SFB 441: Linguistic Data Structures

Digital Text Resources for the Humanities: Legal Issues

Georg Rehm, Andreas Witt

Tübingen University, Germany

Digital Humanities 2007

Session 11 – June 5th, 09:00-10:30

Session Overview

- Language Corpora – Copyright – Data Protection: The Legal Point of View
  Timm Lehmberg, Felix Zimmermann.

- Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System
  Timm Lehmberg, Christian Chiarcos, Erhard Hinrichs, Georg Rehm, Andreas Witt.

- Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections
  Georg Rehm, Andreas Witt, Heike Zinsmeister, Johannes Dellert.

Why “Legal Issues”?

- All phases of handling text resources are inherently concerned with legal issues (e.g., compiling a corpus, or the distribution of a text collection).
- Most of the time, legal issues are overlooked. Or ignored.
- Due to missing jurisprudential expertise and the fear of doing something wrong, researchers tend to keep their data collections under lock and key instead of publishing corpora and other digital resources online.

“Sustainability of Linguistic Data”

SFB 441: Linguistic Data Structures

Joint project between three Collaborative Research Centres.

Hamburg University

SFB 538: “Multilingualism”

- About 60 members of staff
- About 20-30 corpora (mainly spoken language)
- Currently 14 research projects.

Humboldt University Berlin, Potsdam University

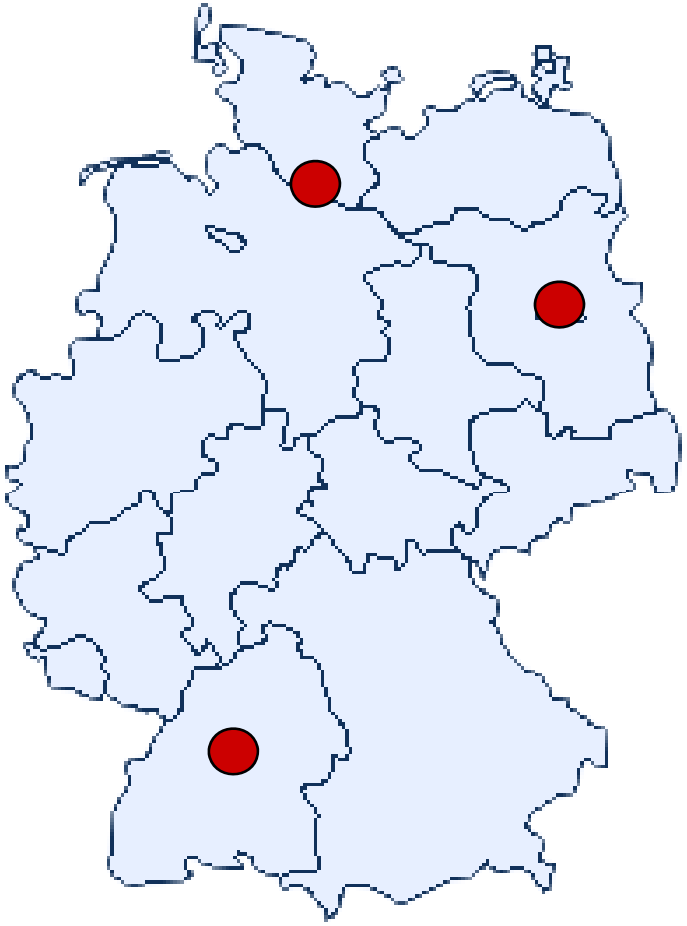
SFB 632: “Information Structure”

- About 60 members of staff
- About 20-30 corpora (mainly written language)
- Currently 14 research projects.

Tübingen University

SFB 441: “Linguistic Data Structures”

- About 60 members of staff
- About 20-30 corpora (mainly written language)
- Currently 16 research projects.



“Sustainability of Linguistic Data”

- The three Collaborative Research Centres created (and still create) a total of about 60–70 linguistic corpora.
  - These corpora took more than 50 person-*years* to create.
- Our goals:
  - to establish a web-platform that contains the corpora created in the three Collaborative Research Centres.
  - to make sure that the corpora can be used in other research scenarios in the future.
  - to prevent other researchers from re-inventing the wheel.
  - to enable other researchers to locate, to explore and to query the corpora in an intuitive way as well as to add corpora themselves.

“Sustainability of Linguistic Data”

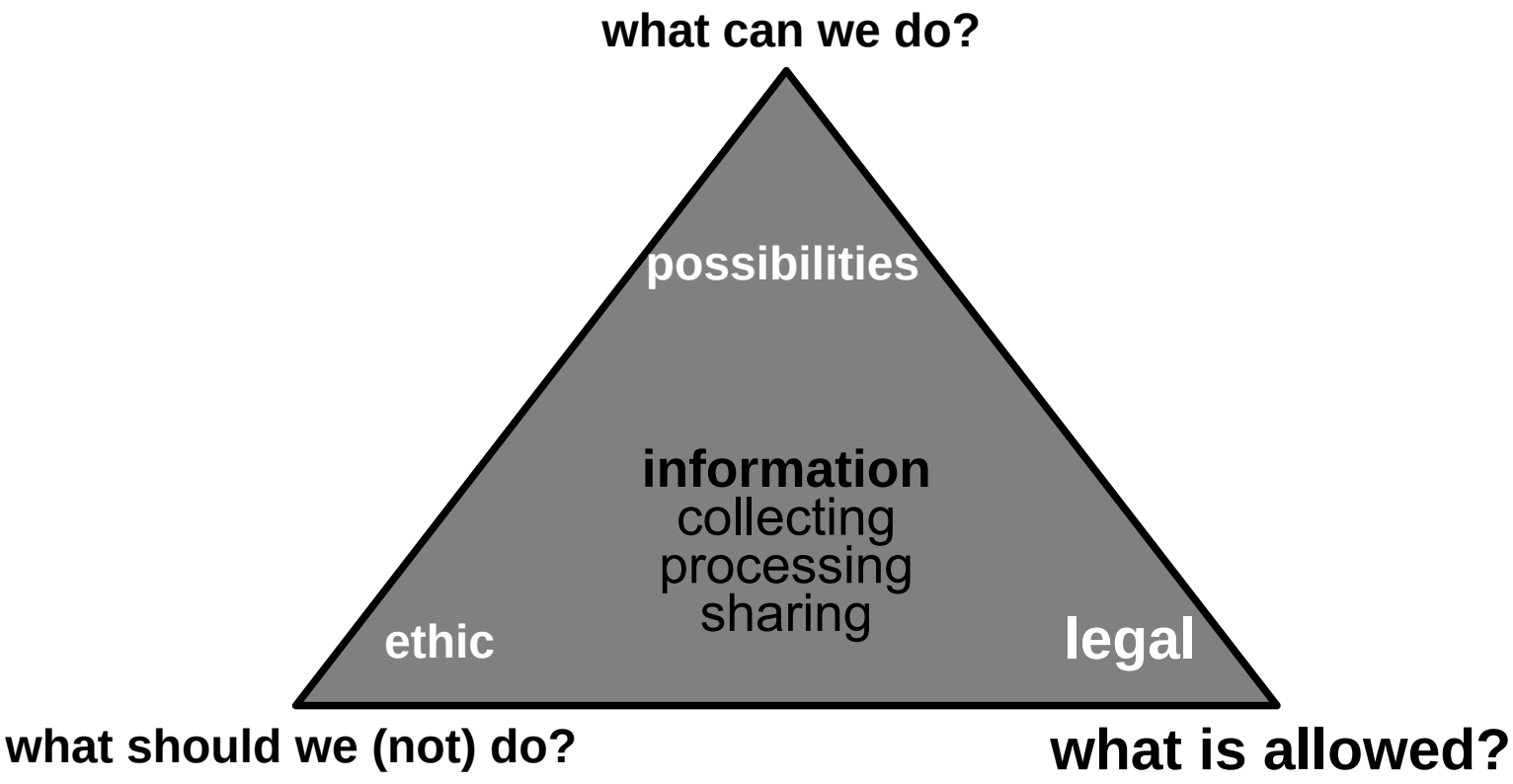
- Important legal questions we are confronted with:
  - What are the terms of distribution of a specific corpus?
  - Are the texts a corpus is based upon copyright-protected? Are we allowed to publish these texts at all?
  - How can we publish a corpus that we aren’t allowed to publish?
  - What kind of user-access scheme do we need for our web-based sustainability platform?
  - …

SFB 441: Linguistic Data Structures

Language Corpora – Copyright – Data Protection: The Legal Point of View

Timm Lehmberg, Felix Zimmermann

Introduction



Introduction

- **Our goal when considering legal aspects is: Improving sustainibility**

- **Figuring out**
  - The objects of legal protection
  - Holders of rights
  - Possibilities for researchers

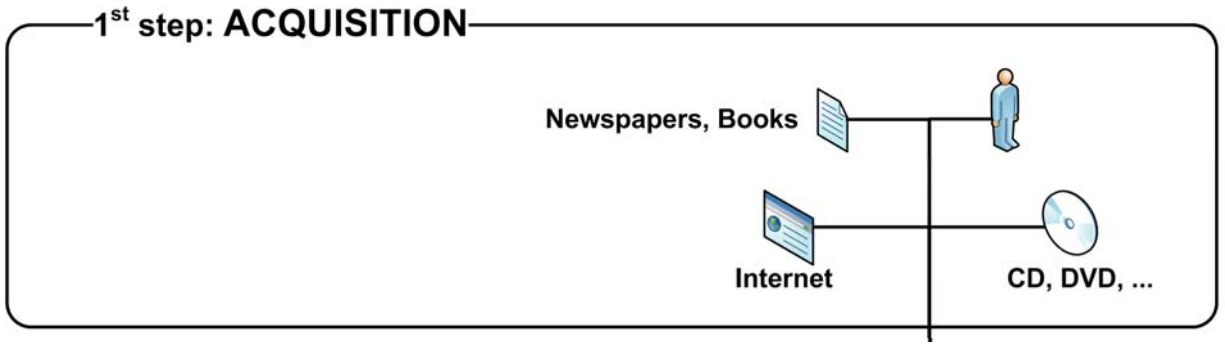
- **Possible effect**
  - Alignment of research methods to the legal situation

Introduction

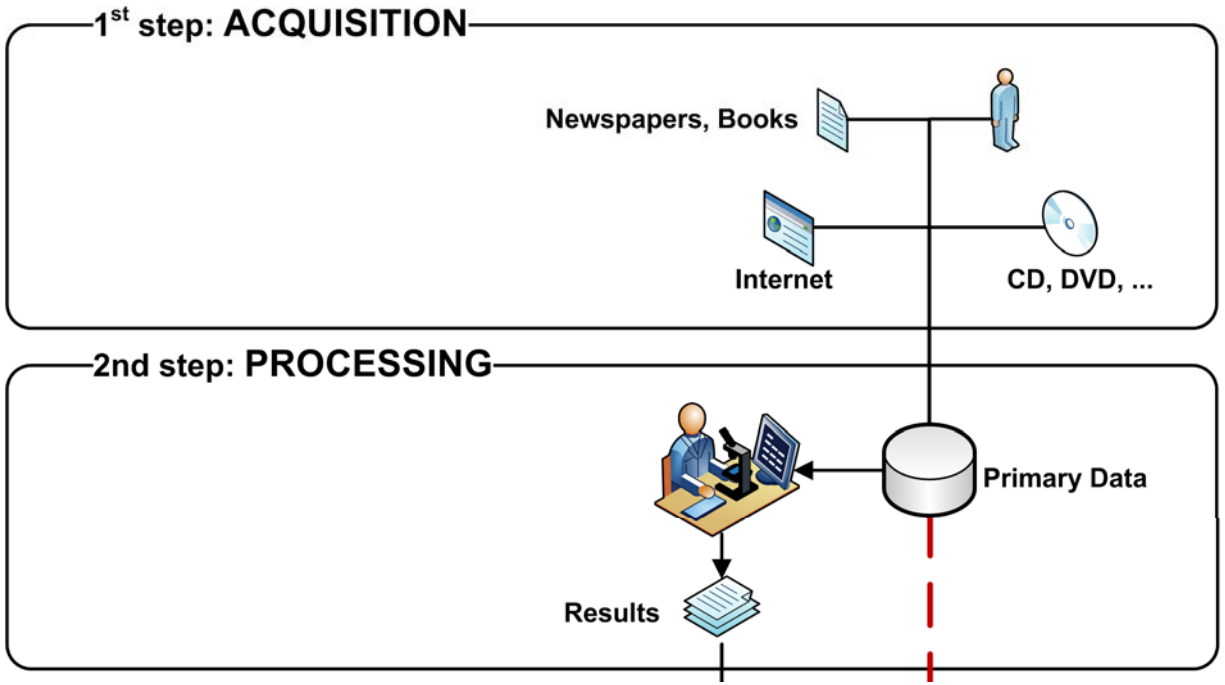
- **Legal scope**
  - Copyright and privacy law
  - Legislation and jurisdiction
  - Legal context
    - national (USA, Germany, ...)
    - supranational (EU)
    - international law (TRIPS, UNO)

Overview: Three Steps

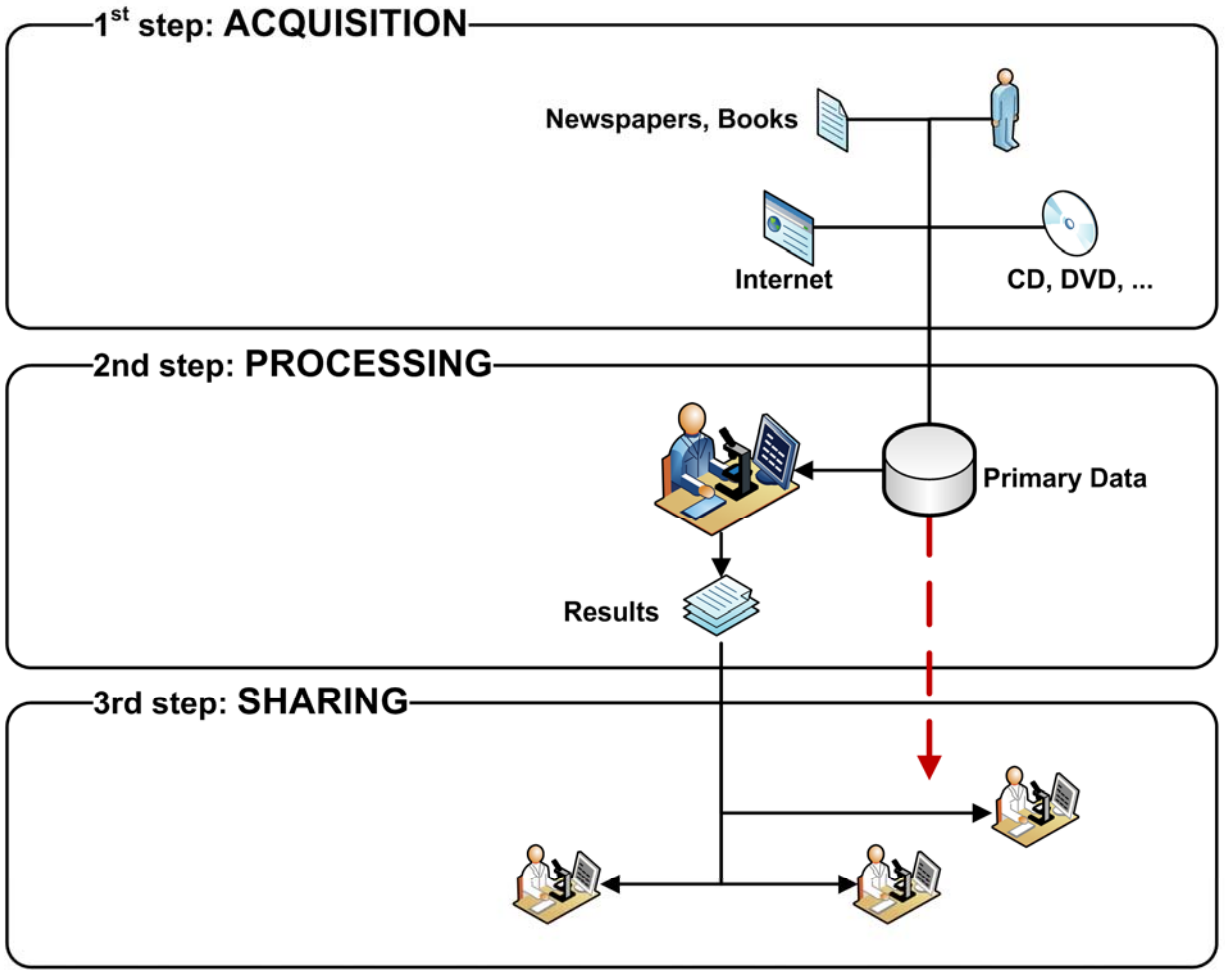
Overview: Three Steps



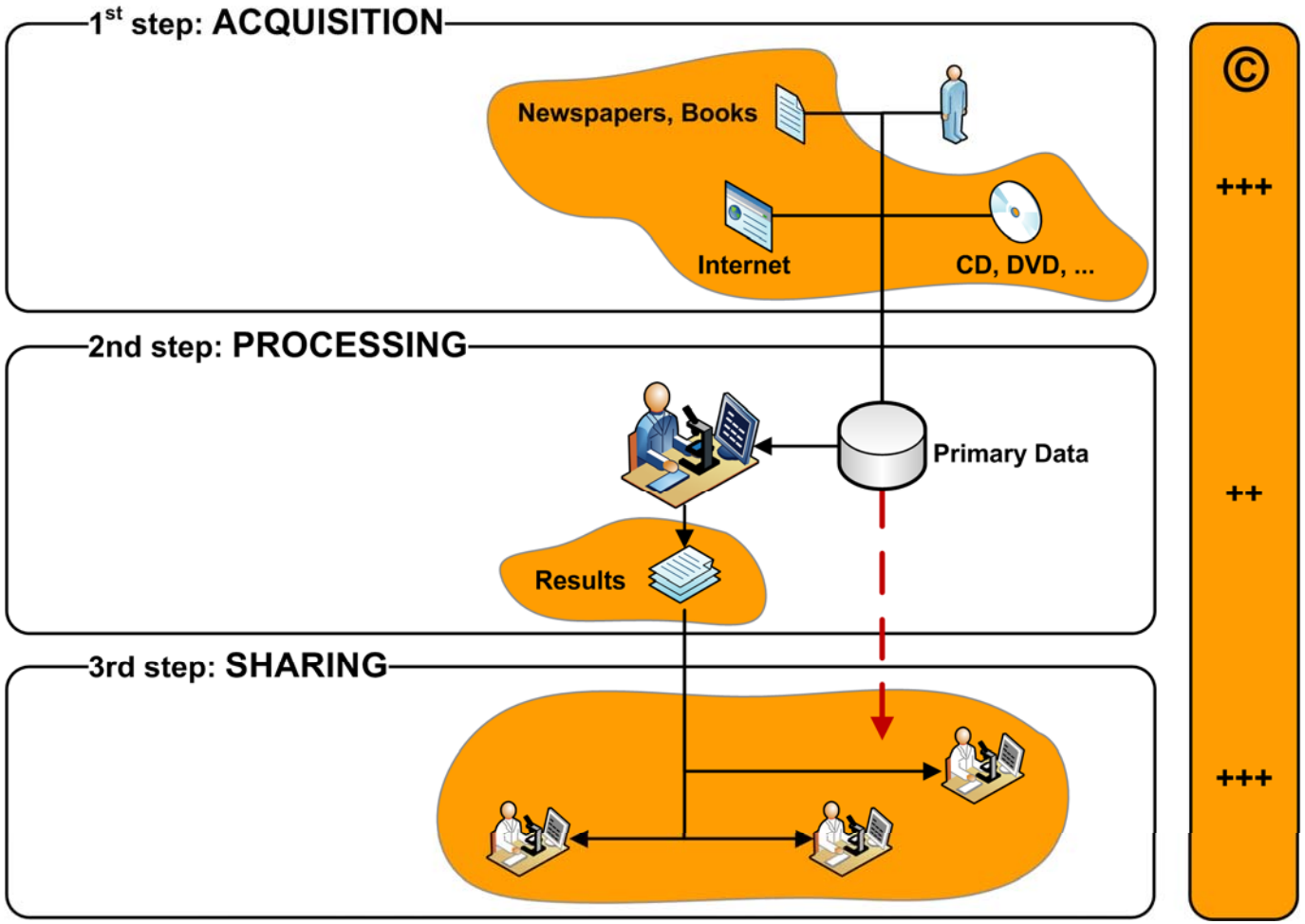
Overview: Three Steps



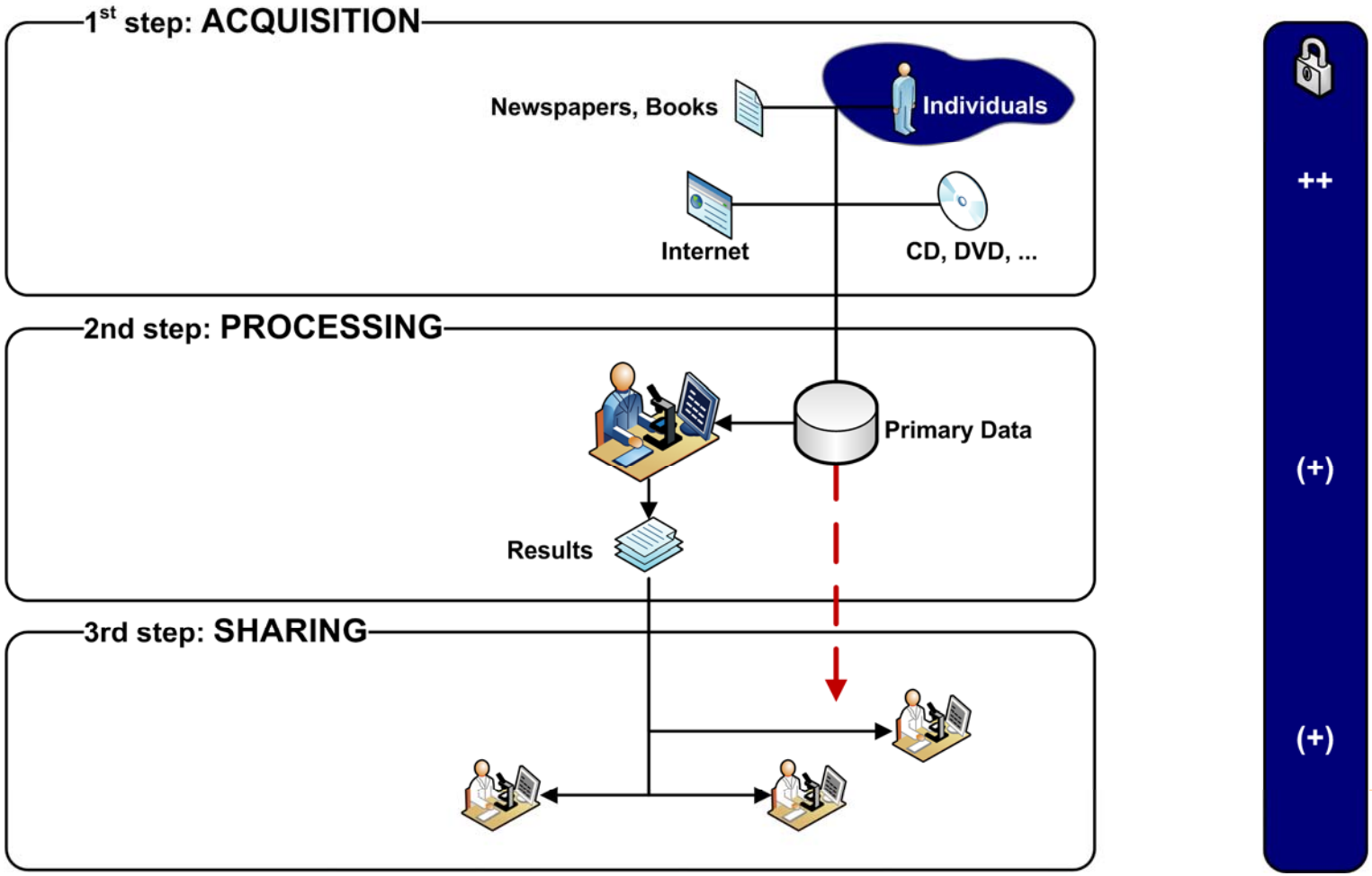
Overview: Three Steps



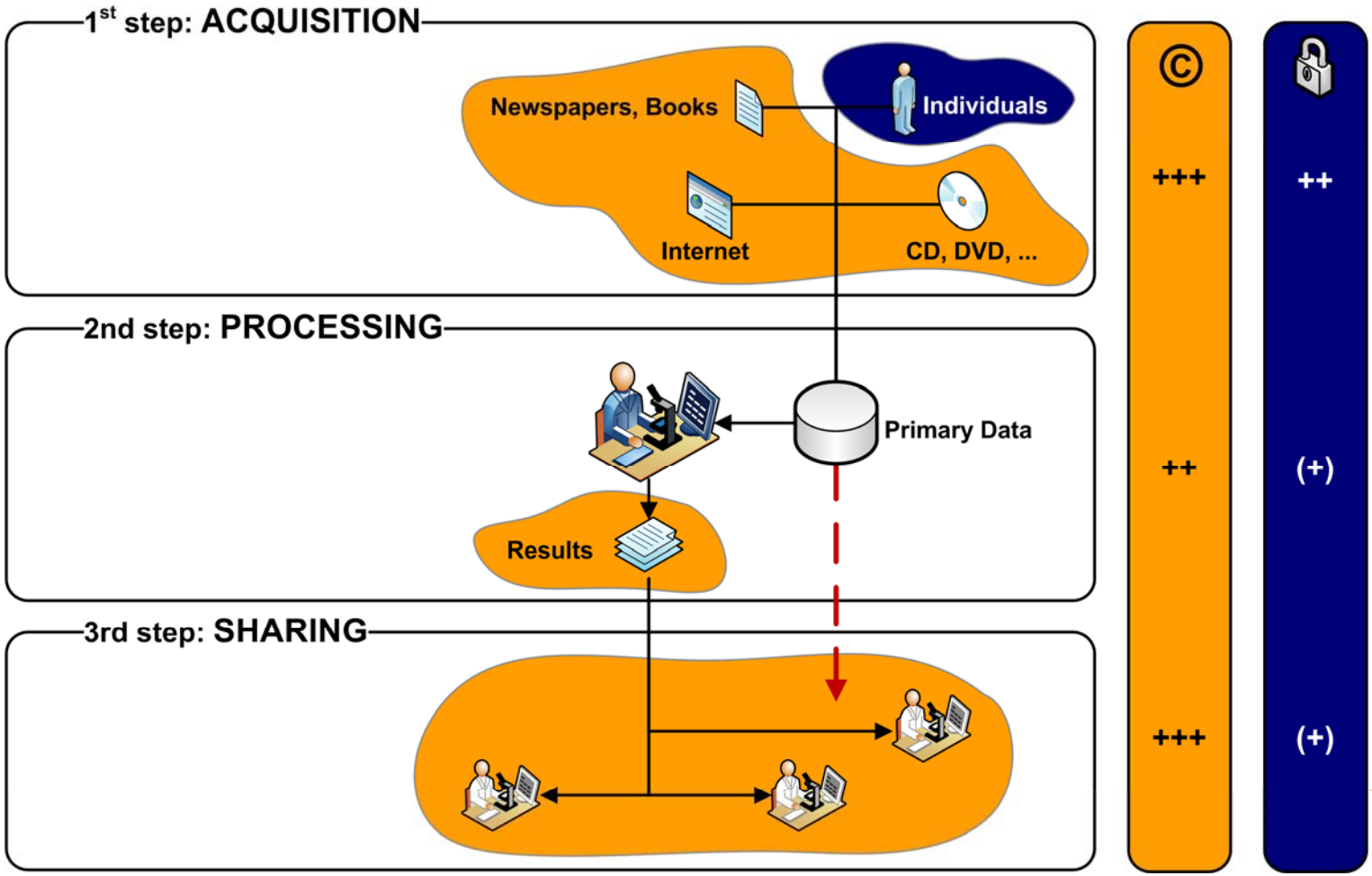
Overview: Three Steps



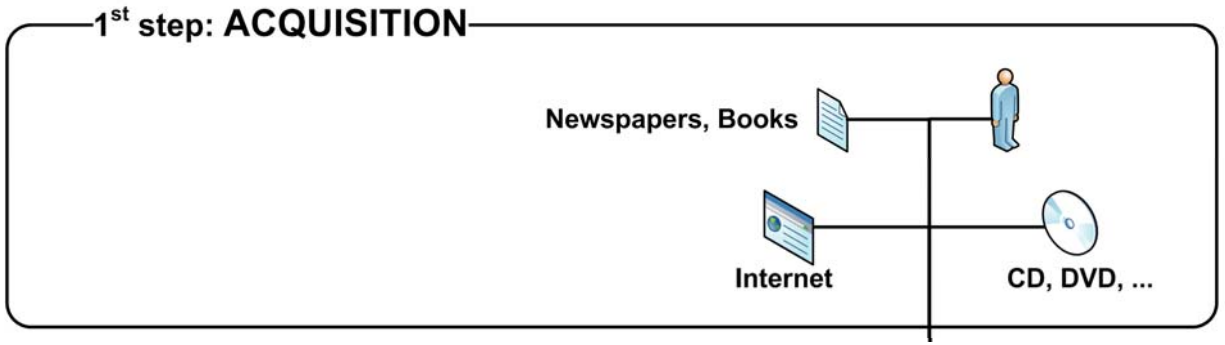
Overview: Three Steps



Overview: Three Steps



Overview: Three Steps



1st Step: Acquisition

- **Legal impact**

Copyright protection

E.g., systematic copying and pasting of texts from newspaper sites without owning a licence

Data protection

E.g., using personal information included in interviews, diaries (...) without asking permission

1st Step: Acquisition

- **Brief introduction into copyright-law (1/3)**
  - Intellectual property right
  - Legal protection of immaterial goods
    -   Financial background
    -  Personal background
  - E.g., literary works, newspaper articles, music, lyrics, databases, ...
  - **Licence agreement is needed**
    - Contract with holder of a copyright or author

1st Step: Acquisition

- **Brief introduction into copyright-law (2/3)**

Constraints for research purposes (USA)

Fair-Use-Doctrine (Legal Defend)

Video: Disney-Mashup

A Fair(y) Use Tale

by Eric Faden

Bucknell University, Pennsylvania

http://www.facstaff.bucknell.edu/efaden/

1st Step: Acquisition

- **Brief introduction into copyright-law (2/3)**

Constraints, e.g., for research purposes

Fair-Use-Doctrine (Legal Defend)

1. Nature of the copyrighted work
2. Amount copied
3. Commercial impact

1st Step: Acquisition

- **Brief introduction into copyright-law (3/3)**

Constraints, e.g., for research purposes

Rights

Art. 5 (3) a) Directive 2001/29/EG

» *small amounts of copyright protected material*

Art. 6 (4) Directive 2001/29/EG

» *right to get a DRM-free copy*

1st Step: Acquisition

- **Brief introduction into privacy law (1/4)**
  - Privacy is the expectation that **confidential personal information that is disclosed** in a private place
  - will **not be disclosed to third parties**
  - when that disclosure causes either
    - **embarrassment or**
    - **emotional distress**

    to a person of reasonable sensitivities

1st Step: Acquisition

- **Brief introduction into privacy law (2/4)**

Acquisition of personal data only ...

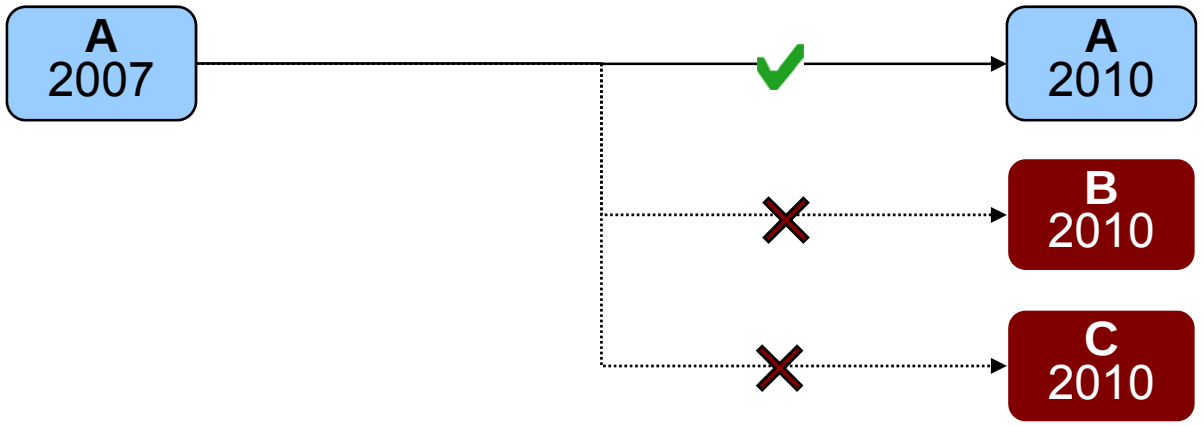
- If provided by law
- In compliance with the person
  - paper or paperless (e.g., via the web)
  - definition of specific **purposes**
  - double confirmation when dealing with ethnic or religious information

1st Step: Acquisition

- **Brief introduction into privacy law (3/4)**

Problems

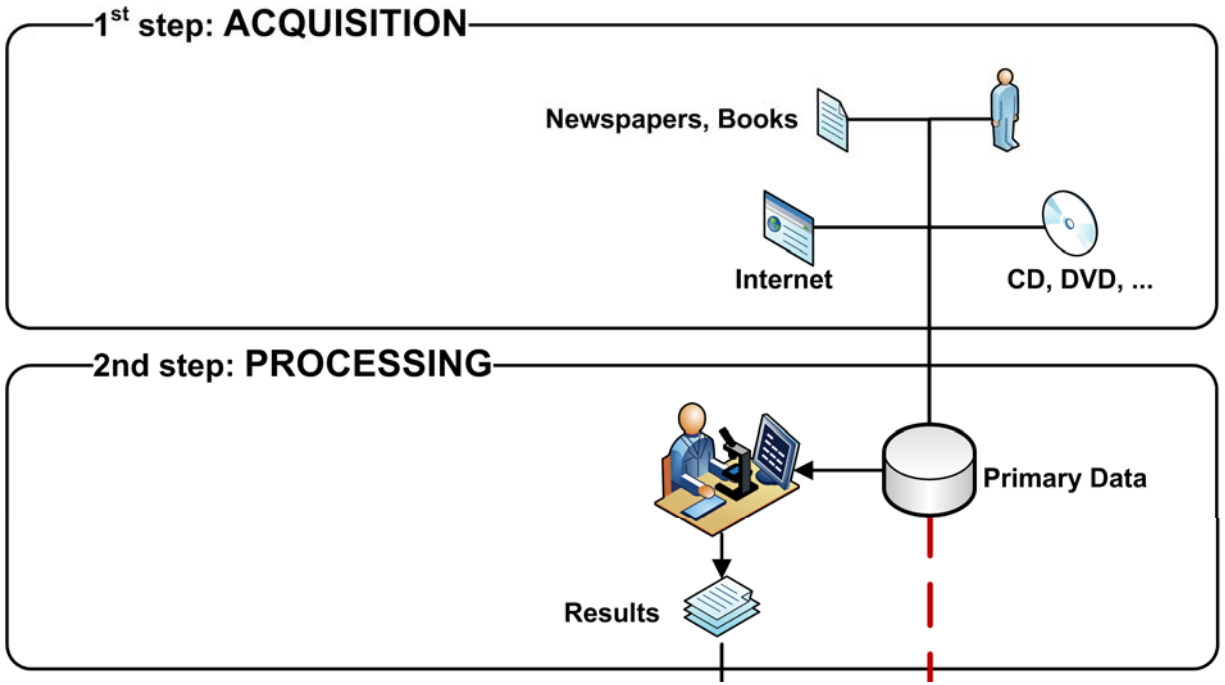
- Compliance for specific purpose given in advance



1st Step: Acquisition

- **Brief introduction into privacy law (4/4)**
  - No privacy problems:
    - Anonymization
    - Pseudonymization (aliases)
  - Economic definition!

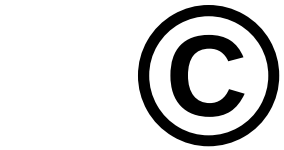
2nd Step: Processing



2nd Step: Processing

- **Processing corpus data e.g.**
  - Annotation
  - Transcription, transliteration
  - ...

2nd Step: Processing

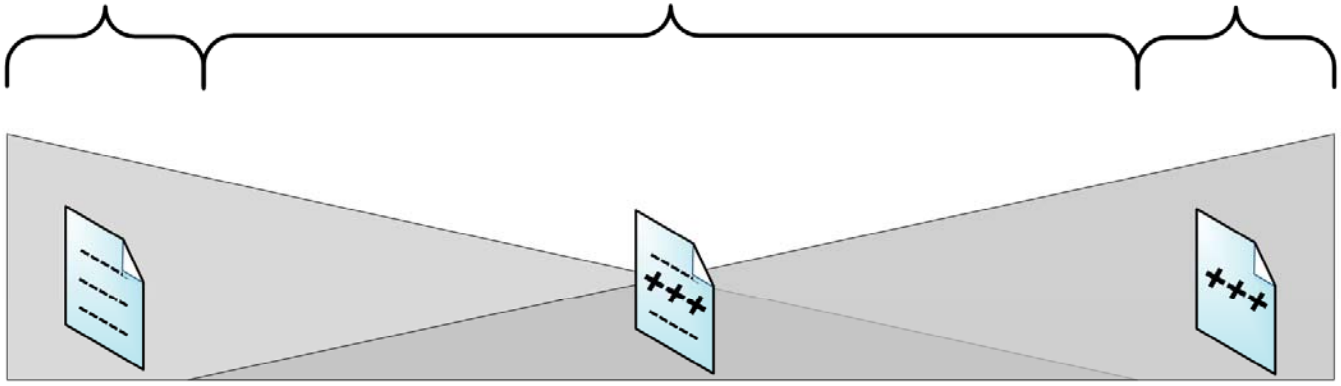


- **Copyright impact of annotation**

owner:
author
of primary data

owner:
authors of
primary data
and **annotation**

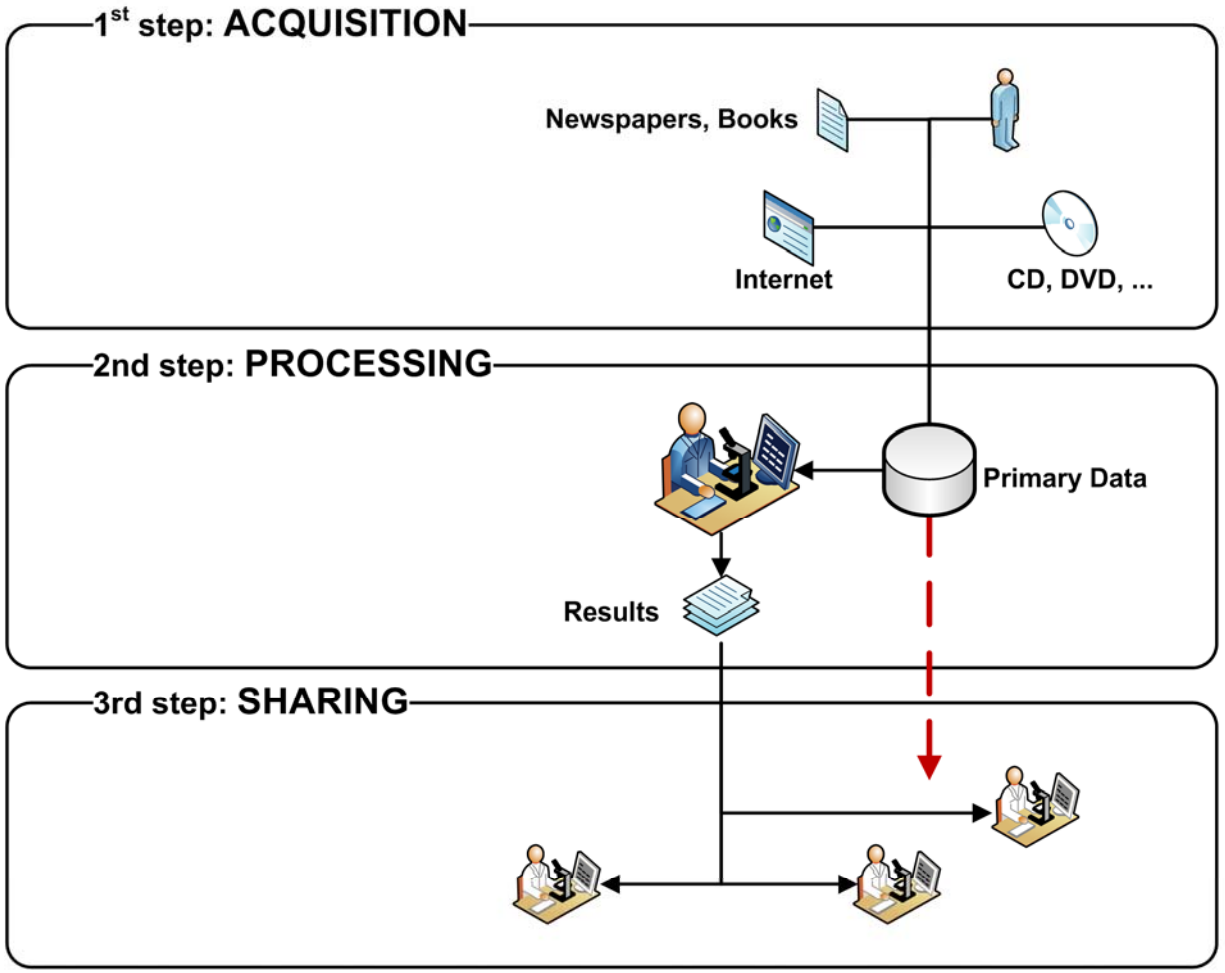
owner:
author
of annotation



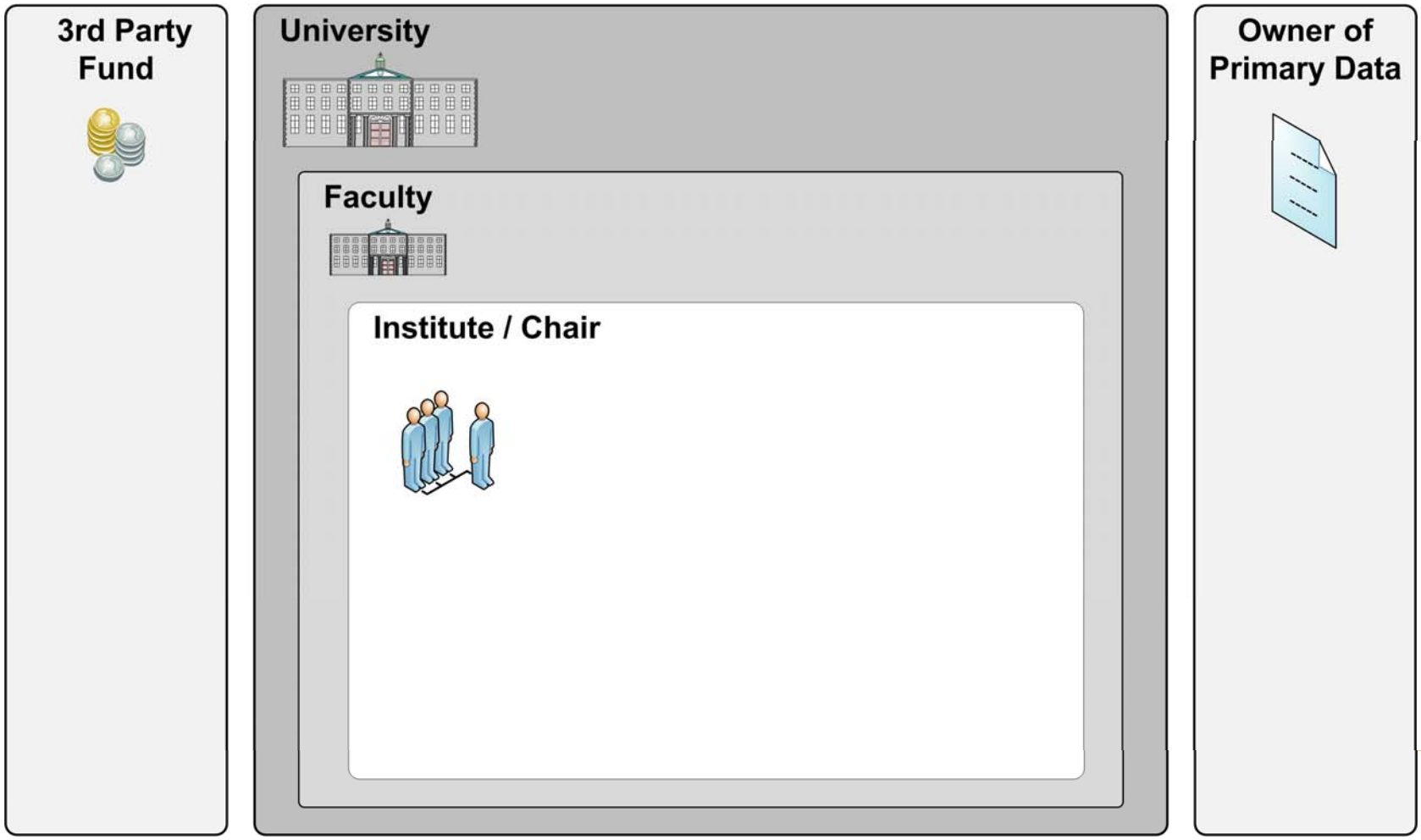
% amount of
primary data

% amount of
annotation

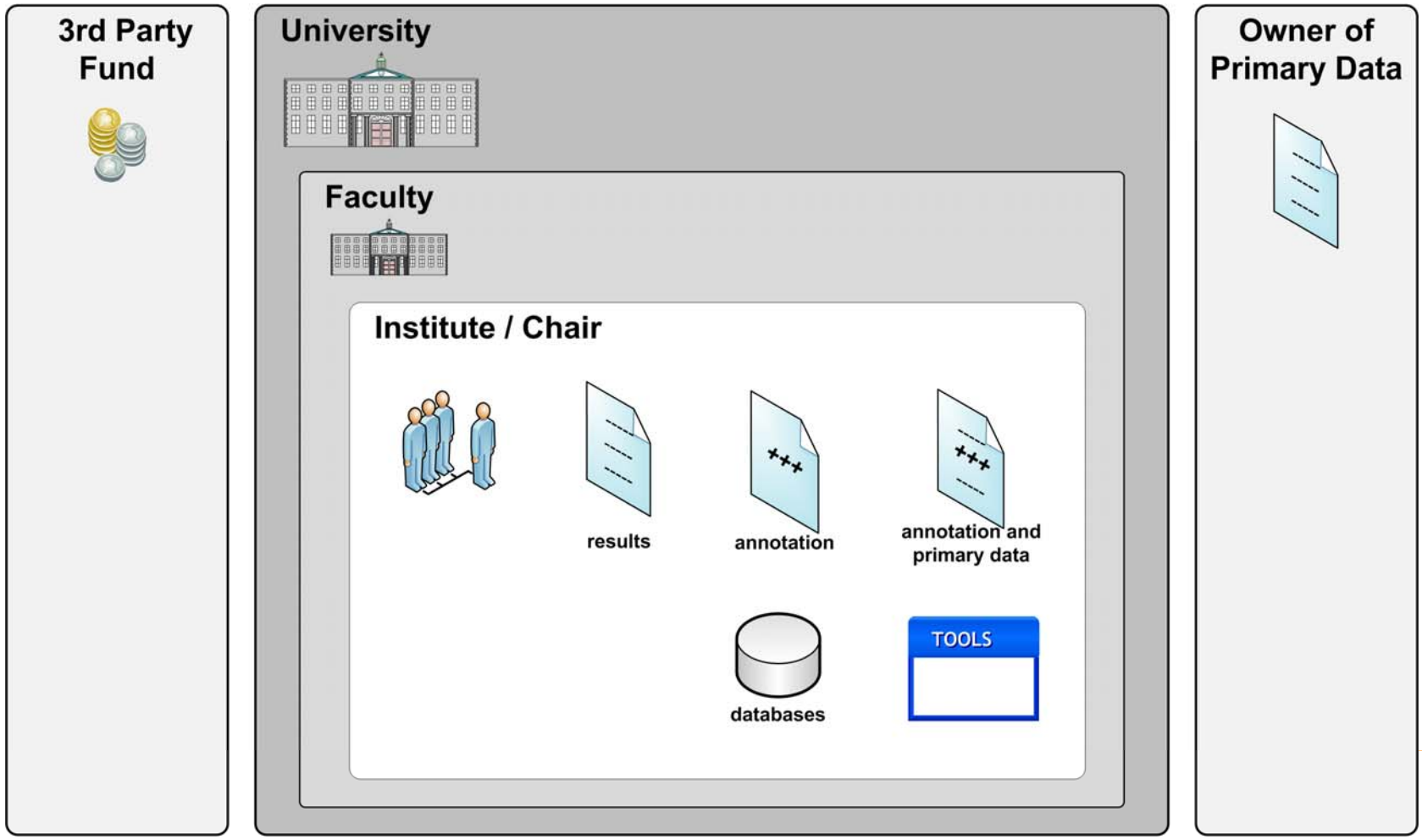
3rd Step: Sharing



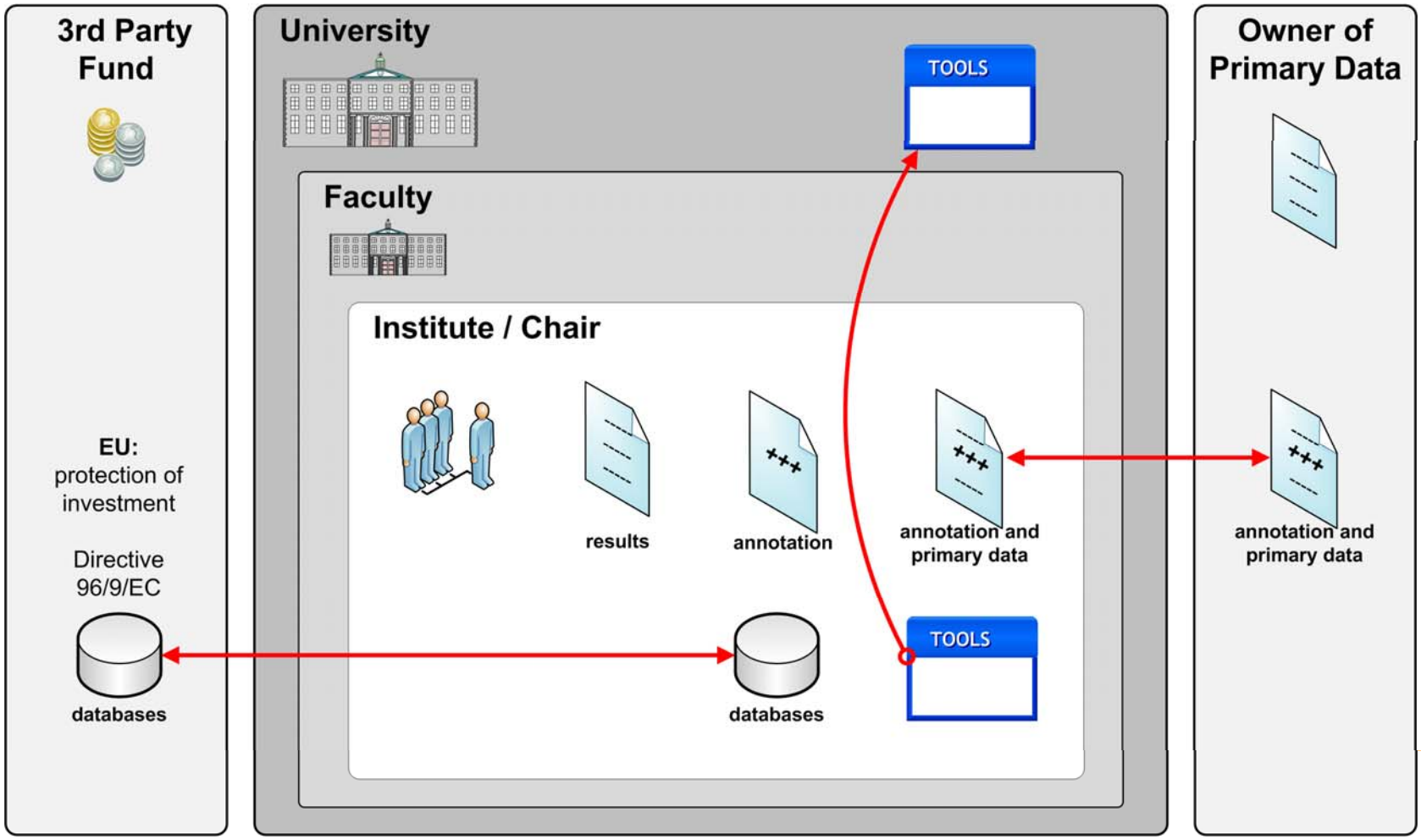
3rd Step: Sharing



3rd Step: Sharing



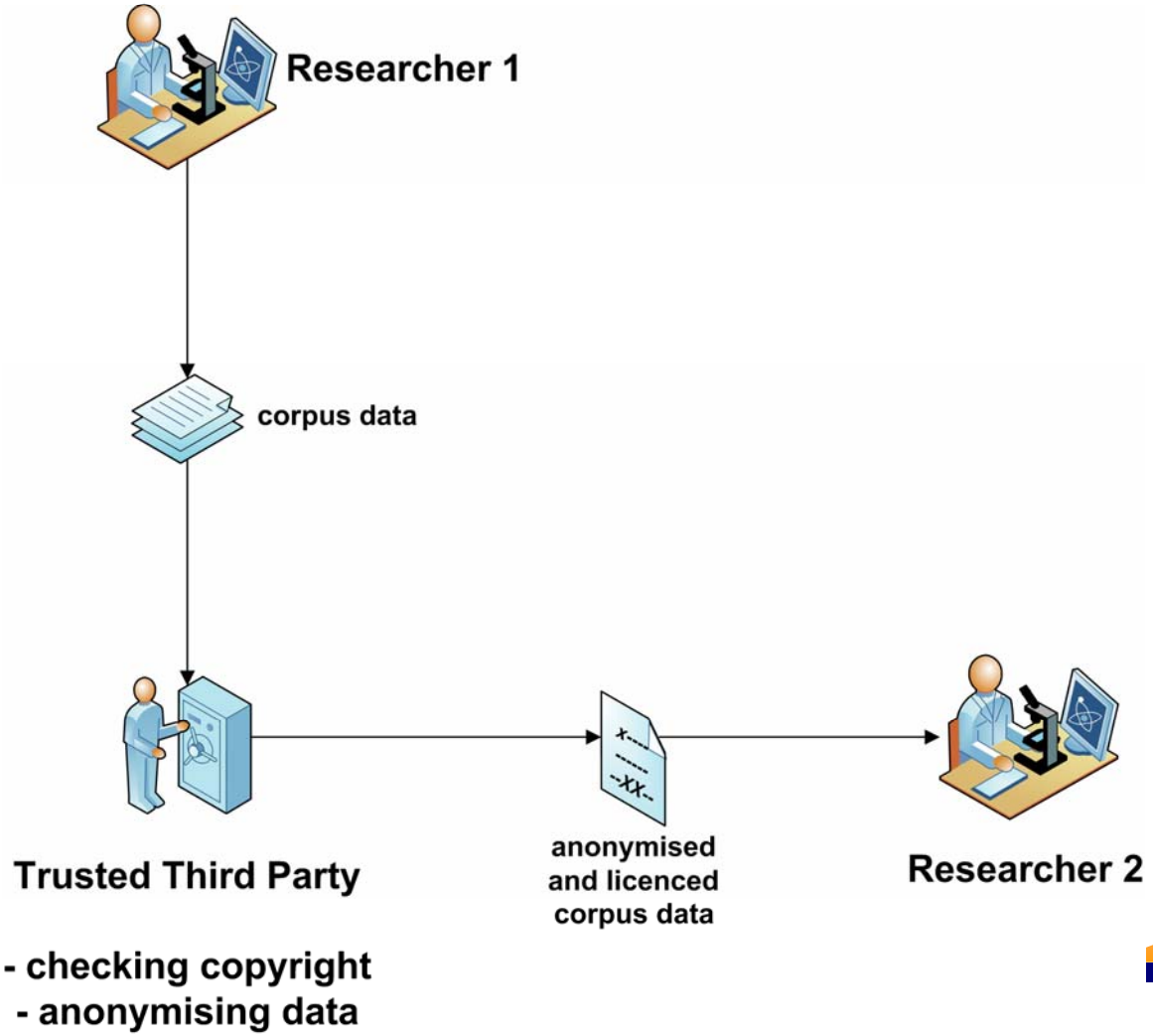
3rd Step: Sharing



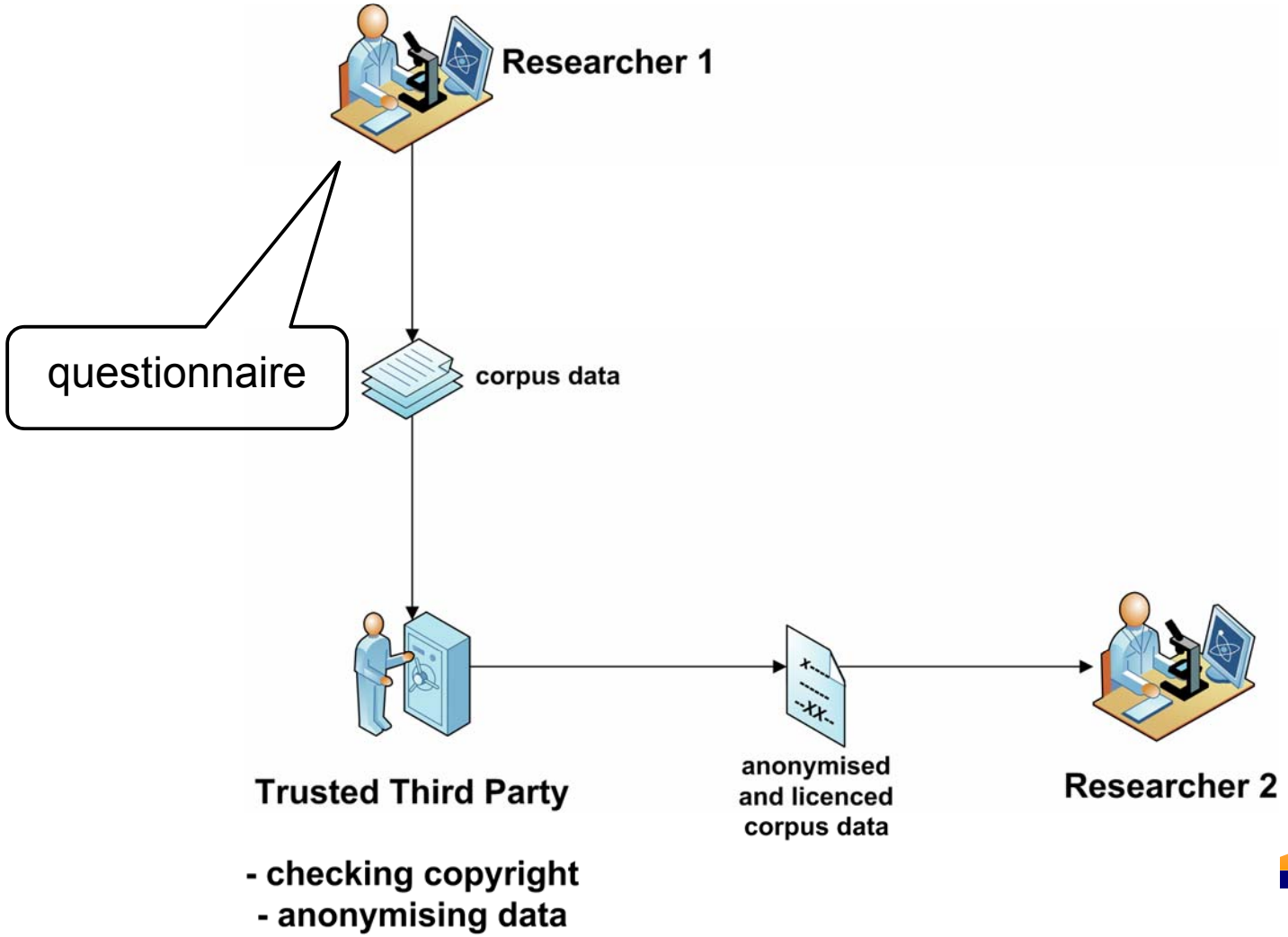
SFB 441: Linguistic Data Structures

How to achieve sustainability of data in line with legal aspects?

Trusted Third Party Model



Trusted Third Party Model



Q/A

Thank you very much! Questions?

SFB 441: Linguistic Data Structures

Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System

Timm Lehmberg, Christian Chiarcos,
Erhard Hinrichs, Georg Rehm, Andreas Witt

Overview

- Goals
- Conceptualization
- Implementation
- Experiences and Results

Goals: General Requirements

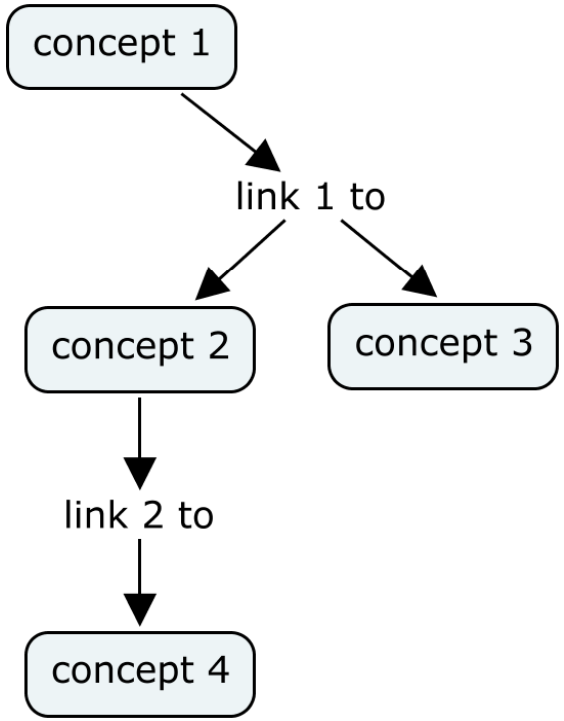
- A centralized acquisition of legally relevant metadata from the three research centers that enables us
  - to get an overview of all existing corpora and data collections
  - to get an overview of possible legal claims
  - to collect further related information that corresponds to other tasks of our project

Goals: Practical Requirements

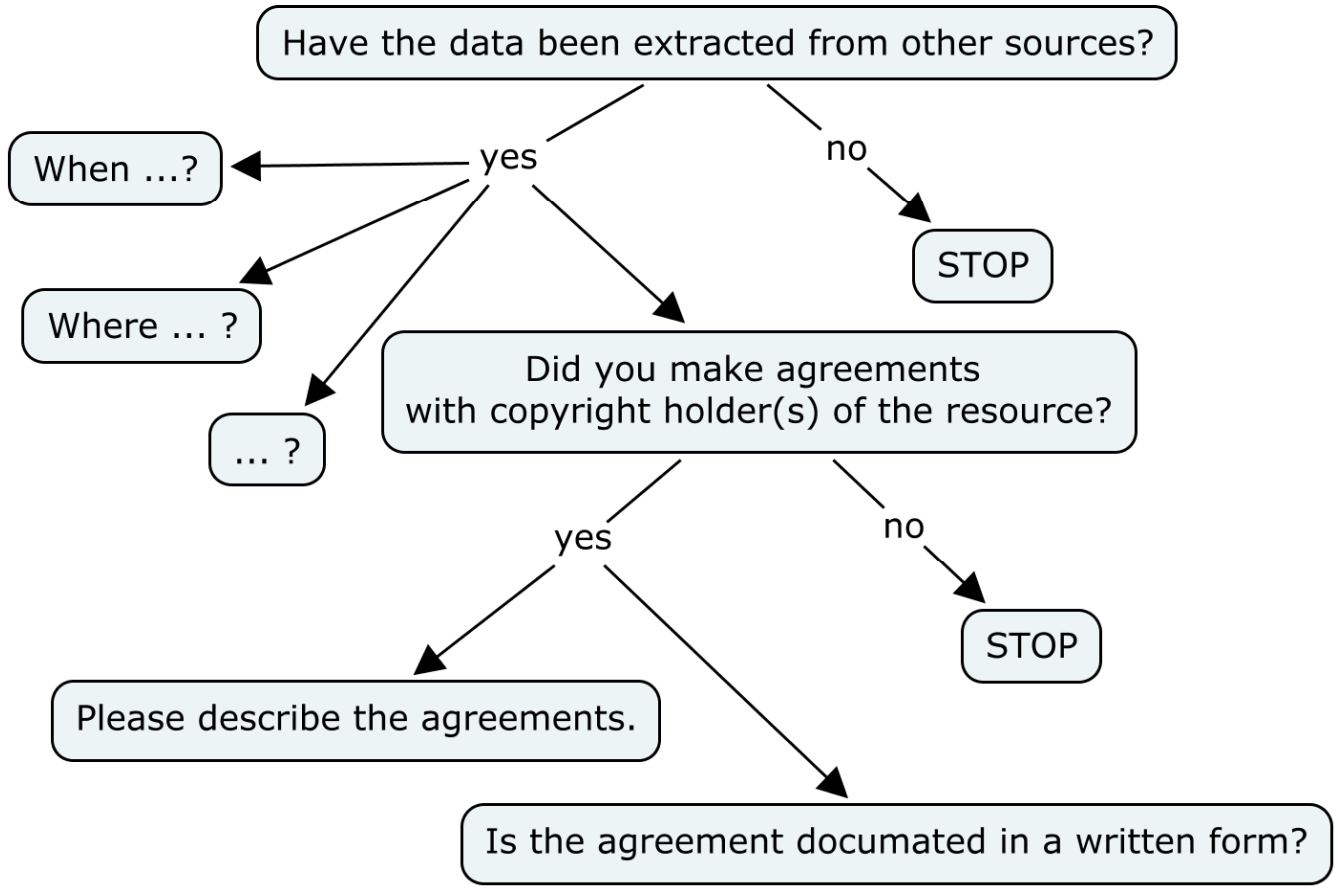
- The system should
  - provide web based and user-friendly interfaces that can be used intuitively
  - accommodate to the structures of projects and data
  - contain questions that are understandable to non-experts

Conceptualization

- Technique of our choice:
  concept mapping with IHMC “CmapTools”



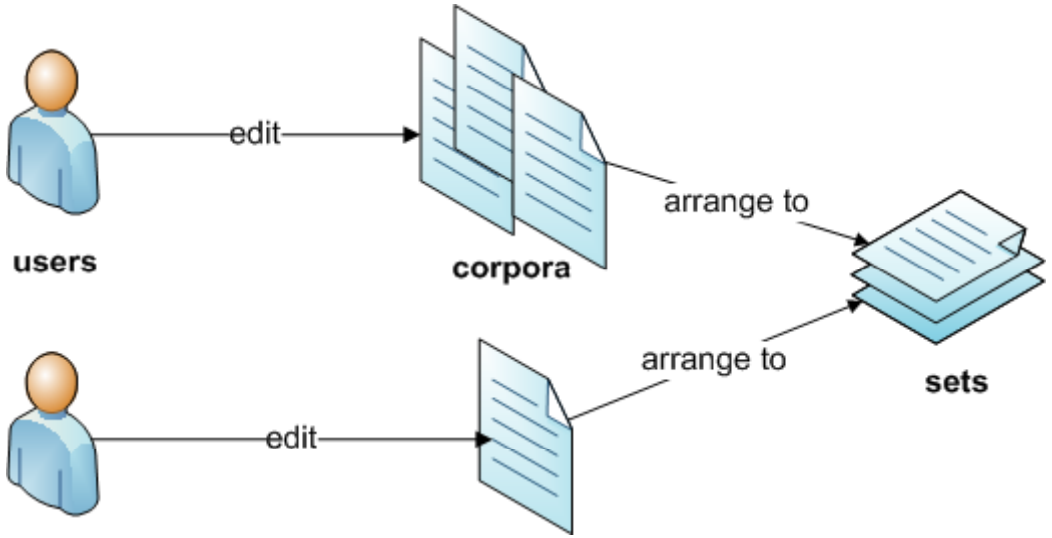
Conceptualization



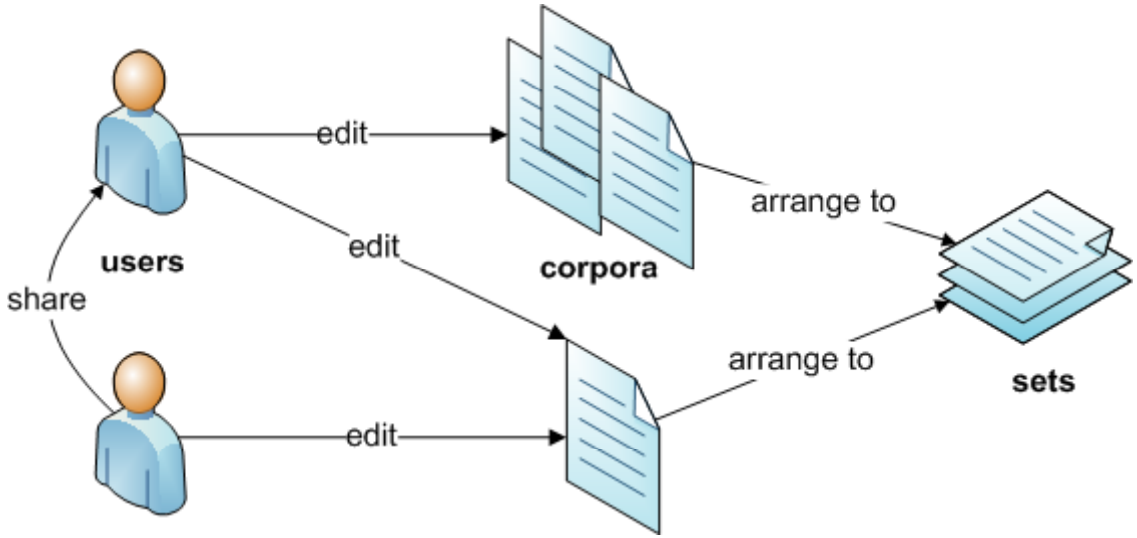
Implementation

- LAMP-environment (Linux, Apache, MySQL, PHP)
  - web accessibility
  - portability
  - access control

Implementation

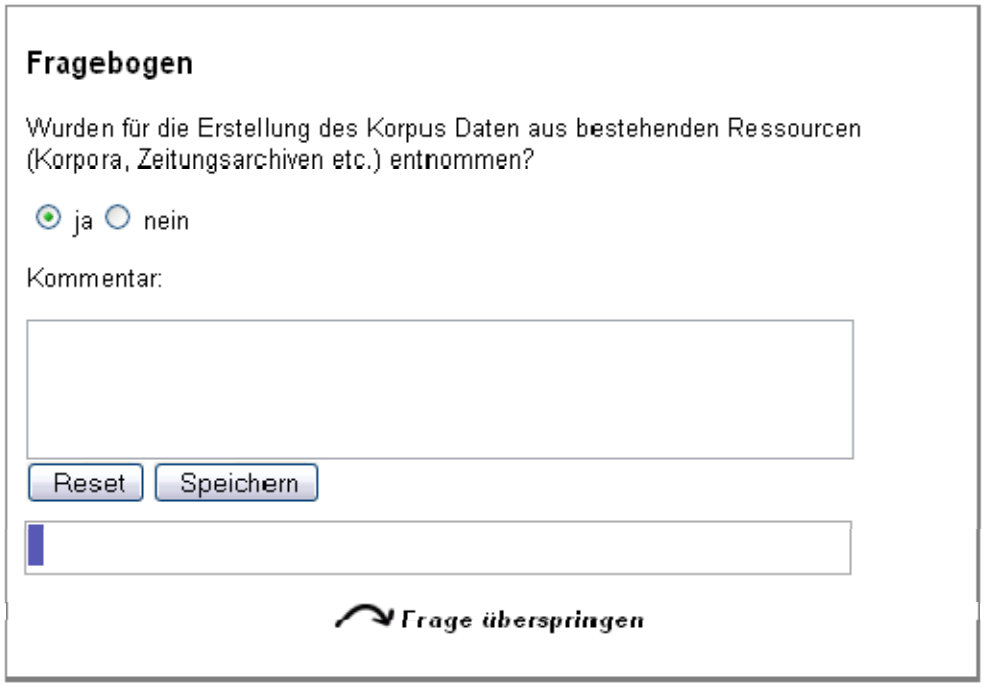


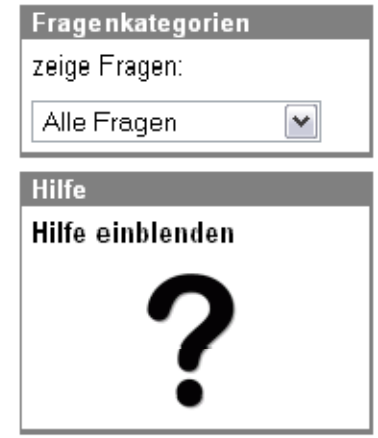
Implementation



Implementation

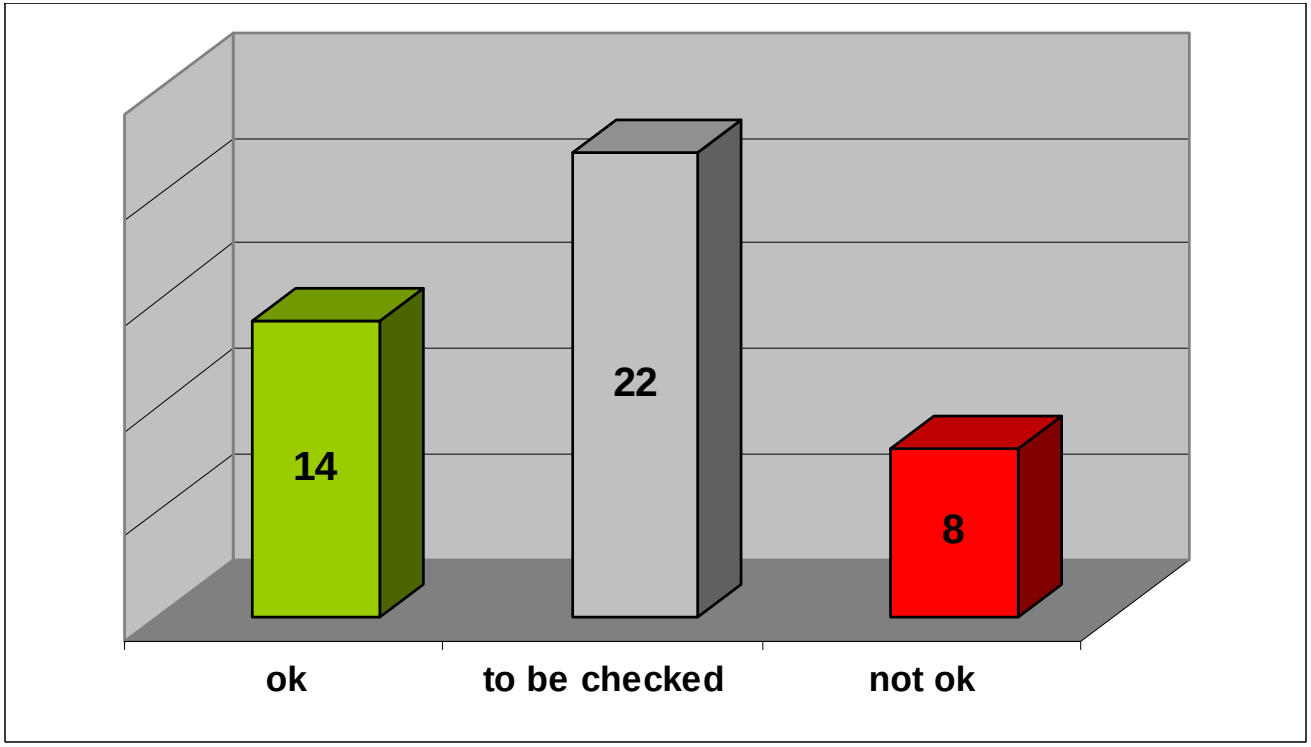
SFB 441: Linguistic Data Structures





Experience / Results

- 44 corpora within two weeks



Experience/Results

- **Example**: The Uppsala Corpus
  A large corpus of Russian:
  - Compiled at the University of Uppsala/Sweden
  - Morphologically annotated at SFB 441 in Tübingen/Germany

Q/A

Thank you very much! Questions?

SFB 441: Linguistic Data Structures

Corpus Masking

Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections

Georg Rehm, Andreas Witt,

Heike Zinsmeister, Johannes Dellert

Overview

- Introduction: The Problem
- Corpus Masking
- The CorpusMasker Tool
- Application Scenarios for Masked Corpora
- Summary and Concluding Remarks

Introduction: The Problem

- There are thousands of linguistic resources (e.g., annotated text collections). Few are freely available.
- *Very* rigid license agreements restrict the distribution of linguistic resources.
- Goal: to bypass rigid licensing restrictions in a legal way, so that we can freely distribute linguistic resources.

Introduction: The Problem

- A linguistic resource has two components:
  a) Source text collection (STC), i.e., one or more source texts (usually acquired by, e.g., web sites or publishing houses).
  b) One or more layers of annotation that refer to linguistic properties of the STC.
- In nearly all cases the STC is a copyrighted property:
  – The *copyright holder* decides if, and under which conditions, the linguistic resource – a crucial part of which is the STC – can be made available to the public or to the research community.

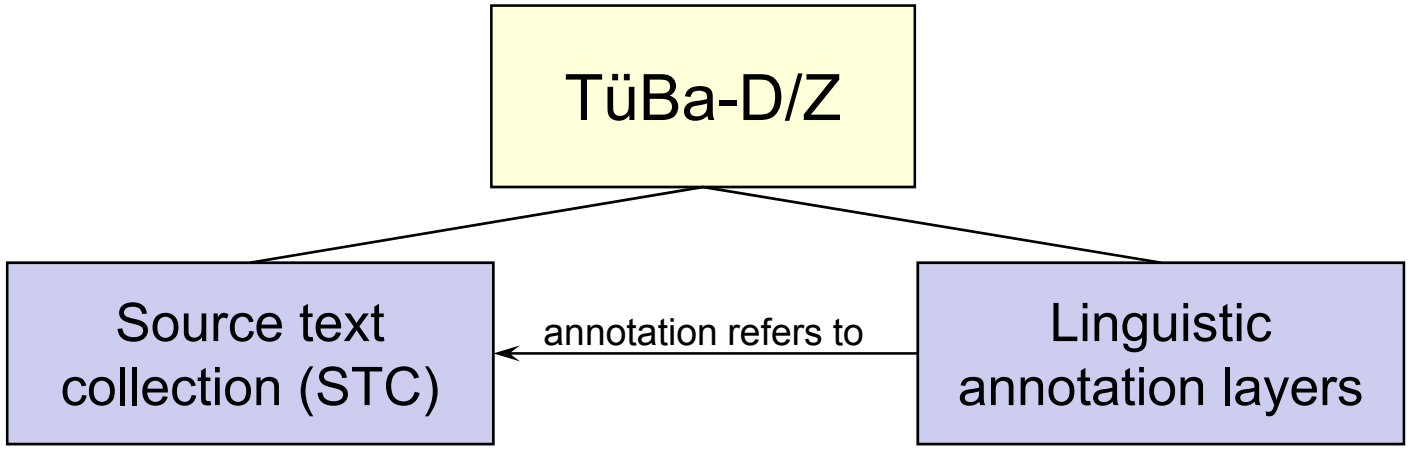
The Problem: Example

- TüBa-D/Z: Tübingen Treebank of Written German
  - Based on a commercially available newspaper CD ROM (archive of all *die tageszeitung* issues published since 1986)
  - Release 3: 27,000 sentences (470,000 tokens)
  - Available in several formats (such as, for example, XML)
  - Linguistic annotation:
    - Parts-of-speech, morphology, syntax, coreference
    - Carried out manually
    - Annotation is extremely time-consuming and expensive

The Problem: Example

- TüBa-D/Z available free of charge for academic purposes.
- How to obtain TüBa-D/Z:
  - Researcher has to sign a license agreement with the Linguistics Department at Tübingen University. This agreement states:
    - The Linguistics Department at Tübingen University is the copyright holder of the annotation.
    - The company contrapress media GmbH is the copyright holder of the STC as published on the CD ROM.
  - Researcher has to sign a statement that certifies that he or she or the institution the person works for has a valid license of the *die tageszeitung* CD ROM.
    - A copy of the CD ROM invoice has to be submitted as proof.

The Problem: Example



- License agreement: *rigid*

  The copyright holder (contrapress media GmbH) determines the terms of distribution of this part of the resource.

- License agreement: *not rigid*

  Tübingen University determines the terms of distribution of this part of the resource.

Overview

- Introduction: The Problem
- Corpus Masking
- The CorpusMasker Tool
- Application Scenarios for Masked Corpora
- Summary and Concluding Remarks

Corpus Masking

- Goal: to bypass rigid licensing restrictions legally (such as the ones that apply to TüBa-D/Z's STC), so that we can freely distribute linguistic resources.
- Approach:
  - We mask/obfuscate/garble the STC beyond recognition, so that the licensing restrictions that apply to the STC no longer hold.
  - We do *not* mask the linguistic annotation layer(s), of course.
- Consequence: the linguistic annotation itself can be made available for free.

Corpus Masking

Veruntreute die AWO Spendengeld? Staatsanwaltschaft muss AWO-Konten prüfen/Flossen 165.000 Mark Sammelgelder für Flutopfer in ein Altenheim in Danzig? Landesvorsitzende Ute Wedemeier: Ein Buchungsfehler. Im Januar hat die Arbeiterwohlfahrt Bremen ihren langjährigen Geschäftsführer Hans Taake fristlos entlassen, nun wird auch der Vorstand der Wohlfahrtsorganisation in den Fall hineingezogen. In einer anonymen Anzeige werden der Bremer Staatsanwaltschaft Details über dubiose finanzielle Transaktionen mitgeteilt. Verantwortlich, so das Schreiben einer Mitarbeiterin der AWO, sei die Landesvorsitzende Ute Wedemeier, die sich jetzt als “Sauberfrau” gebe, “wo doch alle Wissen, wie eng sie mit Taake zusammenhing”. Vorwurf […]

Source text collection (STC)

Access requires license for STC

Corpus Masking

Veruntreute die AWO Spendengeld? Staatsanwaltschaft muss AWO-Konten prüfen/Flossen 165.000 Mark Sammelgelder für Flutopfer in ein Altenheim in Danzig? Landesvorsitzende Ute Wedemeier: Ein Buchungsfehler. Im Januar hat die Arbeiterwohlfahrt Bremen ihren langjährigen Geschäftsführer Hans Taake fristlos entlassen, nun wird auch der Vorstand der Wohlfahrtsorganisation in den Fall hineingezogen. In einer anonymen Anzeige werden der Bremer Staatsanwaltschaft Details über dubiose finanzielle Transaktionen mitgeteilt. Verantwortlich, so das Schreiben einer Mitarbeiterin der AWO, sei die Landesvorsitzende Ute Wedemeier, die sich jetzt als “Sauberfrau” gebe, “wo doch alle Wissen, wie eng sie mit Taake zusammenhing”. Vorwurf […]

Annotation project creates linguistic resouce

Access still *requires license for STC*

Corpus Masking

Annotation project creates linguistic resouce

Access still *requires license for STC*

Corpus Masking

```
<s id="s1"><ntNode><tok><orth>…</orth><pos>VVFIN</pos>
</tok>…<ntNodeCat func="HD">VXFIN</ntNodeCat>
<orth>…</orth><pos func="HD">NN</pos>…
<s id="s2"><ntNode><tok><orth>…</orth><pos>VVFIN</pos>
</tok>…<ntNodeCat func="HD">VXFIN</ntNodeCat>
<orth>…</orth><pos func="HD">NN</pos>…
<s id="s3"><ntNode><tok><orth>…</orth><pos>VVFIN</pos>
</tok>…<ntNodeCat func="HD">VXFIN</ntNodeCat>
<orth>…</orth><pos func="HD">NN</pos>…
<s id="s4"><ntNode><tok><orth>…</orth><pos>…
```

STC is masked.

Access no longer requires license for STC.

Corpus Masking: Examples

- Complete removal of source text
  - The 19 blue houses → ø
- Keep information on word length and upper/lower case
  - The 19 blue houses → Xxx 99 xxxx xxxxxx
- Keep vowel vs. consonant distinction
  - The 19 blue houses → Xxa 99 xxaa xaaxax
- Keep morphological information (affixes)
  - The 19 blue houses → Xxa 99 xxaa xaaxes
- Map words onto randomly generated “words”
  - The 19 blue houses → Wko 99 gboe paexes
- Don’t mask closed word classes (e.g., determiners)
  - The 19 blue houses → The 99 gboe paexes

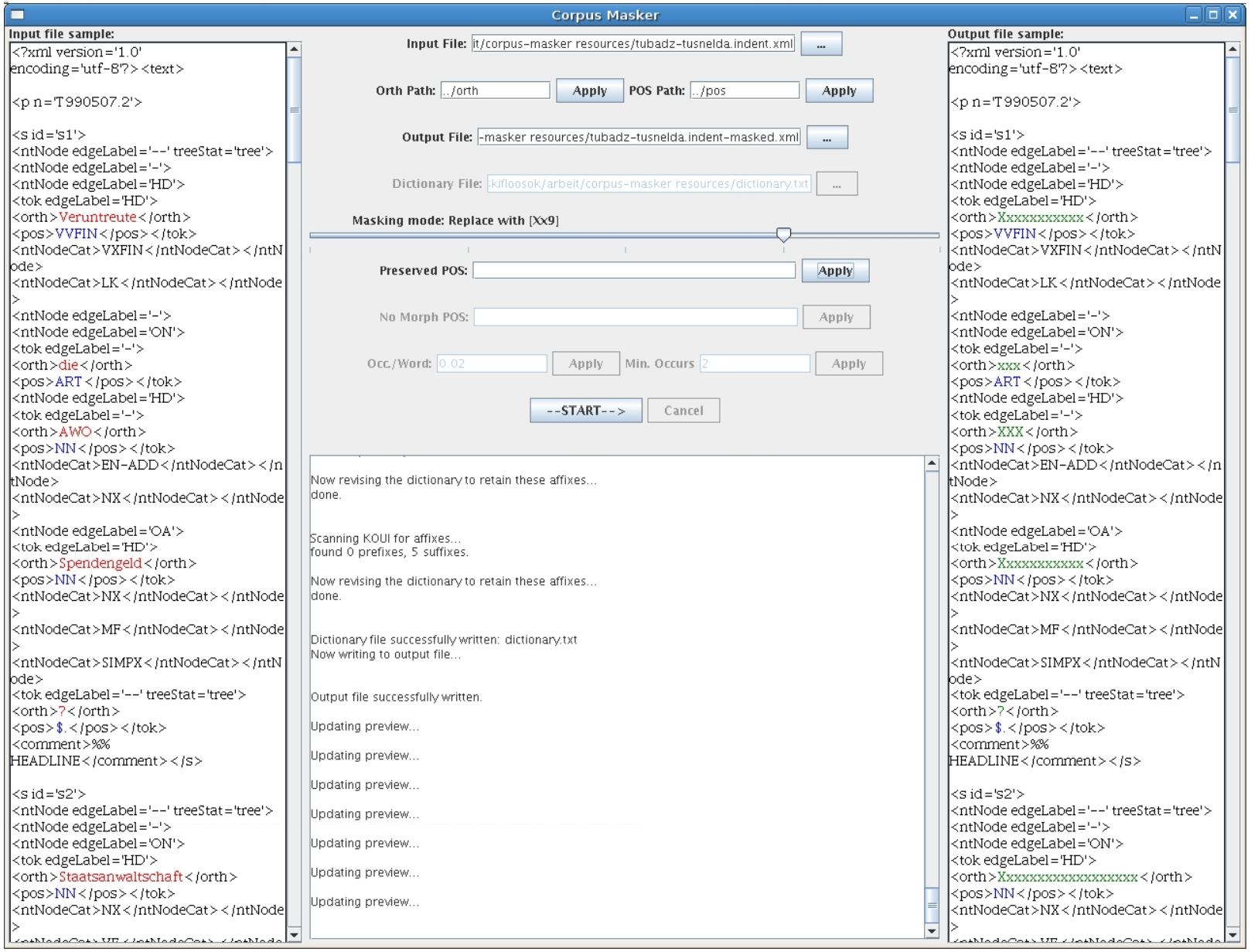
Overview

- Introduction: The Problem
- Corpus Masking
- The CorpusMasker Tool
- Application Scenarios for Masked Corpora
- Summary and Concluding Remarks

The CorpusMasker Tool

- CorpusMasker: tool for masking corpora
- Java
- Library, GUI, command-line tool
- Input: XML file that contains the linguistic resource
- Output: masked version of the linguistic resource
- Tokens to be masked can be specified using XPath expressions (e.g., in TüBa-D/Z, the **`orth`** element).

The CorpusMasker Tool: GUI



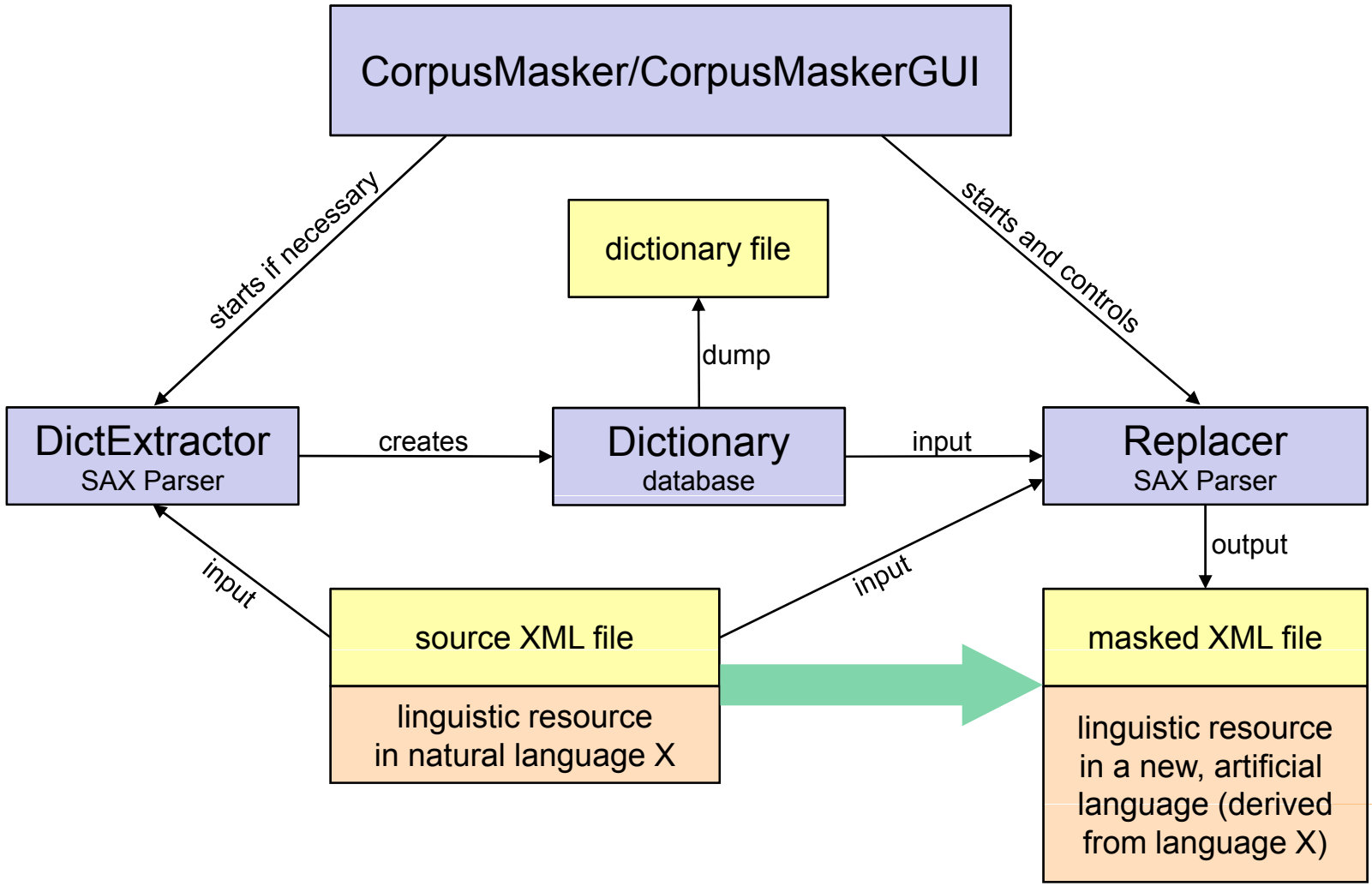
The CorpusMasker Tool: Features

- Dictionary-based corpus masking:
  - CorpusMasker collects all word forms.
  - Every word is mapped onto a randomly generated string.
  - Every word form is replaced by that string.
- Word length can be retained.
- Positions of vowels and consonants in a word can be retained.
- Sentence-initial upper casing can be retained
  ("dort" → "kulp" – "Dort" → "Kulp").
- Character classes are represented in configuration files
  (to enable processing of non-latin character sets, e.g., cyrillic).
- Affix detection and analysis: roots are masked, affixes are kept intact (affixes alone are insufficient to reconstruct or even to interpret the source text)

The CorpusMasker Tool: Affixes

- Affix analysis is based on morphology induction.
- For each POS category and token, all possible prefixes and suffixes are counted and stored in a hash table.
- Selection of affixes using statistical parameters (e.g., how often an affix occurs in a word class with regard to the total number of words in this class)
  Example: Veruntreute → Verildniite
- Approach yields good results for German corpora
- For other languages, parameters might have to be adjusted based on the richness of their morphology.
- Alternative: pre-defined affix catalogues

The CorpusMasker Tool: Architecture



The CorpusMasker Tool: Example

Veruntreute die AWO Spendengeld? Staatsanwaltschaft muss AWO-Konten prüfen/Flossen 165.000 Mark Sammelgelder für Flutopfer in ein Altenheim in Danzig? Landesvorsitzende Ute Wedemeier: Ein Buchungsfehler. Im Januar hat die Arbeiterwohlfahrt Bremen ihren langjährigen Geschäftsführer Hans Taake fristlos entlassen, nun wird auch der Vorstand der Wohlfahrtsorganisation in den Fall hineingezogen. In einer anonymen Anzeige werden der Bremer Staatsanwaltschaft Details über dubiose finanzielle Transaktionen mitgeteilt. Verantwortlich, so das Schreiben einer Mitarbeiterin der AWO, sei die Landesvorsitzende Ute Wedemeier, die sich jetzt als “Sauberfrau” gebe, “wo doch alle wissen, wie eng sie mit Taake zusammenhing”. […]

Original text

Verimpmiite die AFU Sbalkartamb? Steodveljonkwwyijt muss AJU-Bumten pmafen/Fmufjen 153.000 Meng Silnimpumger sar Fmedidwer in ein Angilzuun in Diryug? Lergafwunwaghalge Ake Wuporaaer: Ein Bevyirkfsuymer. Im Vemaur hat die Anguodinseymjeymt Bnalen ihren lirbfezligen Gejvzisgfwuzler Hels Tuipe flujdlos endmessen, nan wird auch der Venwdumb der Wayrjezndwunpomuvikuon in den Fuml hulaulpahogen. In einer alelhlen Amyaube werden der Bnuner Steodveljonkwwyijt Dipiams über depuave famemhuomle Tlorjuppoenen ruktagielt. Verulkjiltlich, so das Schnuuten einer Megolguedolan der AFU, sei die Lergafwunwaghalge Ake Wuporaaer, die sich jibyt als “Suipamwnoo” goge, “wo dach alle wejjen, wie erg sie mit Tuipe zawelrolyolt”. [...]

Masked text

Overview

- Introduction: The Problem
- Corpus Masking
- The CorpusMasker Tool
- Application Scenarios for Masked Corpora
- Summary and Concluding Remarks

Application Scenario I

- What are masked linguistic resources good for?
- Web-based corpus delivery platform:
  - Masked linguistic resources can be made freely available.
  - Potential audience of a corpus can be enlarged substantially.
  - Researchers can examine the corpus' contents without ordering, e.g., the *die tageszeitung* CD ROM first.
  - Enhance the security of copyrighted data:
    - Only masked versions of corpora can be viewed or downloaded by users (e.g., activate CorpusMasker before every single download).

Application Scenario II (Future Work)

- Masked linguistic resources for educational purposes:
  - The masked resource contains, for the most part, random strings and associated POS tags. This fact makes it a valuable resource in the context of teaching computational linguistics and grammar.
  - If the artificial language has a known syntax, rudimentary morphology, but almost meaningless lexical entries, students might be able to concentrate better on developing grammar rules.
  - This idea of “blanking out meaning and semantics” is compatible with Chomsky’s notion of language as processing a set of symbols.

Application Scenario III (Future Work)

- Masked corpora and evaluation of natural language processing (NLP) software:
  - Most NLP tools (taggers, parsers) use statistical *n*-gram language models.
  - These NLP tools can be trained on annotated data.
  - With a masked corpus it is possible to measure the influence syntactic annotations have concerning precision and recall:
    - We can compare the performance of the tool with regard to original, as well as slightly and fully masked corpora.
  - Such an experiment could result in substantial arguments in favour of or against the use of treebanks for training NLP tools.

Overview

- Introduction: The Problem
- Corpus Masking
- The CorpusMasker Tool
- Application Scenarios for Masked Corpora
- Summary and Concluding Remarks

Summary and Concluding Remarks

- Approach for masking linguistic resources (e.g., corpora)
- CorpusMasker:
  - Available very soon at http://www.sfb441.uni-tuebingen.de/c2/
  - Affix analyser will be improved, e.g., by enabling blacklists and whitelists of prefixes and suffixes for specific POS classes.
- Future work:
  - Masked linguistic resources for educational purposes.
  - Masked linguistic resources and evaluation of NLP tools.

An Alternative Approach …

- Distribute the linguistic annotation only (standoff format), i.e., *without* the primary text.
- Provide tools, such as Perl scripts, so that a person who has a license for the *tageszeitung* CD ROM can use
  - the source text collection (that he or she already owns), and
  - the linguistic annotation (provided by Tübingen University), and
  - the tools (provided by Tübingen University as well)

  in order to "build" TüBa-D/Z him- or herself.
- Sounds easy and straightforward *but* is extremely difficult:
  - Segmentation of the STC, tokenisation, sentence boundary detection etc. would have to be integrated into the tools …

Q/A

Thank you very much! Questions?

People

Christian Chiarcos	Potsdam University	**chiarcos@uni-potsdam.de**
Johannes Dellert	Tübingen University	**jdellert@sfs.uni-tuebingen.de**
Erhard Hinrichs	Tübingen University	**eh@sfs.uni-tuebingen.de**
Timm Lehmberg	Hamburg University	**timm.lehmberg@uni-hamburg.de**
Georg Rehm	Tübingen University	**georg.rehm@uni-tuebingen.de**
Andreas Witt	Tübingen University	**andreas.witt@uni-tuebingen.de**
Felix Zimmermann	Hannover University	**felix.zimmermann@stud.uni-hannover.de**
Heike Zinsmeister	Tübingen University	**heike.zinsmeister@uni-tuebingen.de**

http://www.sfb441.uni-tuebingen.de/c2/