

#1. Find concepts in scenario

The assistant began showing the scenario to Isaac. As performed with Dennis, It was suggested the use of **Digital Library** concept to begin searching.

#2. Disambiguate given a target (concept)

The assistant showed the following options according to wikipedia, and ask him to choose only one option.

```
Wikidata item search
Number of results:      10

Results:
1      digital library (Q212805) - library in which collections are stored in electronic media formats and accessible via computers
2      Digital Library for Dutch Literature (Q2451336) - website about Dutch language and Dutch literature
3      Digital Library Federation (Q5275906) - organization
4      Digital Library of Mathematical Functions (Q24534) - online project meant to be a collection of special functions
5      Digital Library of Slovenia (Q3435281) - digital library supported by National and University library of Slovenia
6      Digital Library Perspectives (Q53952043)
7      Digital library of the Lombardy (Q28531719)
8      Digital Library from the Meiji Era (Q11638378)
9      Digital Library of India (Q56480935)
10     Digital Library of the History of Friesland (Q2133797) - digital library
```

figure 2. different contexts where digital library is used.

Isaac chose option 1.

The assistant processed option1 text in wikifier.org

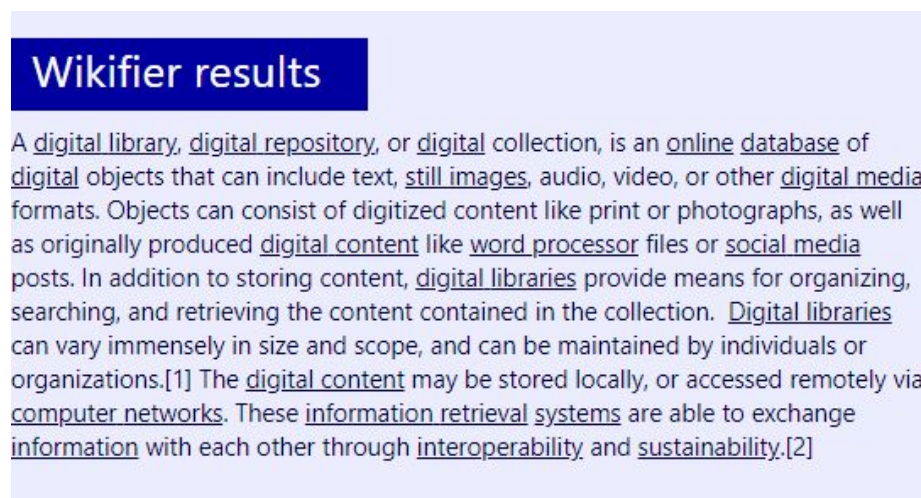


The image shows a screenshot of the Wikifier results page. At the top, there is a blue header with the text 'Wikifier results'. Below the header, the text 'library in which collections are stored in electronic media formats and accessible via computers' is displayed. The words 'electronic media' and 'computers' are underlined, indicating they are identified concepts.

figure 3. finding concepts for the option selected

#3. Process a paragraph-article from wikipedia

The assistant took the main article related to “digital library” in wikipedia and looked for its concepts in wikifier.org



The image shows a screenshot of the Wikifier results page for a Wikipedia article. At the top, there is a blue header with the text 'Wikifier results'. Below the header, a paragraph of text is displayed, with several words underlined to indicate identified concepts. The underlined words are: 'digital library', 'digital repository', 'digital collection', 'online database', 'digital objects', 'still images', 'audio', 'video', 'other digital media', 'formats', 'digitized content', 'print', 'photographs', 'originally produced digital content', 'word processor files', 'social media', 'posts', 'digital libraries', 'organizing', 'searching', 'retrieving', 'computer networks', 'information retrieval systems', 'interoperability', and 'sustainability'.

figure 4. finding concepts in a related wikipedia article

#4. Show concepts according elicitor selections

The assistant showed to Isaac the concepts collected (from fig.3 and fig4.)

library, electronic media, computers, digital library, digital repository, digital, online, database, still images, digital media, digital content, word processor, social media, digital libraries, digital

content, computer networks, information retrieval, systems, information, interoperability, sustainability

The assistant asked Isaac to select 5 keywords

Isaac selected: digital repository, systems, interoperability, social media, digital

#5. Create a corpus using keywords selected by elicitor

The assistant asked to create a query using "digital library" in the query:

Isaac chose initially: "digital library" + "digital repository"

For the query, the results in Github were only 1 projects. The assistant suggested to improve his query.

Isaac chose: "digital library" + systems

GitHub results were 56. The assistant asked to confirm this query

Isaac confirmed

The assistant created the corpus in <http://corpus-retrieval.herokuapp.com/> for the query

The tool returned 724 projects

#6. Finding keywords in Github

The assistant looked the context of 1st keyword: "digital library" in the corpus.

It was found 719 texts

the assistant looked the texts with the 2nd keyword: "systems"

it was found 48 texts / 719

after pruning duplicated texts, assistant got 35 unique texts with both keywords

#7. Processing texts to find GitHub keywords (show top30)

"name" "python" "access" "architecture" "viewer" "kitodo" "project" "tool"
"patron" "redshift" "science" "world" "number" "framework" "data" "wiley"
"zetoc" "team" "note" "time" "jcdl" "tpdl" "programming" "storage"
"core" "notification" "perseus" "collection" "management" "default"

The assistant showed the concepts saying that apparently these are words well related (in order of appearance) to the concept of Digital Library and Systems

The assistant asked Isaac to select 5 keywords:

Isaac selected: jcdl, zetoc, collection, project, kitodo.

note that Isaac took 3 minutes to perform to clarify by himself the terms showed

#8. Create a new query

The assistant showed the selections made by Isaac earlier and asked to create a final query

1g: digital repository, systems, interoperability, social media, digital"

2g: jcdl, zetoc, collection, project, kitodo

Isaac said that he will use only 2nd group as 1st group was already used.

Isaac created: "digital library" jcdl

#9. the ten first projects from the corpus created

1. <https://github.com/napsternxg/awesome-scholarly-data-analysis>
2. <https://github.com/PhilippMayr/NKOS-bibliography>
3. <https://github.com/fhamborg/Giveme5W1H>
4. <https://github.com/machawk1/warcreate>
5. <https://github.com/oduwsdl/MemGator>
6. <https://github.com/WING-NUS/scisumm-corpus>
7. <https://github.com/dbamman/jcdl2017>
8. <https://github.com/ag-gipp/citeplag>
9. <https://github.com/oduwsdl/Reconstructive>
10. <https://github.com/buaaliuming/Resources-for-Scholarly-Big-Data>