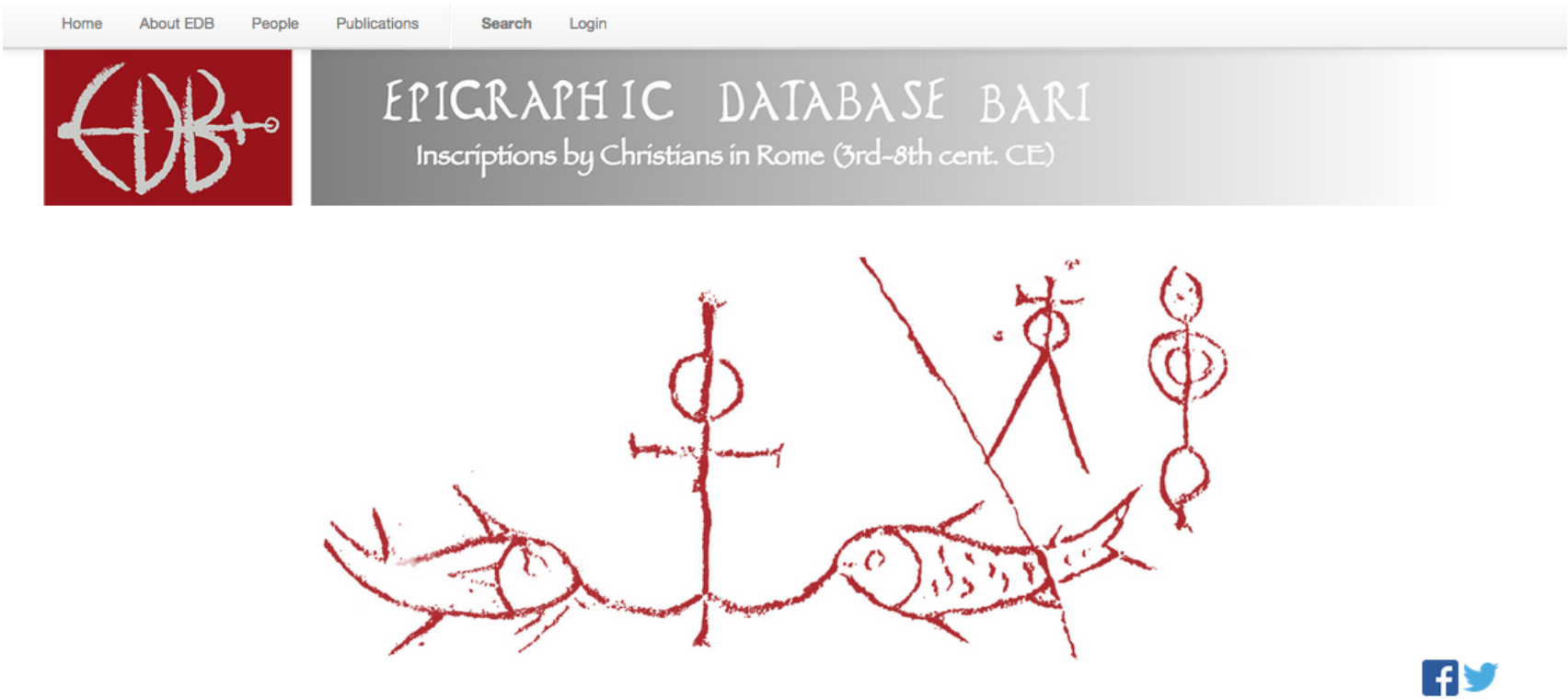
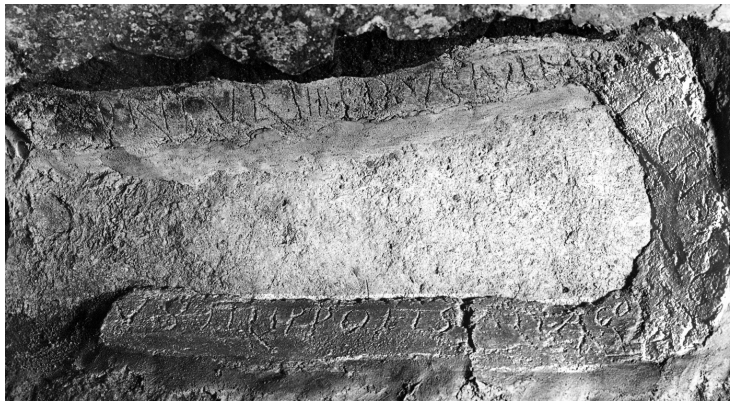
Improving Text-based Search of Inscriptions

Michelangelo Ceci[1], Gianvito Pio[1] and Anita Rocco[2]

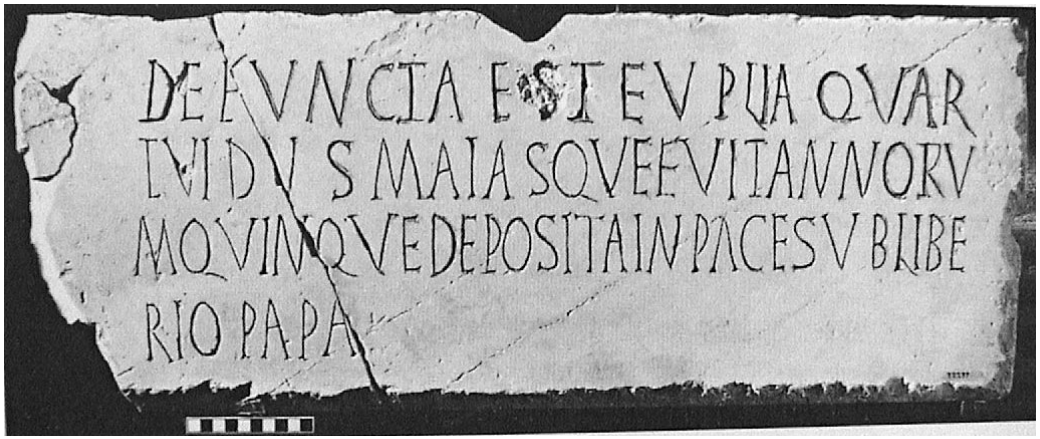
[1]University of Bari - Dip. di Informatica - Via Orabona, 4 - 70125 Bari, Italy
[2]University of Bari - Dip. Di Scienze dell'Antichità e del Tardoantico - Strada Torretta (Citta Vecchia) - 70122 Bari, Italy


Home
About EDB
People
Publications
Search
Login
EPIGRAPHIC DATABASE BARI
Inscriptions by Christians in Rome (3rd-8th cent. CE)
UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO
Fed by the Department of Classics and Late Antiquity - University of Bari (Italy)
Designed by the KDDE Group, hosted at the Department of Computer Science - University of Bari (Italy)
Graphic design by Maria Martinelli - Department of Classics and Late Antiquity
europeana
eagle project

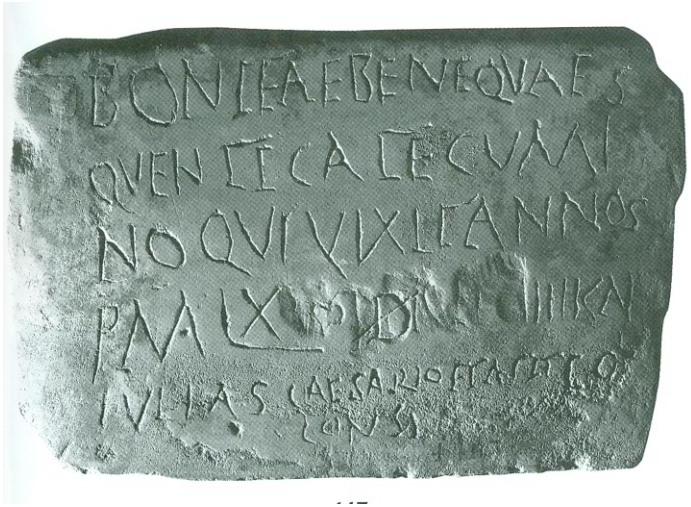
ICVR IX, 25046



ICVR II, 6498

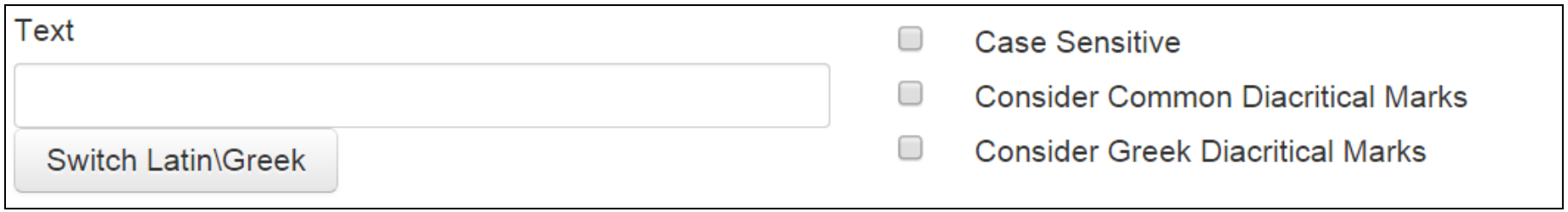


ICVR V, 10852

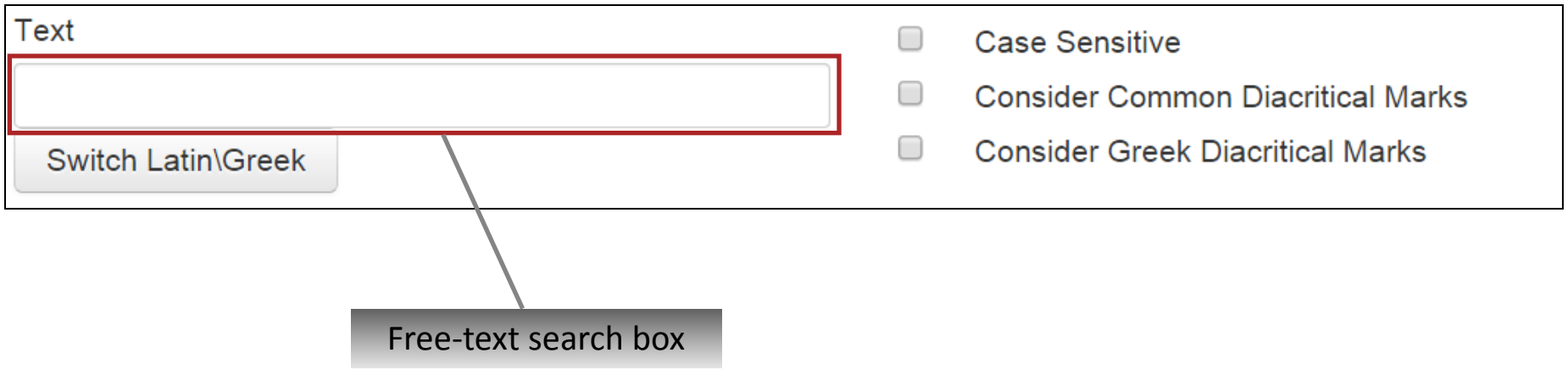


ICVR I, 2812

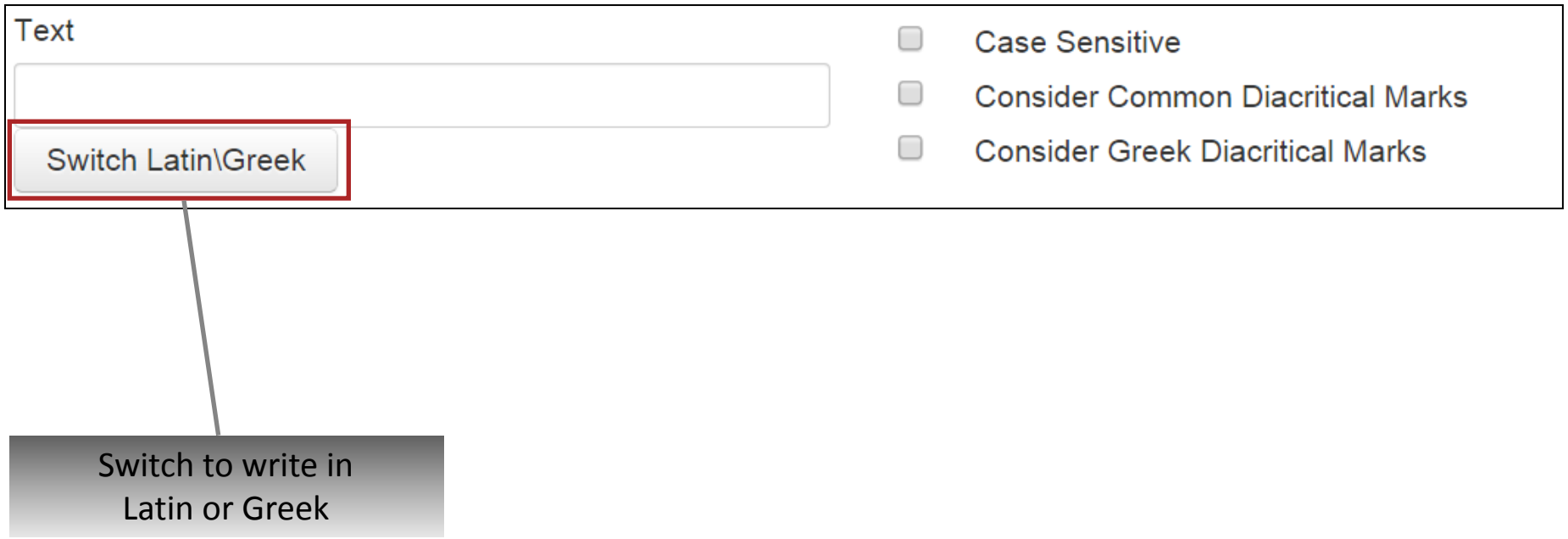
Text-based search in EDB



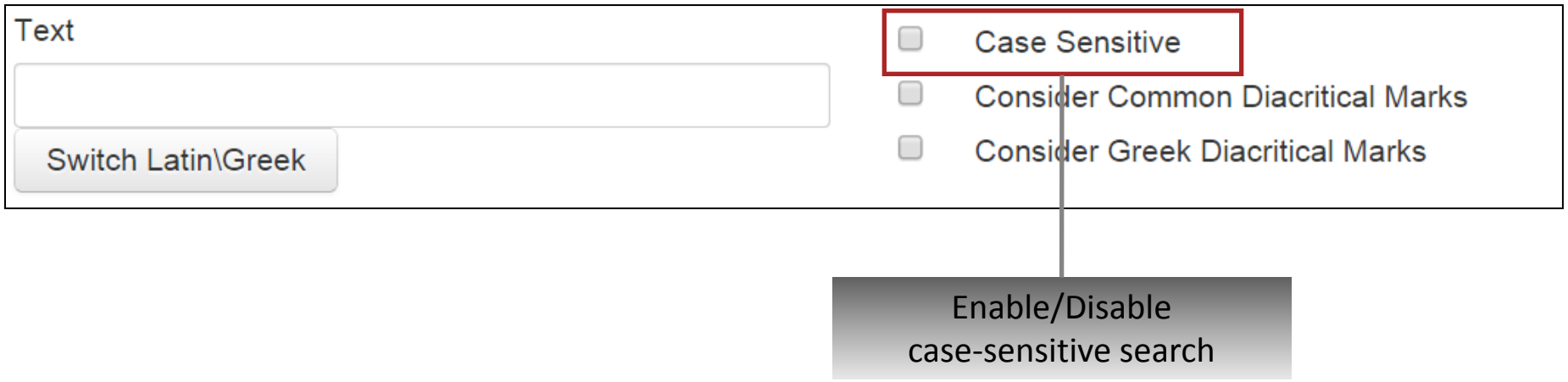
Text-based search in EDB



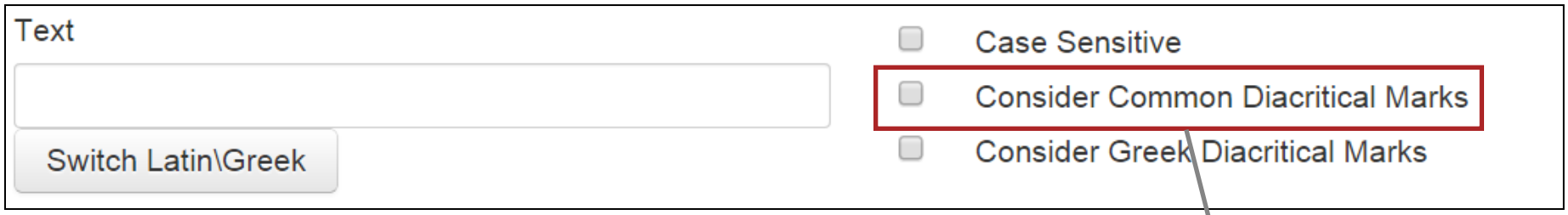
Text-based search in EDB



Text-based search in EDB

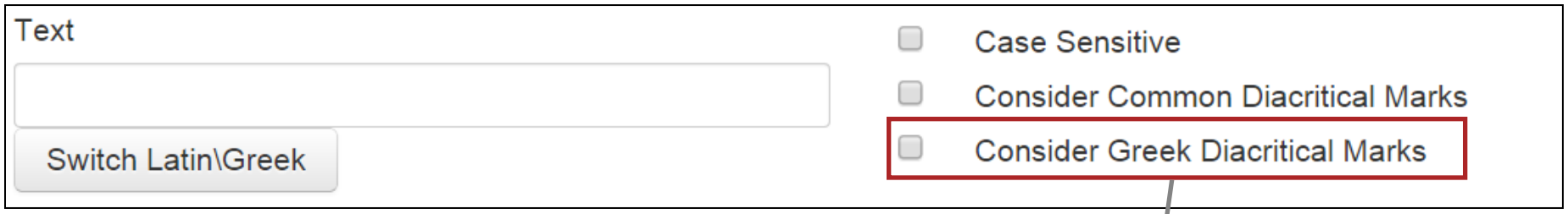


Text-based search in EDB



Consider/Ignore common diacritical marks in the inscriptions

Text-based search in EDB



Consider/Ignore Greek diacritical marks in the inscriptions

Text-based matching

Exact word matching

Query: *quiescit* *"quiescit"*

Output:

ID	ICVR	Bibliography	Pertinence	Conservation	Text
98	10129		Coem. Callisti pars inferior	-	stringe dolor lacrimas quaeris plebs sancta redemptum / levitam subito rapuit sibi regia caeli / dulcia nectario promebat mella canendo / prophetam celebrans placido modulamine senem / haec fuit insontis vita laudata iuventus / invidia infelix tandem conpressa quiescit / nunc paradisus habet sumsit qui ex hoste tropaea Close
303	16001.b		Coem. ss.Marcellini et Petri	-	((crux quadrata)) hic quiescit Romanus / presbiter qui sidit / presbiterio annus XXVI / More
399	14787		Coem. Praetextati	-	[--- qu?]ies(cit) die IIII non(as) octobr(es) / [---] XXXV

All the returned inscriptions contain exactly the term *quiescit*.
Inscriptions containing the terms ***requiescit, conquiescit***, etc. are not returned.

Text-based matching

Exact word matching

The inscription is correctly returned also in presence of diacritical marks (when ignored)

Query: *quiescit* *"quiescit"*

Output:

ID	ICVR	Bibliography	Pertinence	Conservation	Text
98	10129		Coem. Callisti pars inferior	-	stringe dolor lacrimas quaeris plebs sancta redemptum / levitam subito rapuit sibi regia caeli / dulcia nectario promebat mella canendo / prophetam celebrans placido modulamine senem / haec fuit insontis vita laudata iuventus / invidia infelix tandem conpressa quiescit / nunc paradisus habet sumsit qui ex hoste tropaea Close
303	16001.b		Coem. ss.Marcellini et Petri	-	((crux quadrata)) hic quiescit Romanus / presbiter qui sidit / presbiterio annus XXVI / More
399	14787		Coem. Praetextati	-	[--- qu?]ies(cit) die IIII non(as) octobr(es) / [---] XXXV

All the returned inscriptions contain exactly the term *quiescit.*
Inscriptions containing the terms ***requiescit, conquiescit***, etc. are not returned.

Text-based matching

Partial word matching

Query: **quiescit*

Output:

ID	ICVR	Bibliography	Pertinence	Conservation	Text
98	10129		Coem. Callisti pars inferior	-	stringe dolor lacrimas quaeris plebs sancta redemptum / levitam subito rapuit sibi regia caeli / dulcia nectario promebat mella canendo / prophetam celebrans placido modulamine senem / haec fuit insontis vita laudata iuventus / invidia infelix tandem conpressa quiescit / nunc paradisus habet sumsit qui ex hoste tropaea Close
303	16001.b		Coem. ss.Marcellini et Petri	-	((crux quadrata)) hic quiescit Romanus / presbiter qui sidit / presbiterio annus XXVI / mensis X dep(ositus) / X kal(endas) aug(ustas) / [c]ons(ulatu) Seberini v(iri) c(larissimi) c(onsulis) Close
304	16002		Coem. ss.Marcellini et Petri	-	[h]ic requiescit [---] / [t]ituli Eusebi q[ui v]ixit ann(is) [---] / [---] in pace p(ri)d(ie) kal(endas) febr(uarias) [Le]one iun(iore) aug(usto) pr[imum cons(ule)] Close

Returned inscriptions contain the term *quiescit* or other terms ending with the text *quiescit*. Inscriptions containing the terms ***requiescit***, ***conquiescit***, etc are also returned.

Text-based matching

Partial word matching

Query: **quiescit*

Further possibilities:
quescit*
quescit

Output:

ID	ICVR	Bibliography	Pertinence	Conservation	Text
98	10129		Coem. Callisti pars inferior	-	stringe dolor lacrimas quaeris plebs sancta redemptum / levitam subito rapuit sibi regia caeli / dulcia nectario promebat mella canendo / prophetam celebrans placido modulamine senem / haec fuit insontis vita laudata iuventus / invidia infelix tandem conpressa quiescit / nunc paradisus habet sumsit qui ex hoste tropaea Close
303	16001.b		Coem. ss.Marcellini et Petri	-	((crux quadrata)) hic quiescit Romanus / presbiter qui sidit / presbiterio annus XXVI / mensis X dep(ositus) / X kal(endas) aug(ustas) / [c]ons(ulatu) Seberini v(iri) c(larissimi) c(onsulis) Close
304	16002		Coem. ss.Marcellini et Petri	-	(h]ic requiescit [---] / [t]ituli Eusebi q[ui v]ixit ann(is) [---] / [---] in pace p(ri)d(ie) kal(endas) febr(uarias) [Le]one iun(iore) aug(usto) pr[imum cons(ule)] Close

Returned inscriptions contain the term *quiescit* or other terms ending with the text *quiescit*.
Inscriptions containing the terms ***requiescit***, ***conquiescit***, etc are also returned.

Text-based matching

Exact phrase matching

Query: *"quescit in pace"*

Output:

ID	ICVR	Bibliography	Pertinence	Conservation	Text
17806	27417		Coem. et basilica s.Valentini	-	Fl(avio) Seberiano [---] / qui vixit an[nos ---] / d(e)c(essit) XIII kal(endas) aug(ustas) [---] quesc[it in pace?] Close
20962	8930		Tituli qui in coem. Callisti reperti traduntur	-	Grecina Lyconis / quescit in pace ((ramus))

Returned inscriptions contain the exact sequence of terms *quiescit in pace*.
Inscriptions containing the specified terms in any other order are not returned.

Text-based matching

Exact phrase matching

The inscription is correctly returned also in presence of diacritical marks (when ignored)

Query: *"quescit in pace"*

Output:

ID	ICVR	Bibliography	Pertinence	Conservation	Text
17806	27417		Coem. et basilica s.Valentini	-	Fl(avio) Seberiano [---] / qui vixit an[nos ---] / d(e)c(essit) XIII kal(endas) aug(ustas) [--] quesc[it in pace?] Close
20962	8930		Tituli qui in coem. Callisti reperti traduntur	-	Grecina Lyconis / quescit in pace ((ramus))

Returned inscriptions contain the exact sequence of terms *quiescit in pace*.
Inscriptions containing the specified terms in any other order are not returned.

Text-based matching

Partial phrase matching (further possibilities)

- *benemerenti *quiescit*
- **enem* quiescit*
- *“in pace” benemerenti*
- *“in pace” “in sommo”*
- *enem* *quescit “in pace”*
- *etc.*

Dealing with aberrant forms

Some specific characteristics of Late Antique inscriptions should not be lost in the transcription.

Dealing with aberrant forms

Some specific characteristics of Late Antique inscriptions should not be lost in the transcription.

- the so-called **aberrant forms** should not be normalized, if they are grapho-phonetic outcomes of linguistic modifications/evolutions of Latin and Greek.

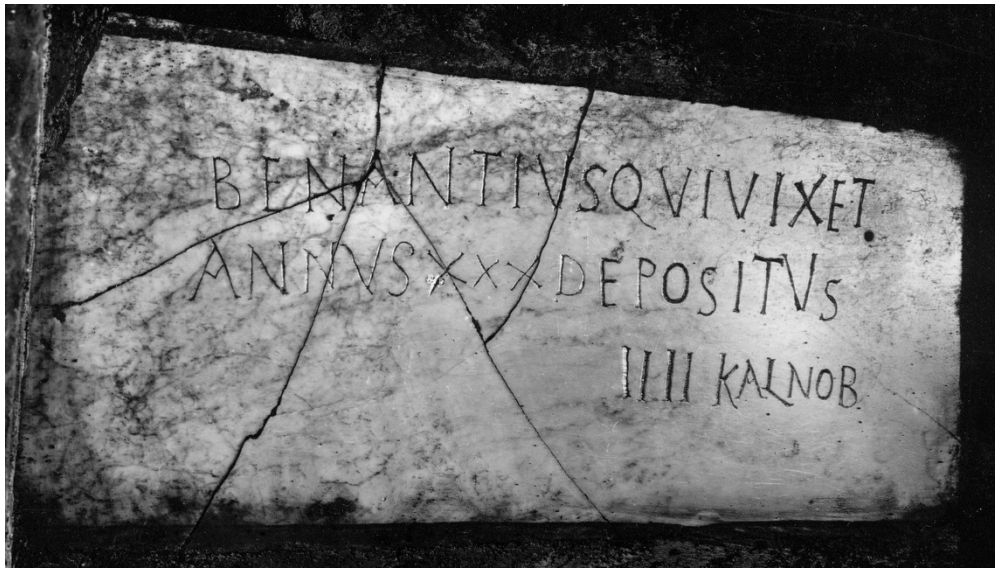
Dealing with aberrant forms

Some specific characteristics of Late Antique inscriptions should not be lost in the transcription.

➡ the so-called **aberrant forms** should not be normalized, if they are grapho-phonetic outcomes of linguistic modifications/evolutions of Latin and Greek.

Challenge: handling the possible presence of aberrant forms of the same term. The query system should be able to **match the query with all the inscriptions containing different spellings** of a word.

Dealing with aberrant forms



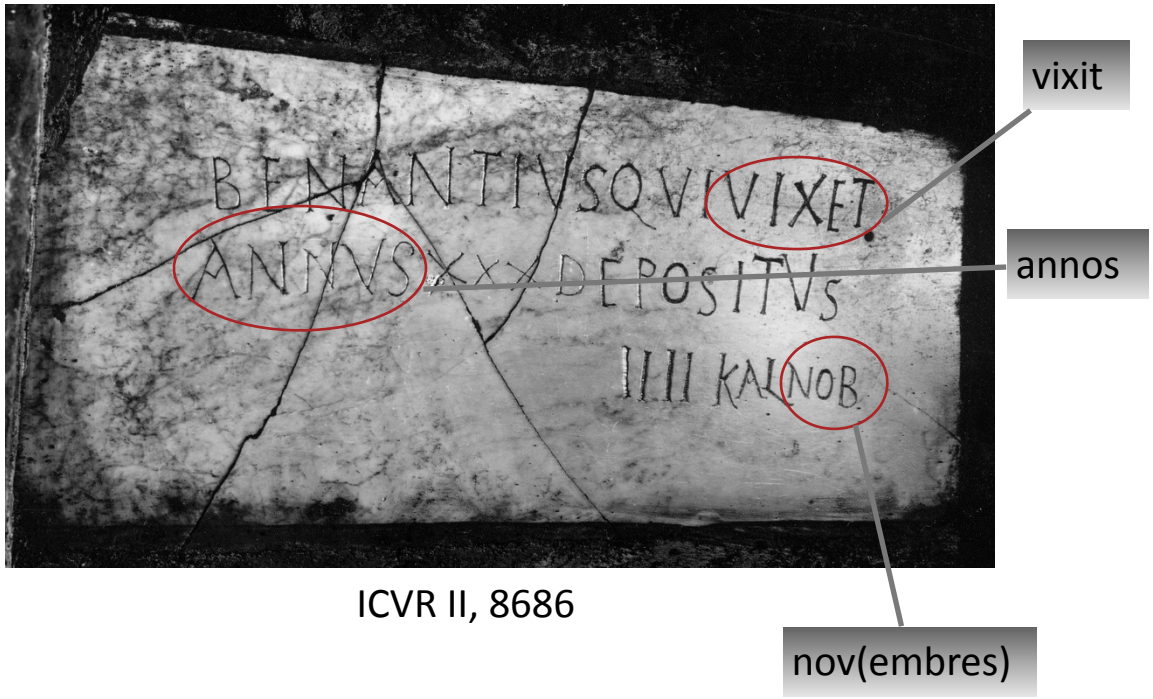
ICVR II, 8686

EDB
Benantius qui vixet
annus XXX depositus
IIII kal(endas) nob(embres)

EDCS
Benantius qui vix<i=E>t
ann<o=V>s XXX depositus
IIII Kal(endas) No<v=B>(embres)

Krummrey - Panciera
Benantius qui vixet (!)
annus (!) XXX depositus
IIII Kal(endas) Nob(embres) (!)

Dealing with aberrant forms



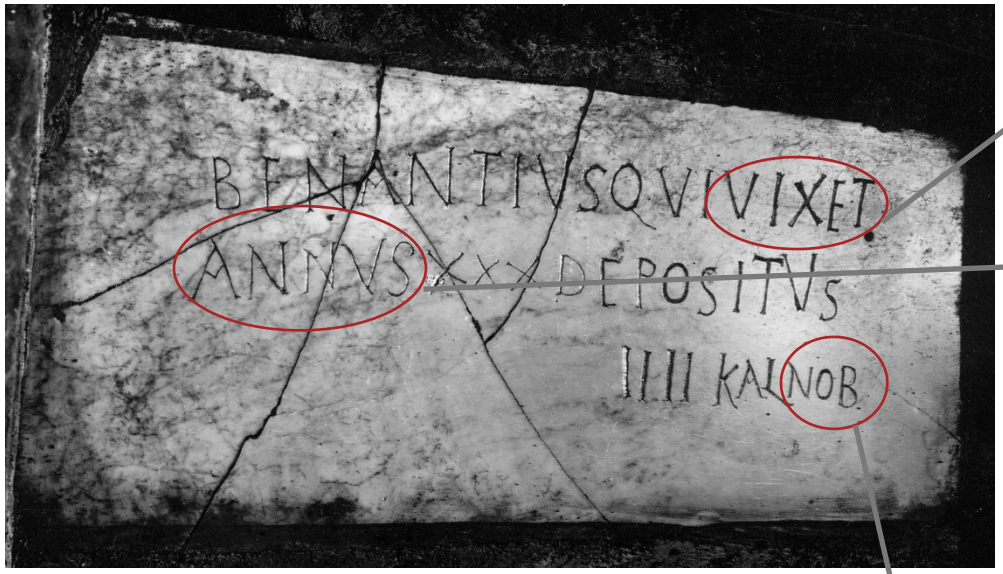
ICVR II, 8686

EDB
Benantius qui vixet
annus XXX depositus
IIII kal(endas) nob(embres)

EDCS
Benantius qui vix<i=E>t
ann<o=V>s XXX depositus
IIII Kal(endas) No<v=B>(embres)

Krummrey - Panciera
Benantius qui vixet (!)
annus (!) XXX depositus
IIII Kal(endas) Nob(embres) (!)

Dealing with aberrant forms



ICVR II, 8686

vixit

annos

nov(embres)

Normalized Text
Benantius qui vixit
annos XXX depositus
IV kal(endas) nov(embres)

EDB
Benantius qui vixet
annus XXX depositus
IIII kal(endas) nob(embres)

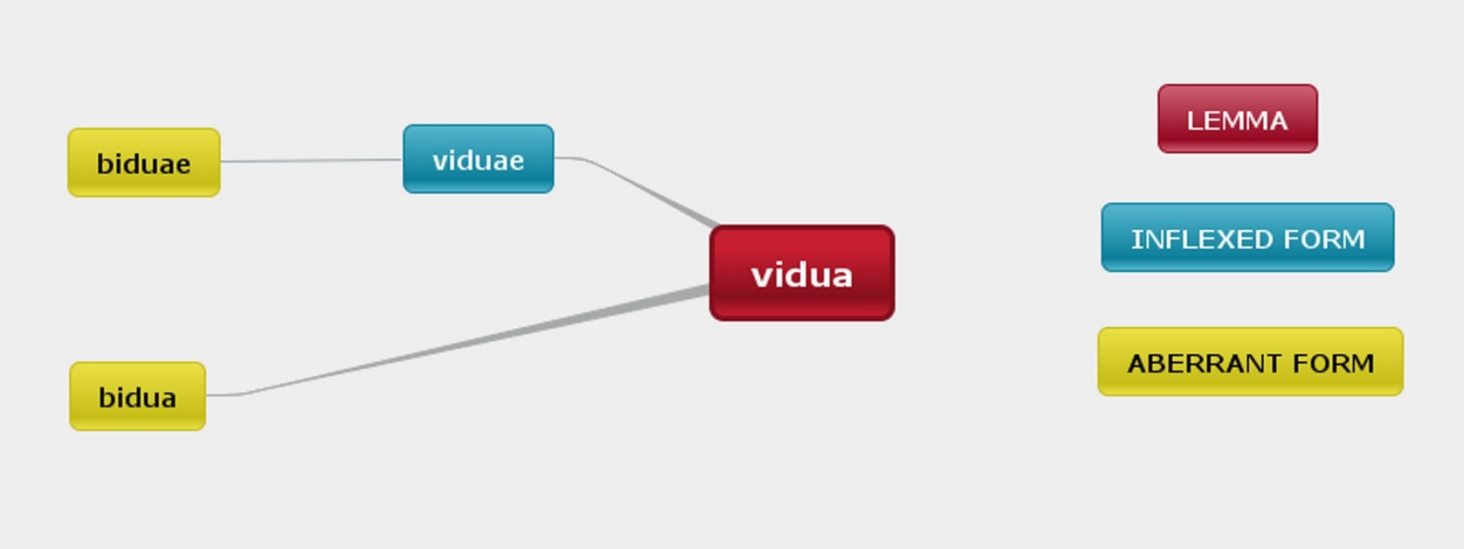
EDCS
Benantius qui vix<i=E>t
ann<o=V>s XXX depositus
IIII Kal(endas) No<v=B>(embres)

Krummrey - Panciera
Benantius qui vixet (!)
annus (!) XXX depositus
IIII Kal(endas) Nob(embres) (!)

Dealing with aberrant forms

vidua ⟷ *bidua*

The query system should consider all the aberrant forms as the same term. Each term is reconducted to its lemma.



Proposed Solution

1. Extraction of **the list of all the terms** from the collection of stored inscriptions.

[1] *www.ilc.cnr.it/lemlat/* for Latin.

Proposed Solution

1. Extraction of **the list of all the terms** from the collection of stored inscriptions.
2. Identification of the **lemma of each term**, by adopting an appropriate lemmatizer[1]. When the tool fails to identify the correct lemma, it is manually specified by experts.

[1] *www.ilc.cnr.it/lemlat/* for Latin.

Proposed Solution

1. Extraction of **the list of all the terms** from the collection of stored inscriptions.
2. Identification of the **lemma of each term**, by adopting an appropriate lemmatizer[1].
   When the tool fails to identify the correct lemma, it is manually specified by experts.
3. A **lemmatized version** of transcriptions is stored, by replacing each term with its lemma.

[1] *www.ilc.cnr.it/lemlat/* for Latin.

Proposed Solution

1. Extraction of **the list of all the terms** from the collection of stored inscriptions.
2. Identification of the **lemma of each term**, by adopting an appropriate lemmatizer[1].
   When the tool fails to identify the correct lemma, it is manually specified by experts.
3. A **lemmatized version** of transcriptions is stored, by replacing each term with its lemma.
4. When a new inscription is stored, **steps 2-3 are executed only for its transcription**.
   If there are new terms that are still not indexed, their lemmas are (automatically or, if necessary, manually) extracted and **a lemmatized version of the transcription is stored**.

[1] *www.ilc.cnr.it/lemlat/* for Latin.

Proposed Solution

1. Extraction of **the list of all the terms** from the collection of stored inscriptions.
2. Identification of the **lemma of each term**, by adopting an appropriate lemmatizer[1].
   When the tool fails to identify the correct lemma, it is manually specified by experts.
3. A **lemmatized version** of transcriptions is stored, by replacing each term with its lemma.
4. When a new inscription is stored, **steps 2-3 are executed only for its transcription**.
   If there are new terms that are still not indexed, their lemmas are (automatically or, if necessary, manually) extracted and **a lemmatized version of the transcription is stored**.

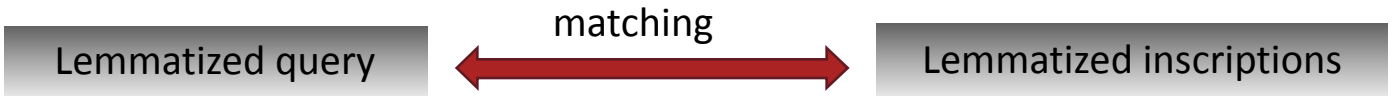
➡ matching between terms and lemmas is **built incrementally** when not-indexed terms appear in a new inscription.

[1] *www.ilc.cnr.it/lemlat/* for Latin.

Proposed Solution

1. Extraction of **the list of all the terms** from the collection of stored inscriptions.
2. Identification of the **lemma of each term**, by adopting an appropriate lemmatizer[1].
   When the tool fails to identify the correct lemma, it is manually specified by experts.
3. A **lemmatized version** of transcriptions is stored, by replacing each term with its lemma.
4. When a new inscription is stored, **steps 2-3 are executed only for its transcription**.
   If there are new terms that are still not indexed, their lemmas are (automatically or, if necessary, manually) extracted and **a lemmatized version of the transcription is stored**.

→ matching between terms and lemmas is **built incrementally** when not-indexed terms appear in a new inscription.



[1] *www.ilc.cnr.it/lemlat/* for Latin.

Conclusions

• Some issues about the text-based search of inscriptions have been analyzed.

• The presence of **common and Greek diacritical marks** is already considered in the query system of Epigraphic Database Bari (EDB).

• **Exact, partial and hybrid matching of single words or of whole phrases** is already available in the text-based search of EDB.

• The presence of aberrant forms can be handled by semi-automatic procedures which allow the query system to **perform matching between lemmas** instead of between the original terms.

Questions?