

DSD

Department  
of Distributed  
Systems

*Máté Pataki*

MTA SZTAKI

## KOPI Protection Instead of Copy Protection

- n **Plagiarism**
- n KOPI Portal
- n How KOPI Works
- n KOPI Protection
- n Future Plans

## Problems

- n Plagiarism is a huge problem at universities
- n There are too many theses even at one university, no one can be familiar with all of them
- n It is not enough to feel that something could be a plagiarism, some proof is needed

## Problems - Existing Systems

- n Watermark or checksum
- n Authorship attribution
- n Open search engines
- n Text comparison
- n Questionnaire
- n Systems with unknown algorithms
- n No system for the Hungarian community

## What we need

- n Detects Partial Overlapping
- n Can not be automatically removed
- n Language independent
- n Can protect proprietary documents
- n One to many comparison
- n Without user intervention
- n Known algorithm

- n Plagiarism
- n **KOPI Portal**
- n How KOPI Works
- n KOPI Protection
- n Future Plans

## The KOPI Project

- n KOPI Online Plagiarism Search and Information Portal – Web based **similarity** and **plagiarism** search service
- n Partner: Monash University, Melbourne
- n Sponsored by the Hungarian Government
- n Developed 2003-2004
- n The Service is freely available to everybody

## The Goal of KOPI

- n Protect digital libraries from illegal copying
- n Help teachers, professors, conference organizers to easily find copied work, and the original source
- n Inform students and authors about plagiarism and citations and the relevant (Hungarian) laws
- n Increase the values of papers, theses by certifying their genuineness

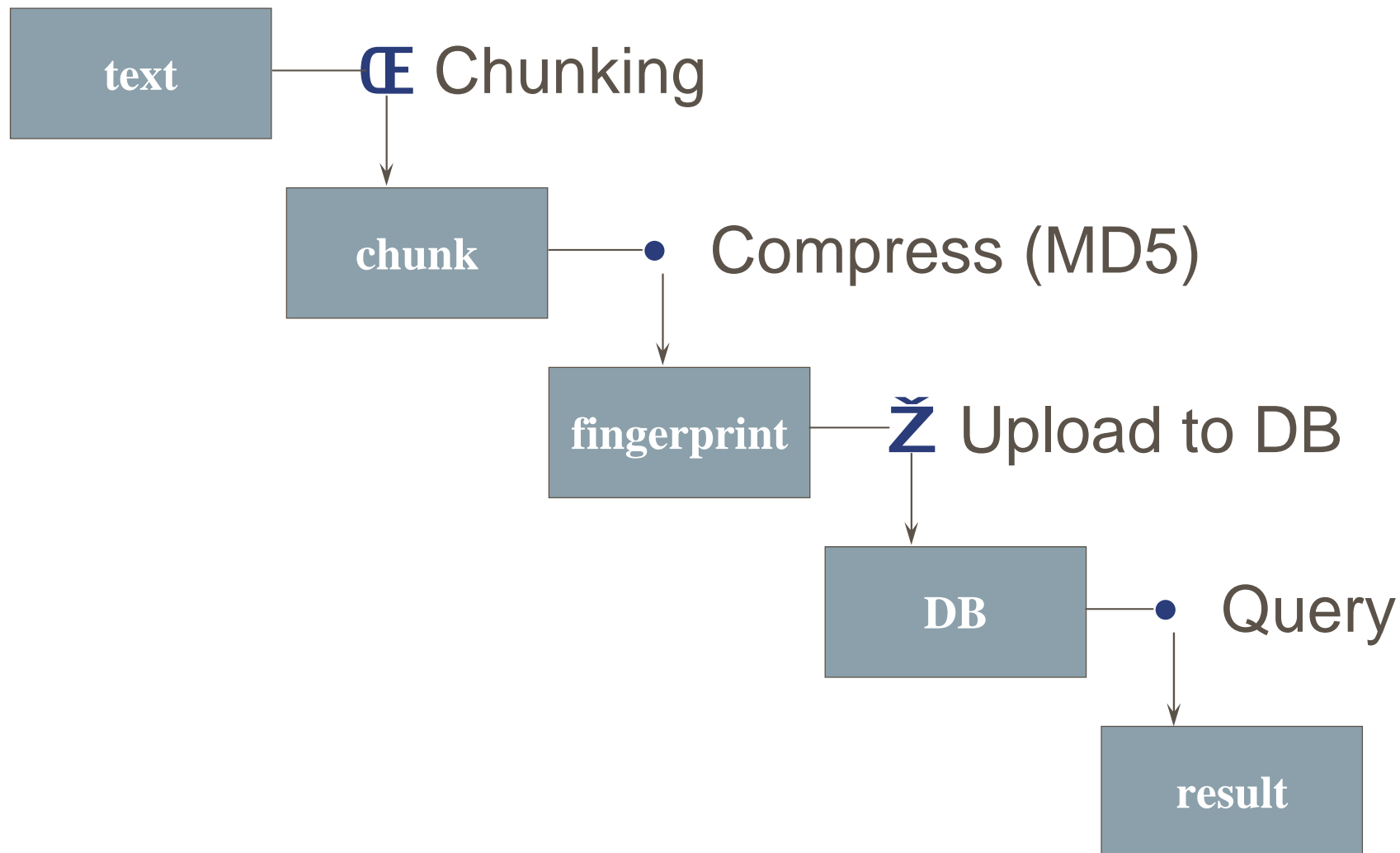


## Plagiarism Search Services

- n Compare uploaded documents to each other
- n Find similar documents on the database of the system:
  - n Within the users own documents
  - n Documents uploaded by others
  
  - n Documents from the Internet
  - n Digital libraries (MEK)
  - n Universities
  - n ...

- n Plagiarism
- n KOPI Portal
- n **How KOPI Works**
- n KOPI Protection
- n Future Plans

## How it works



## Word chunking

- Original  
The goal of the KOPI online Plagiarism Search and Information Portal is to protect documents against plagiarism.
- Word chunking (n=5)  
the goal of the kopi  
online plagiarism search and information  
portal is to protect documents  
...

## Overlapping word chunking

- Original  
The goal of the KOPI online Plagiarism Search and Information Portal is to protect documents against plagiarism.
- Overlapping word chunking (n=5)
  - the goal of the kopi
  - goal of the kopi online
  - of the kopi online plagiarism
  - the kopi online plagiarism search
  - kopi online plagiarism search portal
  - ...

# Compressing fingerprints

## Hash based algorithm (MD5)



- n Input length is not limited
- n Fast
- n The chance of two different texts to have the same MD5 code is small
- n Irreversible
- n Can protect proprietary documents

- n Plagiarism
- n KOPI Portal
- n How KOPI Works
- n **KOPI Protection**
- n Future Plans

# Copy protection

## n Pros

- n Harder to copy it
- n The way of the work can be followed (DRM)
- n More income for authors and sellers

## n Cons

- n Harder to use it
- n Can not totally prevent copying
- n Sometimes for the legal use it must be circumvented
- n It is not always legal to use
- n Personal rights problems (DRM)
- n Hinders the spreading of the work



## Copy Protection for Text Documents

- n PDF, DOC... protection
  - n Can be easily and automatically circumvented
- n Allow only online viewing
  - n Strongly restricts the use
  - n It is harder, but can be circumvented
- n Narrow down the number of authorized users
  - n If once the documents is out of the system...
- n Nothing protects against typing down
  - n Close up into a drawer and leave it there

## KOPI Protection

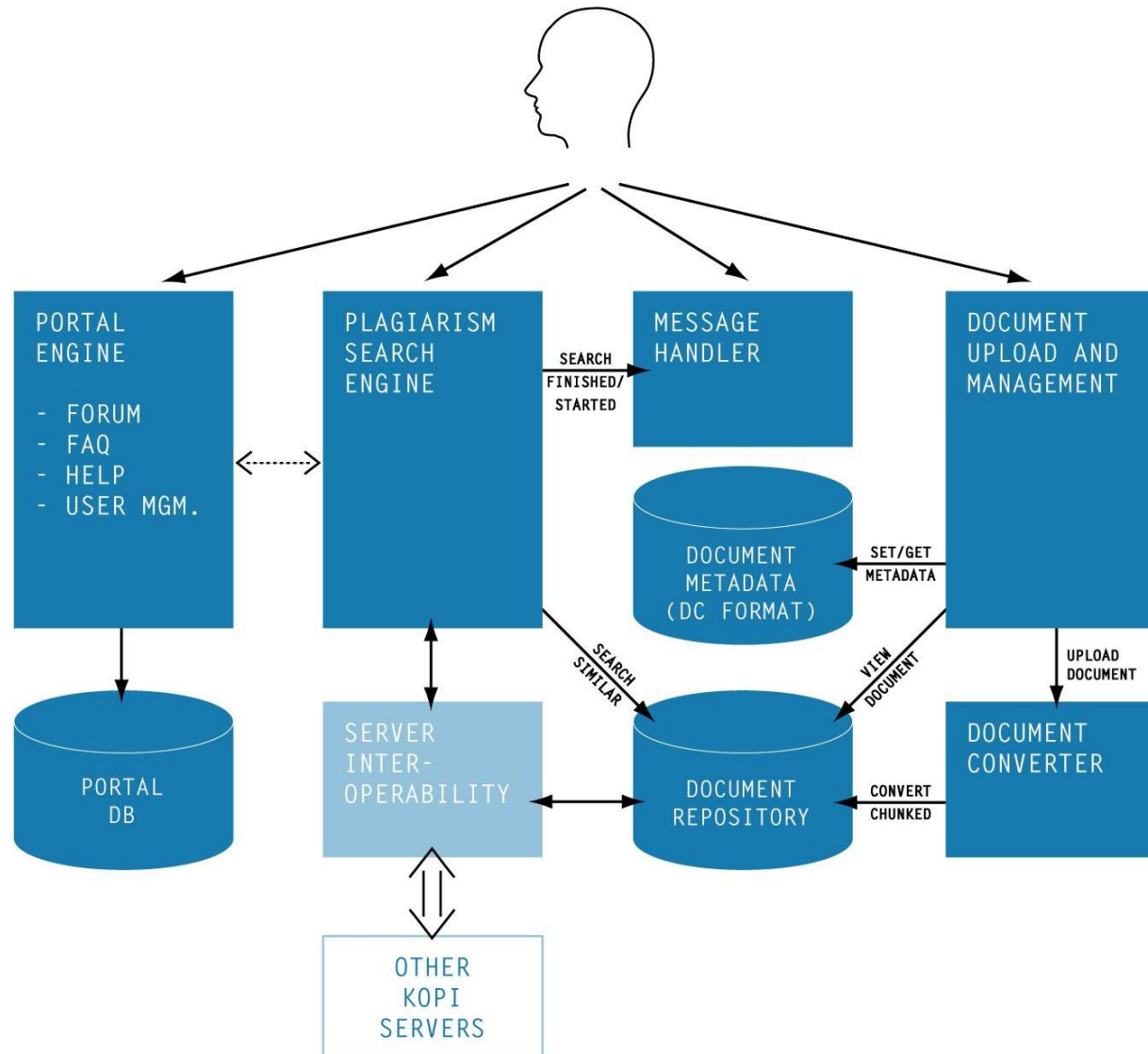
- n Documents uploaded into the KOPI System
  - n Plagiarism can be easily discovered
  - n The sources will also be known
  - n The risk to plagiarize will be too high
  - n Circumventing it is time consuming and can not be done automatically
- n The work can be freely distributed
  - n Must not deal with copy protection
  - n Search engines can index it
  - n More people read it
  - n More people cite from it

- n Plagiarism
- n KOPI Portal
- n How KOPI Works
- n KOPI Protection
- n **Future Plans**

## Future Plans

- n Distributed System
  - n Each university has an own system, but
  - n Their are able to search in the others DB
  - n Secure search with MD5 codes
- n Upload databases
  - n Online and offline databases
  - n Documents found on the Internet
- n Recognizing source codes and programming languages
- n SOAP interface for integrated use of KOPI

# Future Plans



**<http://kopi.sztaki.hu>**

# Thank you for your attention!

**Web:** <http://dsd.sztaki.hu>

**Email:** *Mate.Pataki @ sztaki.hu*