

# Cross-language Information Retrieval in Geoportal Discovery Services



**Xeni Kechagioglou, Michael Lutz,**  
Nicole Ostländer, Hong Cao, Ioannis Kanellopoulos

*Joint Research Centre (JRC), Institute for Environment & Sustainability*

3<sup>rd</sup> Workshop “**Construction of multilingual ontologies  
for Urban Civil Engineering projects**”

20 October 2008, Zaragoza, Spain

- Introduction
- Similarity-based discovery
- Helping the user with the similarity approach
- Conclusion & Future Work

- **Introduction**
  - Setting: INSPIRE & GEMET
  - Approach: Keyword search based on a multi-lingual thesaurus
- Similarity-based discovery
- Helping the user with the similarity approach
- Conclusion & Future Work

- Text-based search
  - ⇒ low recall
  - ⇒ low precision

- Multi-lingual metadata
  - ⇒ support for multi-lingual discovery

Keyword

soil

Organisation name

http://www.inspire-geoportal.eu - Current search crit...

File Edit View History Bookmarks Tools Help

Current search criteria

A B C D E F G H I J K L M

N O P Q R S T U V W X Y Z Other

ArcIMS, area di ablazione, area di accumulo, axl, bilancio di massa, copertura detritica

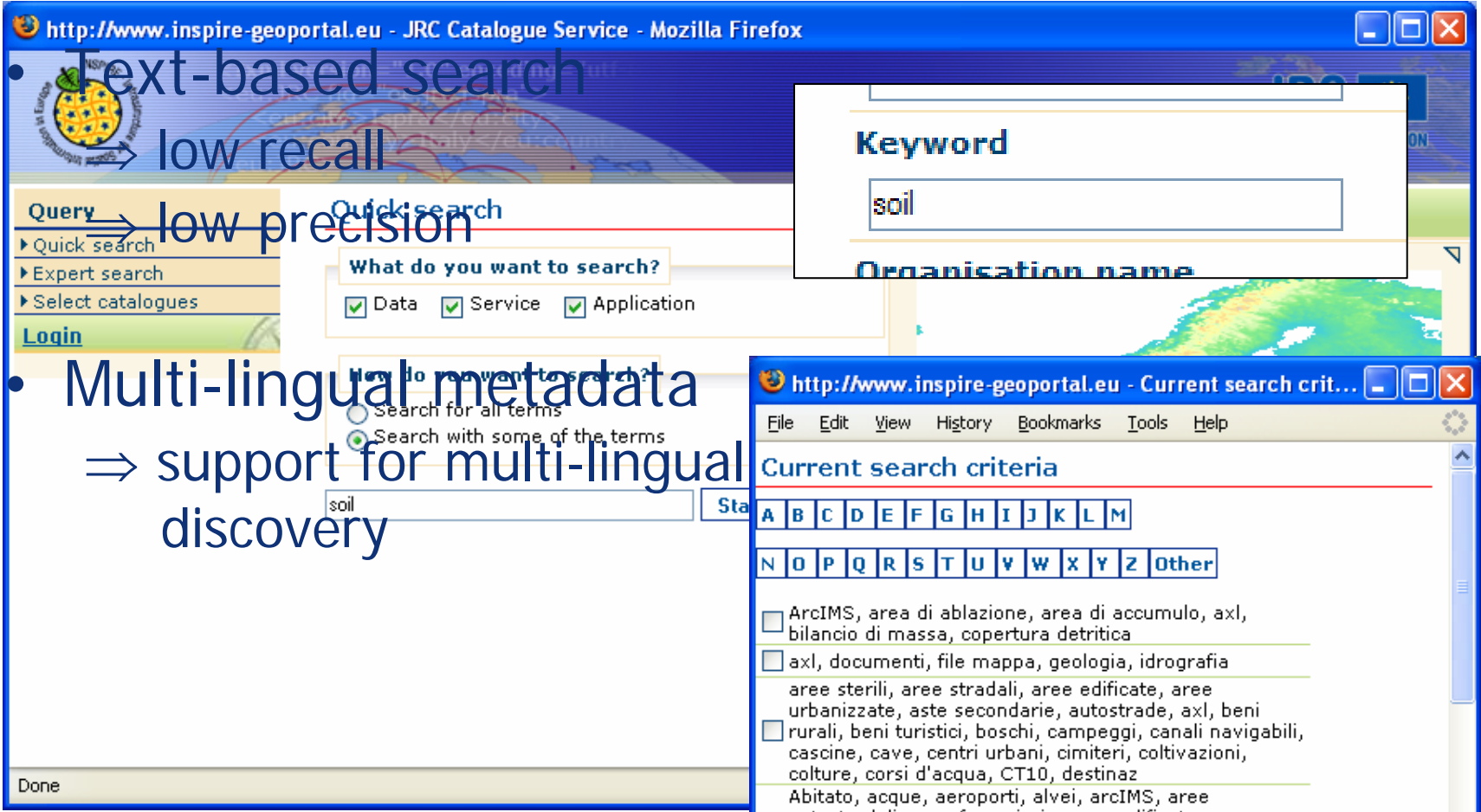
axl, documenti, file mappa, geologia, idrografia

aree sterili, aree stradali, aree edificate, aree urbanizzate, aste secondarie, autostrade, axl, beni rurali, beni turistici, boschi, campeggi, canali navigabili, cascine, cave, centri urbani, cimiteri, coltivazioni, colture, corsi d'acqua, CT10, destinaz

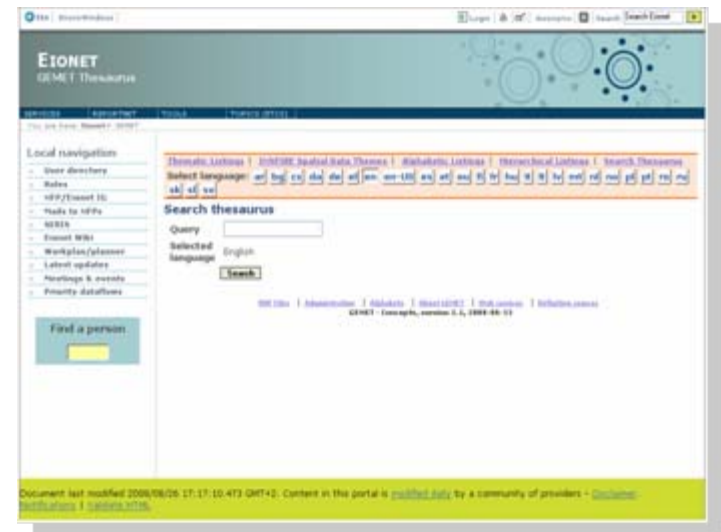
Abitato, acque, aeroporti, alvei, arcIMS, aree autostradali, aree ferroviarie, aree edificate, aree idriche, aree stradali, aree urbanizzate, argini, aste secondarie, autostrade, axl, canali navigabili, centri urbani, corsi d'acqua

Abbazie, abitato, alvei, architettura civile, architettura

Done



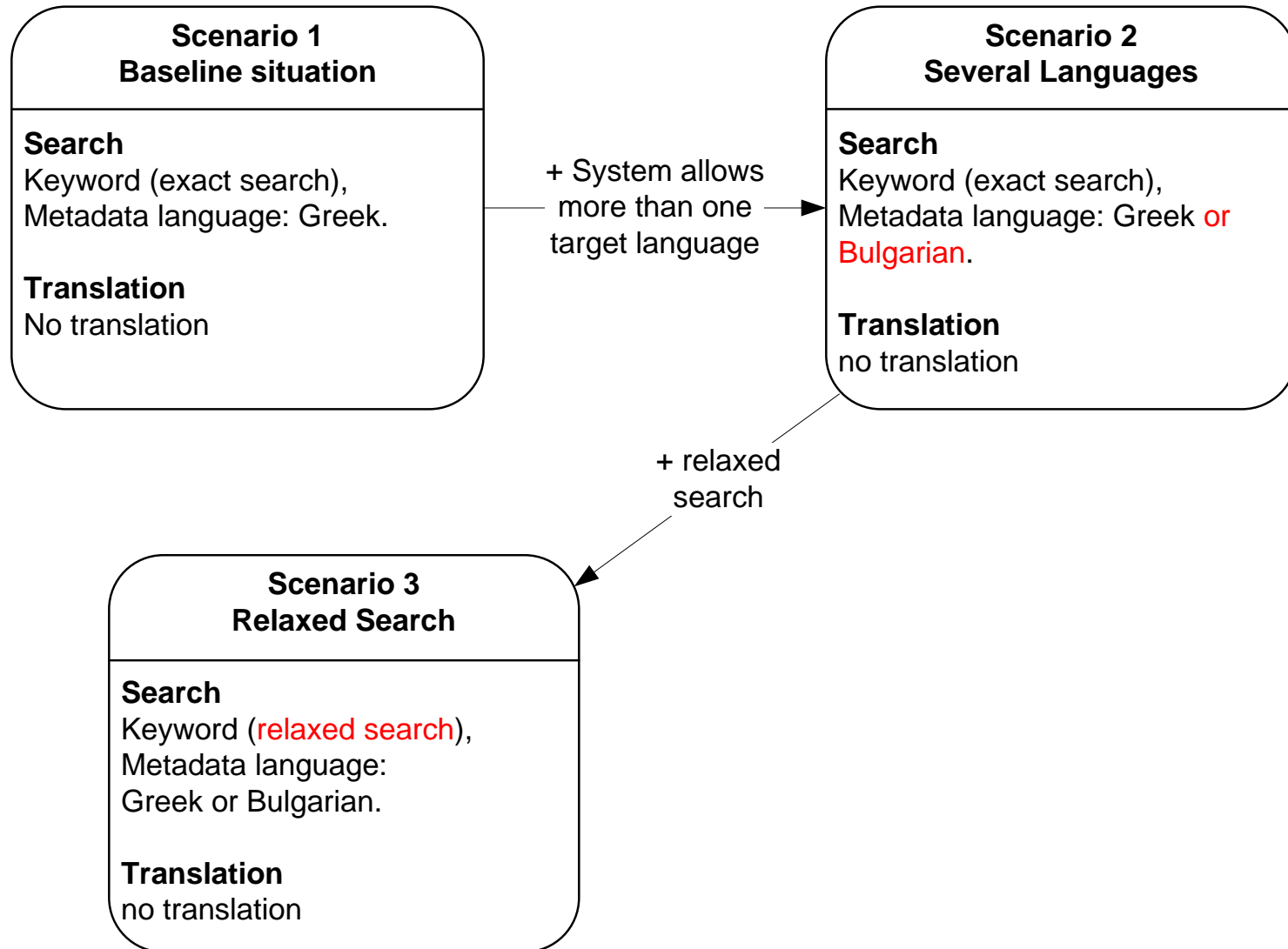
- Approach (Metadata IR)
  - “at least one keyword shall be provided from the General Environmental Multi-lingual Thesaurus (GEMET) describing the relevant spatial data theme ...”
- GEMET
  - Published and managed by EIONET
  - More than 5000 concepts
  - Currently 26 languages
  - Links to INSPIRE spatial data themes
    - ⇒ thematic structure
    - ⇒ additional entry point

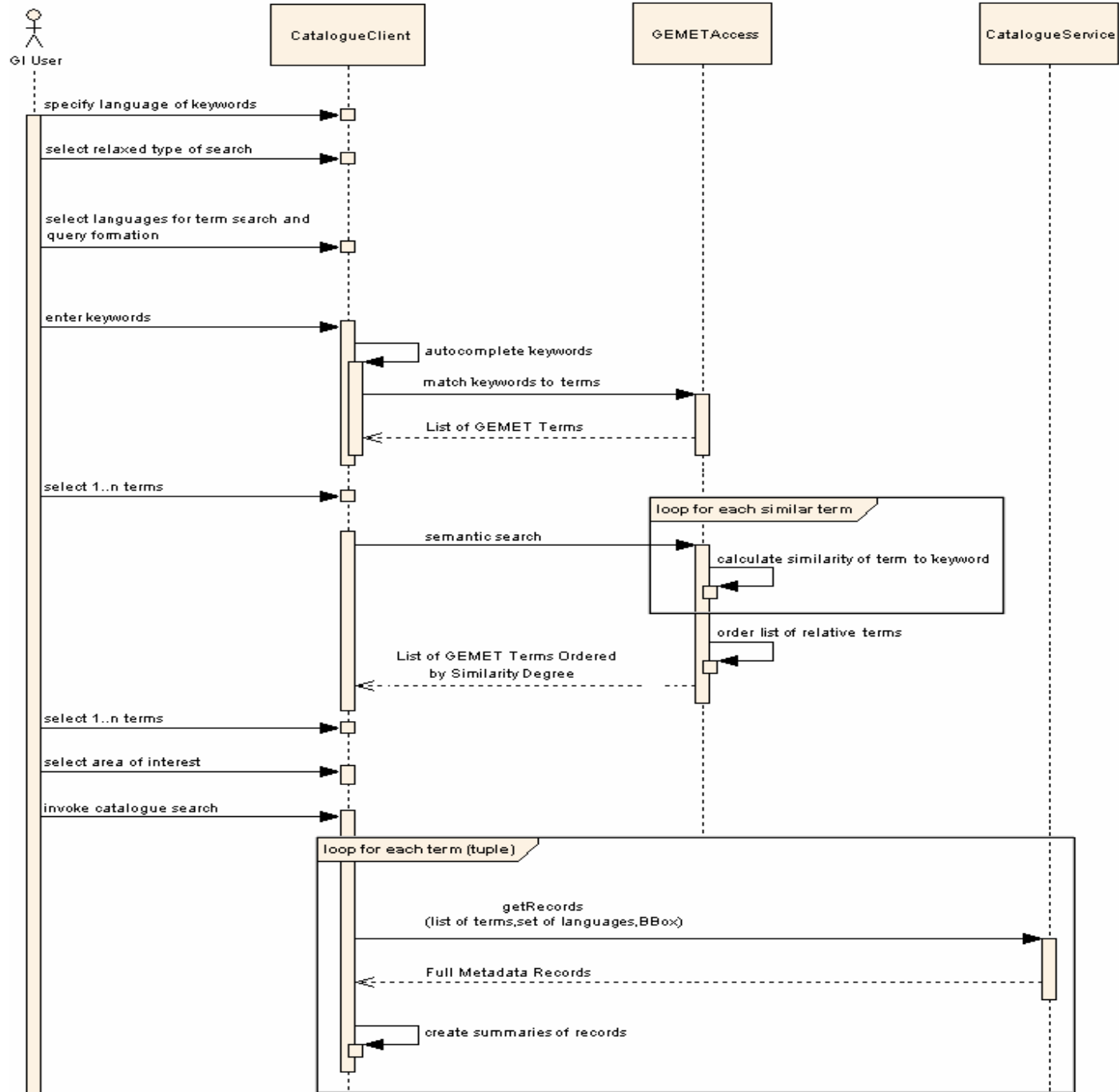


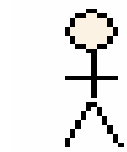
- Increasing precision
  - Using GEMET as a controlled vocabulary
  - Spatial data theme keyword required in metadata ⇒ can be used in discovery
- Increased recall
  - Spatial data themes: all 23 official EU languages  
GEMET terms: some official EU languages missing, but additional ones
  - Links to GEMET for suggesting other GEMET terms

- Use similarity to find (additional) keywords
  - further increase recall
- Use different presentations to help the user with choosing (additional) keywords
  - increase usability

- Introduction
- **Similarity-based discovery**
  - Discovery Scenarios
  - Similarity Measures
  - Demo
- Helping the user with the similarity approach
- Conclusion & Future Work

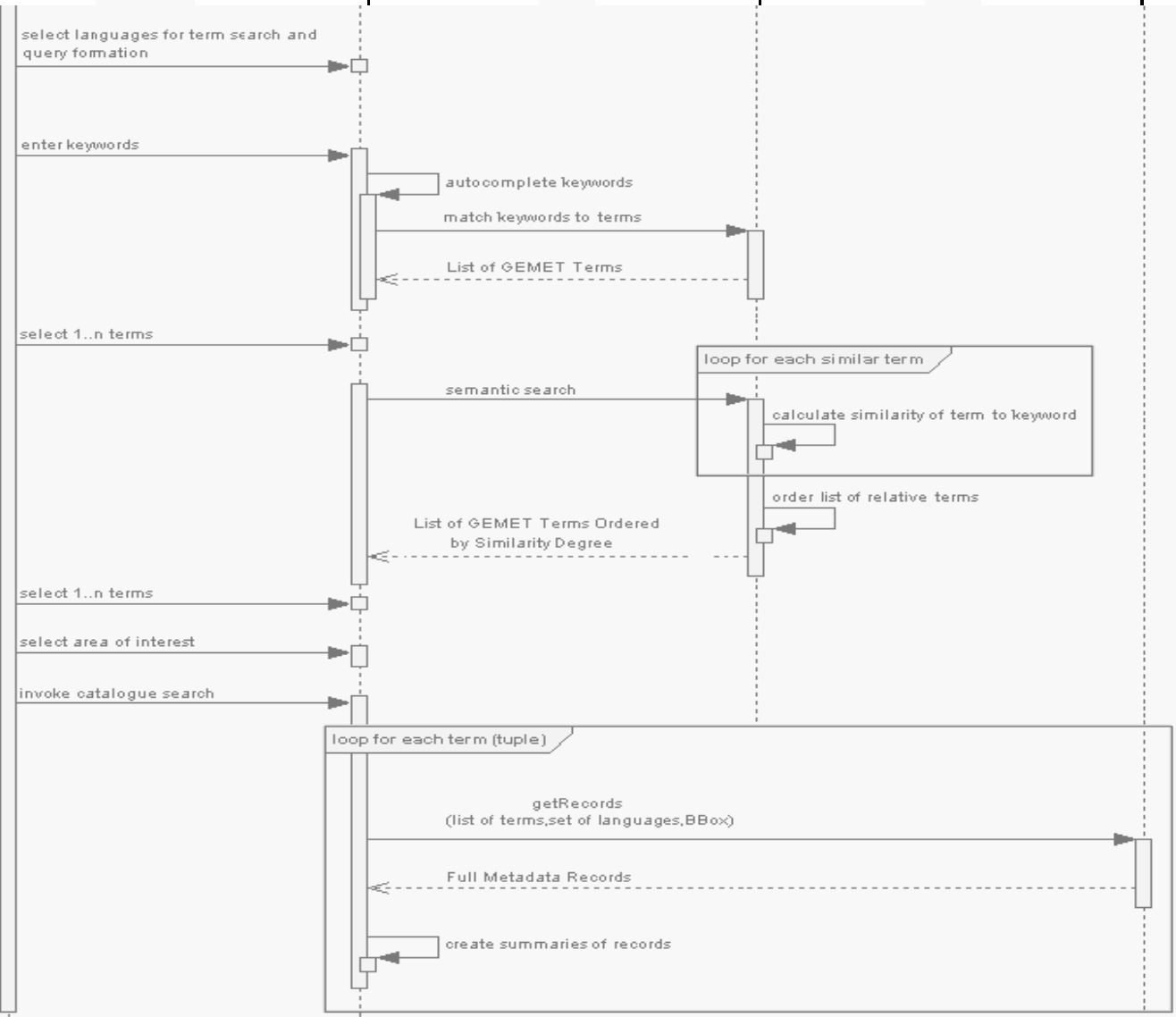
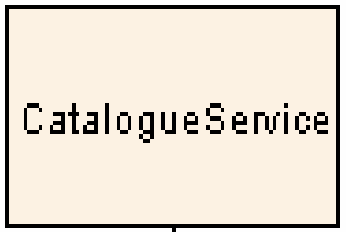
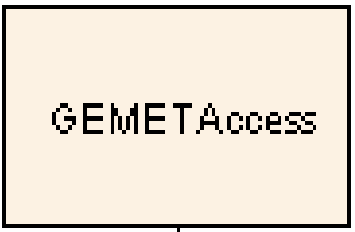


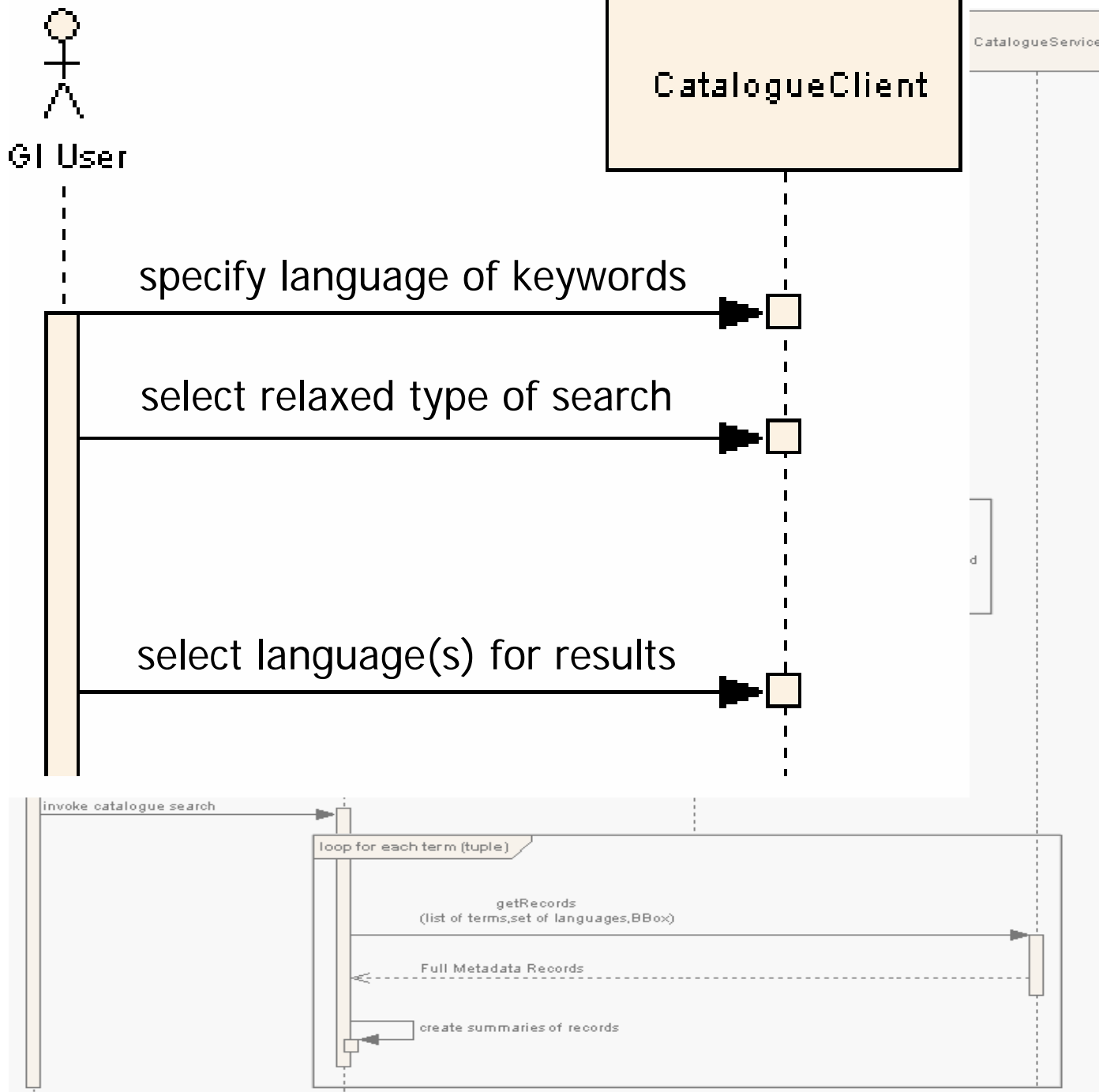


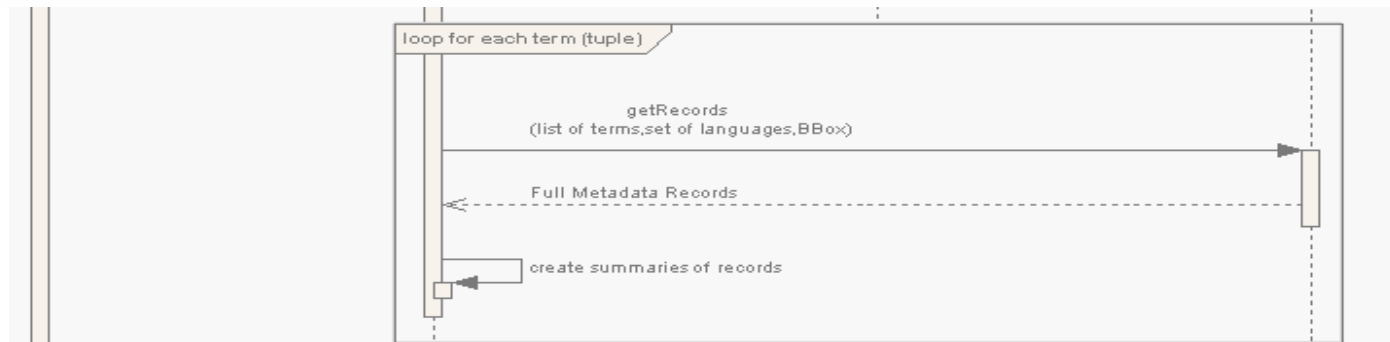
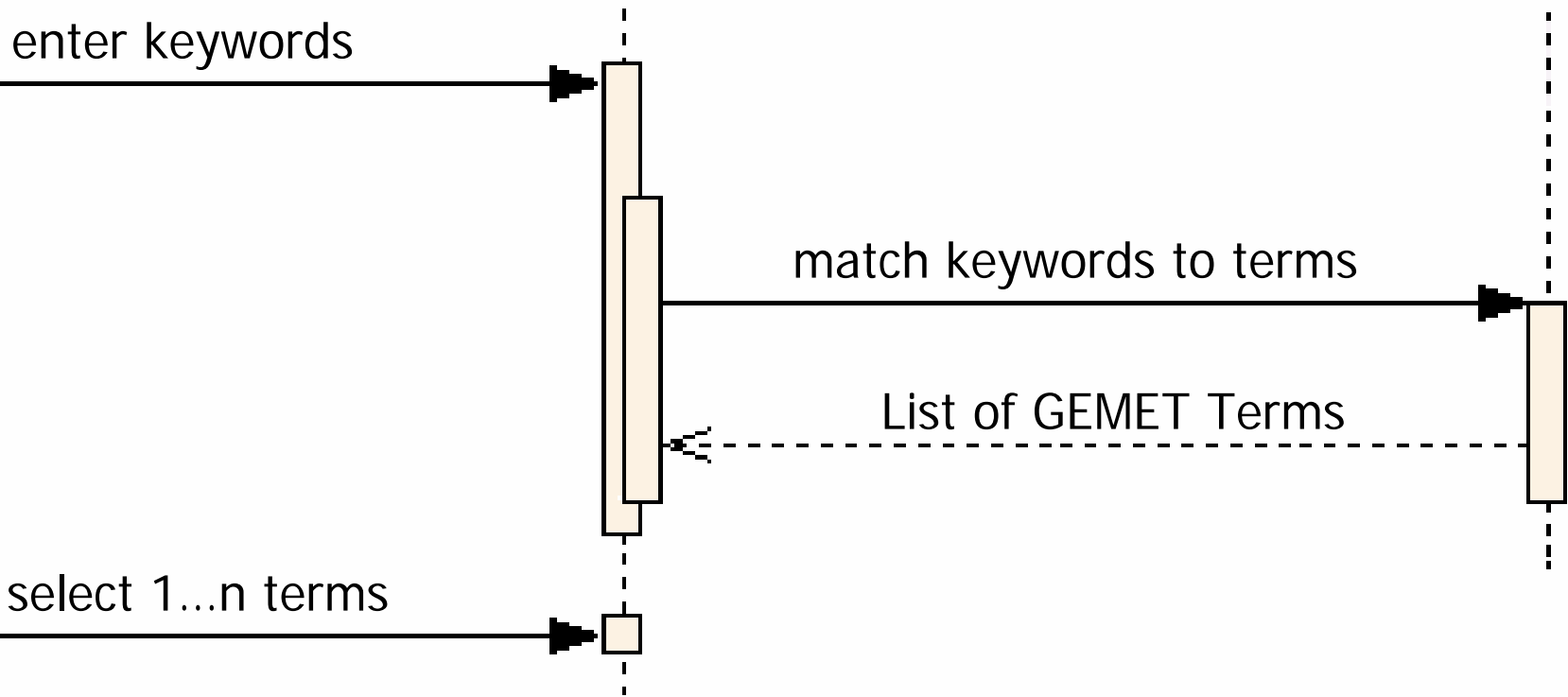
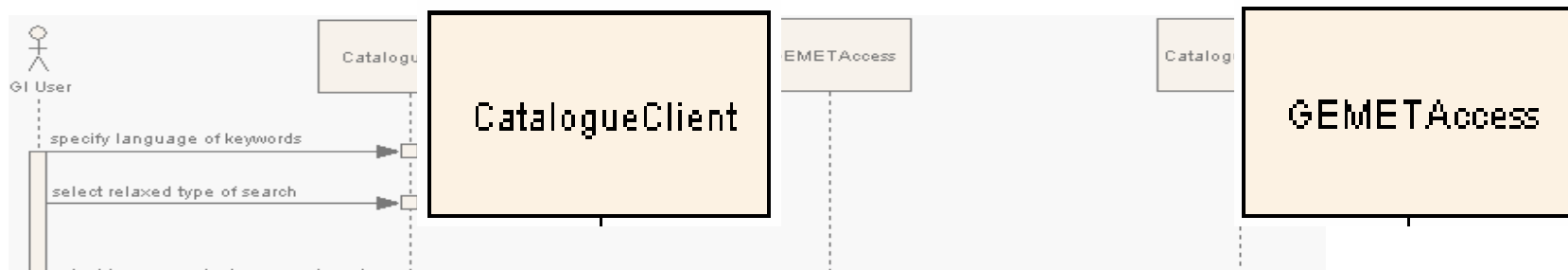
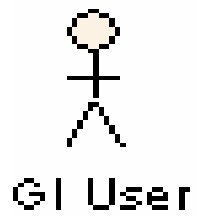


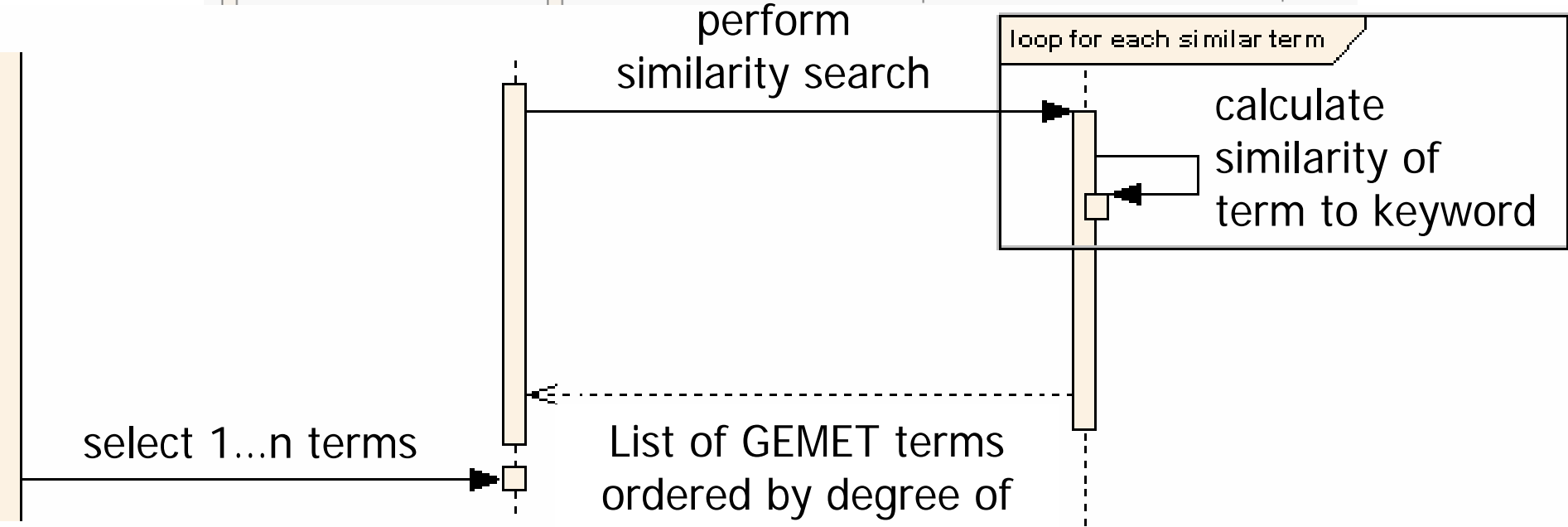
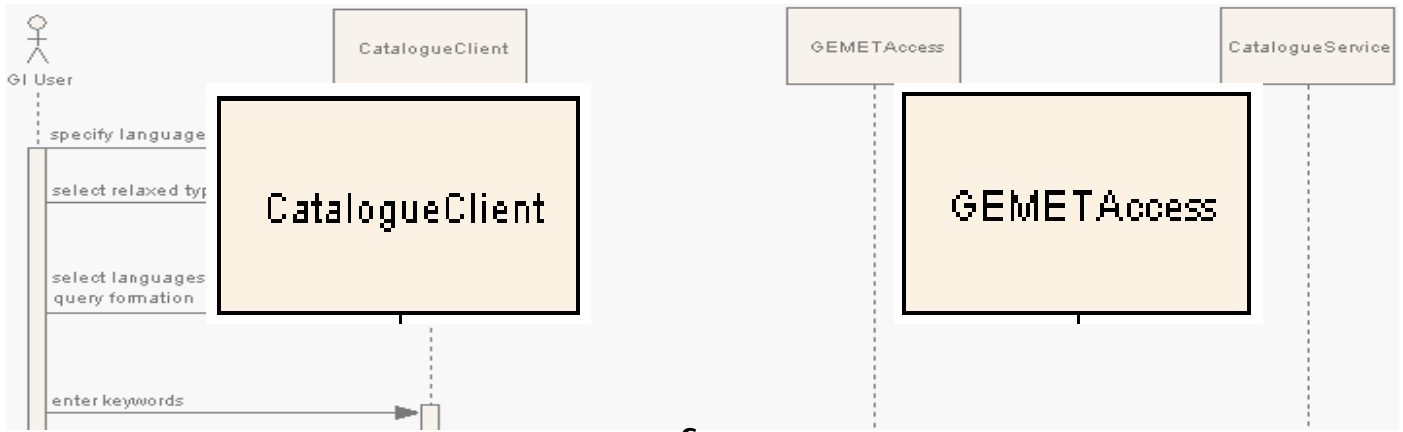
GI User

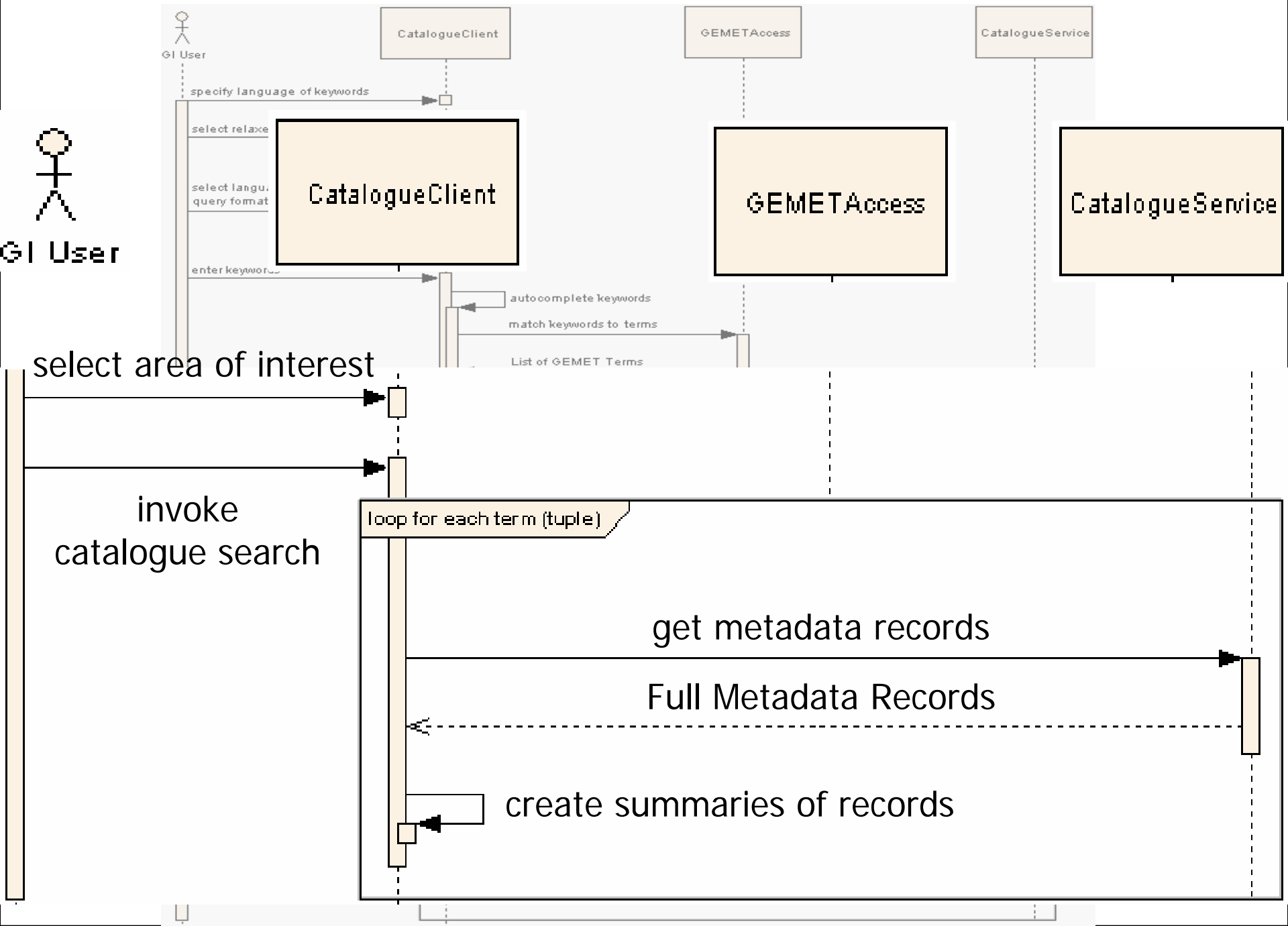
language of keyword  
keyword type of







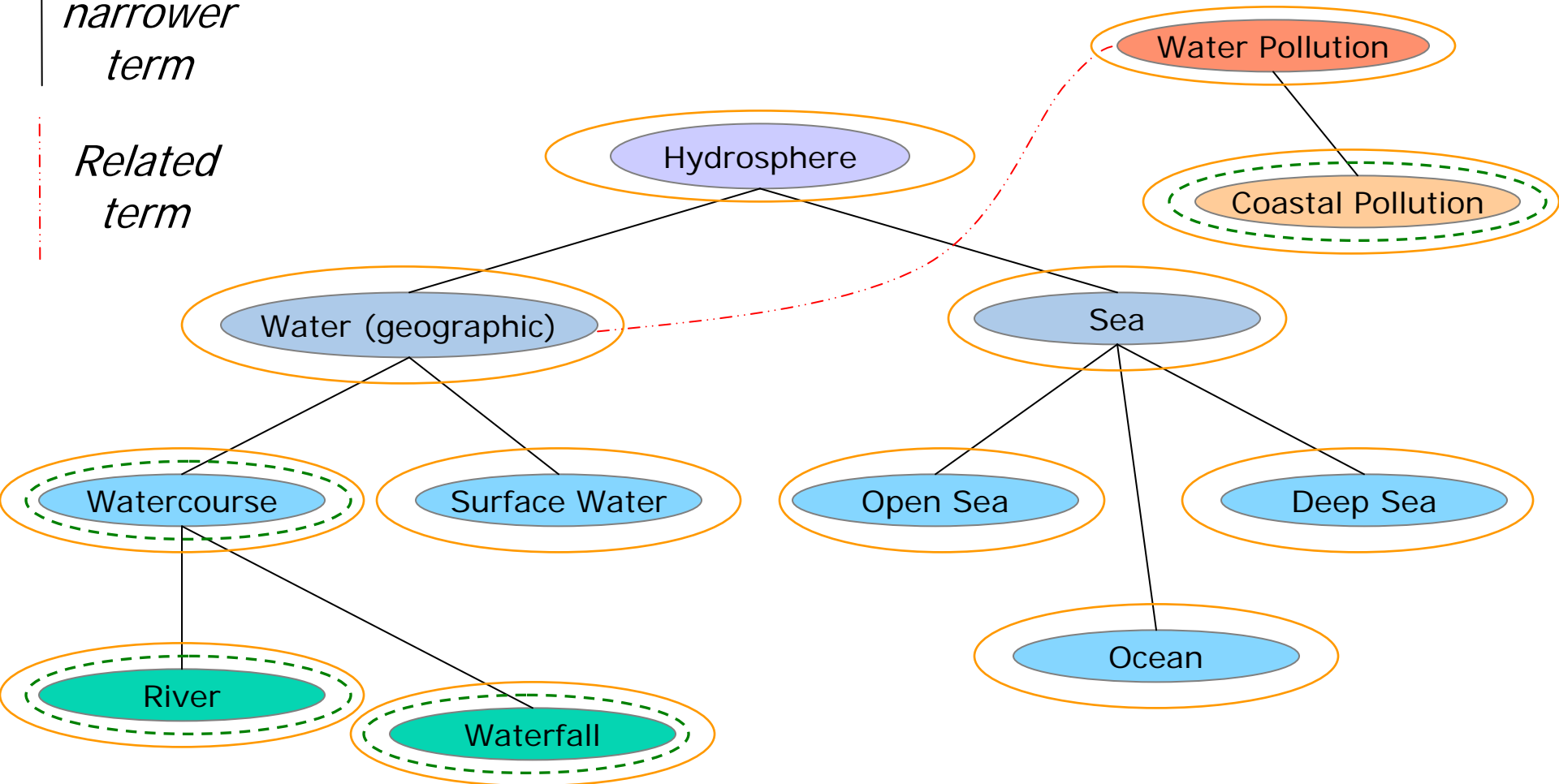




- Distance-based
  - based on the number of links between concepts
- Feature-based
  - based on common and distinct attributes of concepts
- Probabilistic
  - based on the information content of each concept
- Hybrid measure
  - combination of suitable measures

*Broader/  
narrower  
term*

*Related  
term*



- Probabilistic

- Information content (IC) derived from probability of finding the term in a metadata record
- Low probability  $\Rightarrow$  high information content

$$\text{Sim}(x_1, x_2) = \frac{2 * IC(C_3)}{IC(C_1) + IC(C_2)}$$

$C_1, C_2$ : concepts to compare

$C_3$ : most specific concept that subsumes both

- Distance based

- Equal weights for all semantic links
- Concepts deeper in the hierarchy assumed more similar

$$\text{Sim}(x_1, x_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

$N_1, N_2$ : nodes  $C_1/C_2 \rightarrow C_3$

$N_3$ : nodes  $C_3 \rightarrow$  root

- Query expansion using the similarity measure and the multi-lingual aspect of GEMET for the user's keywords

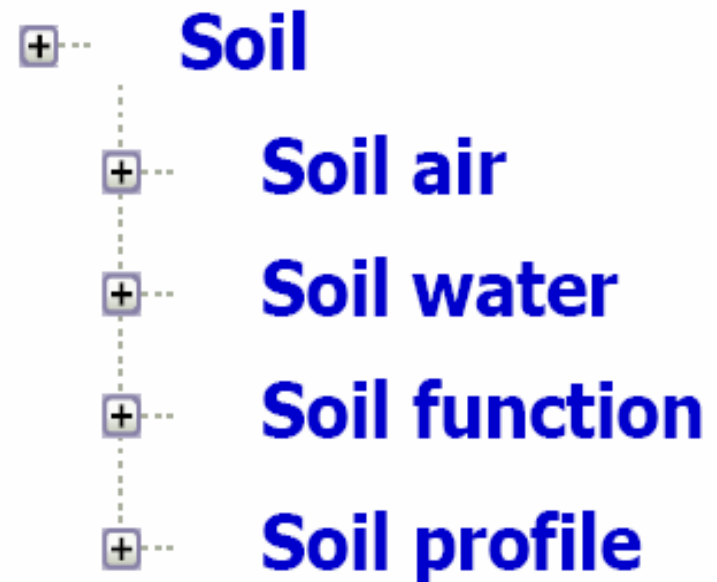
- Measures developed for ontologies (in principle IS-A links)
  - GEMET is a thesaurus
- Only considered hierarchical structure so far
- Multiple hierarchy
  - Currently the shortest paths are used
  - Should there be special treatment of such cases?

- Introduction
- Similarity-based discovery
- **Helping the user with the similarity approach**
  - Problems with similarity-based approach
  - Different presentation approaches based on
    - GEMET structure
    - metadata statistics
- Conclusion & Future Work

- Which similarity measure is suitable?
- Users might need to get used to a certain measure – what does 'similar' mean?
- No guidance on which (combination of) keyword(s) to pick

- Based on
  - GEMET structure
  - Frequency in metadata
  - GEMET structure & frequency in metadata
  - Co-occurrence of terms in metadata
  
- Starting points
  - Free text search term (from GEMET)
    - Similar GEMET terms
  - INSPIRE spatial data themes
    - Linked GEMET terms

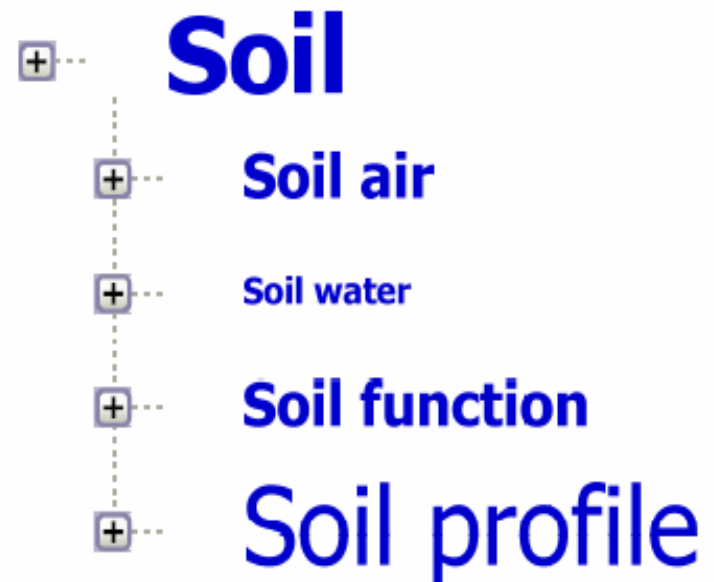
- Using structure of GEMET taxonomy
- Tree representation of GEMET taxonomy of concepts linked to the INSPIRE theme
- The user gets an idea of the (taxonomic) relationships of the selected GEMET terms



- Cloud representation based on frequency in metadata
  - of terms (one language)
  - of concept (in all selected languages)
- Independent of the GEMET taxonomy
- The user gets an idea of the amount of results ⇒ Natural selection of frequently used keywords

agricultural land **landscape**  
soil air soil analysis soil moisture  
**Soil profile** soil water

- Tree (or graph) representation; size of each keyword depends on frequency in metadata
  - of terms (one language)
  - of concept (in all selected languages)
- Combines the benefits of both approaches
- Might be more difficult to apply by inexperienced users



- How often are search term or INSPIRE theme used together with other terms in metadata?
- Cloud representation based on frequency of term pairs in metadata
- Can be used to
  - narrow down search
  - find additional keywords for metadata
- Might be more difficult to apply by inexperienced users



- Similarity search
  - System evaluation by users to compare similarity measures
  - Make use of thematic classification plus associative relations between concepts
  - Sorting of results according to degree of relativity to user's request
- Presentation of terms
  - System evaluation by users to compare different presentation approaches (requires GEMET tagging)
- Machine translation of metadata records

**xeni.kechagioglou@jrc.it**  
**michael.lutz@jrc.it**



**INSPIRE Geoportal**  
the EU portal for Geographic Information

**<http://inspire-geoportal.eu>**



**<http://www.ec-gis.org/inspire/>**