I corpora digitali: dall'obsolescenza tecnologica, alla salvaguardia e alla condivisione

Eva Sassolini

Istituto di Linguistica Computazionale
"Antonio Zampolli"
CNR - Pisa

Obsolescenza

Principali problematiche di recupero di testi conservati in file con formati e codifiche obsoleti:

- codifica dei caratteri spesso non standard
- difformità negli schemi di annotazione (quando presenti)
- formato dei file:
  - forte dipendenza dal sistema di indicizzazione cui i testi erano destinati

Un caso emblematico: il progetto "Digesto"

- Programma di ricerca Traduzione dei *Digesta* di Giustiniano "Lessico Giuridico storia e dommatica" settore disciplinare IUS/18, Dipartimento di Storia e Teoria del Diritto dell'Università di Roma Tor Vergata
- Fattori di complessità nella gestione dei materiali testuali:
  - attività dilatate nel tempo (legate alla ricerca di finanziamenti)
  - ripresa dei lavori dopo intervalli di anni
  - obsolescenza delle tecnologie utilizzate
  - problematiche di codifica legate alla fruizione dei contenuti => prima sistemi di indicizzazione e consultazione proprietari (stand-alone), poi applicazioni web e codifica UNICODE

Soluzioni adottate

- Creazione di un *protocollo* per una corretta codifica dei testi e per il recupero di eventuali annotazioni linguistiche o di altra natura
- Mapping dei testi in XML TEI:
  - conservazione a lungo temine
  - divulgazione e consultazione con sistemi, proprietari/open source, nel pieno rispetto dei diritti dei testi
  - riuso dei testi per funzioni di analisi ed elaborazione del testo

La fasi di lavoro

- Conversione dal formato originario del testo ad uno *intermedio* di interpretazione:
  - Stessa struttura gerarchica delle informazioni
  - Stessa codifica per l'annotazione morfo-sintattica
- Conversione dal formato *intermedio* a quello standard XML TEI con definizione di struttura e tag:
  - Per ogni testo recuperato definizione di:
    - Unità testuale minima
    - Header (metadati)
  - Organizzazione delle informazioni:
    - di alto livello (categorie, genere, ecc.)
    - a livello linguistico (annotazioni morfo-sintattiche)

Dati sull'applicazione del protocollo al "corpus ILC" da recuperare

Testo sorgente	Perc.	Fasi di transizione richieste (FT)	Meta dati
Testo su nastro magnetico	10%	Molteplici FT	Ricerche su materiali cartacei storici di ILC
Testo diviso in più risorse digitali separate	5%	FT>3	Recuperati da schede cartacee di progetto
Testo digitale in formato obsoleto	10%	FT>2	Recuperati da schede cartacee di progetto
Testo digitale con codifica dei caratteri obsoleta	10%	2<FT<3	Recuperati da: - Schede cartacee - Documentazione digitale
Testo digitale	65%	1 FT	Recuperati da documentazione digitale

Azioni di salvaguardia e condivisione

- Preservare i testi attraverso la standardizzazione dei formati di rappresentazione
- Mettere a disposizione della comunità i testi recuperati attraverso:
  a) applicazioni web dedicate all'analisi testuale
  b) popolamento del nodo CLARIN-IT
  c) realizzazione di nuove modalità di fruizione dei dati:
     - uso di tecniche di *visual analytics*
     - sperimentazioni con tecnologie App mobile

a) Piattaforma web ILC

- Piattaforma web per l'acquisizione, la gestione, la conservazione, l'interrogazione e la valorizzazione di importanti archivi testuali
- Offerta di funzionalità per:
  - indicizzazione
  - analisi quantitative del testo
  - concordanze
  - ricerche in testi lemmatizzati
  - ….

b) CLARIN-IT

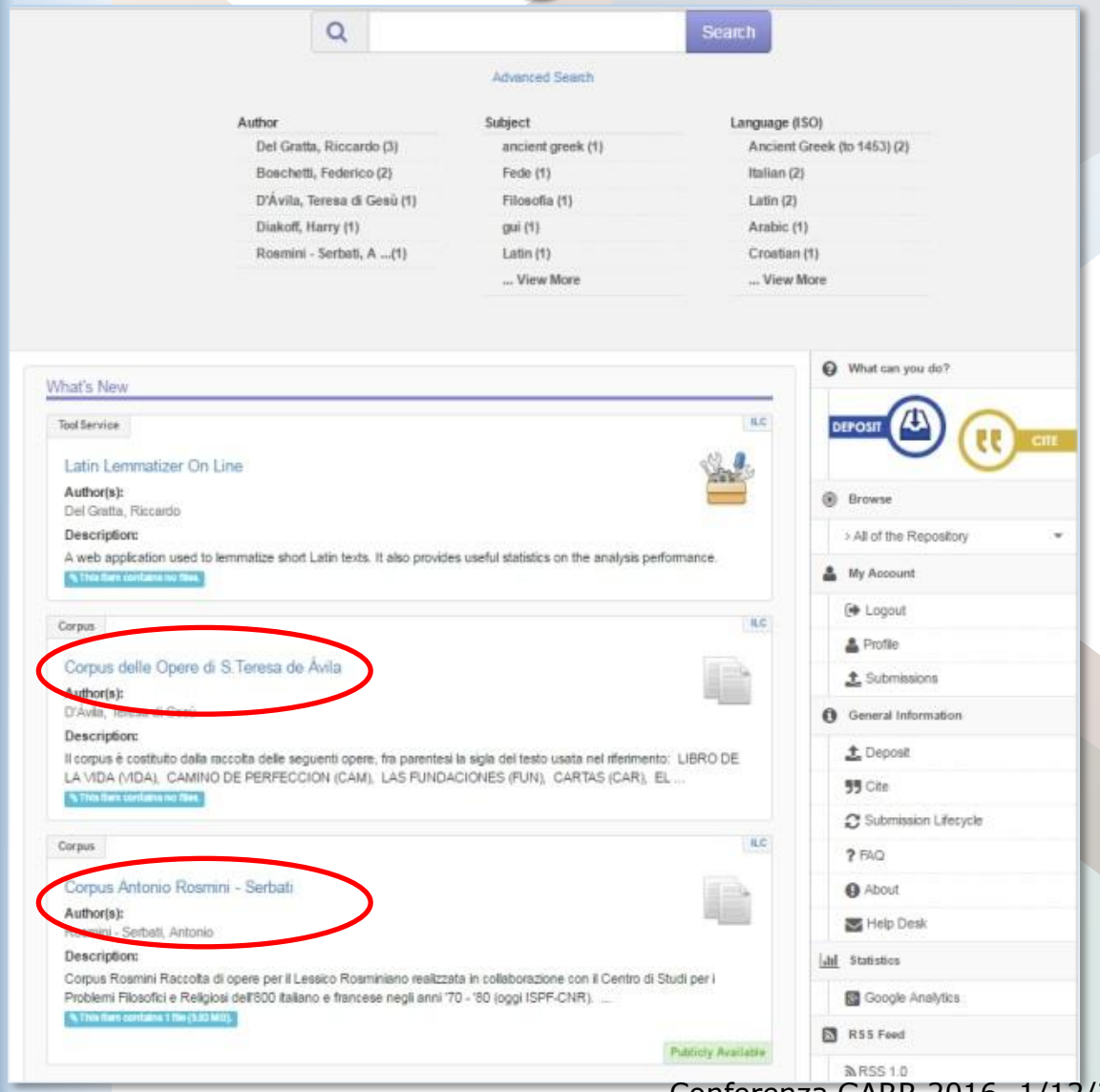
- CLARIN (Common Language Resources and Technology Infrastructure) è un'infrastruttura internazionale che offre soluzioni a lungo termine e servizi tecnologici per la distribuzione, il collegamento, l'analisi e il mantenimento di strumenti e dati testuali/linguistici digitali
- CLARIN-IT è il nodo Italiano di CLARIN e ne condivide l'intera visione

Popolamento del nodo

Sono in corso azioni finalizzate al popolamento del nodo con testi appartenenti al "corpus ILC" recuperato, per esempio:

- Il corpus digitale delle opere di S.Teresa de Ávila, prodotto negli anni '80 e '90, che comprende:
  - Libro De La Vida - Camino De Perfeccion - Las Fundaciones – Cartas - El Castillo Interior – Conceptos – Relaciones – Exclamaciones – Constituciones - Modo De Visitar Los Conventos – Poesias - Apuntes.
- Il corpus delle tragedie di Vittorio Alfieri formato in particolare:
  - Agamennone - Antigone - La Congiura de' pazzi - Merope Maria Stuarda - Oreste Polinice - Virginia - Filippo - Agide - Bruto I - Bruto II - Don Garzia - Mirra - Ottavia Rosmunda – Saul – Sofonisba – Timoleone
- La raccolta dei Dialoghi Italiani (sia morali che metafisici) di Giordano Bruno:
  - La cena de le ceneri - Cabala del cavallo pegaseo - Il candelaio - De la causa, principio e uno - De l'infinito, universo e mondi - Spaccio de la bestia trionfante - De gli eroici furori
- 55 opere del filosofo e sacerdote Antonio Rosmini (corpus Rosmini – Serbati)

Navigare su CLARIN-IT



I testi inseriti nel nodo dell'infrastruttura sono oggi consultabili sul sito insieme ad altri testi e a strumenti che sono stati messi a disposizione della comunità scientifica

CLARIN-IT

c) Da tecniche di *close reading* a quelle di *distant reading*

Oggi le tecnologie informatiche e linguistiche sono mature per consentire analisi ed elaborazioni innovative

- Nuove modalità di rappresentazione dei contenuti testuali:
  - diagrammi temporali, grafici, alberi, mappe, …
  - elaborazione dei testi per l'estrazione di dati utilizzabili in rappresentazioni sintetiche (dati matriciali)

Prima sperimentazione con rappresentazione grafica dei contenuti

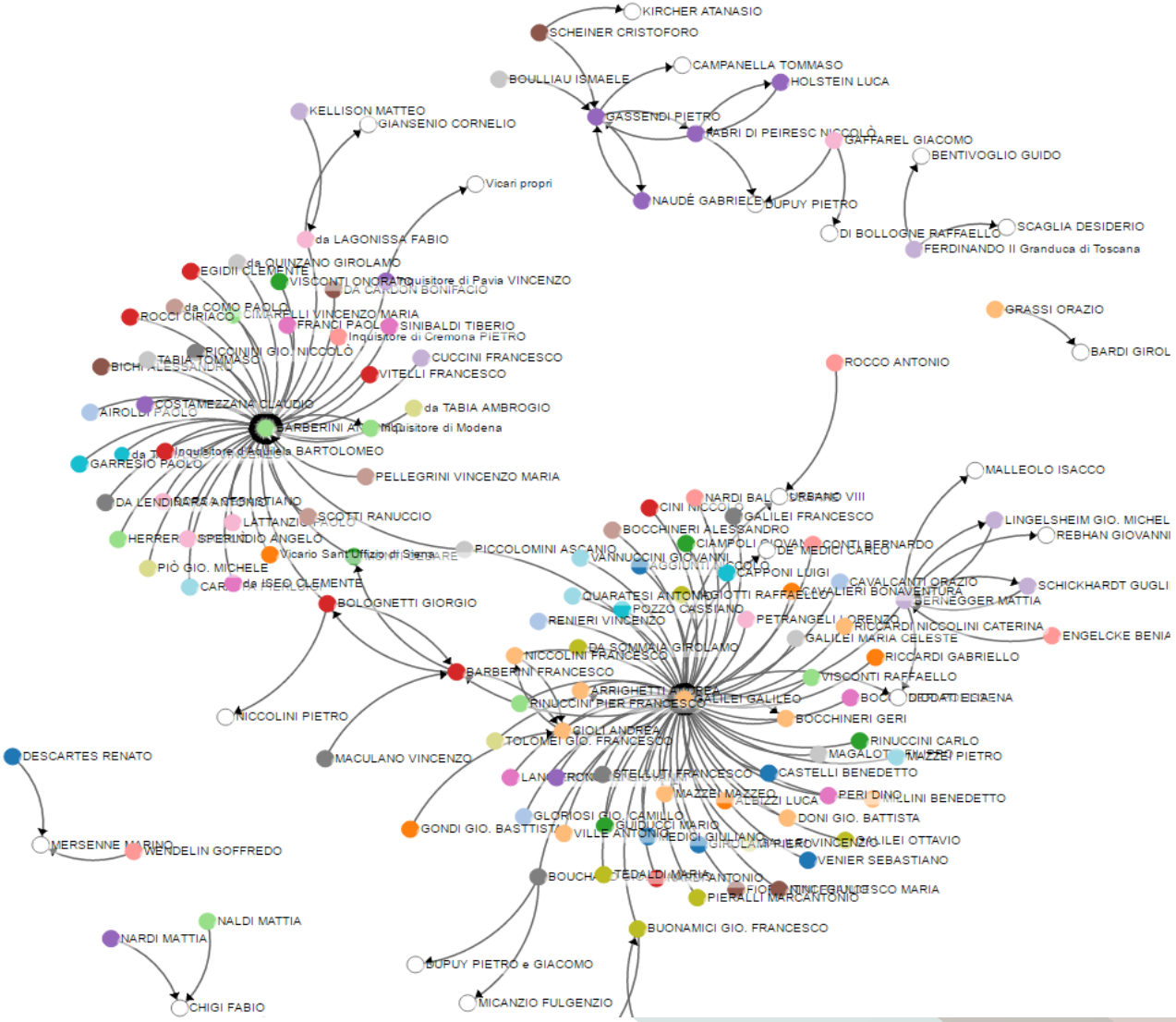
- Costruzione di funzionalità web che mirano ad integrare tecniche di *visual analytics* e funzionalità classiche di *Information Retrieval*
- Realizzare viste e rappresentazioni dei dati che evidenziano le caratteristiche salienti dei dati:
  - rappresentazioni visive interattive dei dati che consentono all'utente di acquisire conoscenze sulla struttura interna dei dati e di individuare eventuali relazioni al loro interno

I testi elaborati per la sperimentazione

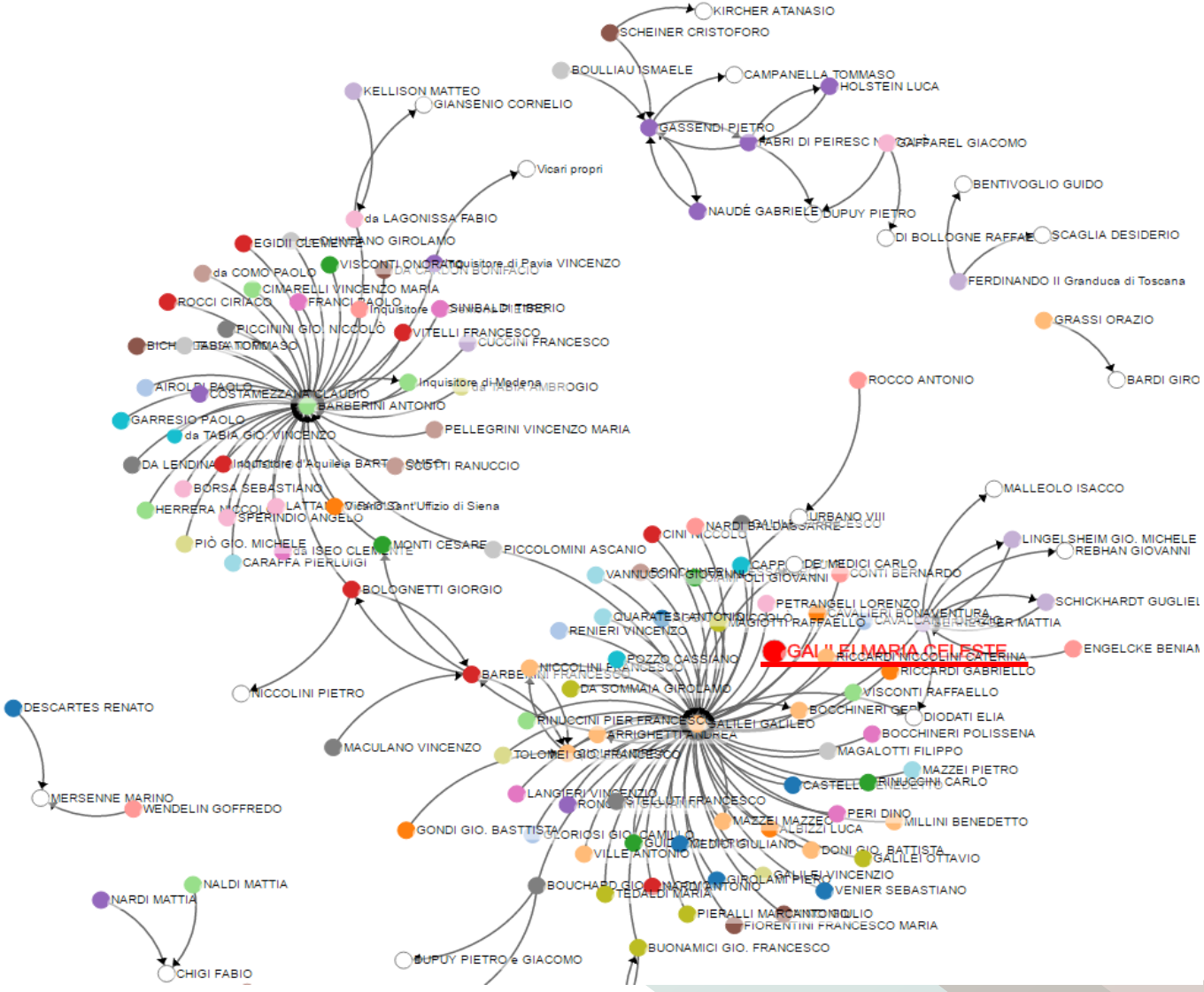
- Carteggio Galileiano comprendente 462 lettere relative ad un arco temporale che va da gennaio 1633 alla fine di dicembre dello stesso anno
- Periodo storico strategico per la presenza di rilevanti eventi storici immediatamente precedenti e/o simultanei:

  ad esempio la pubblicazione del "Dialogo Sopra i Due Massimi Sistemi del Mondo" del 1632 e il successivo processo e condanna dell'autore da parte dell'Inquisizione, nel giugno del 1633
- Corpus in lingua italiana risalente alla prima metà del '600, costituito da comunicazioni personali e scientifiche espresse in un linguaggio prevalentemente informale

Il grafo dal carteggio (1/4)



Il grafo dal carteggio (2/4)

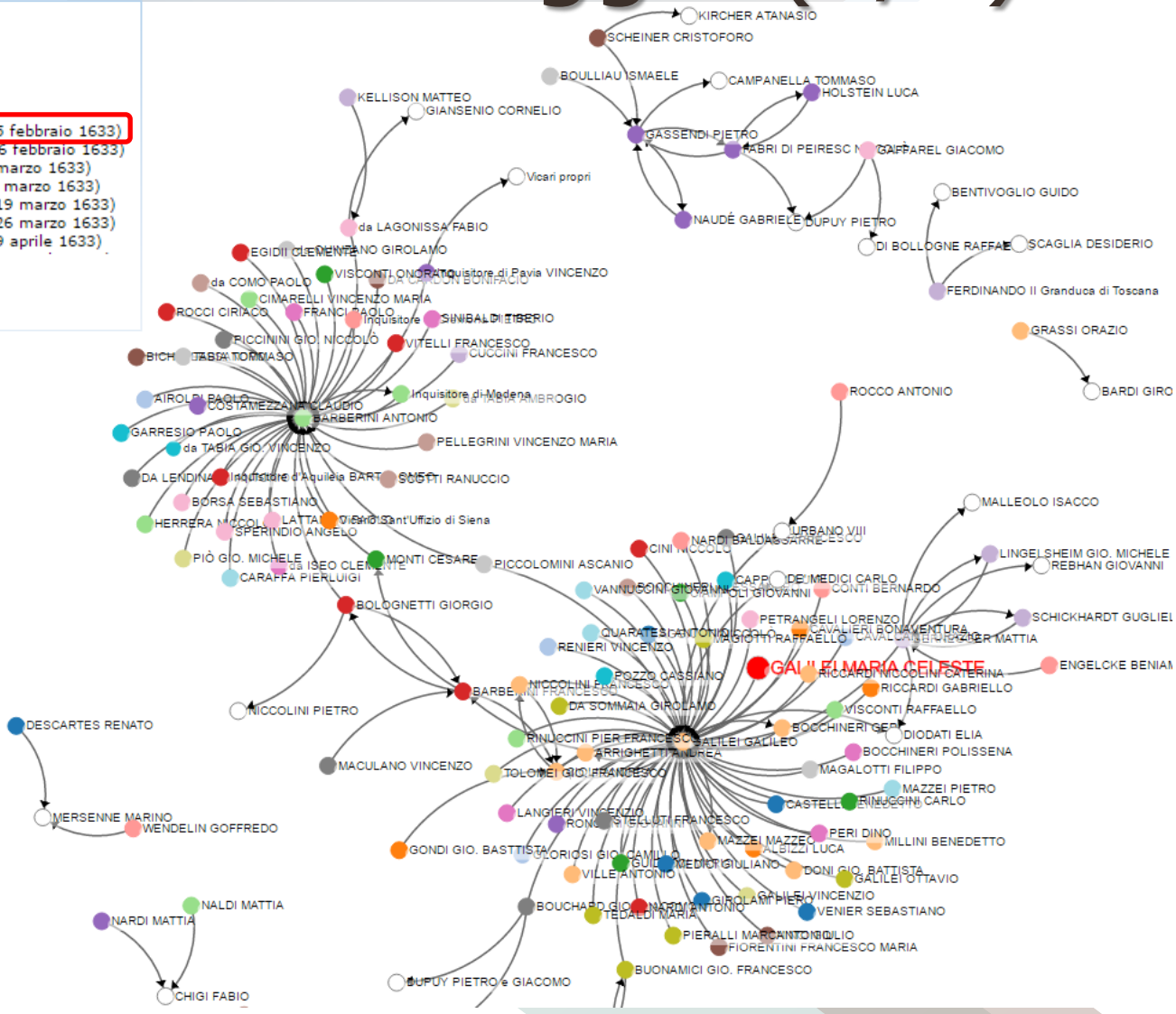


Il grafo dal carteggio (3/4)

mittente: GALILEI MARIA CELESTE

Lista lettere in archivio:

MARIA CELESTE GALILEI a GALILEO [in Roma]. (5 febbraio 1633)
MARIA CELESTE GALILEI a GALILEI [in Roma]. (26 febbraio 1633)
MARIA CELESTE GALILEI a GALILEO in Roma. (5 marzo 1633)
MARIA CELESTE GALILEI a GALILEO in Roma. (12 marzo 1633)
MARIA CELESTE GALILEI a [GALILEO in Roma]. (19 marzo 1633)
MARIA CELESTE GALILEI a [GALILEO in Roma]. (26 marzo 1633)
MARIA CELESTE GALILEI a [GALILEO in Roma]. (9 aprile 1633)



Il grafo dal carteggio (4/4)

mittente: GALILEI MARIA CELESTE

Lista lettere in archivio:

MARIA CELESTE GALILEI a GALILEO [in Roma]. (5 febbraio 1633)
MARIA CELESTE GALILEI a GALILEI [in Roma]. (26 febbraio 1633)
MARIA CELESTE GALILEI a GALILEO in Roma. (5 marzo 1633)
MARIA CELESTE GALILEI a GALILEO in Roma. (12 marzo 1633)
MARIA CELESTE GALILEI a [GALILEO in Roma]. (19 marzo 1633)
MARIA CELESTE GALILEI a [GALILEO in Roma]. (26 marzo 1633)
MARIA CELESTE GALILEI a [GALILEO in Roma]. (9 aprile 1633)

Carteggio - Google Chrome

dbtvm1/VA/GG_Isapi.dll?AZIONE=FULLTEXT&SIGLA=CGG&RIF=28

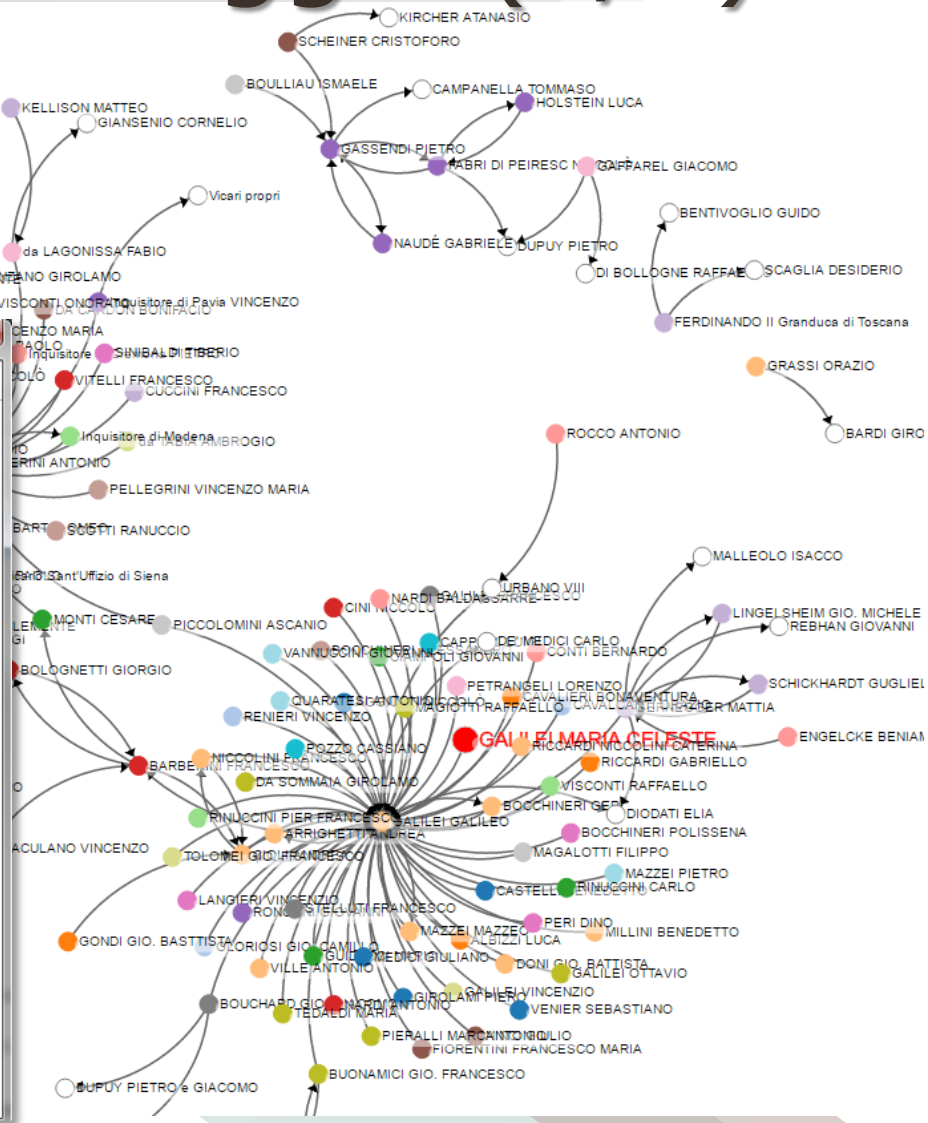
riferimento edizione FAVARO : ***XV.2404*. M.C. GALILEI a GALILEO***

Testo della lettera

\Mittente\ MARIA CELESTE GALILEI a
\Destinatario\ GALILEO [in
\Destinazione\ Roma].
\Luogo\ Arcetri, 5 febbraio 1633.
\Conservazione\ Bibl. Naz. Fir. Mss. Gal., P. I, T. XIII, car. 169. - Autografa.

Molto Ill.re et Amatiss.mo Sig.r Padre, I SSig.ri Bocchineri mi hanno trasmesse tutte le lettere che V. S. ha mandate, delle quali mi appago, sapendo quanto gli sia di fatica lo scrivere. Io non gl`ho scritto fin hora, perchè stavo aspettando l`avviso del suo arrivo a Roma; e quando per l`ultima sua intendo che deve trattenersi tanti giorni in abitazione così cattiva e priva di ogni comodità, ne ho preso grandissima afflizione. Non dimeno sentendo che ella, priva di consolazioni interne et esterne, si conserva sana, mi consolo, e rendo grazie a Dio benedetto, nel quale ho ferma speranza di ottener grazia che V. S. se ne torni qua da noi con quiete d`animo e sanità di corpo. In tanto la prego a star più allegramente che sia possibile; e si raccomandi a Dio, che non abbandona chi in Lui confida. Suor Arcangiola et io stiamo bene, ma non già Suor Luisa, che dal giorno che V. S. si partì in qua, è stata sempre in letto con dolori eccessivi, conforme al suo solito; et a me convenendo star in continuo moto et esercizio per applicargli rimedii e servirla, si porge occasione di sollevar l`animo da quel pensiero che forse troppo l`affliggerebbe per l`assenza di V. S.. Il Sig.r Rondinelli[1] non è ancora venuto a goder la comodità che V. S. gl`ha largita della casa, dicendo che le sue lite non glel`hanno permesso. Ma il nostro Padre confessore non lascia di darvi spesso volta: saluta V. S., et il simile fanno la Madre badessa e tutte le amiche. Suor Arcangiola et io infinitamente e senza intermissione preghiamo Nostro Signore che la guardi e conservi. L`inclusa che gli mando, fu trovata da Gioseppe[2] lunedì, nel luogo dove hanno recapito ordinariamente le sue lettere. Di S. Matteo in Arcetri, li 5 di Febb.o 1633.Di V. S. molto Ill.reFig.la Aff.ma Suor M.a Celeste.
\Fuori\ *Fuori, a tergo della lettera (car. 169t.)*: Al molto Ill.re et Amatiss.mo Sig.r Padre Il Sig.r Galileo Galilei; e *in altro foglio a parte (car. 170t.)*: Al molto Ill. Sig.r Padre mio Oss.mo Il Sig.r Galileo Galilei.Roma.

FRANCESCO RONDINELLI.- [N:1]
Garzoncello al servizio di GALILEO.- [N:2]



La sperimentazione con tecnologia App

App per smartphone e tablet con sistema operativo Android, sviluppata nell'ambito del progetto:

"Censimento e schedatura di complessi di architettura moderna e contemporanea in Liguria" (MIBACT per la Liguria, Regione Liguria e Dipartimento DSA di Scienze per l'Architettura dell'Università degli Studi di Genova)

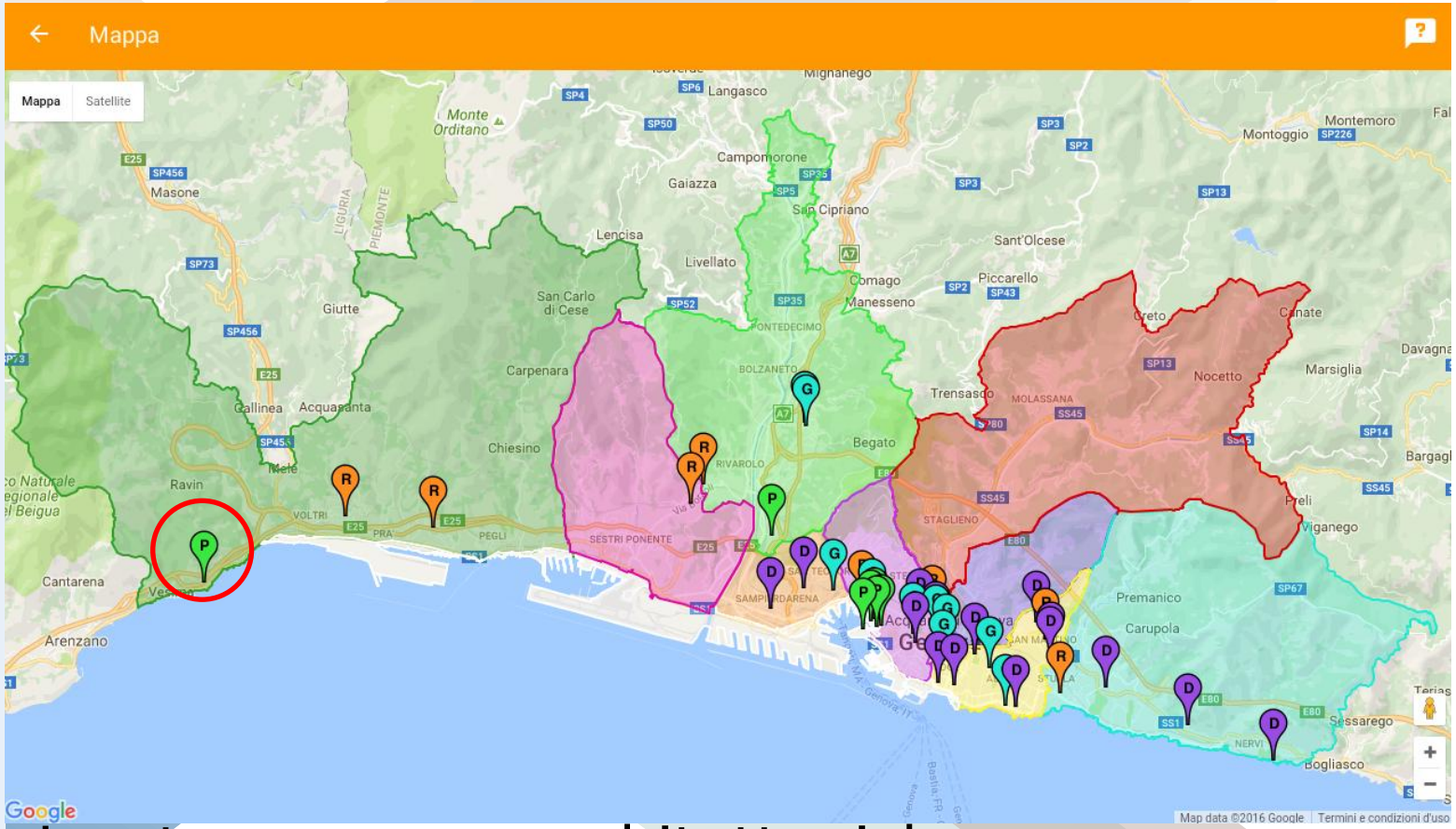
Caratteristiche dell'applicazione:

- Consente all'utente di costruire personali itinerari di visita, selezionando gli edifici d'interesse
- Accesso ai dati attraverso diverse chiavi di interrogazione:
  - luogo, progettista, periodo storico, tipologia o ricerca libera
- Permette la geo-localizzazione degli edifici presenti
- Utilizza i servizi di Google Maps APIs per la generazione di mappe e percorsi

LIGURARCH900 (1/4)

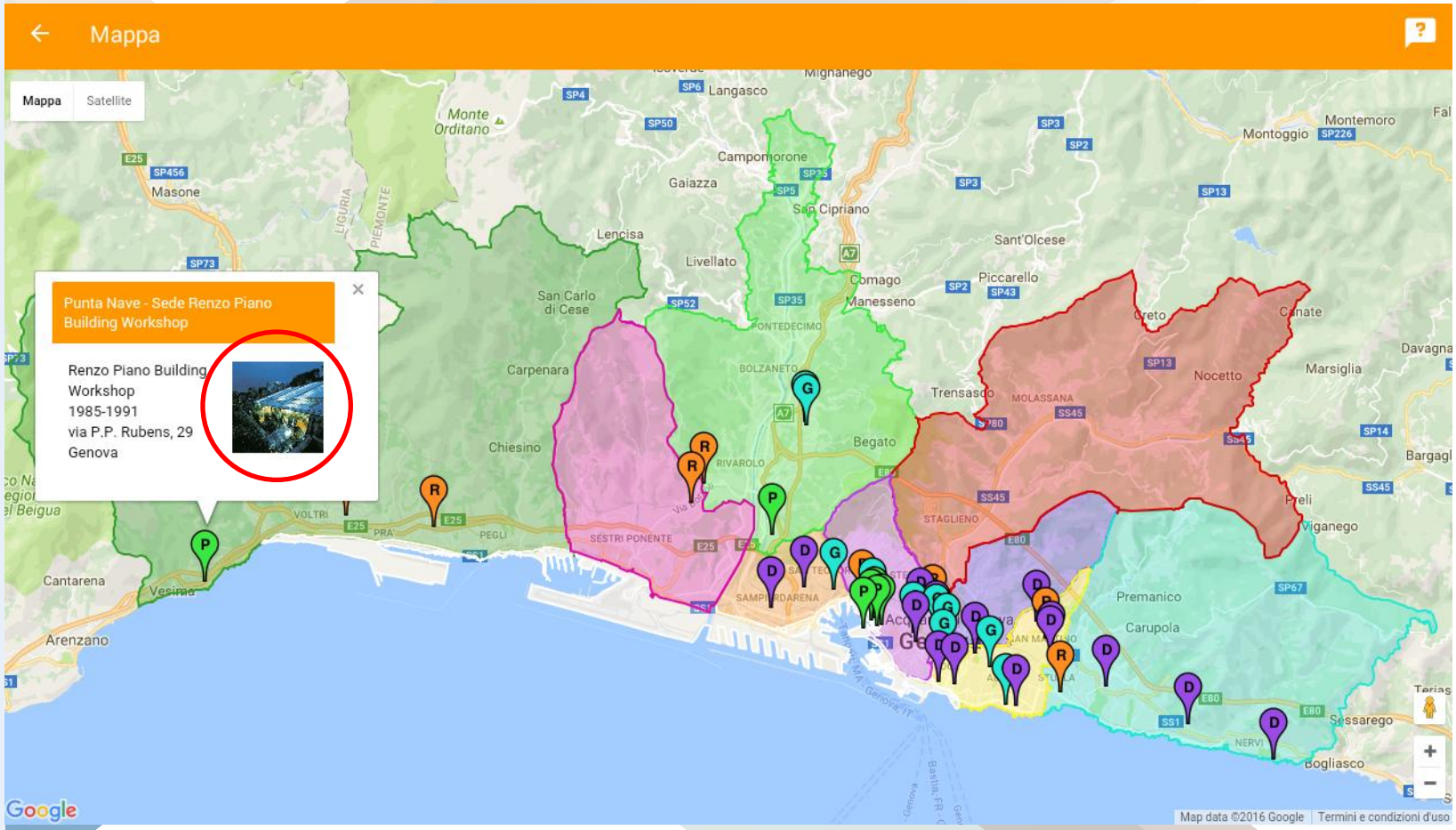
distribuzione sul territorio delle opere architettoniche identificate per progettista

LIGURARCH900 (2/4)



Le stesse opere architettoniche con identificazione del municipio di appartenenza

LIGURARCH900 (3/4)



Informazioni associate allo specifico marker che identificano l'architettura

LIGURARCH900 (4/4)

Testo descrittivo dell'architettura selezionata

Conclusioni

- L'obsolescenza tecnologica è un tema stringente non solo in ambito scientifico:
  - definire un protocollo di recupero per corpora e testi di valore storico-culturale è ormai indispensabile
- La salvaguardia dei testi offerta dall'infrastruttura europea CLARIN è una risposta importante
- La condivisione si può realizzare attraverso:
  - applicazioni web dedicate per l'analisi dei testi
  - ricerche federate all'interno dell'infrastruttura CLARIN
  - nuovi scenari applicativi indirizzati ad una platea web tecnologicamente avanzata:
    - esigenza di una maggiore diffusione di una cultura digitale che non esaurisca il suo compito all'interno delle comunità scientifiche, ma che sia in grado di adeguarsi all'evoluzione delle tecnologie e delle modalità di fruizione dei contenuti

Prospettive e punti critici

Le sperimentazioni hanno prodotto risultati incoraggianti e mettono in evidenza che:

- una maggiore quantità di materiali testuali consente un'interazione più profonda tra la modalità di consultazione classica e quella sintetica e visuale

Per contro:

- un miglior sfruttamento di queste nuove tecnologie richiede una consolidata competenza nell'estrazione di dati matriciali dai testi

- quando i testi sono poco strutturati e presentano annotazioni linguistiche di più livelli sono difficili da trattare con queste modalità