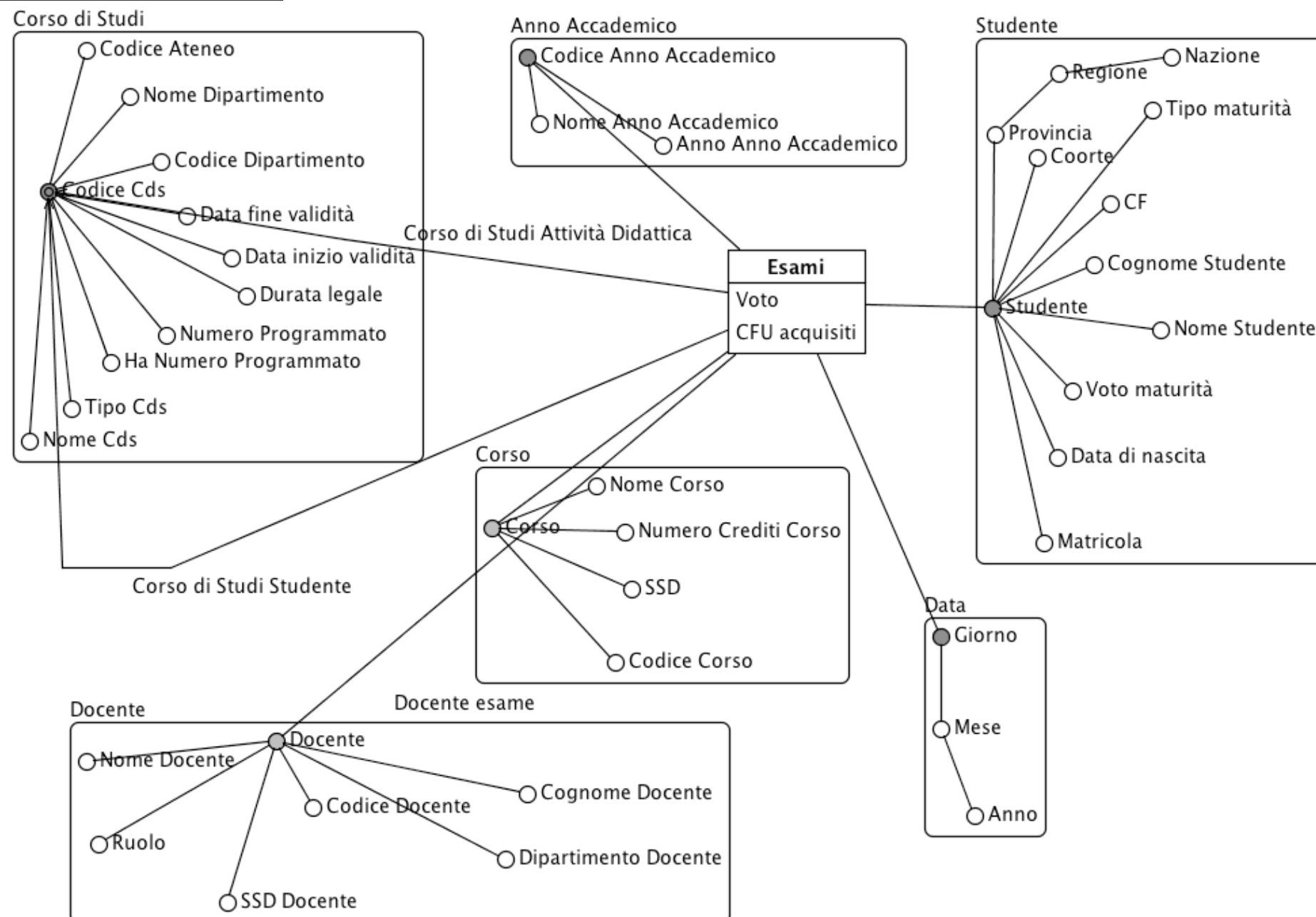
- AA COORTE
- IMM TOT
- N STUD INCORSO
- N STUD RIP
- PCT 0
- PCT FINO 10
- PCT FINO 20
- ...
- PCT OLTRE 100

RC.4.1



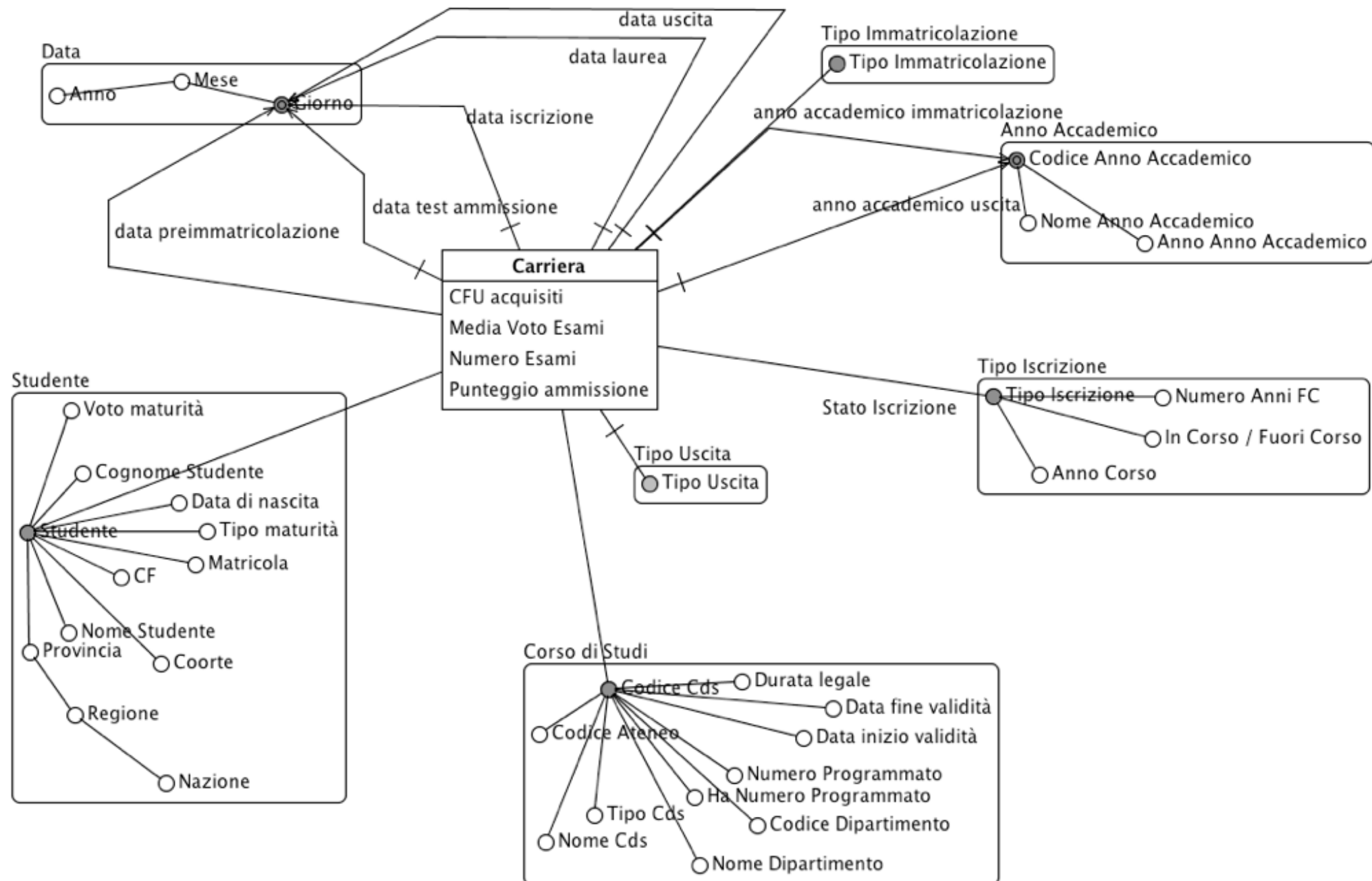
- AA COORTE
- IMM TOT
- N STUD INCORSO
- N STUD RIP
- PCT 0
- PCT FINO 10
- PCT FINO 20
- ...
- PCT OLTRE 100

RC.4.1



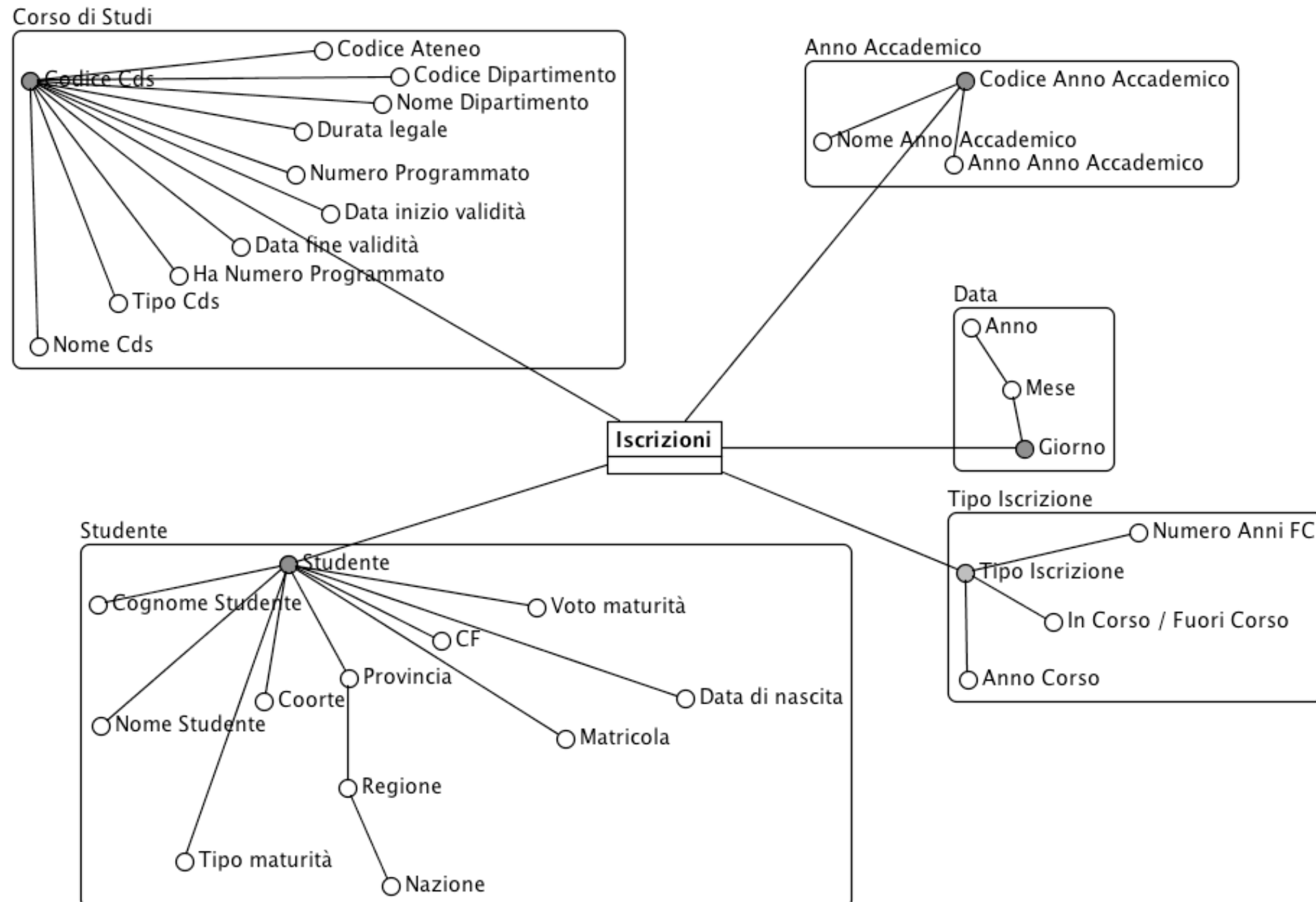
- AA COORTE
- IMM TOT
- N STUD INCORSO
- N STUD RIP
- PCT 0
- PCT FINO 10
- PCT FINO 20
- ...
- PCT OLTRE 100

RC.4.1



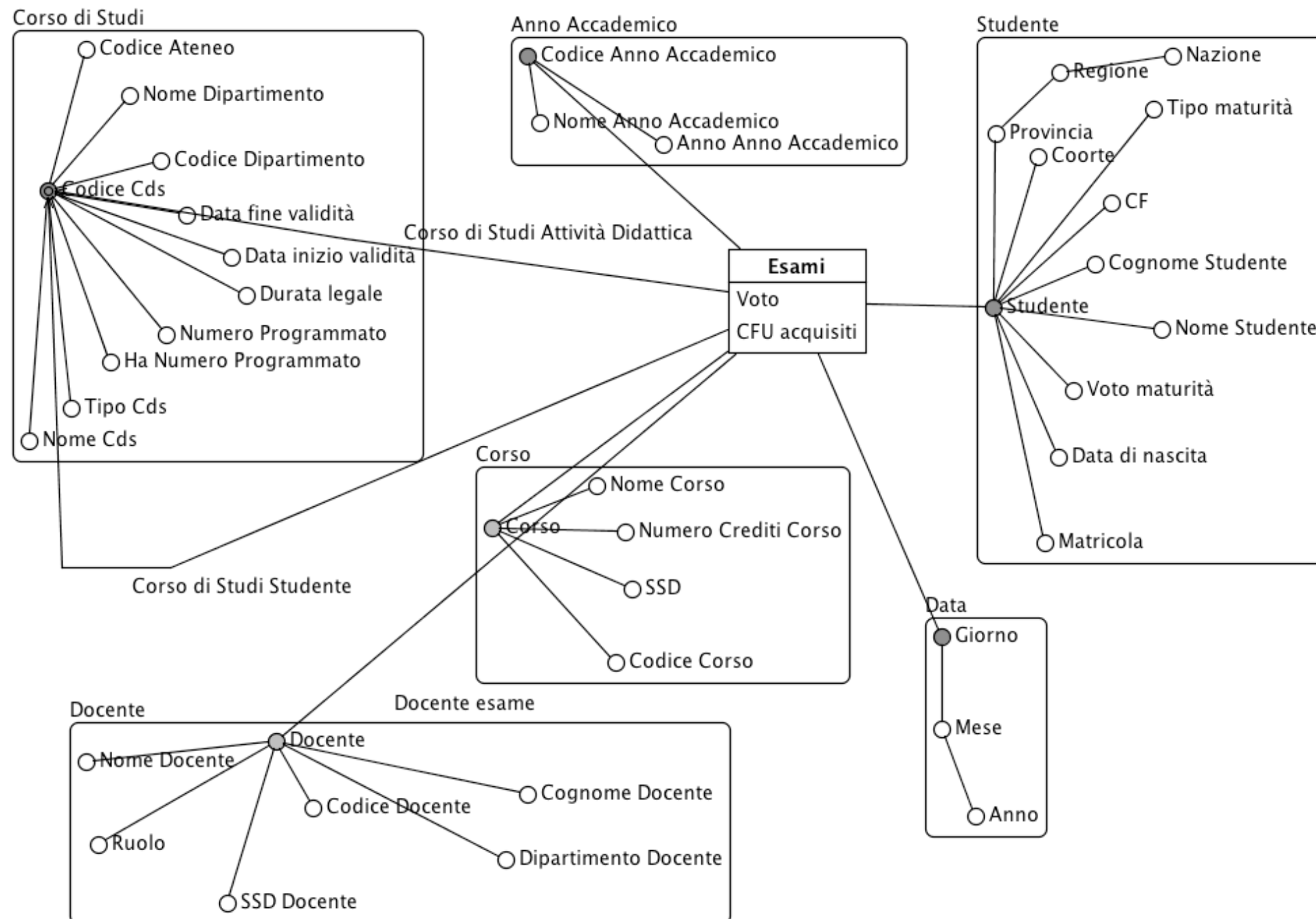
- AA COORTE
- IMM TOT
- N STUDENTI SOPRAVV
- MEDIA CFU STESSO CDS II ANNO
- MEDIA CFU IN CORSO
- MEDIA CFU RIPETENTI

RC.4.2



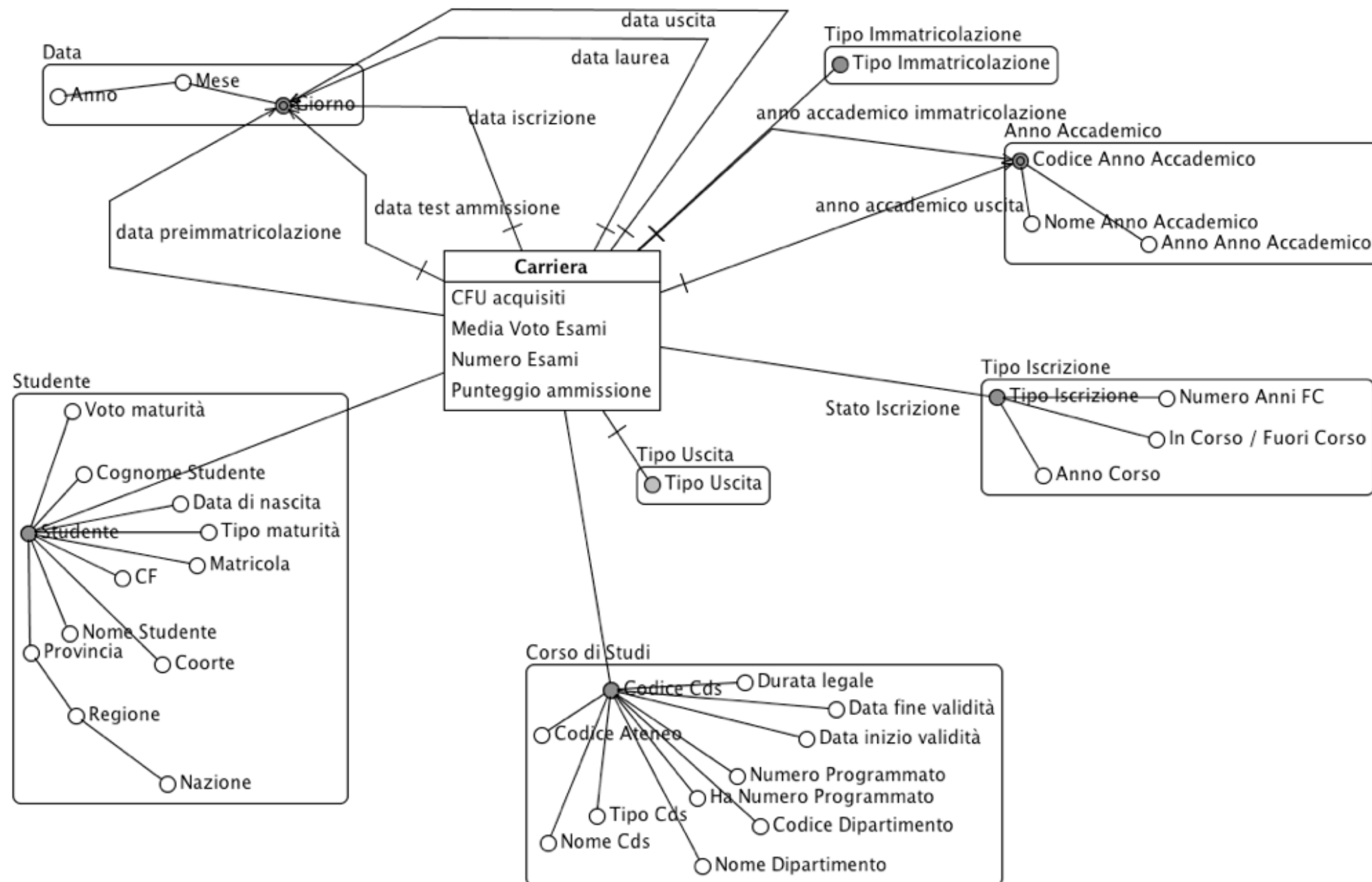
- AA COORTE
- IMM TOT
- N STUDENTI SOPRAVV
- MEDIA CFU STESSO CDS II ANNO
- MEDIA CFU IN CORSO
- MEDIA CFU RIPETENTI

RC.4.2



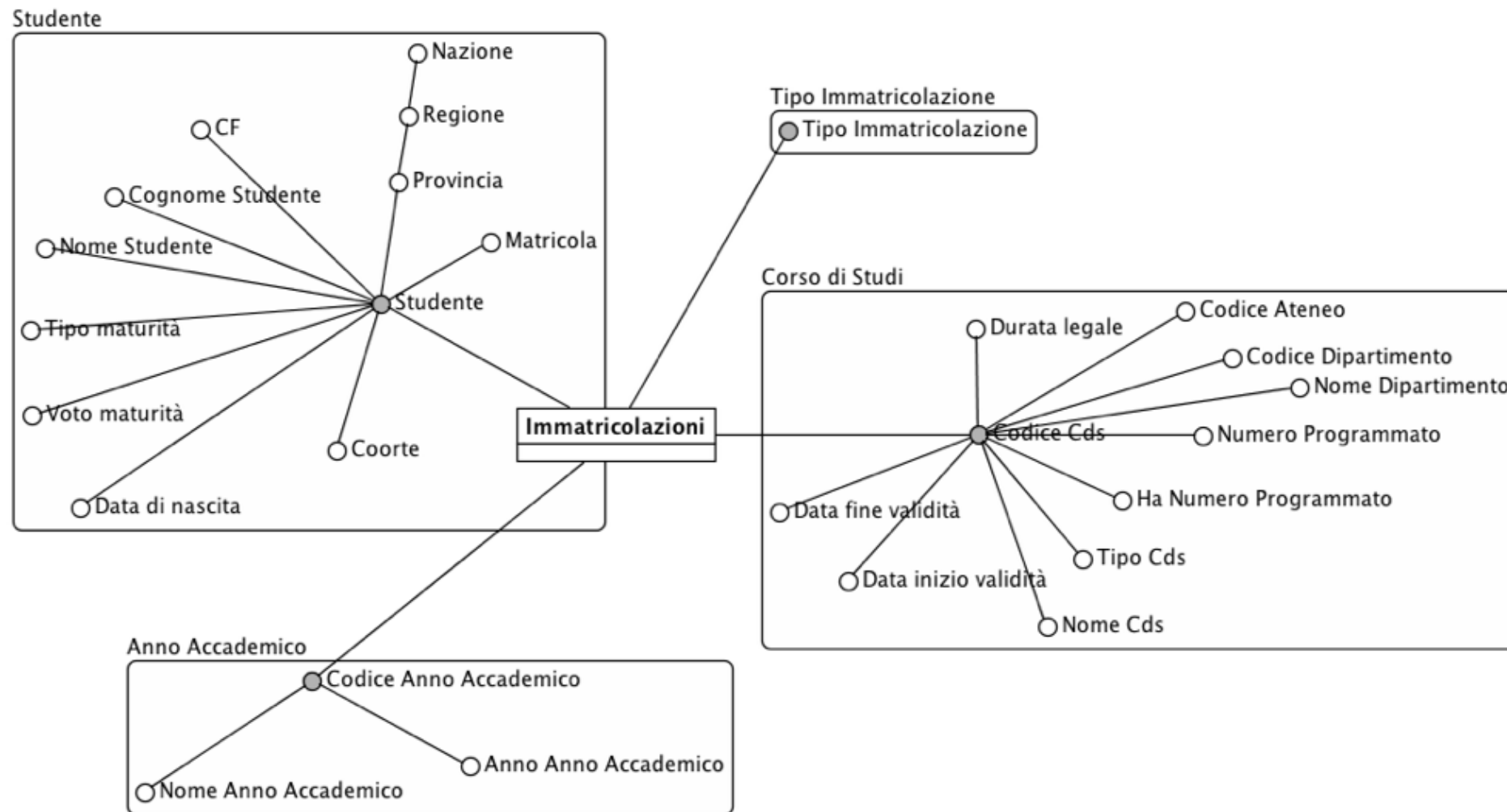
- AA COORTE
- IMM TOT
- N STUDENTI SOPRAVV
- MEDIA CFU STESSO CDS II ANNO
- MEDIA CFU IN CORSO
- MEDIA CFU RIPETENTI

RC.4.2



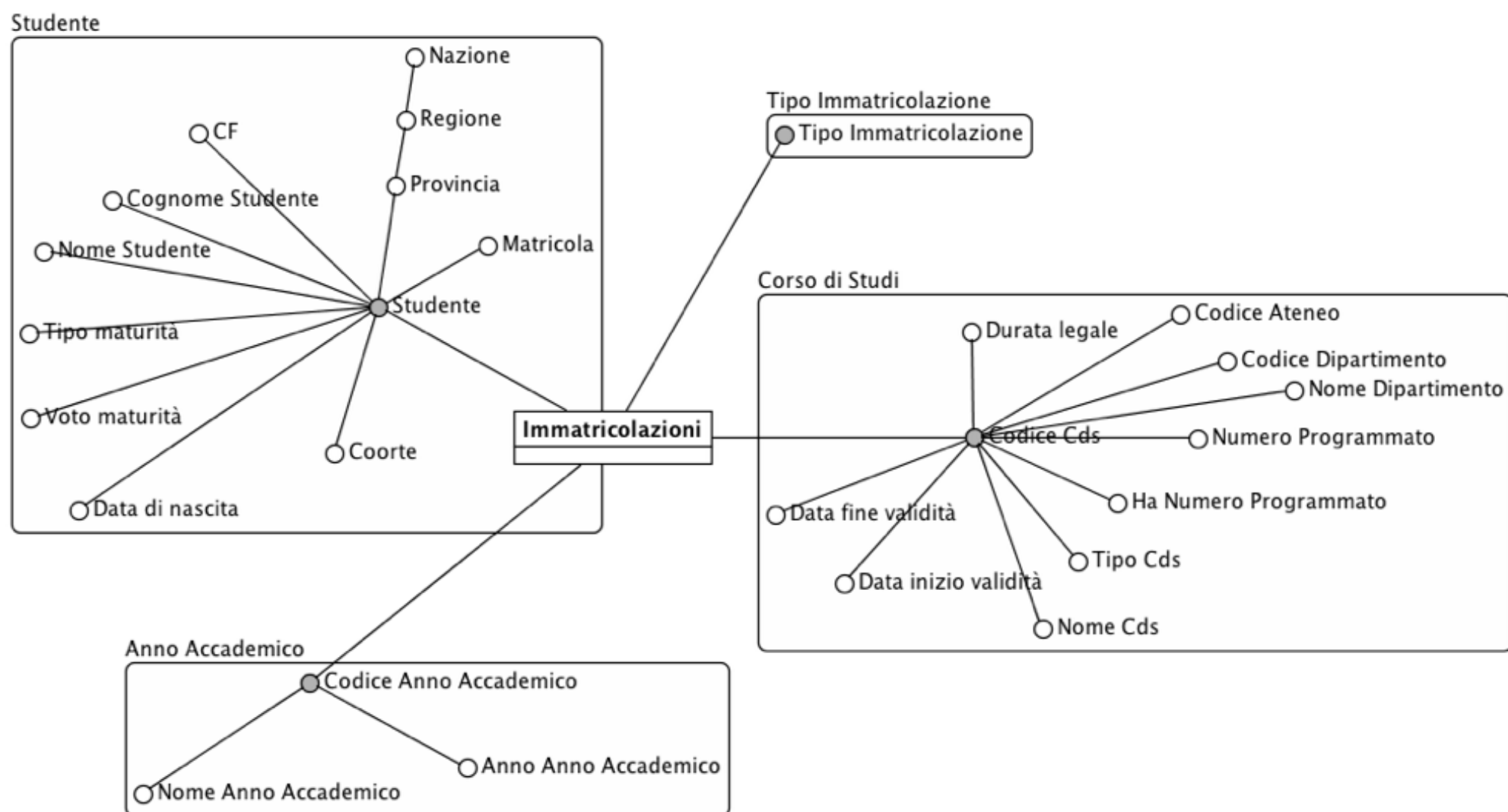
- AA COORTE
- PCT CLASSICA
- PCT SCIENT
- PCT TECNICA
- PCT PROF
- PCT MAGIST
- PCT LING
- PCT ART
- PCT STRANIERA
- PCT NON DISP

RC.5.1



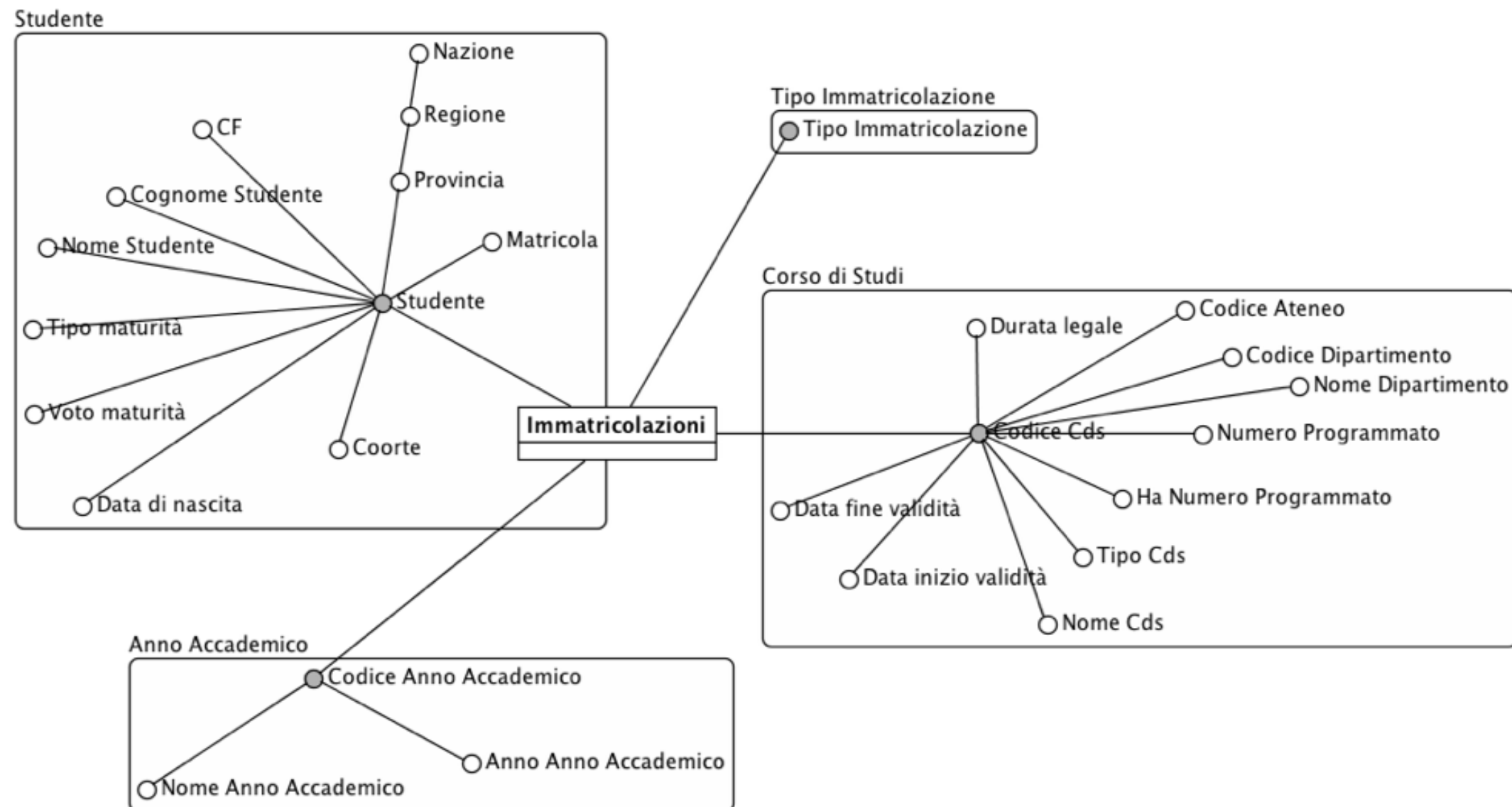
- AA COORTE
- IMM TOT
- PCT 60 70
- PCT 71 80
- PCT 81 90
- PCT 91 100
- PCT NON DISP
- PCT STRANIERA

RC.5.2



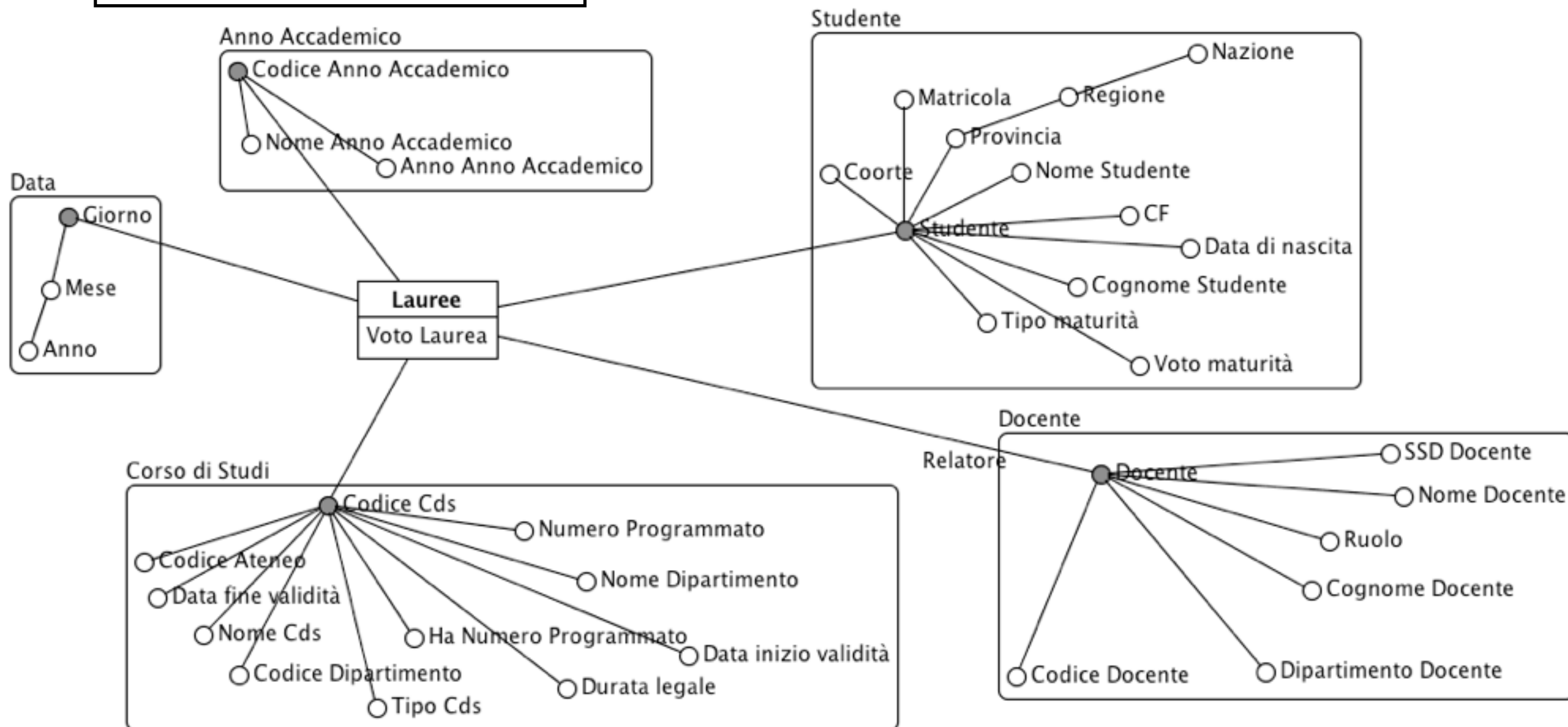
- AA COORTE
- IMM TOT
- PCT 66 90
- PCT 91 100
- PCT 101 105
- PCT 106 110
- PCT NON DISP
- PCT TIT STRANIERO

RC.6



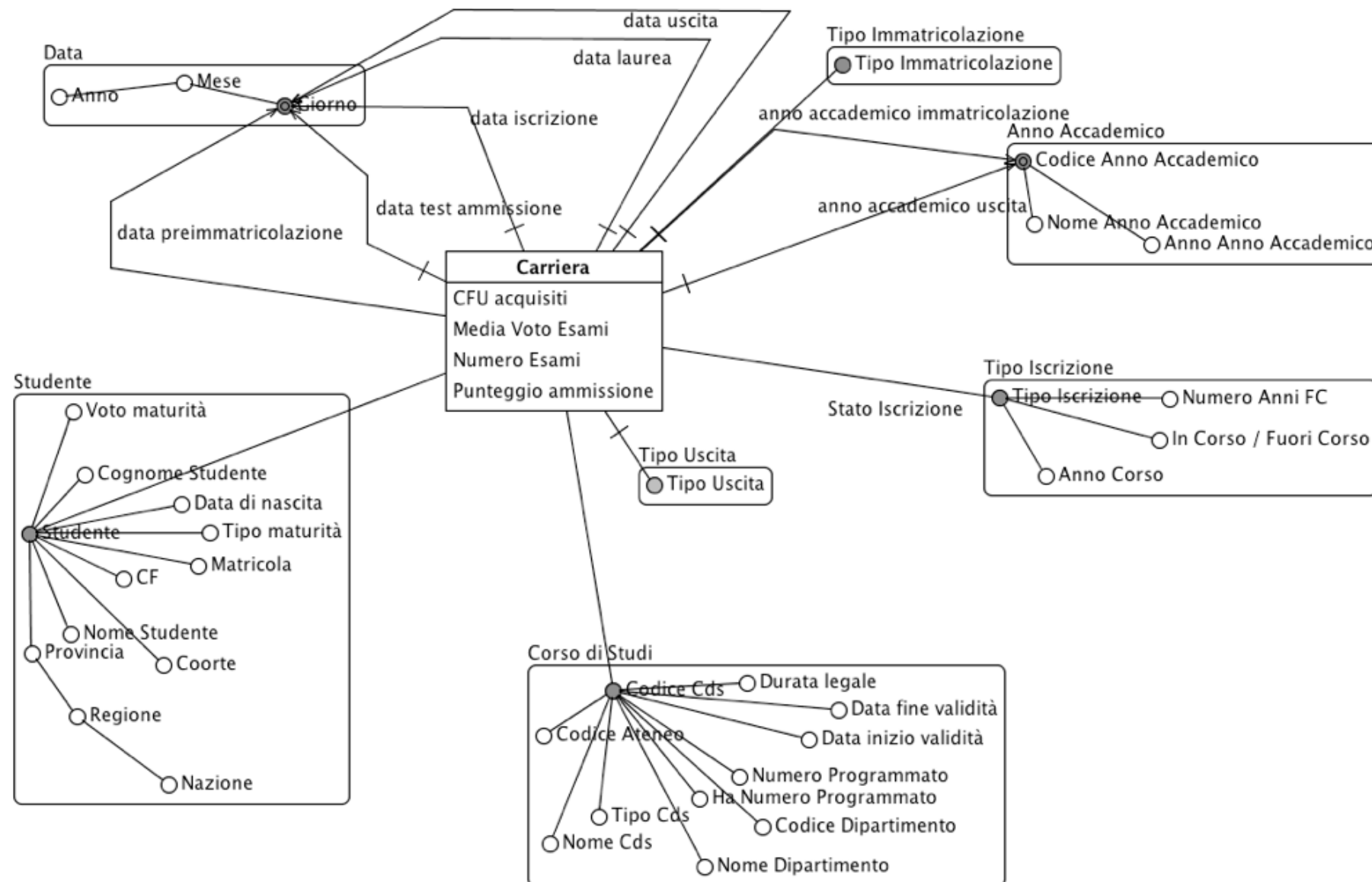
- AA COORTE
- IMM TOT
- PCT 66 90
- PCT 91 100
- PCT 101 105
- PCT 106 110
- PCT NON DISP
- PCT TIT STRANIERO

RC.6



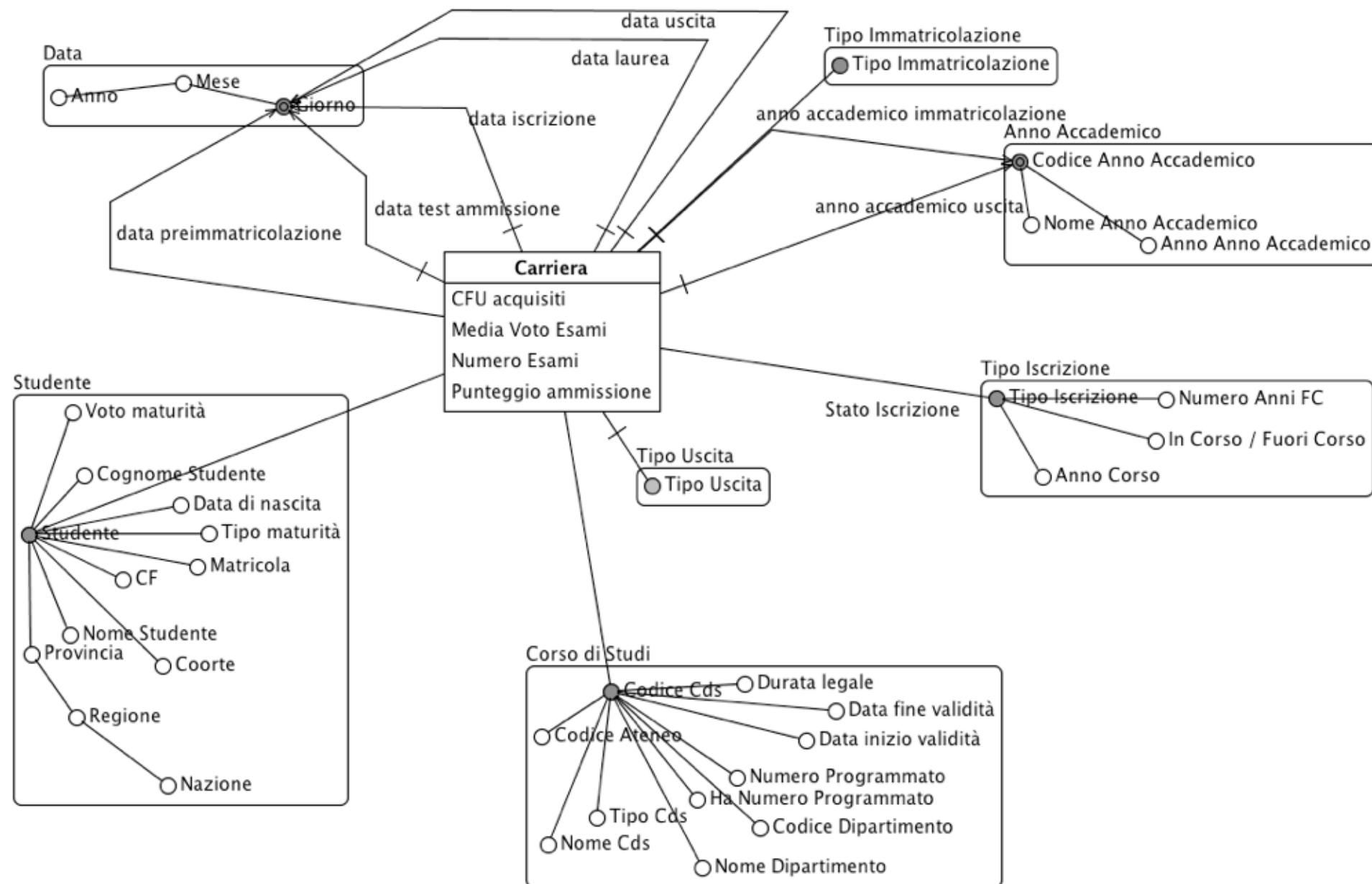
- AA COORTE
- IMM TOT
- LAU TOT
- PERC IN CORSO:
- PERC 1 ANNO FC
- PERC 2 ANNI FC:
- PERC 3 ANNI FC O PIÙ
- MEDIA DURATA CARR CDS
- MEDIA DURATA CARR STR
- MEDIA VOTO

RC.7



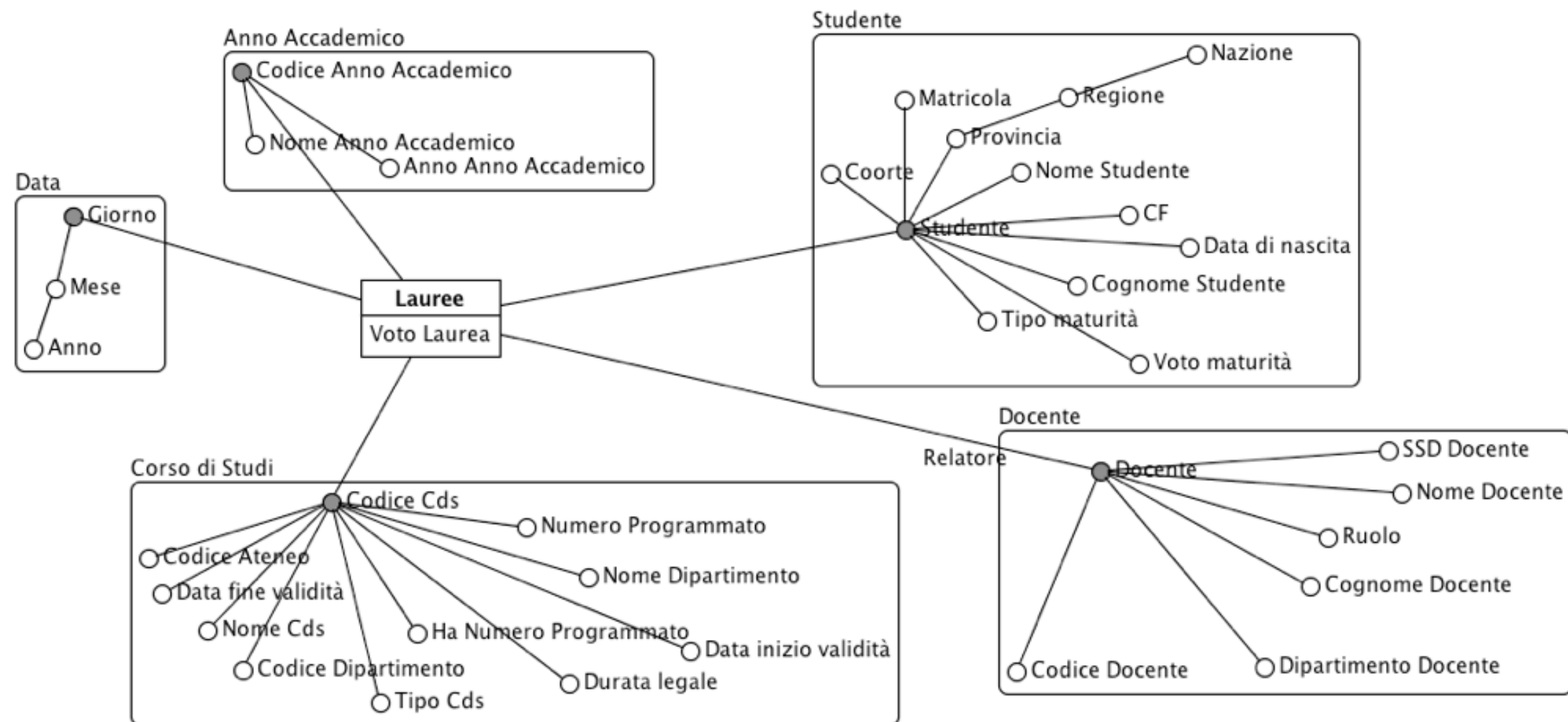
- AA
- N STUDENTI
- MEDIA CFU GENERALE
- MEDIA CFU 1 ANNO
- MEDIA CFU 2 ANNO
- MEDIA CFU 3 ANNO
- MEDIA CFU 4 ANNO
- MEDIA CFU 5 ANNO
- MEDIA CFU STR

RC.8



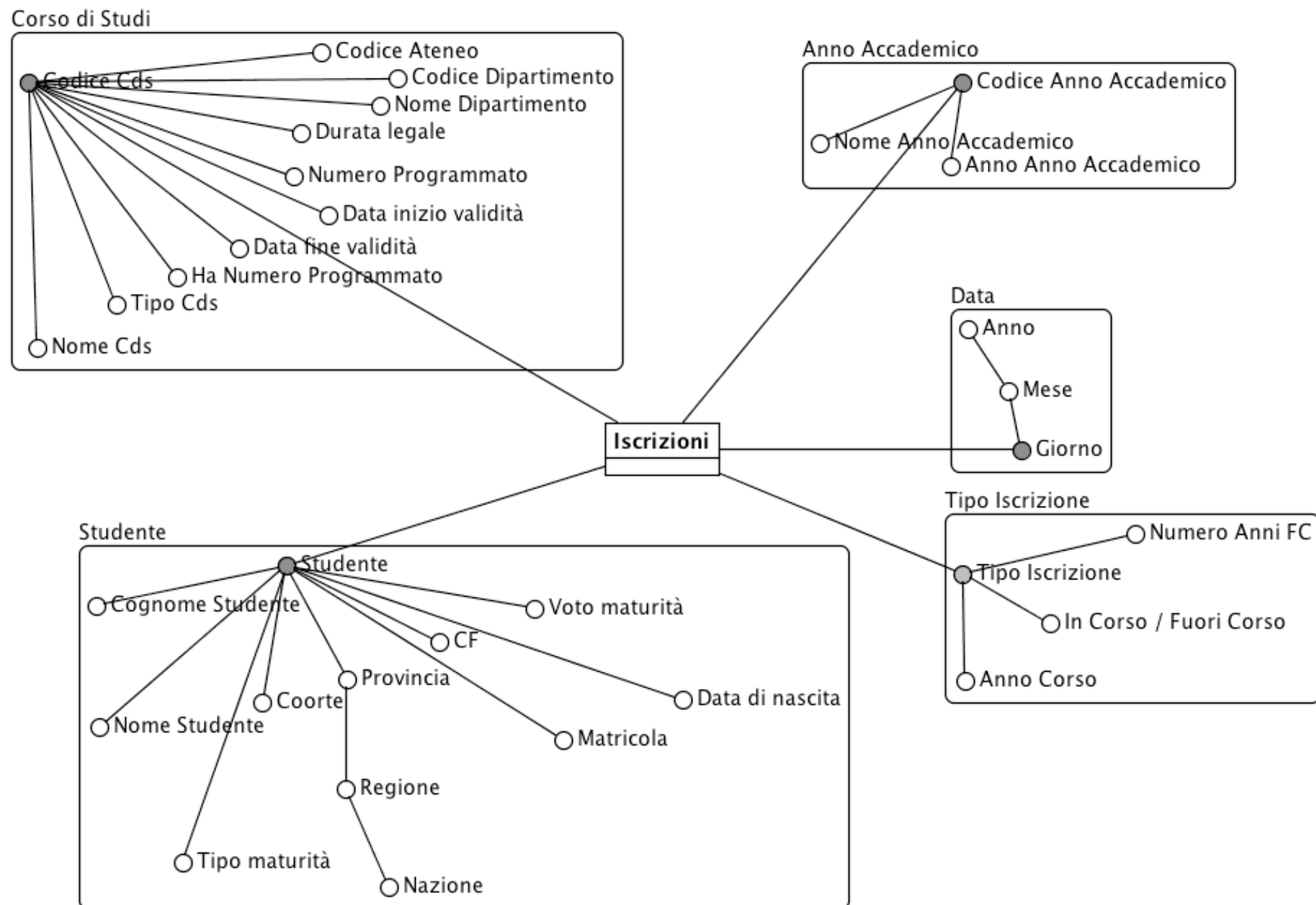
- AA LAUREA
- TOT LAU
- PCT NO IMM
- PCT DOPO 1 ANNO
- PCT DOPO 2 ANNI
- PCT DOPO 3 ANNI

RC.9



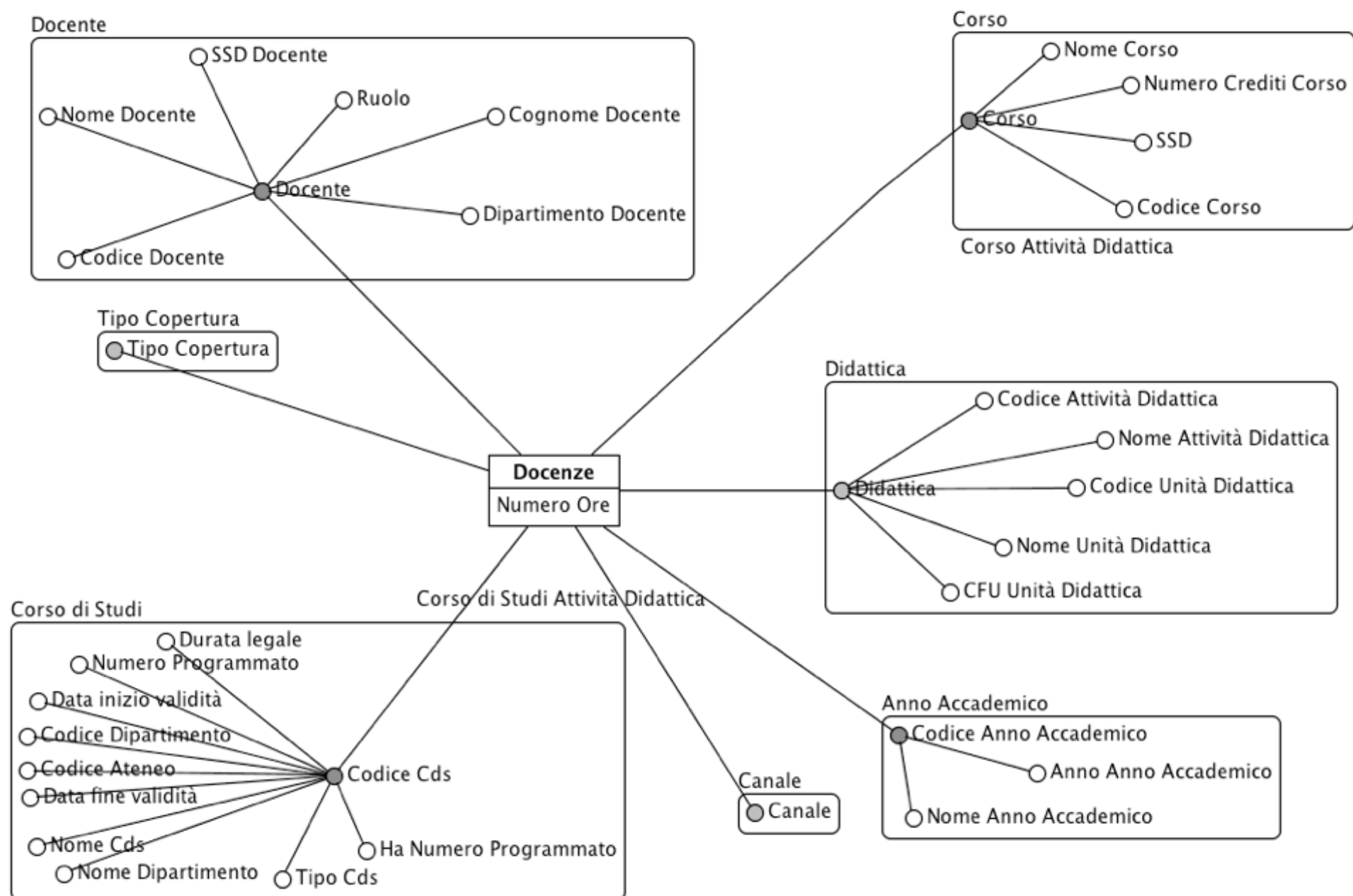
- AA LAUREA
- TOT LAU
- PCT NO IMM
- PCT DOPO 1 ANNO
- PCT DOPO 2 ANNI
- PCT DOPO 3 ANNI

RC.9



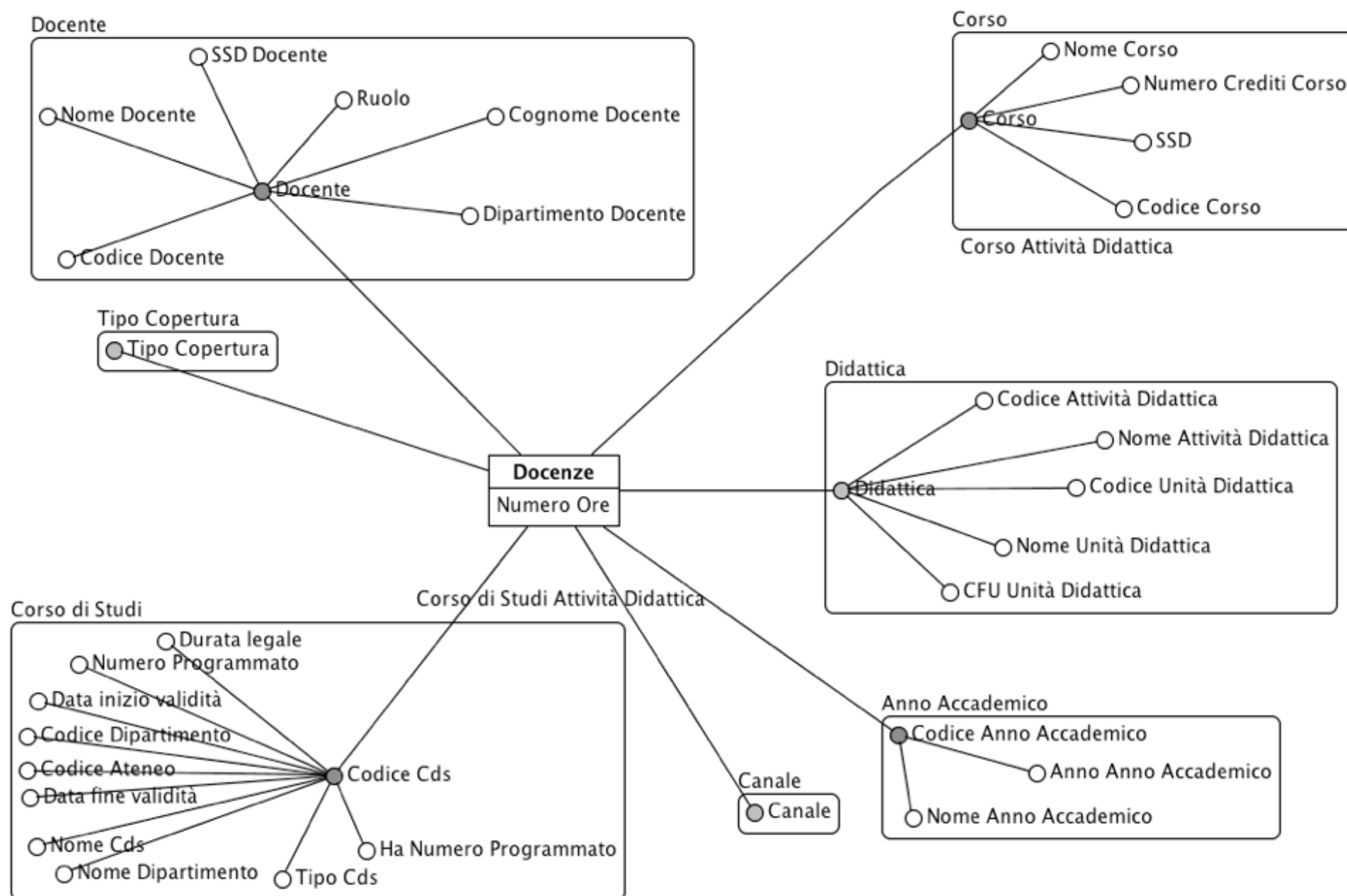
- **DOCENTE**
- **RUOLO**
- **SSD DOCENTE**
- **IMPEGNO**

RI.I.I



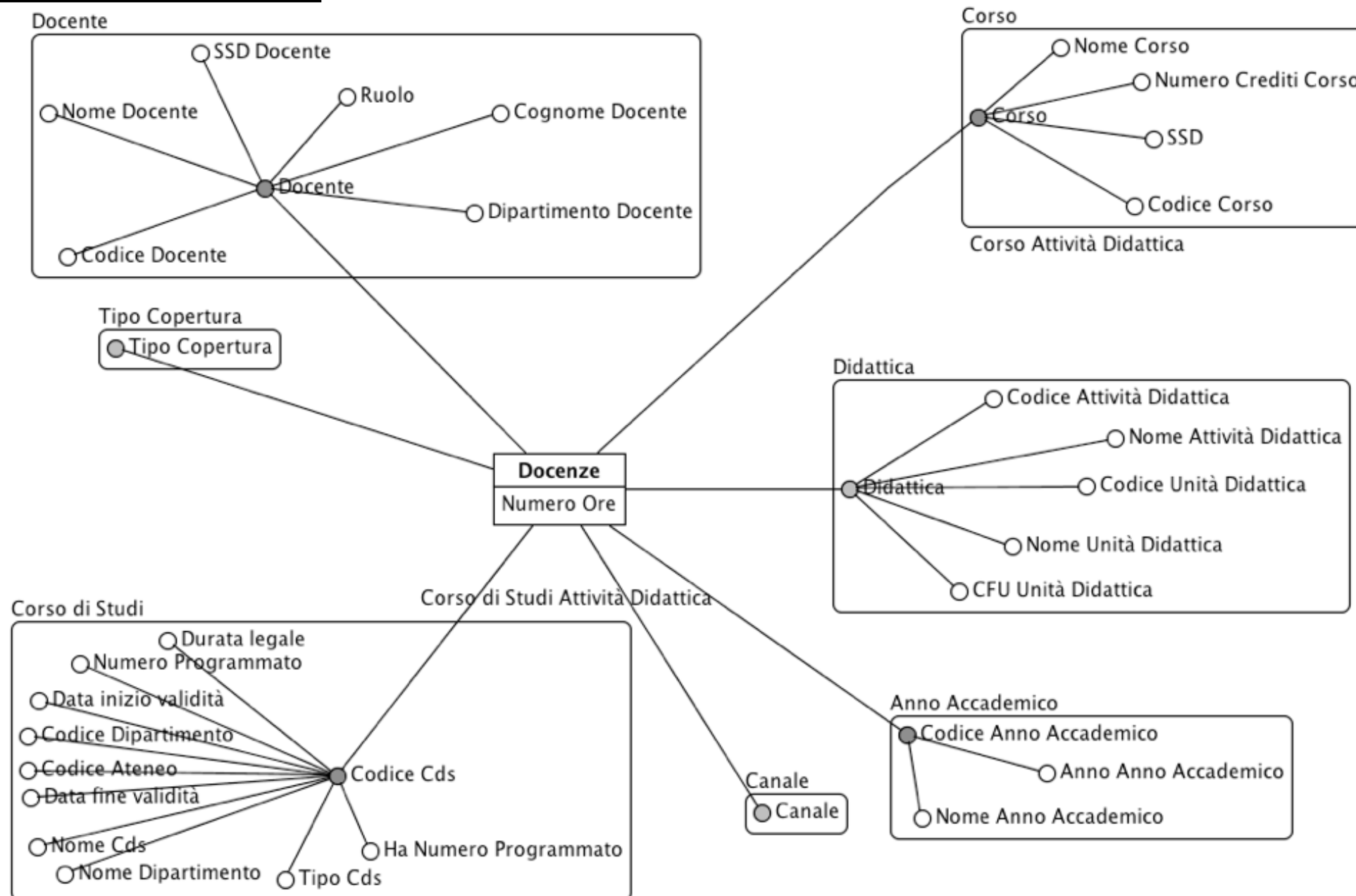
- DOCENTE
- RUOLO
- IMPEGNO

RI.1.2



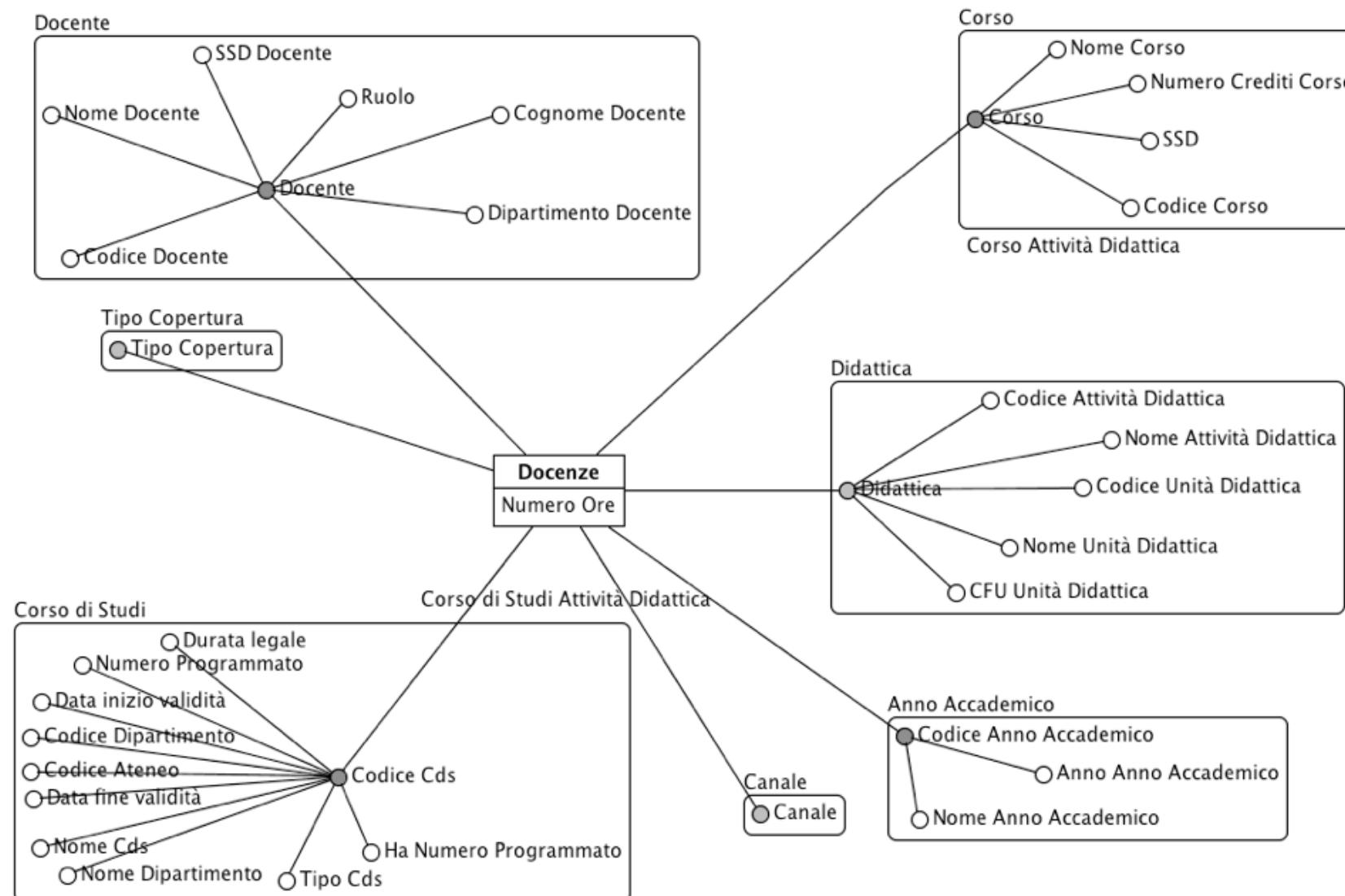
- AD DES
- AD COD
- UD DES
- PARTIZIONAMENTO
- SSD
- SSD DOCENTE
- DOCENTE
- COPERTURA
- CFU AD
- CFU UD
- ORE LEZIONE FRONTALE
- CONDIVISIONE

RI.2



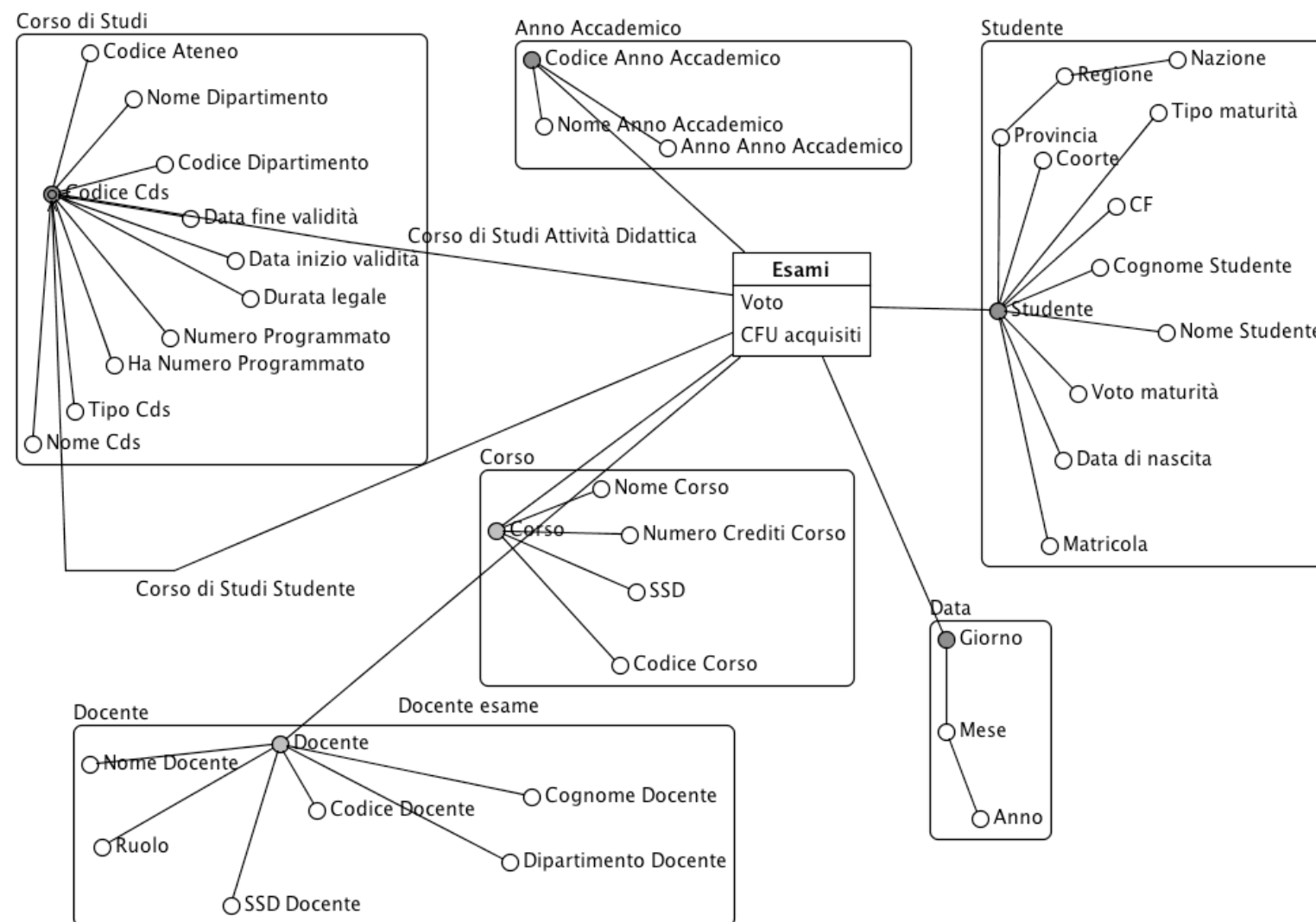
- AD
- AD COD
- DOCENTE
- STESSO CDS
- 2011
- 2012
- 2013
- 2014

RI.3



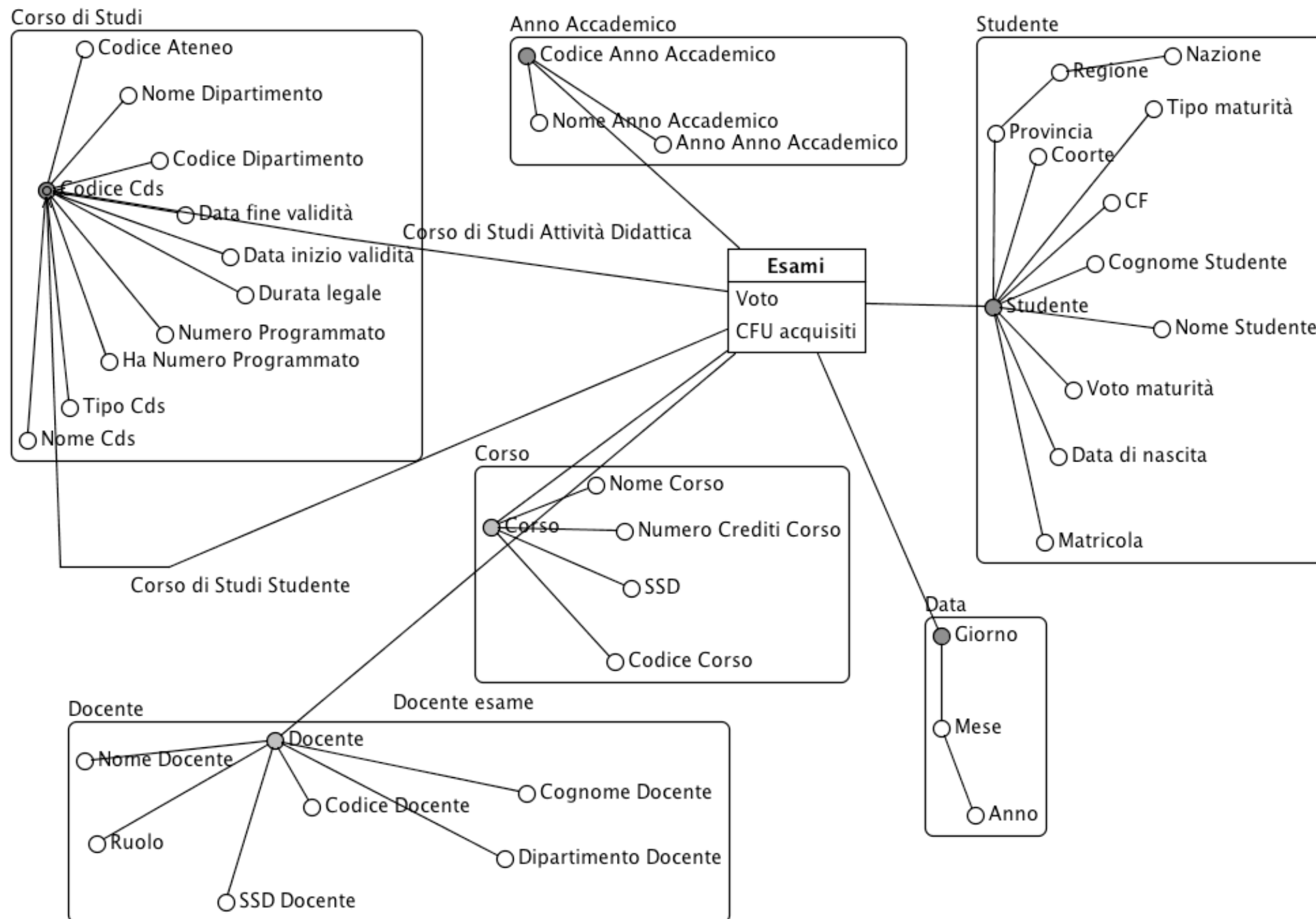
- AD
- AD COD
- DOCENTE
- STESSO CDS
- 2011
- 2012
- 2013
- 2014

RI.3



- VOTO
- PERCENTUALE
- NUMERO ESAMI

RI.4.1



- VOTO ECTS
- PERCENTUALE
- VOTO MASSIMO
- VOTO MINIMO

RI.4.2

