

# Automated subject indexing

Testing of Annif software for Italian language

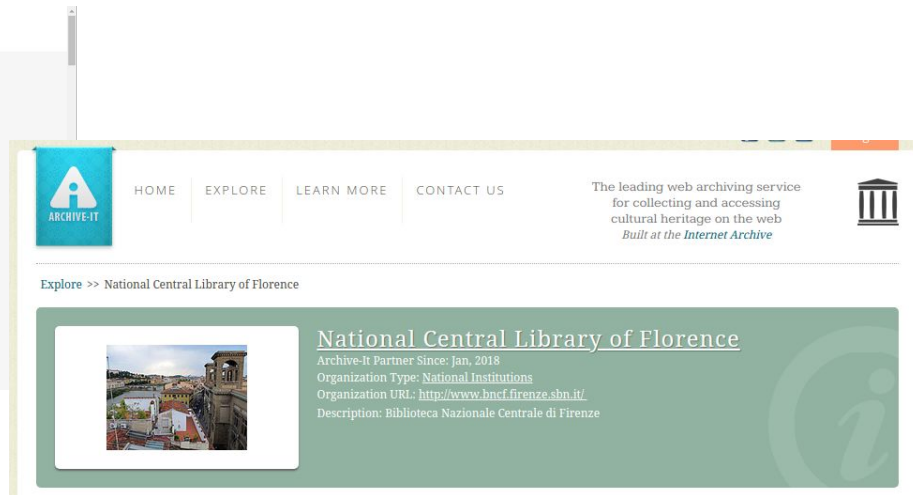
Lorenzo Gobbo



# Context overview



The screenshot shows the website of the Biblioteca Nazionale Centrale di Firenze (BNCFT). The header includes the logo and navigation links: BIBLIOTECA, RISORSE, ATTIVITÀ, and AMMINISTRAZIONE. The breadcrumb trail reads: home > Servizi > Magazzini Digitali. The main heading is 'Magazzini Digitali'. The main text states: 'Magazzini Digitali è il servizio nazionale di conservazione e accesso ai documenti digitali di interesse culturale, curato dalla Biblioteca nazionale centrale di Firenze (BNCFT), in collaborazione con la Biblioteca nazionale centrale di Roma (BNCR) e la Biblioteca nazionale Marciana di Venezia (BNM)'. A sidebar on the right contains the following text: 'Le origini del progetto e il contesto normativo', 'Dal progetto al servizio', 'Deposito, conservazione e accessibilità delle risorse', 'Dati e stato del servizio', 'NBN - National Bibliography Number', and 'Contatti'.



The screenshot shows the Archive-IT website. The header features the Archive-IT logo and navigation links: HOME, EXPLORE, LEARN MORE, and CONTACT US. A tagline reads: 'The leading web archiving service for collecting and accessing cultural heritage on the web Built at the Internet Archive'. The breadcrumb trail is: Explore >> National Central Library of Florence. The main content area has a green background and features a photograph of the National Central Library of Florence. The text on this page includes: 'National Central Library of Florence', 'Archive-It Partner Since: Jan. 2018', 'Organization Type: National Institutions', 'Organization URL: <http://www.bncf.firenze.sbn.it/>', and 'Description: Biblioteca Nazionale Centrale di Firenze'.

Magazzini Digitali: <https://www.bncf.firenze.sbn.it/biblioteca/magazzini-digitali/>

Web archiving service, National Central Library of Florence: <https://www.bncf.firenze.sbn.it/biblioteca/web-archiving/>

National Central Library of Florence web archiving collections, Archive-it: <https://archive-it.org/home/BNCF>

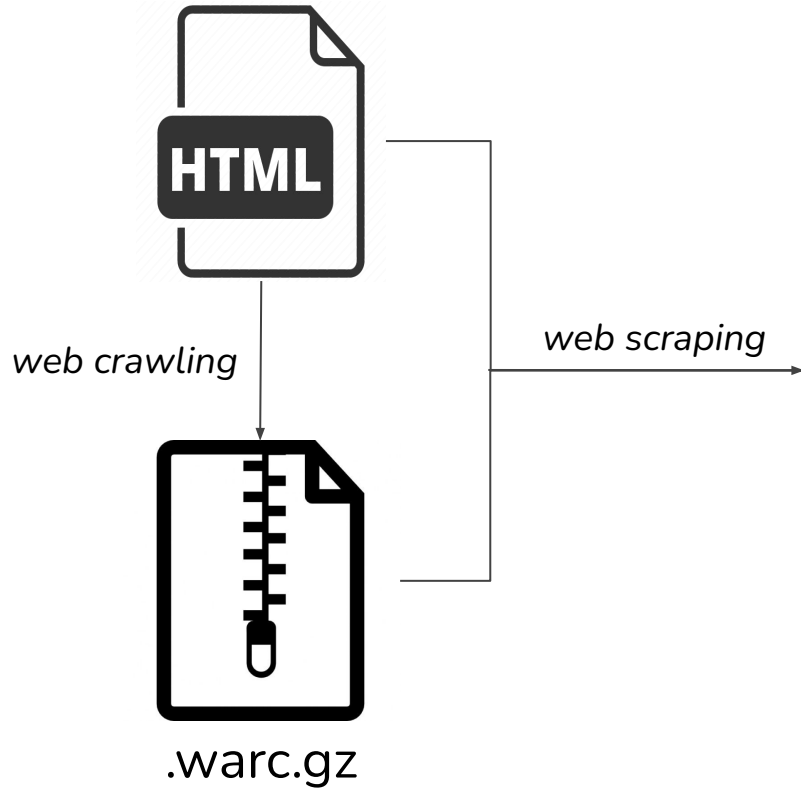
## Large amounts of resources to archive and manage:

since 2018, the National Central Library of Florence has collected 42,488,965 of documents

## HTML pages contain resources of several types and formats\*:

- text (HTML, PDF, JSON, ...);
- images (PNG, JPG, GIF, BITMAP, ...);
- audio (MP3, AAC, WMA, ...);
- video (MOV, MP4, AVI, ...);

\* To ensure [accessibility and archivability](#) each web resource should include an alternative textual description of the content



**Descriptive metadata (Dublin Core)\*:**

`\\dc:title`

`\\dc:publisher`

`\\dc:creator`

`\\dc:date`

...

...

`\\dc:description` → `\\dc:subject`

\* Dooley, Jackie, and Kate Bowers. 2018. Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Dublin, OH: OCLC Research. <https://doi.org/10.25333/C3005C>.



# Annif

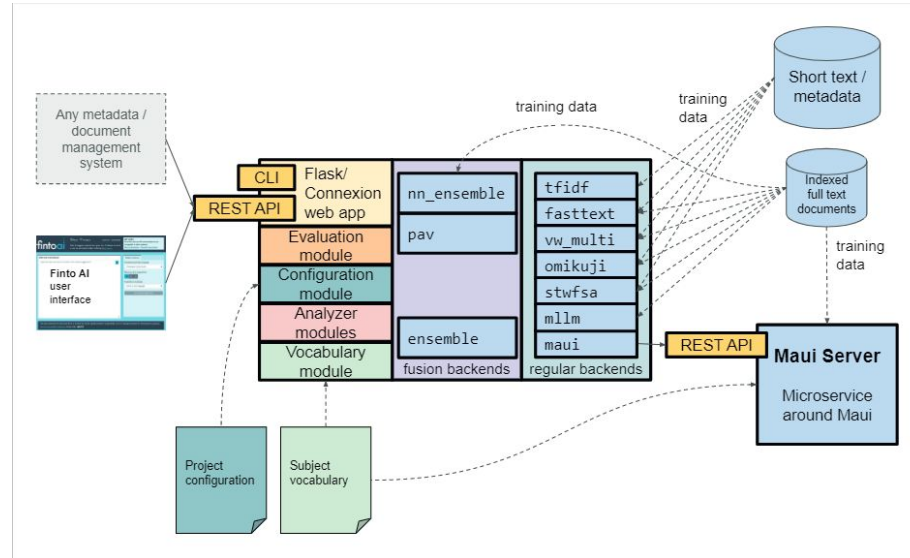
Developed by the National Library of Finland, Annif is a multi-algorithm automated subject indexing tool for libraries, archives and museums.\*

Based on AI and Machine Learning technology and natively developed for English and Finnish languages, Annif is independent of the indexing vocabulary.

**Source:** <https://github.com/NatLibFi/annif>

**Annif website:** <http://annif.org/>

# Annif architecture



**Source:** Suominen, O., Inkinen, J., & Lehtinen, M. (2022). *Annif and Finto AI: Developing and Implementing Automated Subject Indexing*. *JLIS.It*, 13(1), 265–282. <https://doi.org/10.4403/jlis.it-12740>



# Prerequisites

1. Training data set:

Bibliografia Nazionale Italiana (2018-2022)

2. Tokenizer with italian language support:

Natural Language Toolkit (NLTK)

3. Vocabulary:

Thesaurus of the Nuovo Soggettario di Firenze



# Task schedule

Division of the trial into **4 phases**:

**Phase 1.** creation of 3 different training data set [**100, 1000, 10000 titles**], algorithms training by keywords only (entity names excluded) and use of Thesaurus of the Nuovo Soggettario as vocabulary;

**Phase 2.** Increase of the training data set [**30000 titles**], inclusion of external authority files as *VIAF*, *GeoNames* and *Wikidata* in the *Thesaurus*, algorithms training by keywords (entity names included);

**Phase 3.** Inclusion of abstracts in the training data set [**30000 titles + abstracts (25559)**], parameters configuration setup for each algorithm;

**Phase 4.** Graphic User Interface setup for web browser use, definition of the presentation style of algorithms and results.





# Algorithm approaches

## 1. Associative approaches:

TF-IDF, fastText\*, Omikuji

## 2. Lexical approaches\*:

MLLM, STWFSA, YAKE, SVC

## 3. Fusion approaches\*:

Ensemble, PAV, Neural Network

\* Work with a SKOS/RDF file as vocabulary



# Vocabulary

2 types of vocabulary supported:

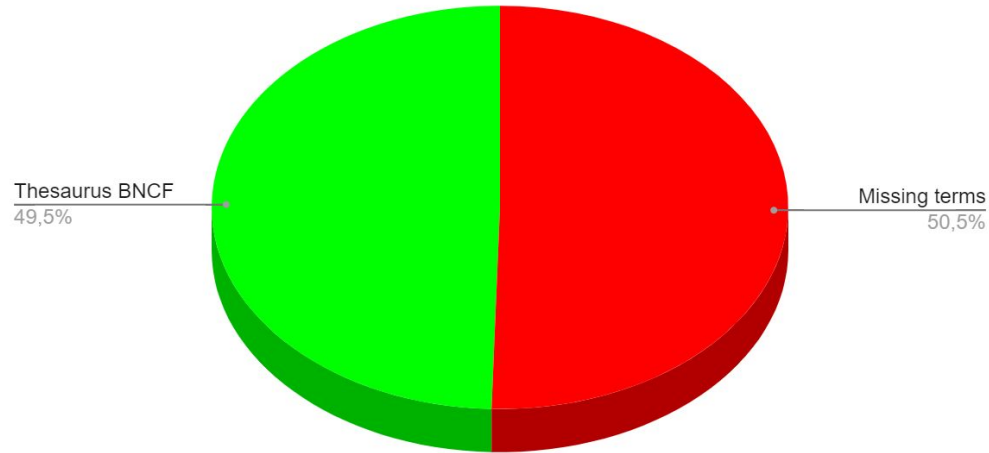
```
<http://purl.org/bncf/tid/1>      Lingua slovacca
<http://purl.org/bncf/tid/10>   Lingue slave meridionali
<http://purl.org/bncf/tid/100>  Letterature semitiche orientali
<http://purl.org/bncf/tid/10000> Cataloghi per materia
<http://purl.org/bncf/tid/10000> Cataloghi per materie
<http://purl.org/bncf/tid/10001> Grafica e politica
<http://purl.org/bncf/tid/10002> Trombossani
<http://purl.org/bncf/tid/10002> Tromboxani
<http://purl.org/bncf/tid/10004> Cataloghi editoriali
<http://purl.org/bncf/tid/10007> Cataloghi nominali
<http://purl.org/bncf/tid/10007> Cataloghi per autore
<http://purl.org/bncf/tid/10007> Cataloghi per autori
<http://purl.org/bncf/tid/10007> Cataloghi per autori e titoli
```

TSV: pairs of URIs / labels

```
<http://purl.org/bncf/tid/10> dc:date "2004-11-09" ;
a skos:Concept ;
skos:broader <http://purl.org/bncf/tid/3> ;
skos:closeMatch <http://d-nb.info/gnd/4120345-8> , <http://data.bnf.fr/ark:/12148/cb119818499> , <http://id.loc.gov/authorities/names/n79-043111> ;
skos:editorialNote "FONTE: Treccani.it (voce: Sloveno); PT; LdL (voce: Sloveno); WebDewey(IT); Wikipedia(IT)" ;
skos:inScheme <http://purl.org/bncf/tid/Thes> , <http://purl.org/bncf/tid/ThesCF15> ;
skos:narrower <http://purl.org/bncf/tid/11484> , <http://purl.org/bncf/tid/163> , <http://purl.org/bncf/tid/265> ;
skos:notation "491.81"^^<http://dewey.info> ;
skos:prefLabel "Lingue slave meridionali"@it ;
skos:related <http://purl.org/bncf/tid/18> .

<http://purl.org/bncf/tid/100> dc:date "2004-11-09" ;
a skos:Concept ;
skos:broader <http://purl.org/bncf/tid/95> ;
skos:editorialNote "FONTE: WebDewey(IT)" ;
skos:inScheme <http://purl.org/bncf/tid/Thes> , <http://purl.org/bncf/tid/ThesCF1> ;
skos:narrower <http://purl.org/bncf/tid/99> ;
skos:notation "892.1"^^<http://dewey.info> ;
skos:prefLabel "Letterature semitiche orientali"@it ;
skos:related <http://purl.org/bncf/tid/92> .
```

SKOS: structured list of concepts and associated labels  
(RDF/XML, Turtle or N-Triples format)

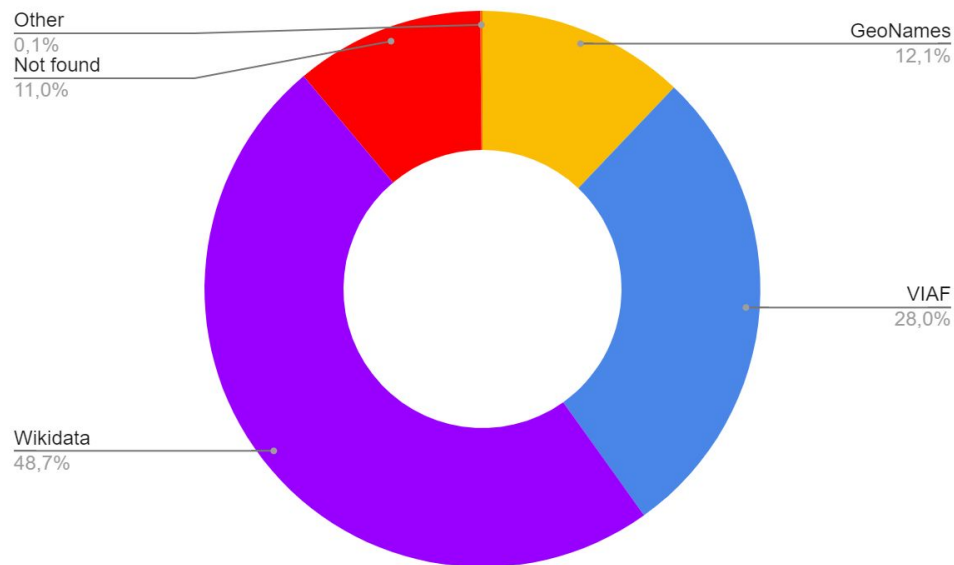


**7018** terms out of **13906** used in Subject heading strings are not included in the Thesaurus of the Nuovo Soggettario

## Vocabulary integration with external authority files (TSV format)\*:

**VIAF:** 1968 terms;  
**GeoNames:** 849 terms;  
**Wikidata\*:** 3420 terms;  
**Altro:** 9 terms;  
**Mancanti:** 772 terms (TM-n)

**BNCF-it-v1:** 235.398 terms



\* Wikidata URIs scraper: <https://github.com/logo94/wikidata-URIs-scraper>



# Training data set

Reduction of Subject heading strings to keywords/single terms:

a [connettore] b  $\rightarrow$  a - b

x [di] a  $\rightarrow$  a

x [:] a  $\rightarrow$  a

Conversion of the training data set\*:

Phase 1: kw100, kw1000, kw10000

Phase 2: kwn30000

Phase 3: abs30000



\*Annif-corpus-toolkit: <https://github.com/logo94/Annif-corpus-toolkit>



# Evaluating indexing quality

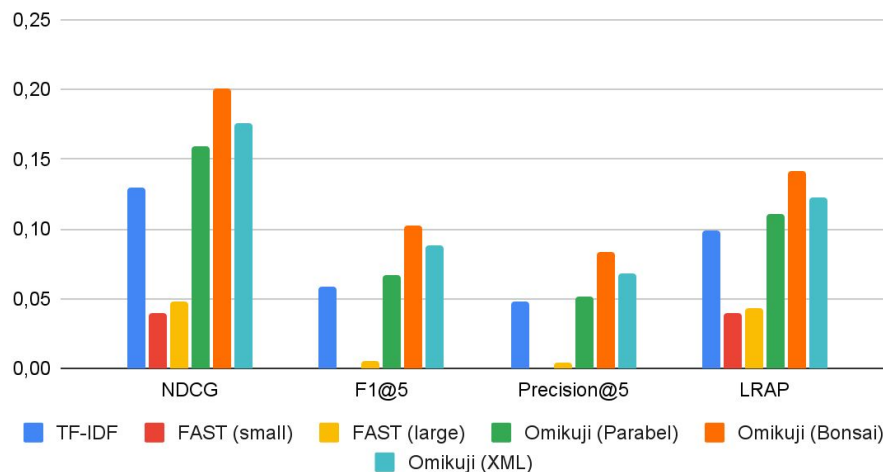
Results evaluation for a single document (precision and recall):

```
<http://purl.org/bncf/tid/9088> Insegnamento 371.102 0.4342578649  
<http://purl.org/bncf/tid/5713> Stranieri 305.90691 0.3481715023  
<http://purl.org/bncf/tid/5074> Lingua italiana 450 0.3481715023  
<http://purl.org/bncf/tid/6845> Acqua 553.7 0.2913397252  
<http://purl.org/bncf/tid/4935> Teorie 0.2802742421  
<http://purl.org/bncf/tid/1576> Informatica 004 0.2726948261  
<http://purl.org/bncf/tid/13140> Scuole secondarie 373 0.2726948261  
<http://purl.org/bncf/tid/4968> Algebra 512 0.2726948261  
<http://purl.org/bncf/tid/5750> Diritto 340 0.2548523843  
<http://purl.org/bncf/tid/15070> Atti di congressi 0.2505477964
```

## Results evaluation for a collection of documents:

Precision (doc avg):	0.07200000000000001
Recall (doc avg):	0.23799999999999996
F1 score (doc avg):	0.10696503496503496
Precision (subj avg):	0.00042512288267305064
Recall (subj avg):	0.000529677048336814
F1 score (subj avg):	0.00044849469087908297
Precision (weighted subj avg):	0.23845303148581837
Recall (weighted subj avg):	0.29508196721311475
F1 score (weighted subj avg):	0.251111739212361
Precision (microavg):	0.072
Recall (microavg):	0.29508196721311475
F1 score (microavg):	0.11575562700964628
F1@5:	0.14725396825396828
NDCG:	0.23088839053196938
NDCG@5:	0.21536902785405243
NDCG@10:	0.23088839053196938
Precision@1:	0.12
Precision@3:	0.14666666666666667
Precision@5:	0.124
LRAP:	0.17333026264775106
True positives:	36
False positives:	464
False negatives:	86
Documents evaluated:	50

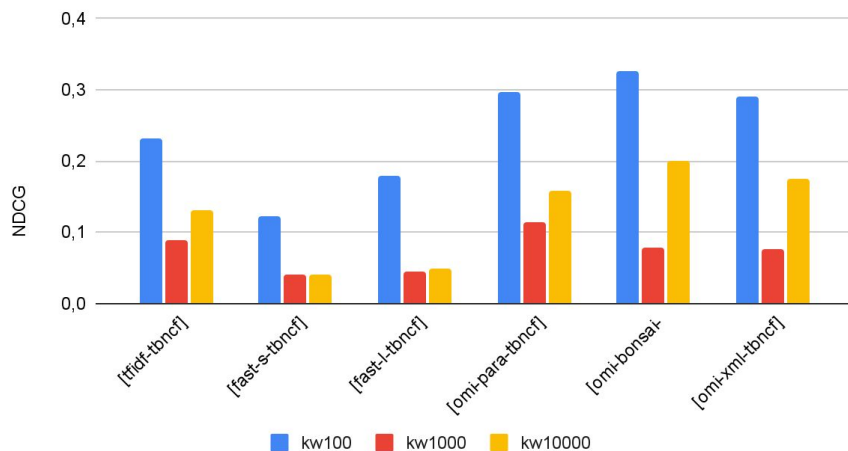
## Metrics comparison



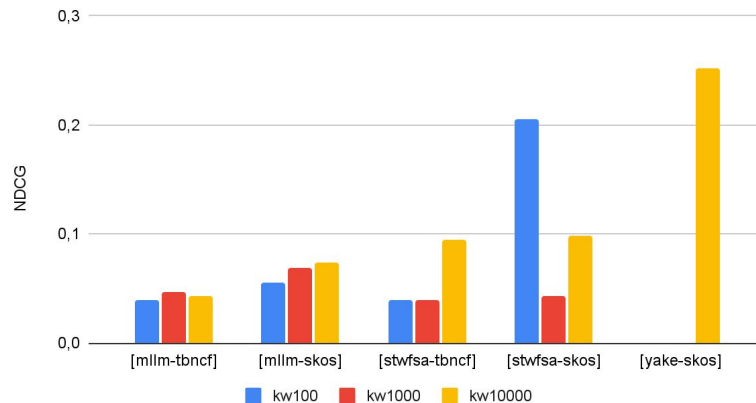


# Results: Phase 1

## Associative algorithms comparison



## Lexical algorithms comparison

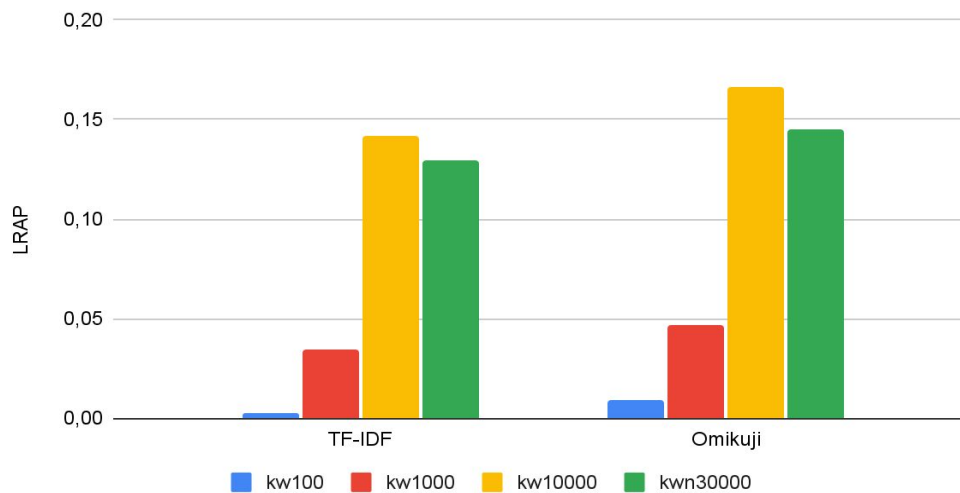






## Results: Phase 2

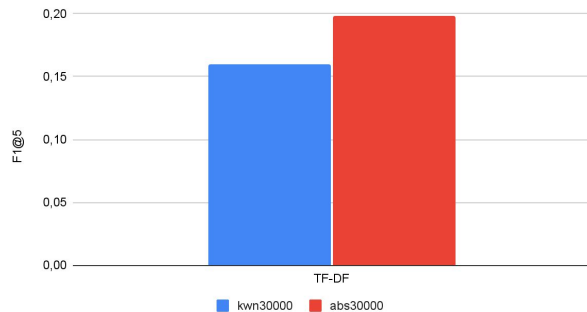
Associative algorithms (TSV vocabulary)



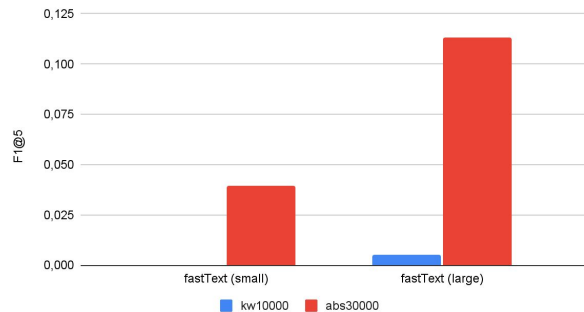


# Results: Phase 3

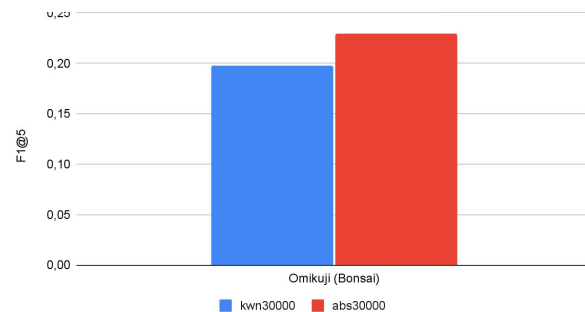
TF-IDF



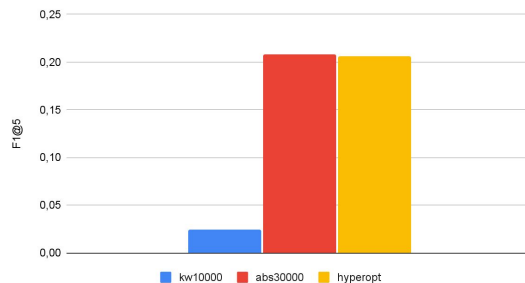
fastText



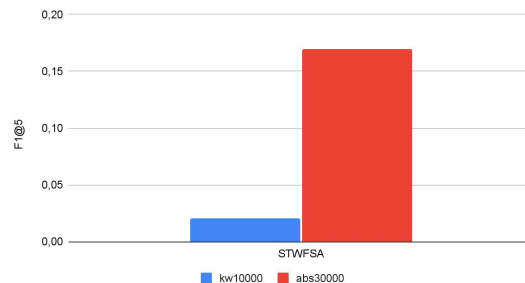
Omikuji



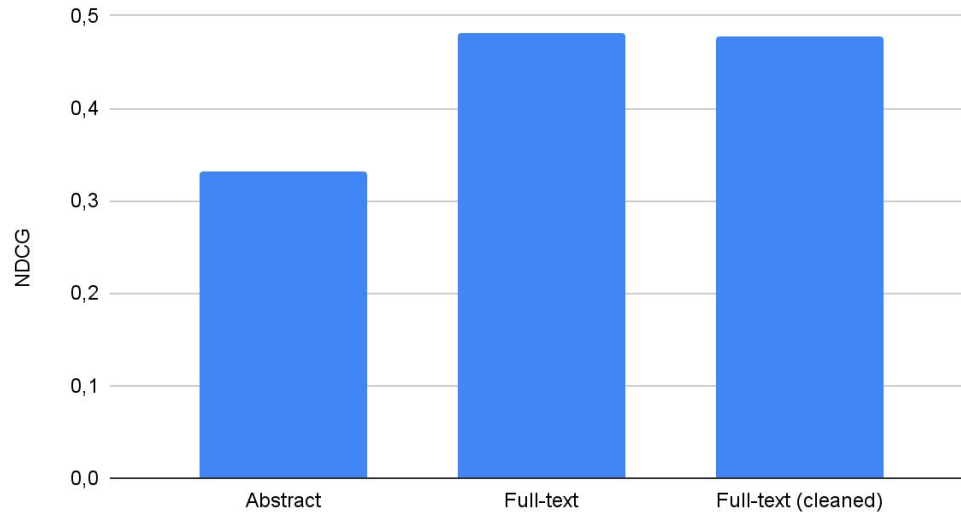
MLLM



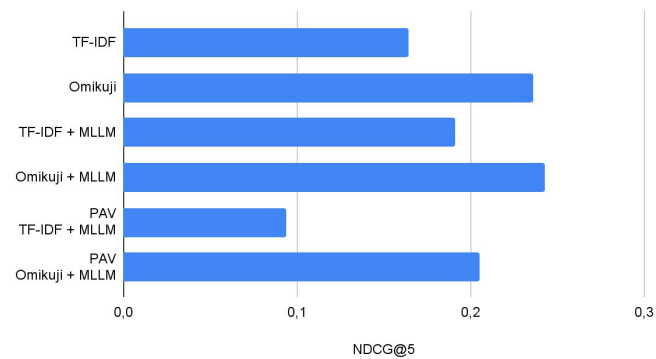
STWFSA



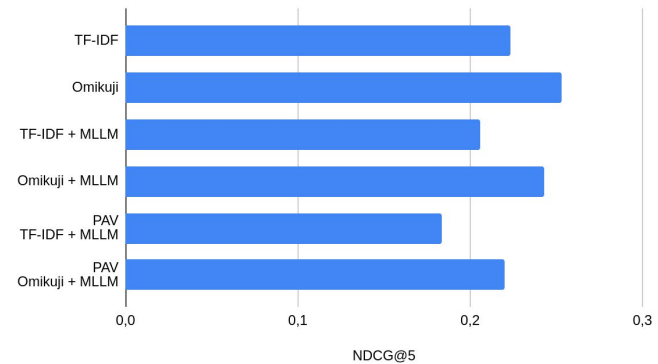
## Quality based on training text length



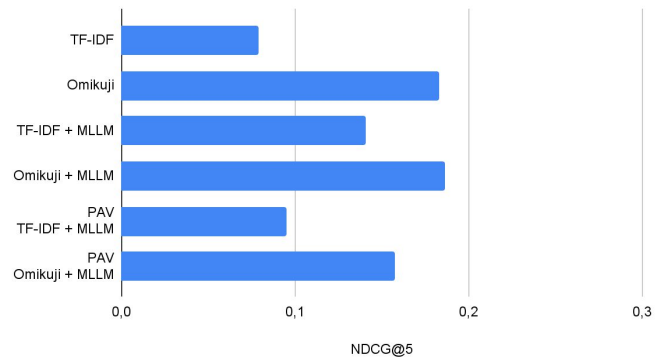
## Titles



## Abstracts



## Full-text



# Results: Phase 4

## Indicizzazione automatica per soggetto

Copiare qui il testo da elaborare...



SELEZIONE PROGETTO

Verifica copertura

NUMERO SOGGETTI

1 3 5 10

Otteni soggetti →

SELEZIONE PROGETTO

Verifica copertura ▼

Verifica copertura

Titolo / Abstract / Articolo

Full-text

CDD III Livello

Ottieni soggetti →

**Reliability check** → suggests a DDC main class, the confidence score indicate the reliability of the suggestion for that specific discipline [SVC]

**Title / Abstract / Article** → for short/medium length text inputs [Ensemble: Omikuji + TF-IDF (tsv)]

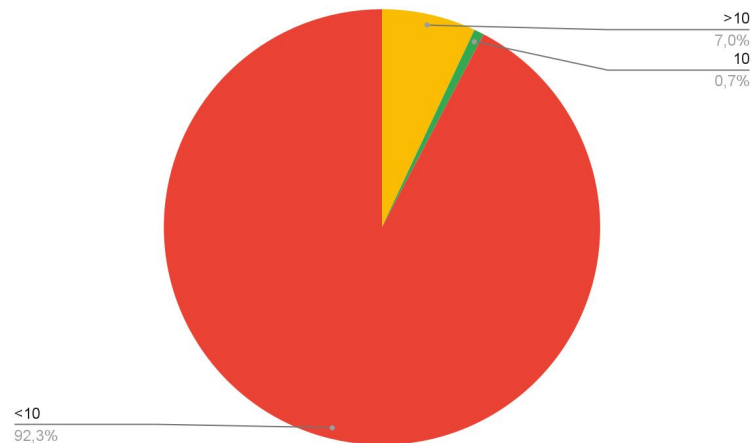
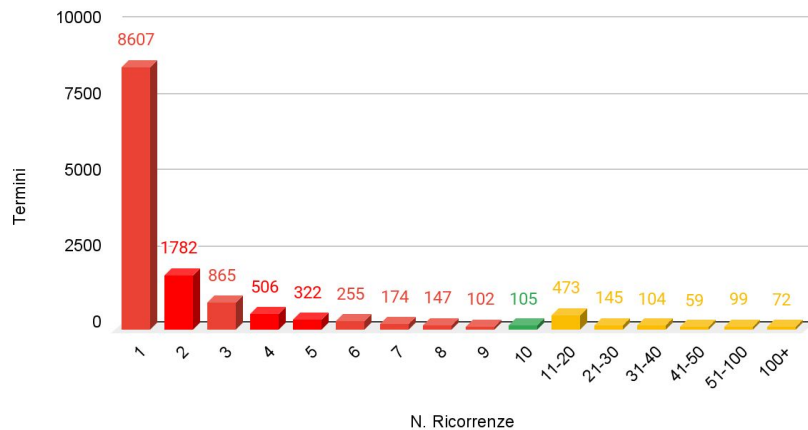
**Full-text** → for long text inputs [Ensemble: Omikuji + MLLM (tsv)]

**DDC** → suggests a list of subjects and related DDC notation [Ensemble: Omikuji + MLLM (skos)]

# Conclusions

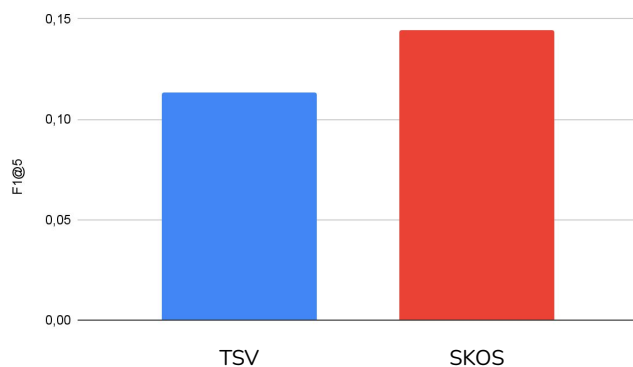
- The quality of the training data set is more relevant than its quantity, to achieve high confidence scores every term of the vocabulary SHOULD appear at least 10 times in the training data set. The current state is:

Numero ricorrenze per termine

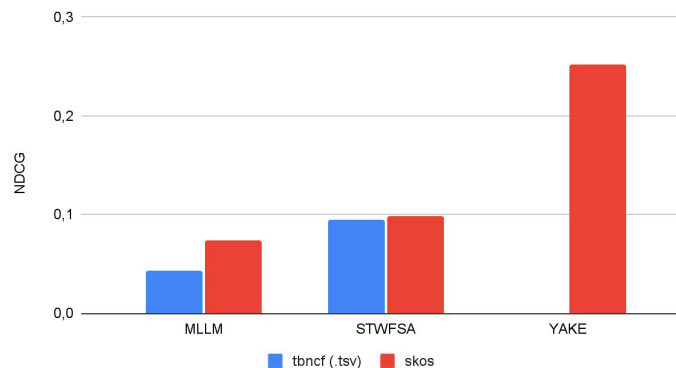


- Lexical and fusion algorithms are projected to work with a SKOS vocabulary

Vocabularies comparison: fastText



Vocabularies comparison: lexical algorithms



Without a complete and structured authority file, it is possible to run only associative algorithms such as TF-IDF, Omikujii, or a fusion between the two;



- An improvement in results can be achieved through a twofold intervention: on one hand by adding subjects to existing titles, on the other by adding other titles in a reasoned way;
- The normalization of the training data set as well as improving automated indexing results allows to fix and optimize the existing metadata;
- The development of an automatic indexing system presents itself as a library and organizational challenge, not a technological one;
- Evolution of cataloging work: from execution to supervision



# Roadmap

## Training data set:

- Adding subjects and verifying the correctness of abstracts
- Reasoned addition of titles
- Clustering of subjects
- Periodic publication of updated versions of the training data set

## Vocabulary:

- Inclusion of SBN authority file for entity names
- Translation of GeoNames english labels to italian language
- Stable URIs association to terms non included in the vocabulary or in external authority files
- Vocabulary conversion from TSV to SKOS (Turtle, XML/RDF, N-Triples)

**Thank you!**

The background is a solid teal color. It features several faint, semi-transparent graphics: a large donut chart in the upper right, a smaller pie chart to its right, a bar chart in the bottom right corner with four bars of increasing height, and several other small pie charts scattered throughout the right side of the image.