

## Introduction to the online hands-on tutorial

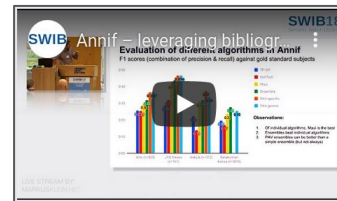
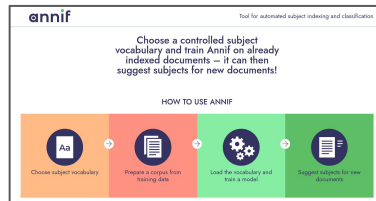


Leibniz-Informationszentrum  
Wirtschaft  
Leibniz Information Centre  
for Economics

1

## Understand what Annif is

Study the website [annif.org](https://annif.org),  
watch a presentation about it,  
or read the LIBER Quarterly [paper](#).



Annif: DIY automated subject indexing using multiple algorithms

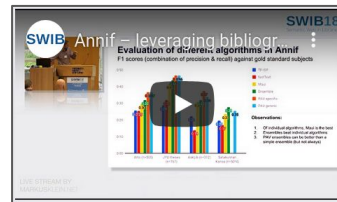
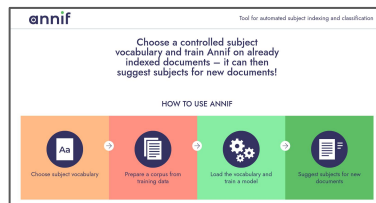
**Abstract**

Manually indexing documents for subject-based access is a labor-intensive process. We propose using metadata gathered from bibliographic databases to train algorithms that assist librarians in their work. We have developed *annif*, an open-source tool and environment for automated subject indexing. After training it with a subject vocabulary and existing metadata, *annif* can be used to suggest subject headings for new documents. We have tested *annif* with different document collections including scientific papers, old scanned books and contemporary e-books. Old papers from an "old-fashioned" service, Project Gutenberg, and the archives of a local newspaper. The results of analyzing scientific papers and current books have been promising, while other types of documents have proved to be more challenging. The current version is based on a combination of existing natural language processing and machine learning tools. By combining multiple approaches and relying on source algorithms, *annif* can build on the strengths of individual algorithms and adapt to different settings. With built-in support for regular subject indexing and classification processes especially for electronic documents as well as collections that otherwise would not be indexed at all.

1

## Understand what Annif is

Study the website [annif.org](https://annif.org), watch a presentation about it, or read the LIBER Quarterly [paper](#).



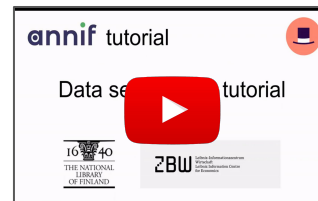
**Annif: DIY automated subject indexing using multiple algorithms**  
Artem Chepur [@artemche](#)

**Abstract**  
Manually indexing documents for subject-based access is a labor-intensive process. We propose using metadata gathered from bibliographic databases to train algorithms that assist librarians in their work. We have developed Annif, an open-source tool and environment for automated subject indexing. After making it with a subject vocabulary and training corpora, Annif can be used to suggest subject headings for new documents. We have tested Annif with different document collections including scientific papers, old newspaper books and contemporary e-books. Our paper focuses on "old-fashioned" services: Periodicals, and the archives of a local newspaper. The results of analyzing scientific papers and current books have been promising, while other types of documents have proved to be more challenging. The current version is based on a combination of existing natural language processing and machine learning tools. By combining multiple approaches and relying on open source algorithms, Annif can build on the strengths of individual algorithms and adapt to different settings. With Annif, we expect to improve subject indexing and classification processes especially for electronic documents as well as collections that otherwise would not be indexed at all.

2

## Complete this hands-on tutorial

Watch the videos, install Annif, and complete the exercises as far as you can, on your own time.



**Exercise 2: Set up and train a TFIDF project**

Annif requires you to set up one or more projects before you can use it. A project is a set of configuration settings and (usually) some data files, such as a trained machine learning model. A project is identified by a project id, which is typically a short string such as "you-first".

Projects are defined in a configuration file called `projects.cfg`. Annif looks for a `projects.cfg` file in the current working directory unless you tell it otherwise using either the `-c` command line option or the `ANNIF_PROJECTS` environment variable. For the rest of this tutorial, we will be using a working directory.

In this lesson, we will set up the simplest kind of Annif project, which uses a TFIDF model that needs to be trained on example documents (and/or metadata records before it can be used).

You need to choose which data set you want to use: either the `youfirst` data set from the National Library of Finland or the `low-dow` data set from ZBW. Either one will work, but you need to do something.

**1. Create a `projects.cfg` file**

Use a text editor to create a `projects.cfg` file within the Annif tutorial directory.

If you use the `youfirst` data set, use the following contents:

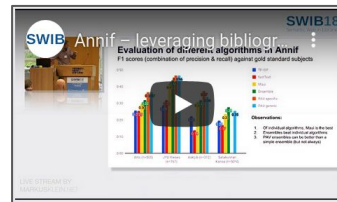
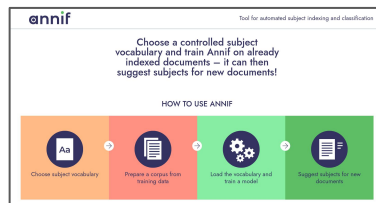
```
[you-first-v01]
name=youfirst-v01
language=fi
```

you  
are  
here

1

## Understand what Annif is

Study the website [annif.org](https://annif.org),  
watch a presentation about it,  
or read the LIBER Quarterly [paper](#).



# AmniDi: Automated subtyping in indexing using multiple algorithms

Author: [Dora Soutamou](#)

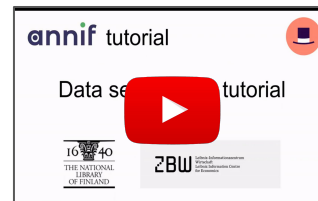
## Abstract

Manually building lexicons for a subject-based access is a labor-intensive process. We propose using automated subtyping (ontological) databases to reduce the manual effort. We build an ontology for the AmniDi lexicon, an open source tool and subsumers for automated subtyping. After testing it with a subject vocabulary and existing ontologies, AmniDi can be used to assign subject headings to the ontology. We have used AmniDi to generate different document collections including scientific subjects, old standard books and contemporary books, (old) plain text and "old alternative" texts. We found Wikipedia, and the AmniDi ontology, to be the most useful. We found that AmniDi and standard books have been maintaining, while other services of the datasets have not been in place through the current version. We found as a combination of existing natural language processing and machine learning tools, in order to improve the application and existing subject-based algorithms, and can be built on the strength of existing subject-based access to different versions. With AmniDi, we expect to improve subject heading and classification processes required by electronic documents as well as the manual effort needed to build a subject-based access.

2

## Complete this hands-on tutorial

Watch the videos, install Annif, and complete the exercises as far as you can, on your own time.



### Exercise 2: Set up and train a TFIDF project

`Anet` requires you to set up one or more **projects** before you can use it. A project is a set of configuration settings and (usually) some data files, such as a trained machine learning model. A project is identified by a **project id**, which is typically a short string such as "you-stuff-en".

Projects are defined in a configuration file called `projects.cfg`. `Anet` looks for a `projects.cfg` file in the current working directory unless you tell it otherwise using either the `-p` command line option or the `ANNET_PROJECTS` environment variable. For the rest of this tutorial, we will use a sample directory:

In this lesson, we will set up the simplest kind of Annot project, which uses a TFIDF model that needs to be trained on example documents and/or metadata records before it can be used.

You need to choose which data set you want to use: either the [yso-rii](#) data set from the National Library of Finland or the [ste-zlw](#) data set from ZSW. Either one will work, but you need to be consistent.

Use a text editor to create a `projects.cfg` file within the `Ans3Tutorial` directory.

```
[yos-tiff-se]
name=YOS TIFF project
```

3

### Join an online session (optional)

In the online sessions, you can ask questions, get help and discuss what you've learned. Registration required.



**you  
are  
here**

# Annif-tutorial GitHub repository

the main resource for this tutorial

NatLibFi / **Annif-tutorial**

Unwatch ▾

10

★ Star

6

🍴 Fork

2

<> Code

🔔 Issues

🔗 Pull requests

🎬 Actions

📁 Projects

📖 Wiki

🛡 Security

📈 Insights

⚙ Settings

🔑 master ▾

🔗 3 branches

🏷 0 tags

Go to file

Add file ▾

📄 Code ▾



**osma** minor wordsmithing

586b131 2 days ago

🕒 189 commits

📁 data-sets	Delete metadata for 2 fulltext docs that have been removed from Ec...	2 months ago
📁 exercises	minor wordsmithing	2 days ago
📁 presentations	Add a slide about Omikujii; other updates	6 months ago
📄 .dockerignore	Move Dockerfile to master branch	6 months ago
📄 .gitignore	Ignore mauidata directory	10 months ago
📄 Dockerfile	Build tutorial image based on Annif 0.49; add "tutorial" image tag	21 days ago
📄 LICENSE.txt	Add CC By 4.0 International license	10 months ago
📄 README.md	Add instructions for cloning/downloading the repo	9 months ago

## About



Instructions, exercises and example data sets for Annif hands-on tutorial

📖 Readme

📄 CC-BY-4.0 License

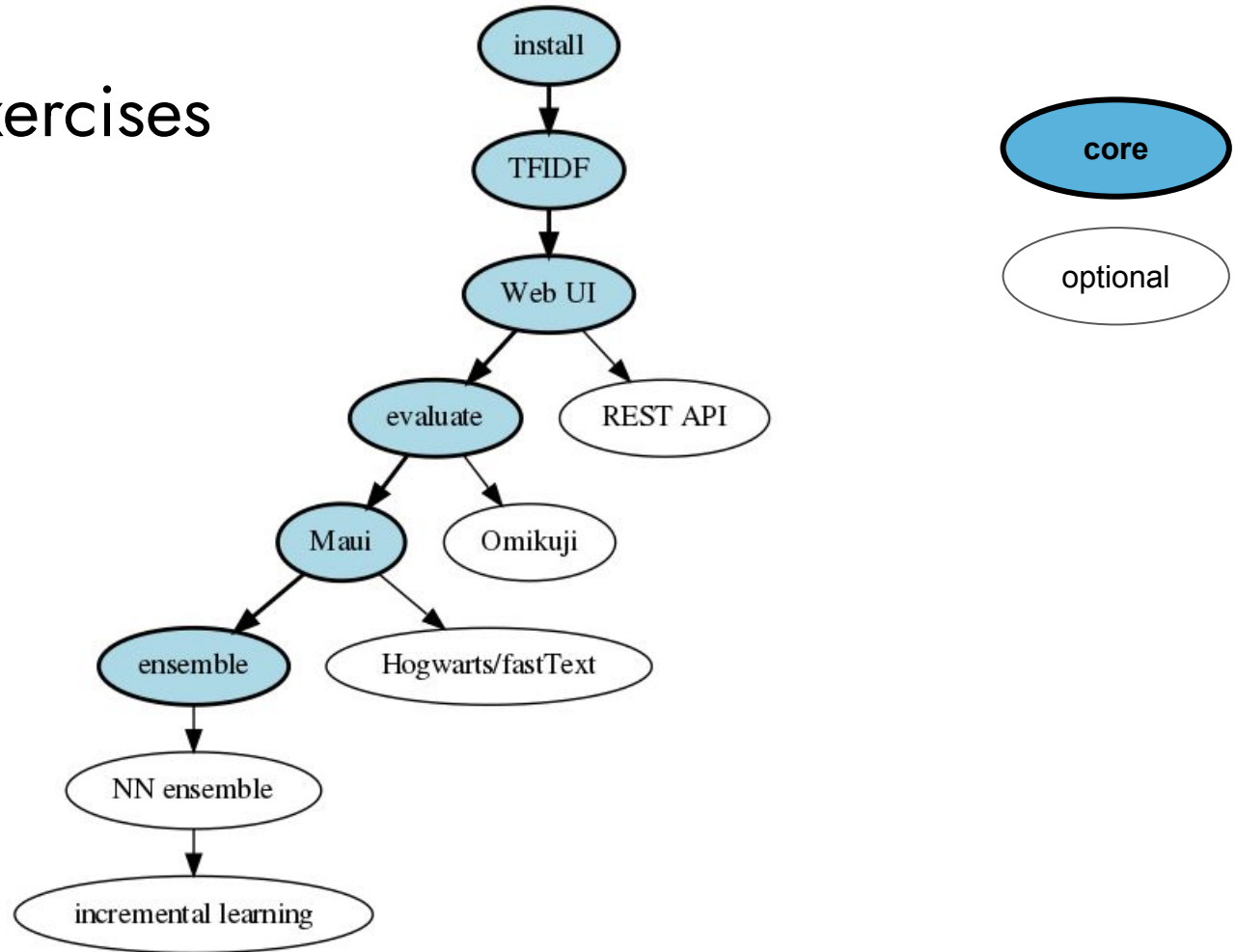
## Releases

No releases published  
[Create a new release](#)

## Packages

No packages published  
[Publish your first package](#)

# Overview of exercises



# Annif installation types

## VirtualBox install

**Recommended** for most people, as it's the easiest way of getting Annif running so you can work on the tutorial exercises.

Need to install VirtualBox software - available for Windows, macOS and Linux

## Docker install

**If you know Docker**, this is a good way of getting Annif set up, with all the dependencies included in a pre-built container.

Need to install Docker software - available for Windows, macOS and Linux

## Linux local install

**If you're an experienced Linux user** and used to working with Python packages, a local install allows maximum flexibility.

Needs Python 3.6, 3.7 or 3.8 and support for virtual environments.

# Accessing Annif

**Command line interface**

- setup and administration
- training models
- testing and evaluating models
- bulk indexing of documents

**Web user interface**

- interactive testing of models

**REST API**

- integrating Annif services to other systems



# Apply Annif on your own data!



Choose subject vocabulary



Prepare a corpus from  
training data



Load the vocabulary and  
train a model



Suggest subjects for new  
documents