Source Code Search System Using The Knowledge Framework of The Semantic Web

Kosuke HOSOZAWA *Takeshi OGIHARA

The Graduate School of Science and Technology
Kobe University

Software Reuse

- Most of researches on software reuse have focused on the way of reusing software made in a closed organization.

↕

- The number of free software offered on the network is increasing.
- We want to get useful software on the net.

Search for Software

- Applying ordinary Web search to source codes, we cannot get satisfying result.
- The vocabulary contained in the source code is scarce.

search word: `sorting`

words in the source code:

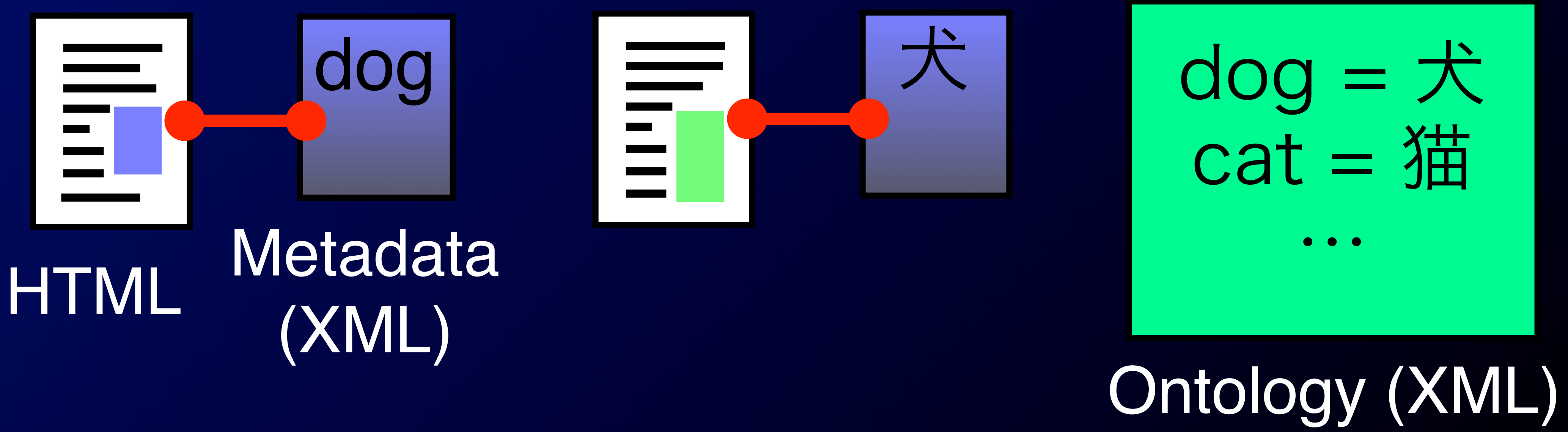
```
i j swap quicksort p q data ndata
arr pivot loopflag last val xx
```

Our Approach

- S4 (Semantic based Source code Share and Search) system is proposed.
  - It is assumed that the source codes are on the other sites.
  - S4 system makes automatically the metadata of the source codes in advance.
  - The source codes are classified referring to the ontologies in advance.
  - Ontologies are also used to search for related words in the source codes.

Semantic Web

- Machine-readable XML data (metadata) is attached to each HTML document to represent the contents of the document.
- Ontology is used to show the relation among the words in the vocabulary.



dog

HTML

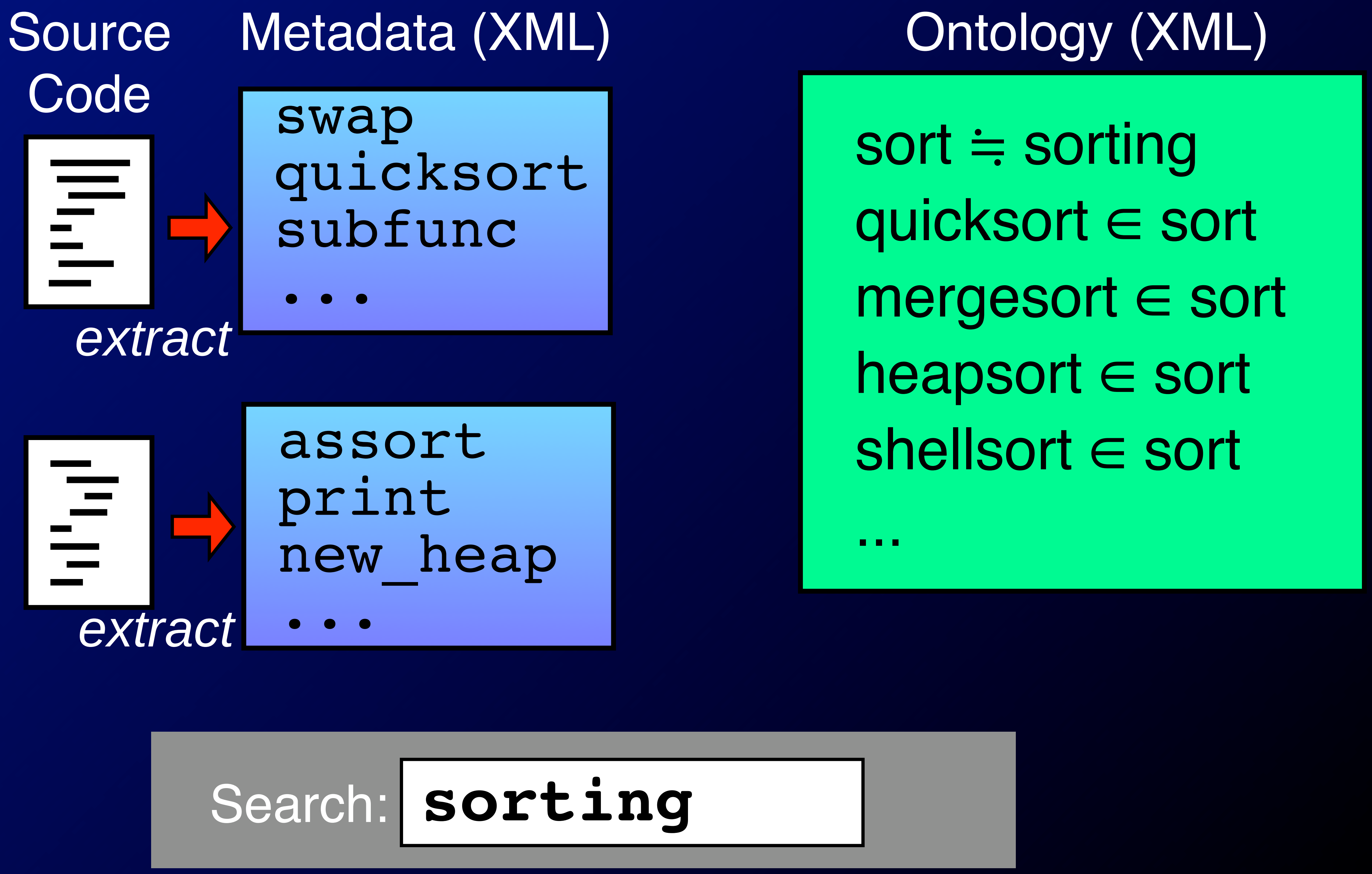
Metadata (XML)

犬

dog = 犬
cat = 猫
…

Ontology (XML)

Conception of S4 System



Source Code

Metadata (XML)

swap
quicksort
subfunc
...

extract

assort
print
new_heap
...

extract

Ontology (XML)

sort ≒ sorting
quicksort ∈ sort
mergesort ∈ sort
heapsort ∈ sort
shellsort ∈ sort
...

Search: **sorting**

Two Types of Metadata

- Source Code Metadata
  - has the filename of the source code,
  - has identifiers of functions, structures, variables, etc. used in the source code.
- Relational Metadata
  - shows the relation between a source code metadata and the ontologies.

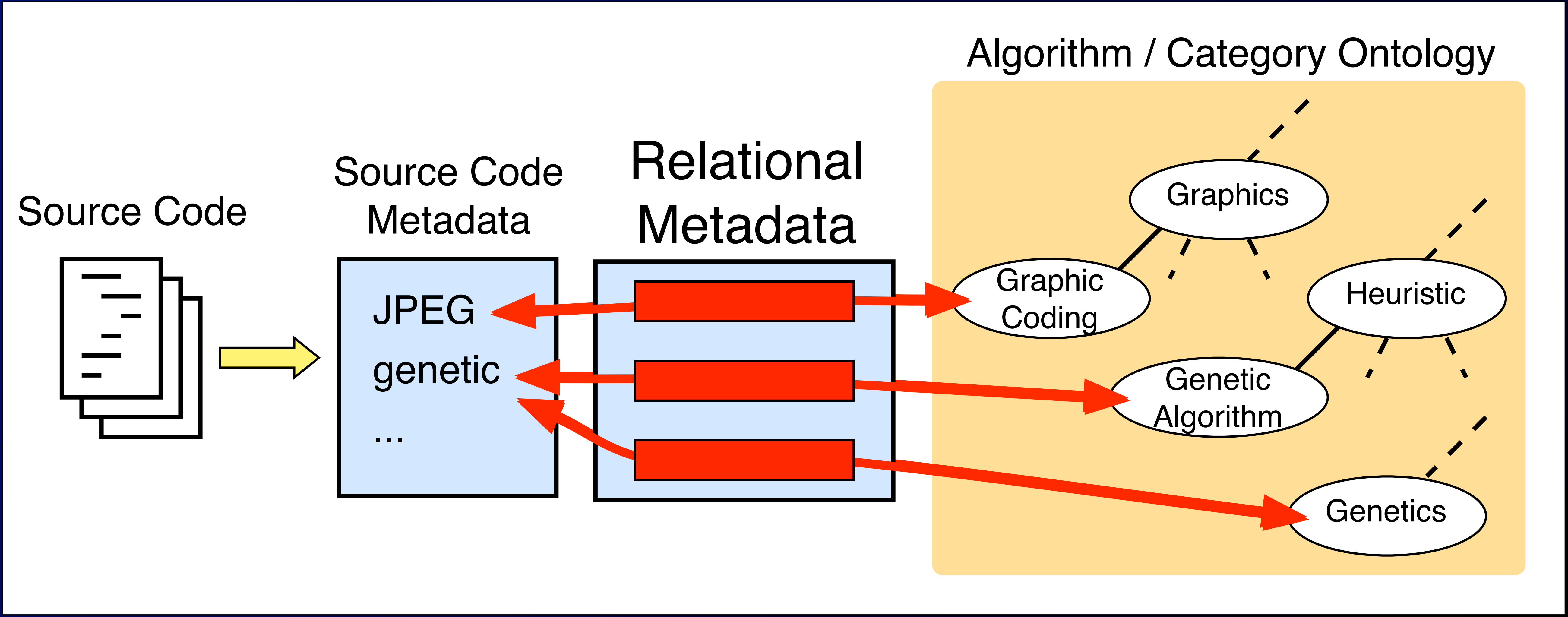
Source Code Metadata

- The information of the element names (filename and the identifiers in source code) is described.
- Element names would serve as a key showing what the source code is.

Software name:	The filename of the source code.
Function name:	The identifier of a function.
Structure name:	The identifier of a structure.
Member name:	The identifier of a member in a structure.
Variable name:	The identifier of a variable.
Array name:	The identifier of an array.

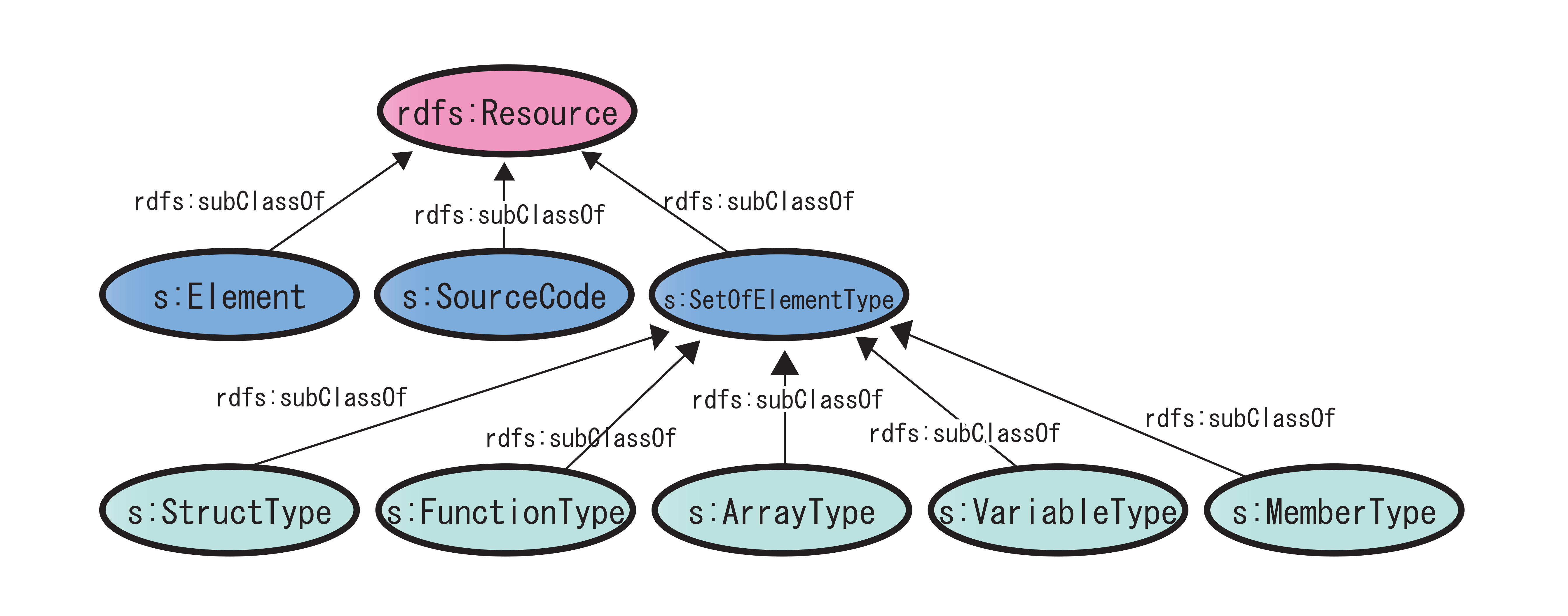
Relational Metadata

- The relation between the source code metadata and the ontologies is described.



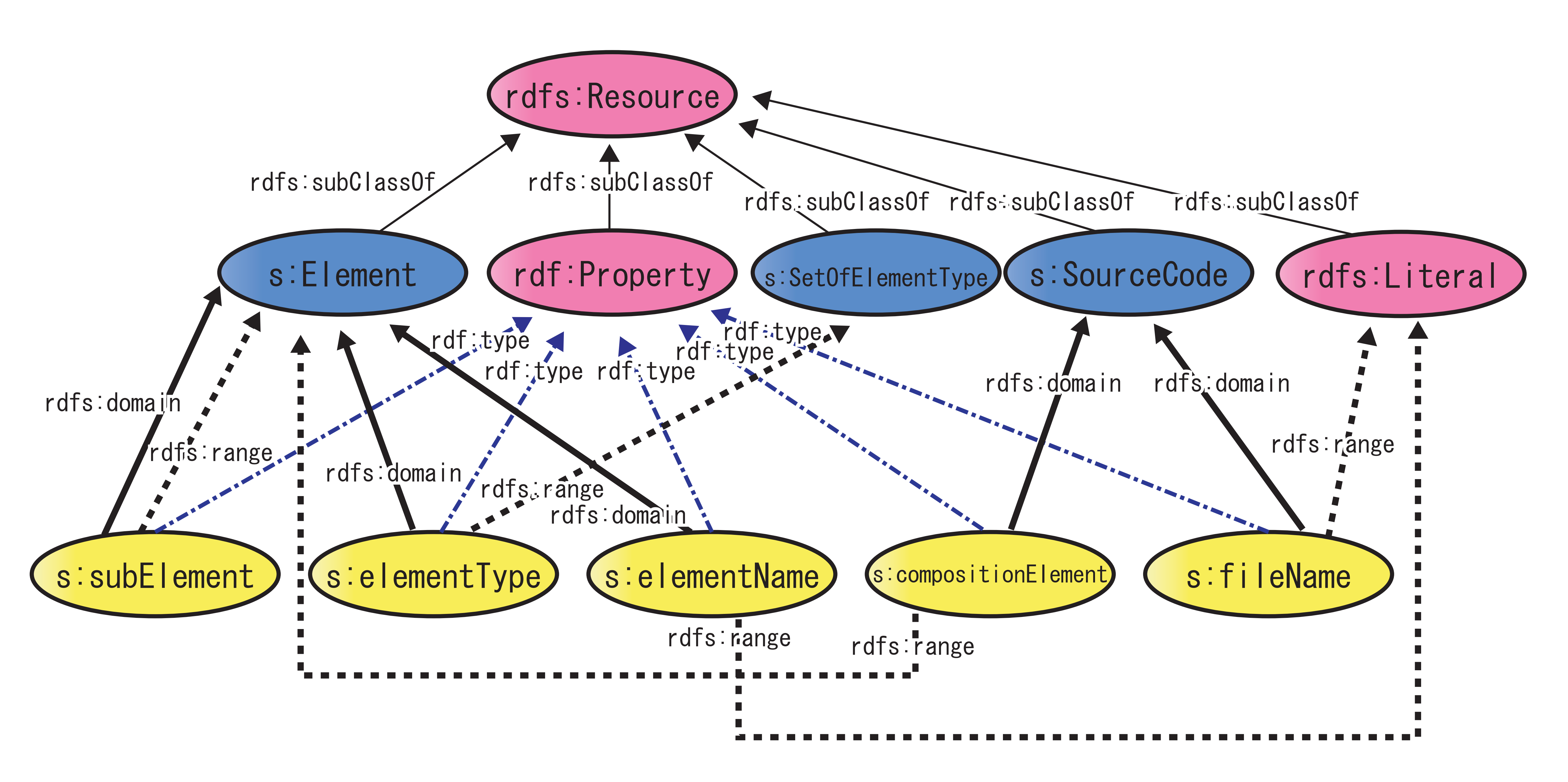
RDF Schema for Metadata (1)

- Class Definition



RDF Schema for Metadata (2)

- Properties and Classes



Ontology

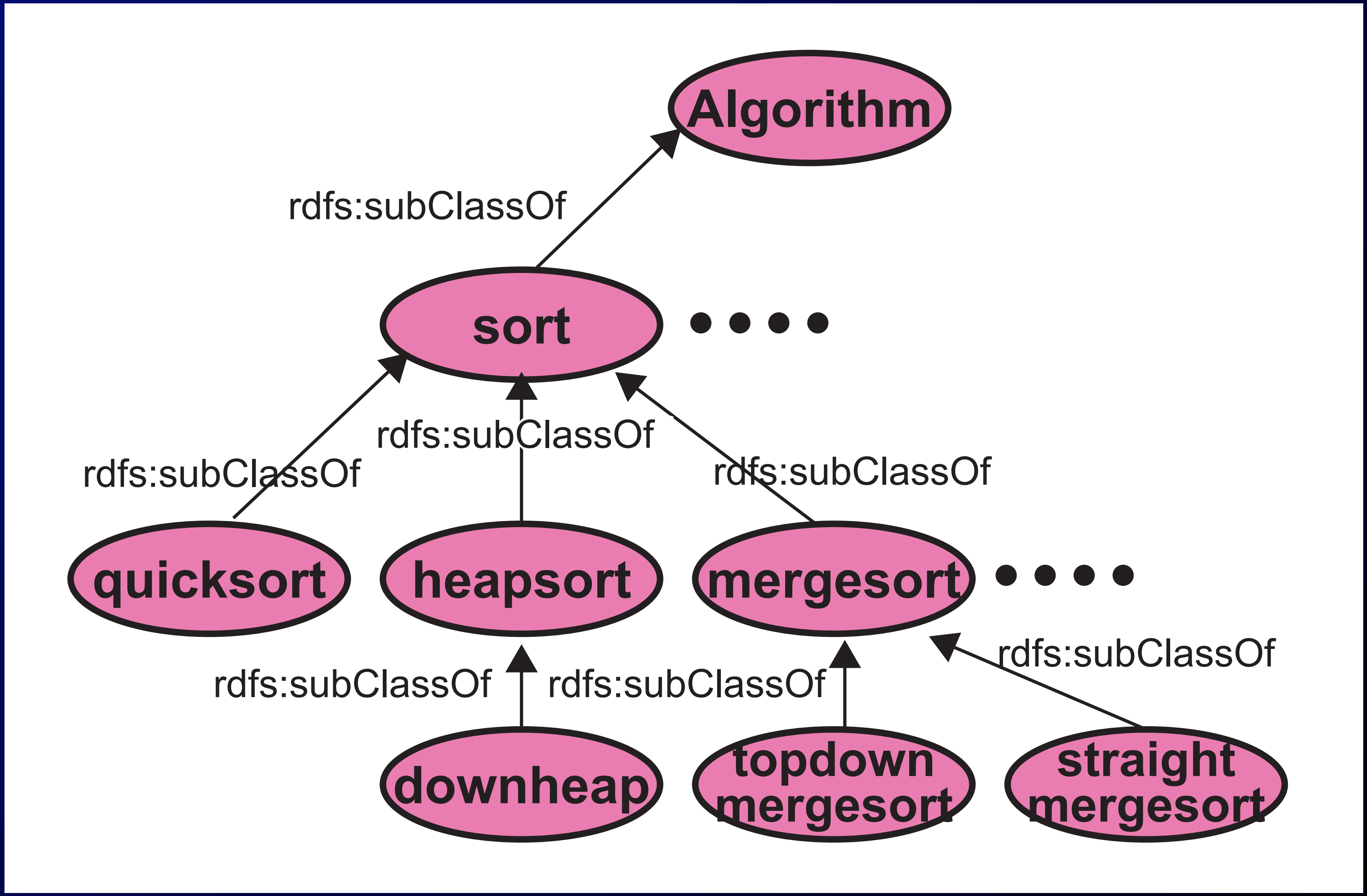
- Ontologies are described in OWL.
- S4 system uses 3 ontologies:
  - Synonym ontology
  - Algorithm ontology
  - Category ontology

Synonym Ontology

- Defines the synonyms of words:
  - used in the source codes,
  - supposed to be used as search words,
  - used in the algorithm / category ontology.
- With the synonym ontology, words which have the same meaning are treated equally.

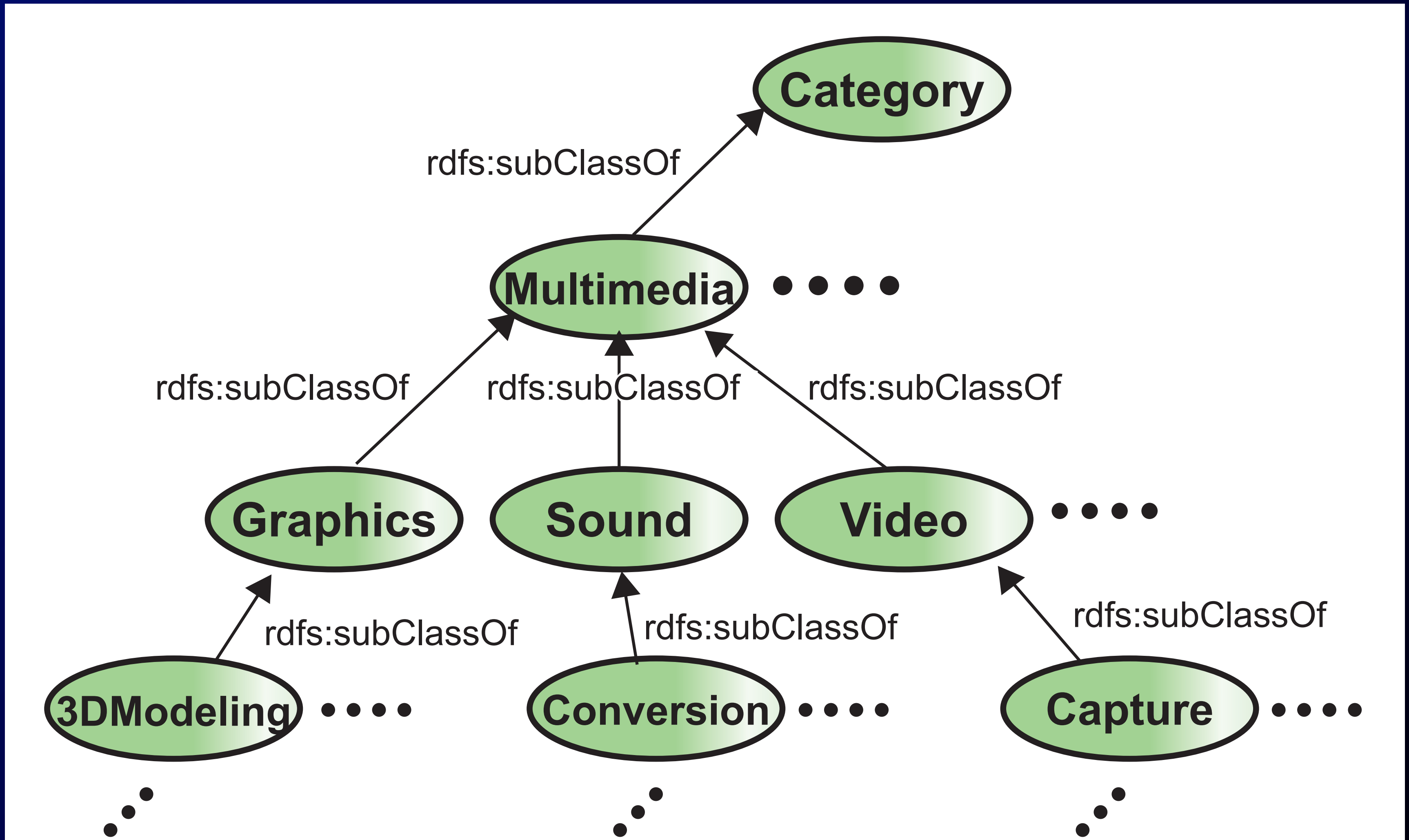
Algorithm Ontology

- Defines the kinds and the relation of the words which represent the algorithms.

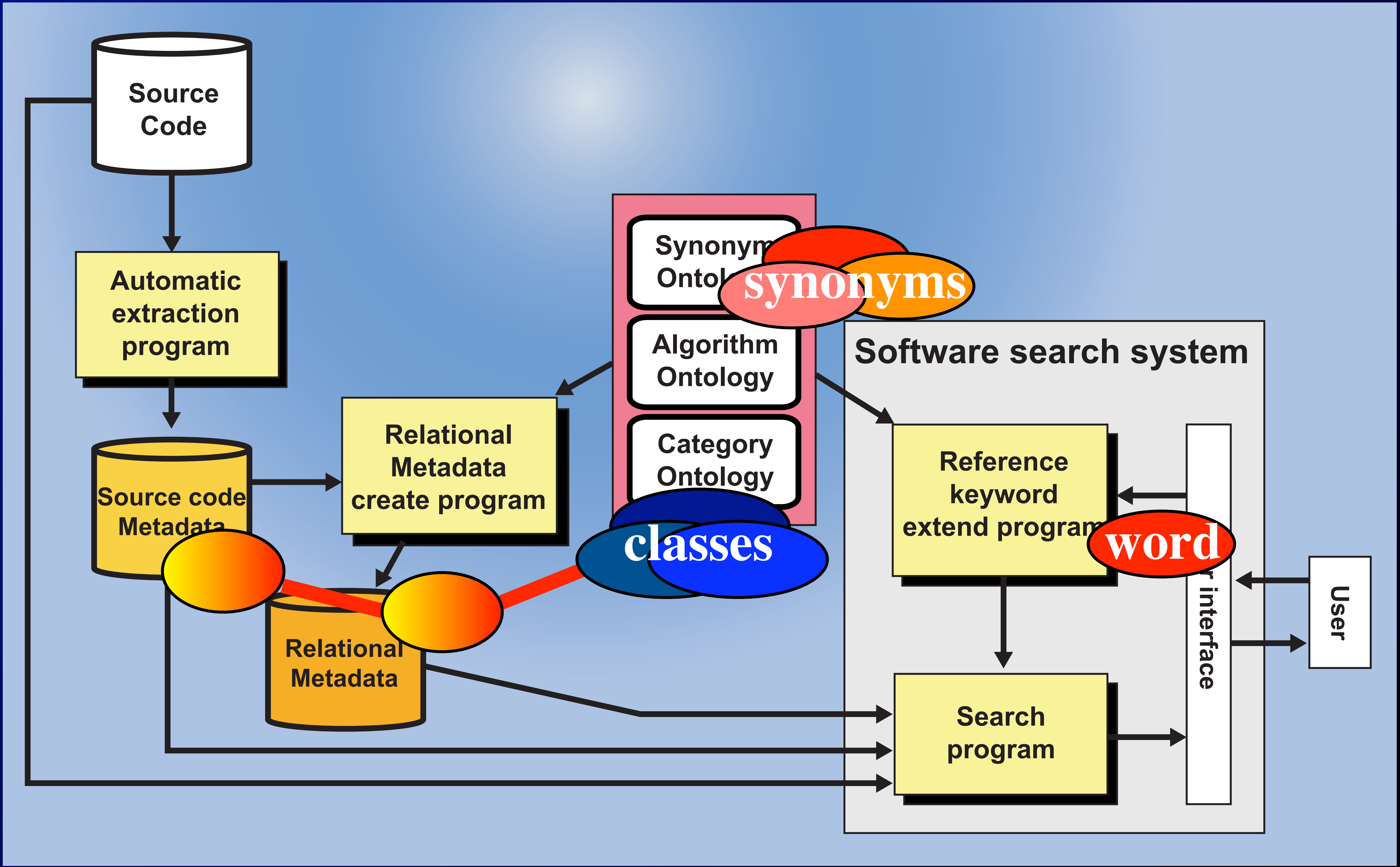


Category Ontology

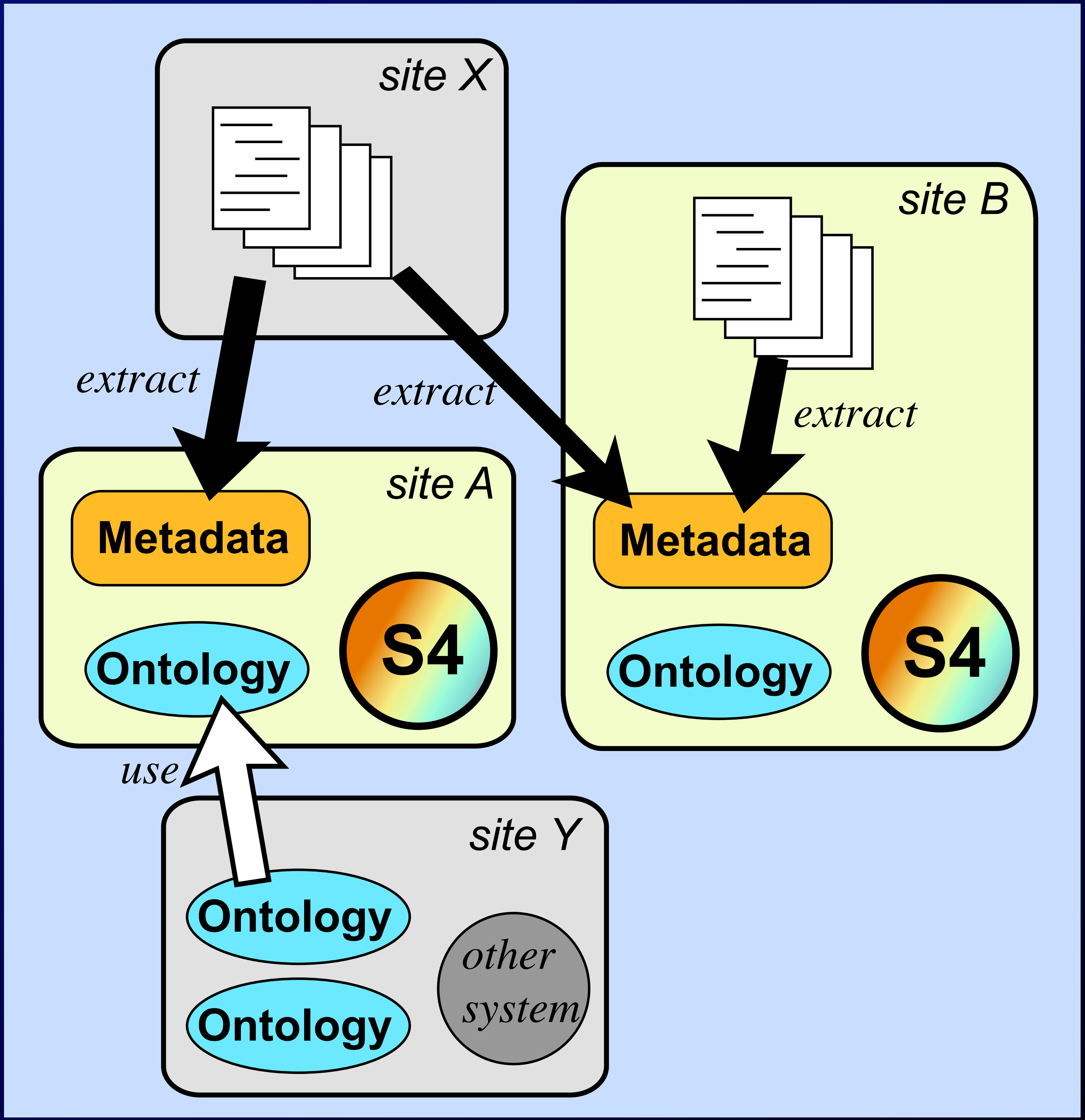
- Classifies the application domain of software.
- Referred to judge what kind of application the source code is.



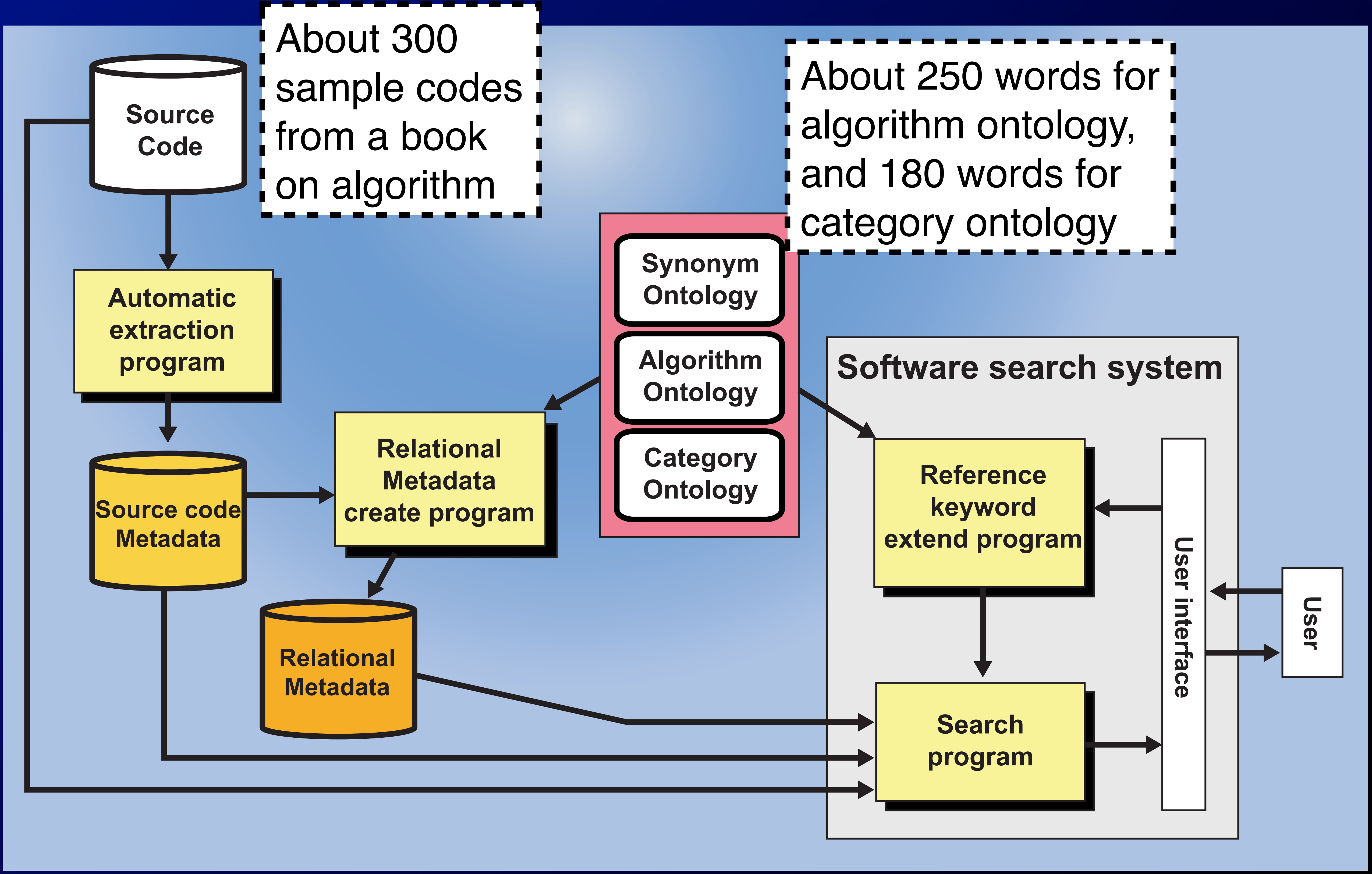
System Overview



Metadata and Ontologies on the Net



Experimental System



About 300 sample codes from a book on algorithm

About 250 words for algorithm ontology, and 180 words for category ontology

Source Code

Automatic extraction program

Source code Metadata

Relational Metadata create program

Relational Metadata

Synonym Ontology

Algorithm Ontology

Category Ontology

Software search system

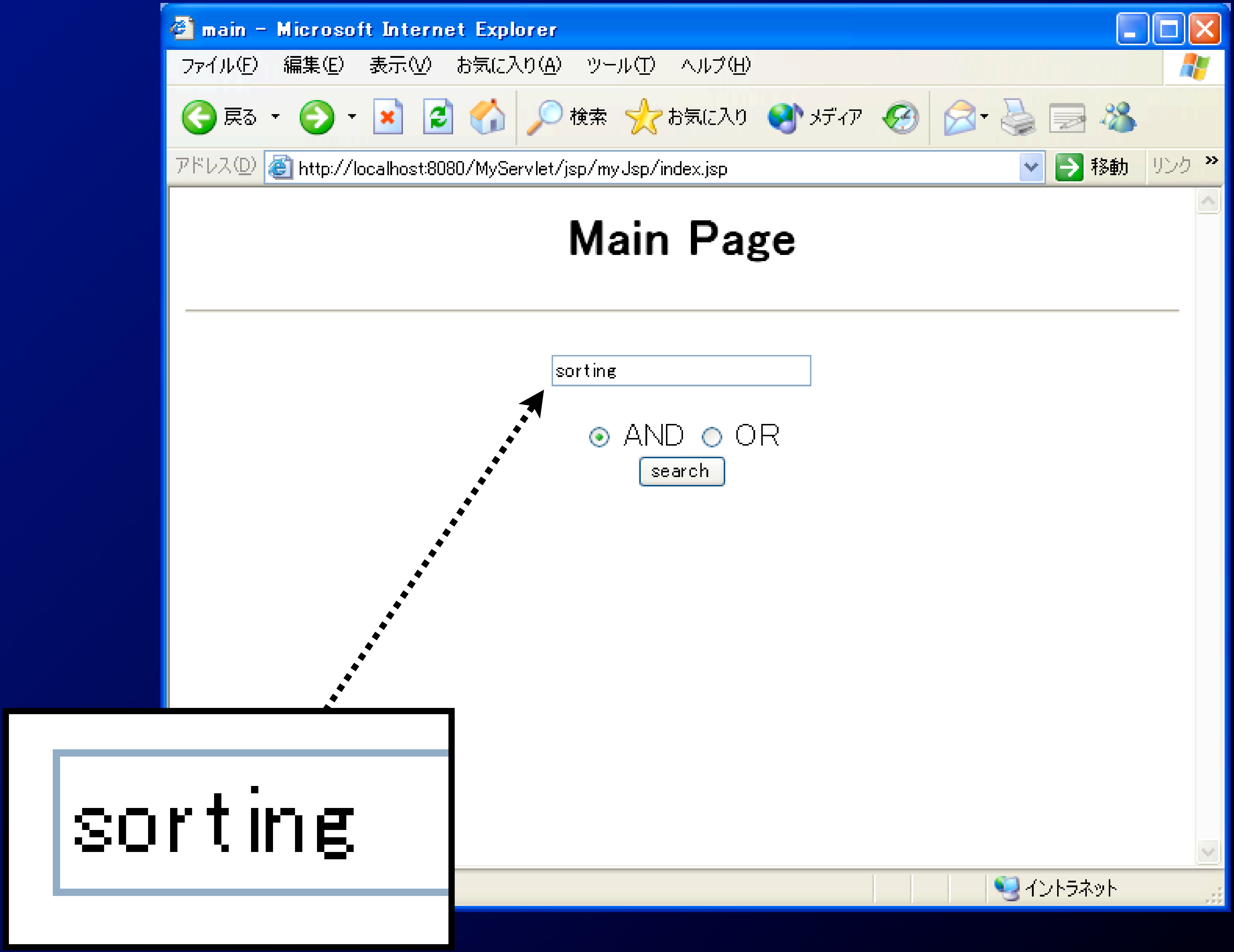
Reference keyword extend program

Search program

User interface

User

An Example (1)



main - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
メディア
アドレス(D)
移動
リンク
イントラネット

http://localhost:8080/MyServlet/jsp/myJsp/index.jsp

Main Page

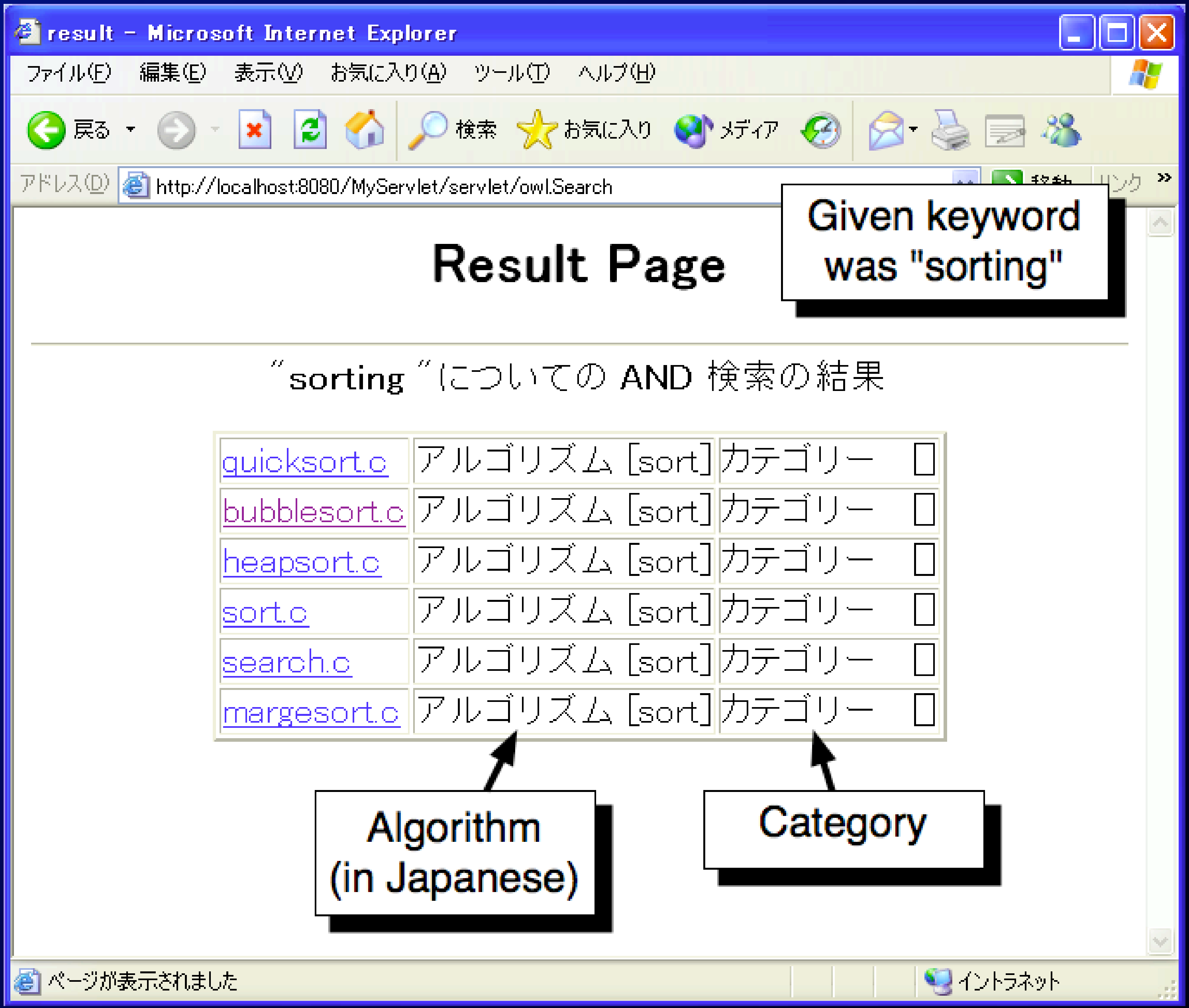
sorting

● AND ○ OR

search

sorting

An Example (2)


result - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
メディア
アドレス(D)
移動
リンク

http://localhost:8080/MyServlet/servlet/owl.Search

Result Page

“sorting “についての AND 検索の結果

quicksort.c	アルゴリズム [sort]	カテゴリー	□
bubblesort.c	アルゴリズム [sort]	カテゴリー	□
heapsort.c	アルゴリズム [sort]	カテゴリー	□
sort.c	アルゴリズム [sort]	カテゴリー	□
search.c	アルゴリズム [sort]	カテゴリー	□
margesort.c	アルゴリズム [sort]	カテゴリー	□

ページが表示されました

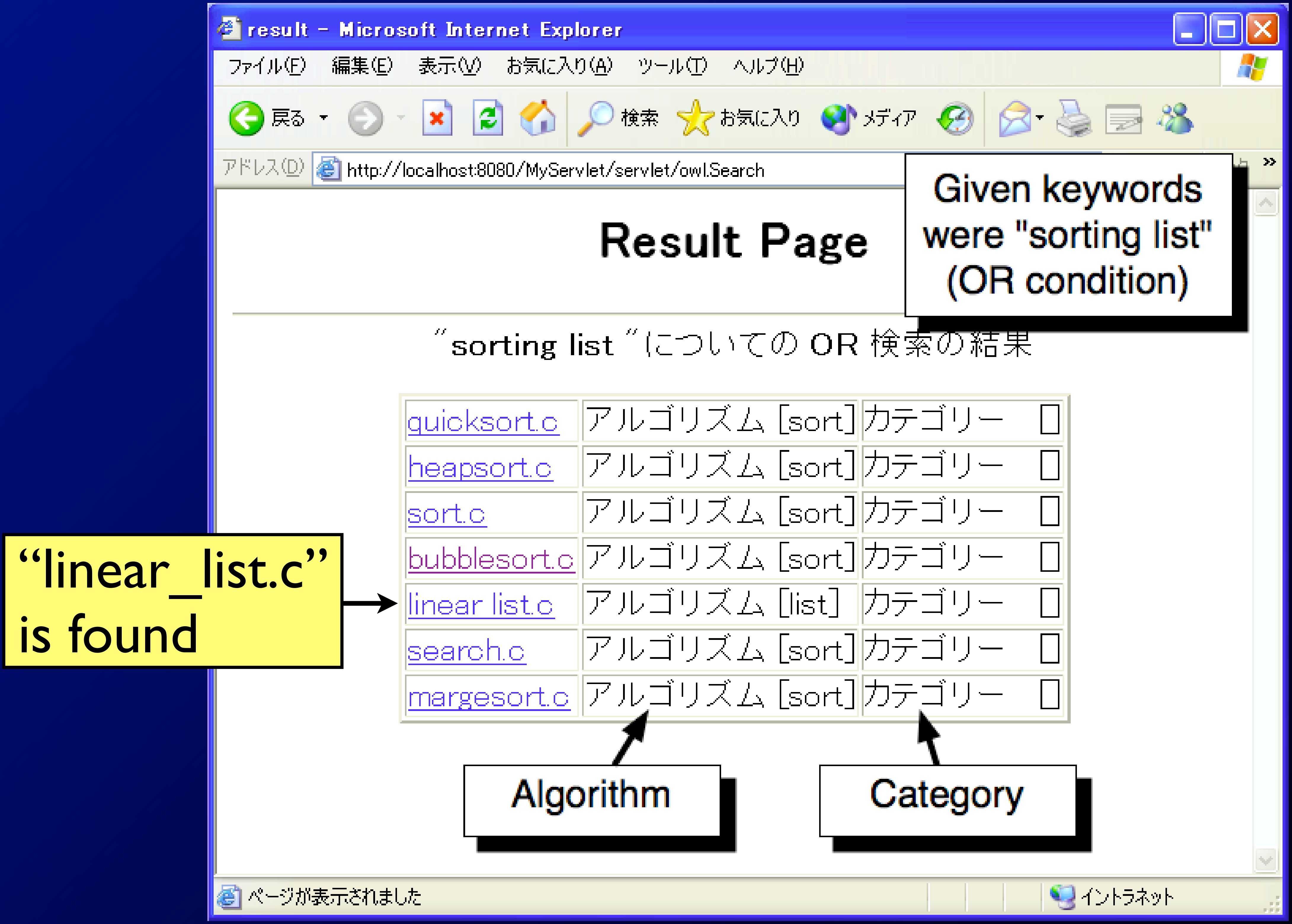
イントラネット

Given keyword was "sorting"

Algorithm (in Japanese)

Category

An Example (3)



http://localhost:8080/MyServlet/servlet/owl.Search

Result Page

“sorting list ”についての OR 検索の結果

quicksort.c	アルゴリズム [sort]	カテゴリー	□
heapsort.c	アルゴリズム [sort]	カテゴリー	□
sort.c	アルゴリズム [sort]	カテゴリー	□
bubblesort.c	アルゴリズム [sort]	カテゴリー	□
linear list.c	アルゴリズム [list]	カテゴリー	□
search.c	アルゴリズム [sort]	カテゴリー	□
margesort.c	アルゴリズム [sort]	カテゴリー	□

Given keywords were "sorting list" (OR condition)

“linear_list.c” is found

Algorithm

Category

Conclusion

- Source code search system S4 is proposed.
- S4 has the metadata and the ontologies, which are the data form proposed for the Semantic Web.
- S4 automatically makes the metadata of the source codes, and classifies them referring to the ontologies.
- S4 searches for the source code which includes the word with the similar meaning to the specified search word.

Currently Working on...

- Creation and expansion of the Ontologies.
- Development of the efficient search engine.
- Application of S4 system to the actual free / open sources.