

INEL

Eine Infrastruktur zur Dokumentation indigener nordeurasischer Sprachen

Beata Wagner-Nagy, Hanna Hedeland, Timm Lehmberg (Universität Hamburg)
Michael Rießler (Universität Freiburg)

INEL

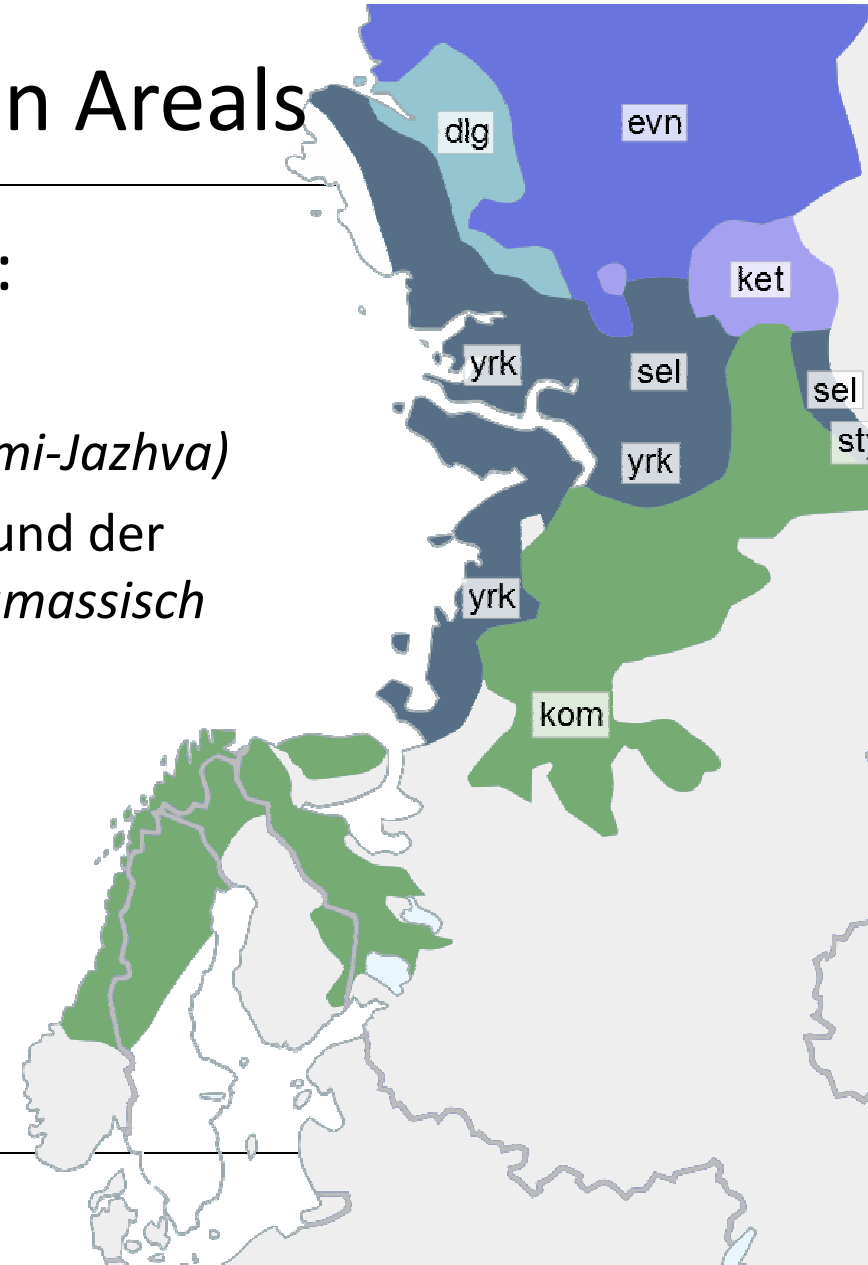
Grammatical Descriptions, Corpora, and
Language Technology
for Indigenous Northern Eurasian Languages

Die Sprachen des nordeurasischen Areals

- zwei Sprachfamilien + isolierte Sprache:
 - *Uralisch*
 - *Komi* (2 Varietäten: *Komi-Permjakisch* und *Komi-Jazhva*)
 - *Nenzisch* (2 Varietäten: Dialekten der Tajmyr- und der Kanin-Halbinsel), *Selkupisch* (3 Varietäten), *Kamassisch*
 - *Altaisch*
 - *Dolganisch*, *Evenkisch* (Norddialekt),
Sibirisch-Tatarisch
 - *Ket*

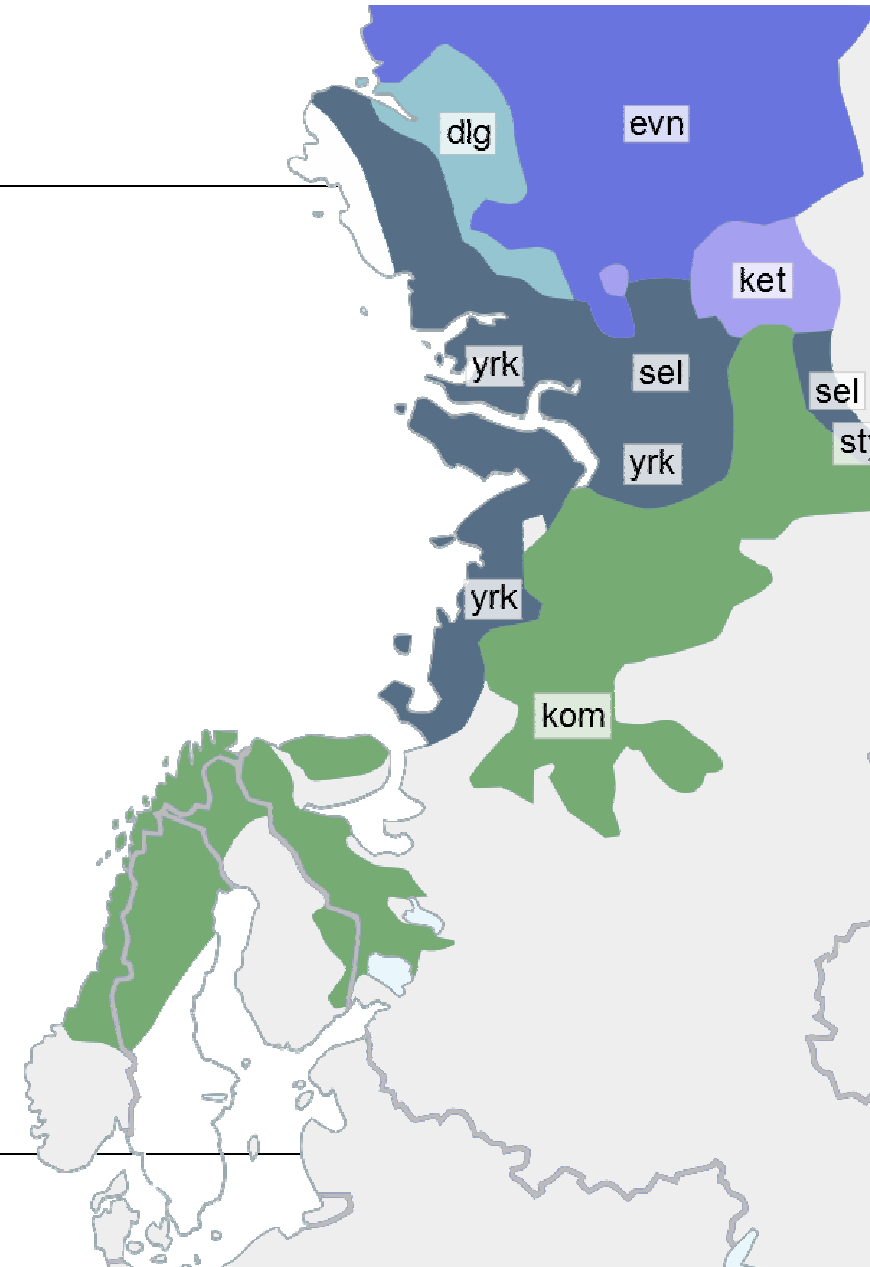
Die Sprachen des nordeurasischen Areals

- zwei Sprachfamilien + isolierte Sprache:
 - *Uralisch*
 - *Komi* (2 Varietäten: *Komi-Permjakisch* und *Komi-Jazhva*)
 - *Nenzisch* (2 Varietäten: Dialekten der Tajmyr- und der Kanin-Halbinsel), *Selkupisch* (3 Varietäten), *Kamassisch*
 - *Altaisch*
 - *Dolganisch*, *Evenkisch* (Norddialekt), *Sibirisch-Tatarisch*
 - *Ket*



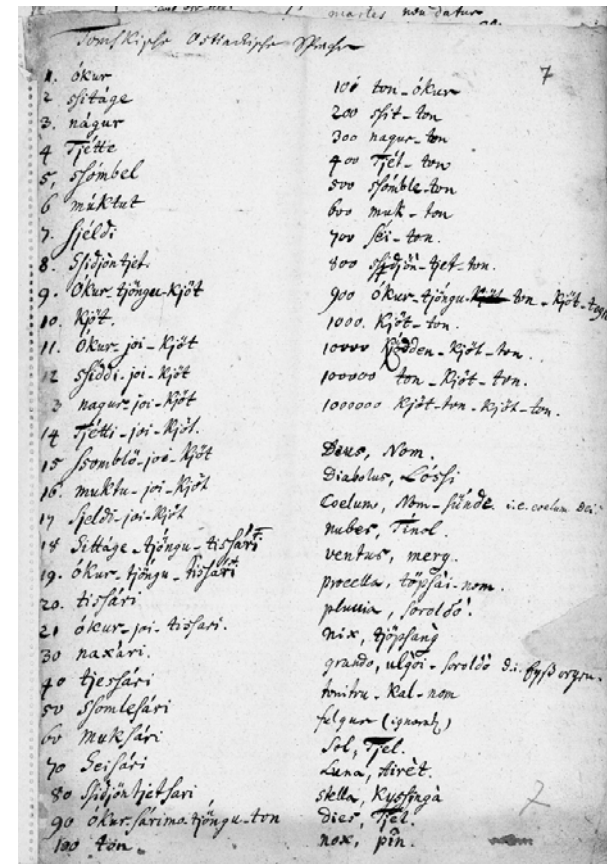
Die Sprachen im Projektfokus

xas: Kamassisch †
ket: Ketisch
sel: Selkupisch
kom: Komi
yrk: Tundra-Nenzisch
dlg: Dolganisch
evn: Nord-Ewenkisch
sty: Sibirisch-Tatarisch



Stand der linguistischen Erfassung

- Wortlisten vom 18. Jh.
- Textsammlungen und sprachwissenschaftliche Beschreibungen einiger Sprachen vom 19. Jh
- Textsammlungen, Grammatiken von 20. Jh

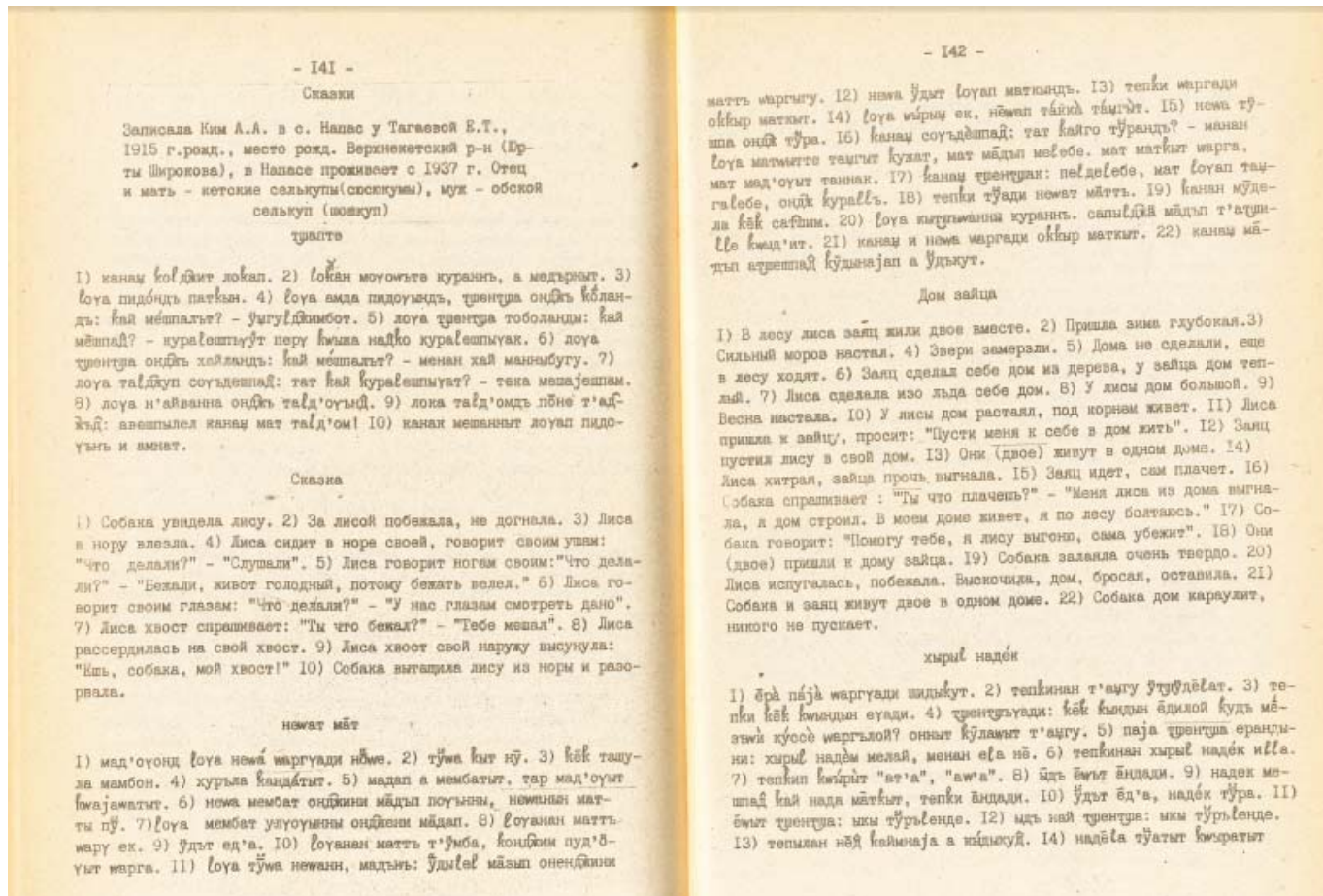


Materialsammlungen

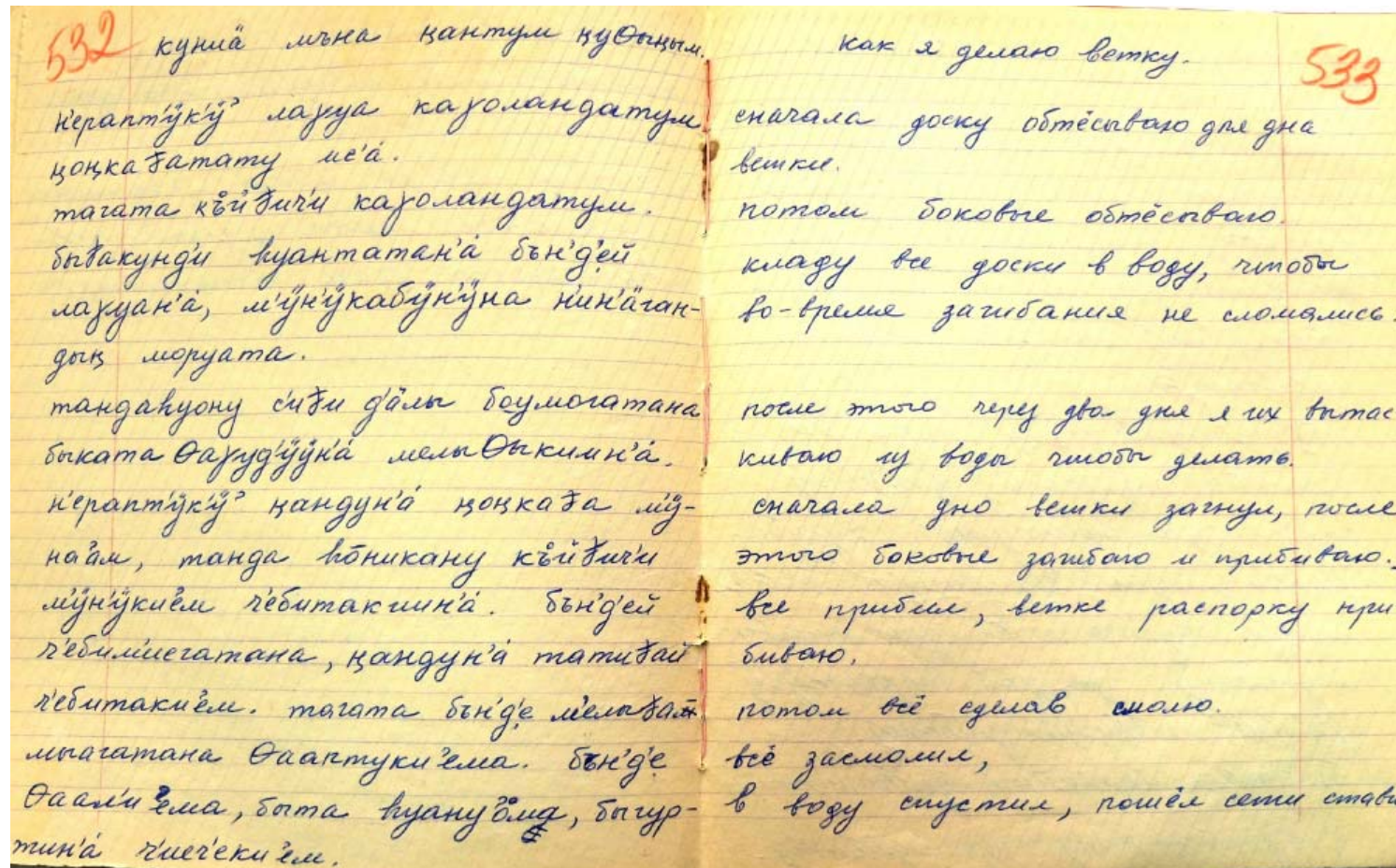
- große Mengen an Sprachdaten in den Archiven und in Veröffentlichungen
 - Daten mit oder ohne Aufnahmen
- Archive in Tomsk, St.Petersburg, Moskau, Hamburg, Tartu



Rohdaten



Rohdaten



Rohdaten

кол'дасе 263 (Тюмур - в
Узбеки
Колташев)

№5) гудин ал'дига

Фриёна Федоровна Тоболэ жини
1911г. рондони, рондиси Тибетакс,
жини с'квалити, 43 лет. Топур и шибет в Тюмуре
Колба малат т'анга -

1. 'евак'ва баркык'ва охыр ~~ед~~ Эд'овин
жини били в о'сноб с'рение

№6) ит'а кал'дига ит'ин ал'дига
сит с' марерто. мар с' ситом
изг

ка поккурку в'анк'догагой.
сери гавыр поили.

3. 'адмак'вий ваттогын.
изг по с'оме.

4. 'монт' лёд'агой. кв'еба к'а
до о'сера доили в
анд' монт' о'ид'агой
бин



Kuzmina Archiv

vorhandene Grammatiken

- von unterschiedlichem Niveau
- meistens russischsprachig

Урок 1.

Тема: Личные формы глаголов неопределенного времени единственного числа лица.

Рассмотрим предложения.

<i>Мань тодж'.</i>	Я пришел (пришла).
<i>Пыдар тон.</i>	Ты пришел (пришла).
<i>Пыда то.</i>	Он (она) пришел (пришла).

Это простые нераспространенные предложения, состоящие из подлежащих, выраженных местоимениями *мань* «я», *пыдар* «ты», *пыда* «он (она)», и сказуемых, выраженных глаголом *то(сь)*¹ «прийти» в личной форме (1, 2, 3-го лица ед. числа).

Как видно из примеров, глагол *тось* изменяет свою форму в зависимости от того, кто является действующим лицом: если действующее лицо *мань* «я», мы прибавляем окончание *-дж'*, если *пыдар* «ты» — *-н*; если действующее лицо *пыда* «он (она)», окончания не прибавляем.

Составим такие же предложения с глаголами *илесь* «жить», *манзарась* «работать», *тоходанась* «учиться».

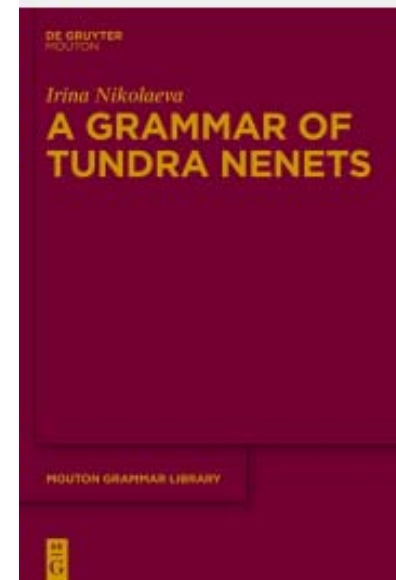
Образец: <i>Мань иледж'.</i>	Я живу.
<i>Пыдар илен.</i>	Ты живешь.
<i>Пыда иле.</i>	Он (она) живет.

vorhandene Grammatiken

- angeblich korpusbasiert, aber die Beispiele sind im Korpus nicht identifizierbar

Sentence (30) was spontaneously produced when the speaker and the addressee looked at several moving cars. The speaker wanted to refer to one of these cars and used the word 'car' with the 2nd person possessive affix, although the car did not belong to the addressee. The reason for using the possessive suffix is that the speaker intended to call the attention of the addressee to the car. In other words, the car is 'yours' because 'I am talking to you about it'.

(30) *siər-^oq, taki^o mašina-r^o xəya*
look-IMP.2SG that car-2SG go
'Look, that car of yours left.'



vorhandene Online-Ressourcen

Text ID: NENETS_01_23
Language: Tundra Nenets
Genre: Text

0:12 2:30

Speaker: Olga Taybarey

Nenets, IGT, English

← Prev Same → Next

Search this transcript

Search Reset

▶ 0:00	S'em ^o ya ^o nan ^o mən ^o əb yəŋk ^o n'am ^h . family.LOC.POSS.1SG PRON.1SG one excessive.PTC.IMPF.1SG In my family I'm the eleventh.
▶ 0:06	N'a ^o r n'aw ^o woy ^o na ^o na xaec ^o . three friend.POSS.1SG war.LOC die.PST.SUBJ.3PL Three brothers of mine died in the war.
▶ 0:10	Woy ^o na ^o na xaec ^o da, (bolyše) n'ic ^o tūq. war.LOC die.PST.SUBJ.3PL EMPH () not.PST.SUBJ.3PL come.CONNEG And they died in the war. they didn't come (any more).
▶ 0:14	(Один), əroy ^o b ^o ratuw ^o ran'enoy ^o [ŋe ^o], ran'enoy ^o ŋe ^o tūŋas ^o . () one.MOD brother.POSS.1SG wounded.ESS wounded.ESS come.FREQ.PST.SUBJ.3SG (One), one brother of mine got wounded, he was coming as wounded.

Problemfelder: Bedarfslage

- die existierende Grammatiken sind meistens nicht korpusbasiert
- die Transkriptionen sind sehr unterschiedlich
- Sprachaufnahmen und Texte sind nicht verknüpft
- Texte sind meistens nicht glossiert, nicht annotiert

Anforderungen

- **Auffindbarkeit** von Ressourcen und Dokumenten
→ ?
- **Referenzierbarkeit**
→ ?
- **Zugänglichkeit** existierender sprachlichen Ressourcen
→ ?
- **Durchsuchbarkeit und Benutzbarkeit** der Ressourcen
→ ?

Anforderungen

- **Auffindbarkeit** von Ressourcen und Dokumenten
 - Einsatz generischer Metadatenformate
- **Referenzierbarkeit**
 - Vergabe von Persistent Identifiern
- **Zugänglichkeit** existierender sprachlichen Ressourcen
 - Authentifizierungsstruktur und offene Lizenzmodelle
- **Durchsuchbarkeit und Benutzbarkeit** der Ressourcen
 - generische Datenformate und intuitive GUIs

Vor 18 Jahren ...

HIAT-Dos

F1: 1 Pos: 1 Transkriptname:DEMOTXT0 HIAT-DOS 2.2

```
Bezugszeile <
Kontrollzeile =
```

Der Kommentar zum ersten Sprecherblock steht in diesen Zeilen daneben.

Kommentare sind mit der Taste <F2> erreichbar und zu verlassen.

```
> Dies ist die Zeile für "Intonation" des 1. Sprechers.
[15] Hier steht die "verbale Kommunikation" des 1. Sprechers.
N1 Dies ist die erste der 3 NVK-Zeilen des 1. Sprechers.
N2 Außer zur Fixierung der "nonverbalen Kommunikation"
N3 ist in NVK-Zeilen Platz für Kommentare zum Transkript.
>
[25] Das Beenden des Programms wird mit der < Esc >-Taste
N1 eingeleitet - anschließend < Q > (für Quitt) benutzen.
N2 Hilfetexte sind mit der < F1 >-Taste zu haben.
N3 Sie enthalten weitere Hinweise zur Programm-Benutzung.
>
[34] Zur rechtsliegenden, nachfolgenden Fläche --->>>
N1 kann mit der < F6 >-Taste gesprungen werden.
N2 <<<--- Zur vorhergehenden Fläche (sofern es sie gibt)
N3 kann mit der < F5 >-Taste gesprungen werden.
>
[44] Für schnelle Bewegungen sind Tastenkombinationen der
N1 < Strg >-Taste mit den üblichen Cursor-Tasten verfügbar.
N2 Der Wechsel in einen anderen Sprecherblock erfolgt mit
N3 den Tasten < Bild rauf > bzw. < Bild runter > .
```

Vor 18 Jahren ...

Shoebox

HIAT-Dos

F1: 1 Pos: 1 Transkriptn

Bezugszeile < |
Kontrollzeile = |

Der Kommentar
zum ersten
Sprecherblock
steht in diesen
Zeilen daneben.

> Dies ist
15 Hier steh
N1 Dies ist
N2 Außer zur
N3 ist in NV

>
25 Das Beend
N1 eingeleit
N2 Hilfetext
N3 Sie entha

>
30 Zu
N1 ka
N2 <<<--- Zu
N3 ka

>
34 Für schne
N1 < Strg >-
N2 Der Wechsel in einen anderen Sprecherblock erfolgt mit
N3 den Tasten < Bild rauf > bzw. < Bild runter > .



Vor 18 Jahren ...

Shoebox

HIAT-Dos

Fl: 1 Pos: 1 Tra

Bezugszeile
Kontrollzeile

Der Kommentar
zum ersten
Sprecherblock
steht in diesen
Zeilen daneben.

Kommentare sind
mit der Taste
<F2>
erreichbar und
zu verlassen.

- > D
- 15 H
- N1 D
- N2 A
- N3 L
- >
- 25 D
- N1 e
- N2 H
- N3 S
- >
- SU
- N1 <
- N2 <
- N3 >
- >
- S4 F
- N1 < Strg >-Taste mit den üblichen Cursor-Tasten verfügb
- N2 Der Wechsel in einen anderen Sprecherblock erfolgt m
- N3 den Tasten < Bild rauf > bzw. < Bild runter > .

Lexicon.db:1

Lexicon.db:2

vs		bini
vg	Vx	kin
ve	vsd	n
vd	vps	wife
vg	vge	isteri
vx	vgn	Counterpart
vx	vif	laki
vx	viv	husband
vx	vle	suami
vx	vln	Voc Counterpart
vx	vlf	paitua
vx	vve	

Text.db

Word Perfect

WordPerfect 11 [C:\Dokumente und Einstellungen\Graf\Meine Dateien\Handschriftenforschung im Internet.wpd]

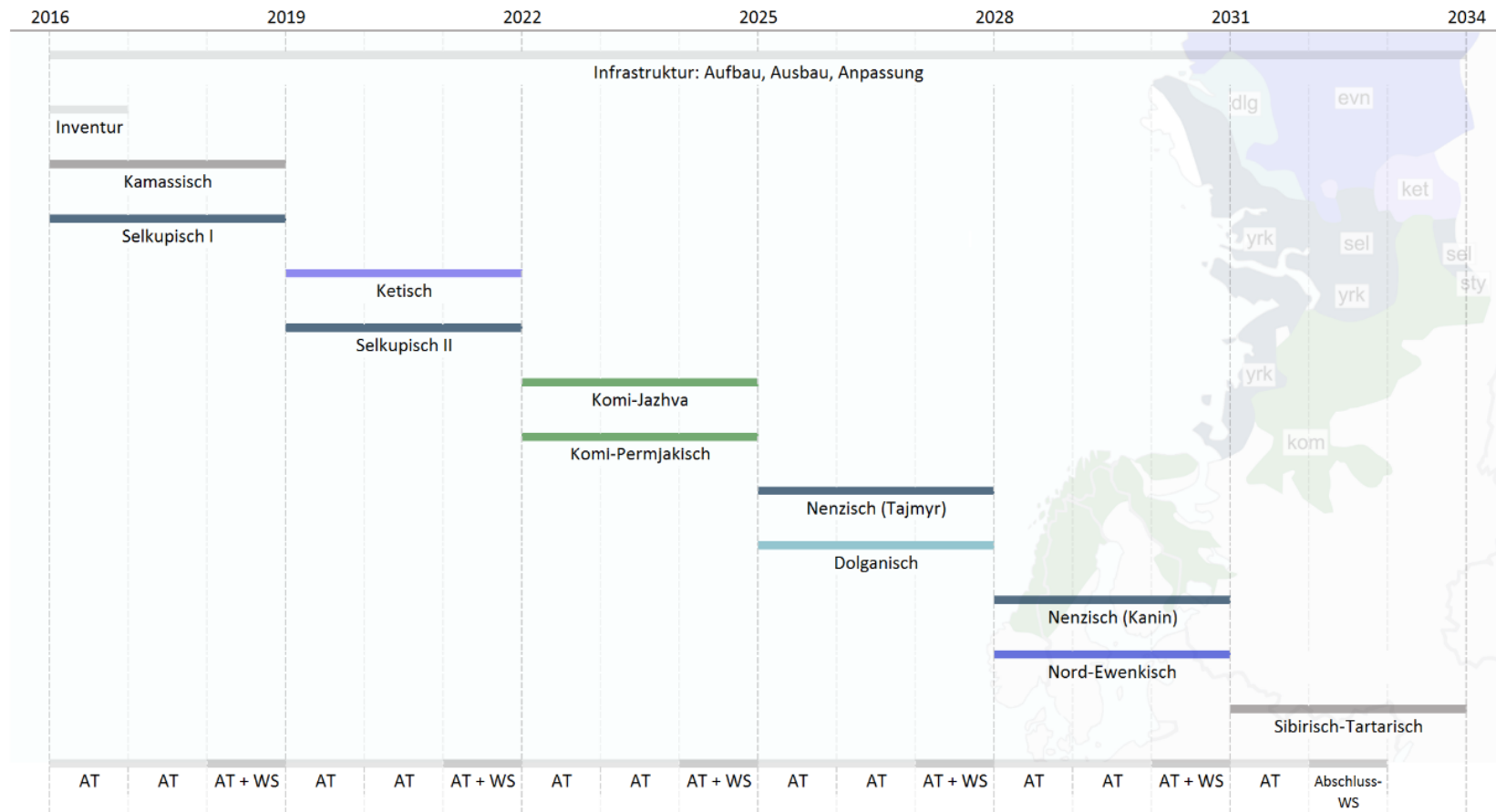
Handschriftenforschung im Internet - ein Blick zurück

Von Klaus Graf

Als ich im April 1997 meine persönliche Homepage "Stadt Adelt Region" ins Netz stellte (damals noch auf dem "Hilfpage" der Universität Erlangen-Lands), war von Anfang an eine Informationsseite "Handschriftenforschung im Internet" enthalten. Im Herbst 1997 fand eine Einführung in die Abfragepraxis der später abgeschalteten DEL-Datenbank "Handschriften des Mittelalters", die ich als größtenteils Recherche-Werkzeug empfand. Einleitetend machte ich auf die beachtlichen Informationsangebote der Wiener Handschriftenkommission aufmerksam: "Österreich legt die Führung: Wer sich für handschriftliche, vornehmlich mittelalterliche Handschriften interessiert, ist bei der Suche nach deutschsprachigen Angeboten gar beraten, nicht zuletzt nach Wien zu wenden".

Im Juni 1997 gelang mir ein wichtiger Fund am Bildschirm. Dank der Erschließung der Materialien des "Handschriftenarchivs" der Berliner Akademie im Registerform wurde ich auf Einträge zur Handschrift 64 der Hochbibliothek Sigmaringen aufmerksam, die ich als Zweitveröffentlichung der sogenannten Faksimile von Schlettstadt, nur von Einch-Riesenschmied 1874 ediertes lateinisches Exemplar umgab, erkannte. Im November 1997 konnte ich verwenden, wurde Autopsie durch Frau Henner gegeben hatte. Die Handschrift war auch nur ein weiteres Autograph Wilhelm Werners von Zimmern, sondern enthält auch weitere unbekanntere Geschichten über mit dem Namen Faksimile verbundenen Textsammlung. Im Rahmen einer Hausarbeit später einer Magisterarbeit hat der damalige Freiburger Student Stefan Geoghegan die Handschriften eingehend ausgewertet. Er konstatierte, dass die Geschichte nicht von Rudolf von Schlettstadt, sondern von dem nur 1300 Jahre später am 10. Januar 1874 in Karlsruhe veröffentlichten, von dessen Forschungsergebnissen ist (seit 2008) nur der 2003 gedruckte Aufsatz von Henner online. Er ist bezeichnend, dass Henner zwar nicht als Faksimile, aber darauf verzichtet, die Seite "Handschriftenforschung im Internet" zu verlinken. Ob die Freiburger Magisterarbeit von Georges, die nur ein Skizze ausschnitt zugänglich ist, irgendein in seiner beachtlichen Informationsangebote der Wiener Handschriftenkommission aufweist.

Arbeitsprogramm



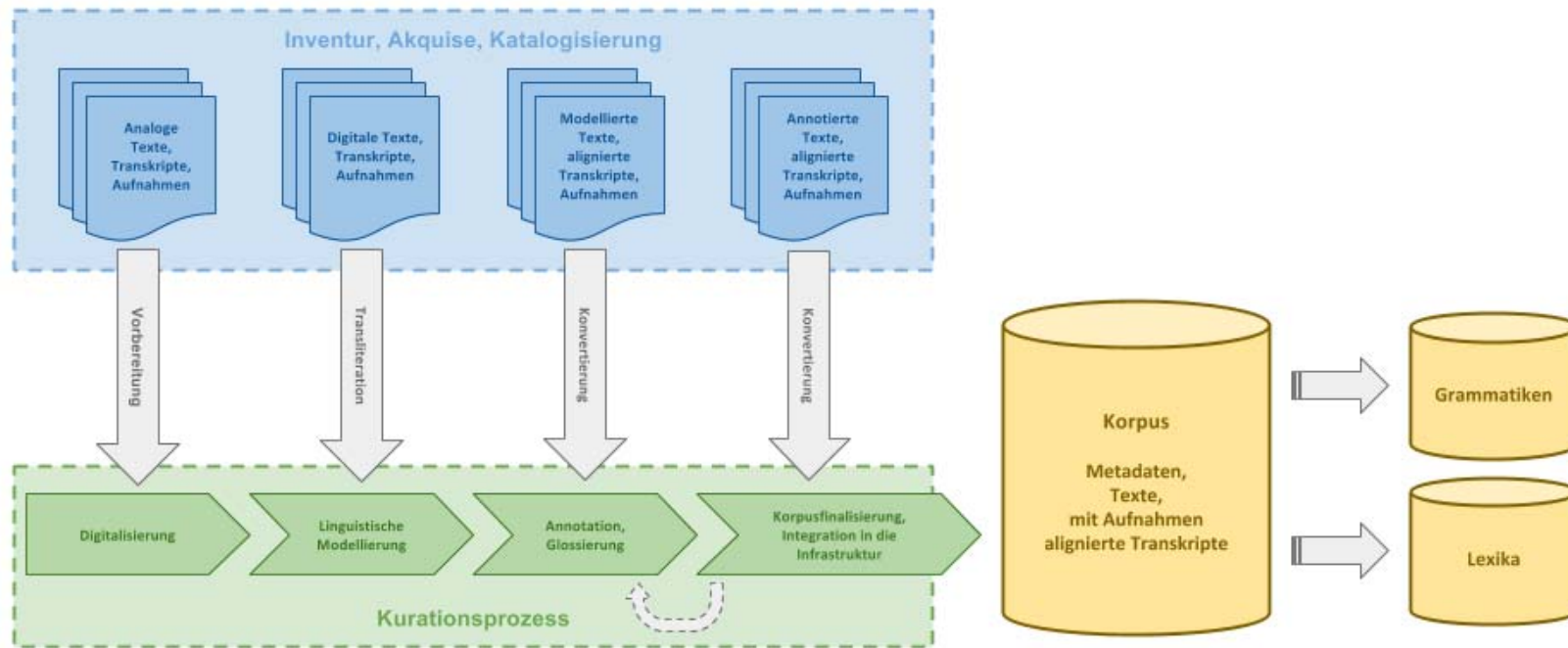
Arbeitsprogramm

	1. Jahr	2. Jahr	3. Jahr
Arbeitspaket 1 Korpusaufbau			
Modul 1 Inventur			
Akquise/Erhebung			
Modul 2 Digitalisierung			
Modul 3 Modellierung			
Modul 4 Annotation			
Modul 5 Korpusfinalisierung			
Arbeitspaket 2 Infrastruktur und Best Practices			
Arbeitspaket 3 Evaluation und Dissemination			
Workshops			
Abschlussbericht			

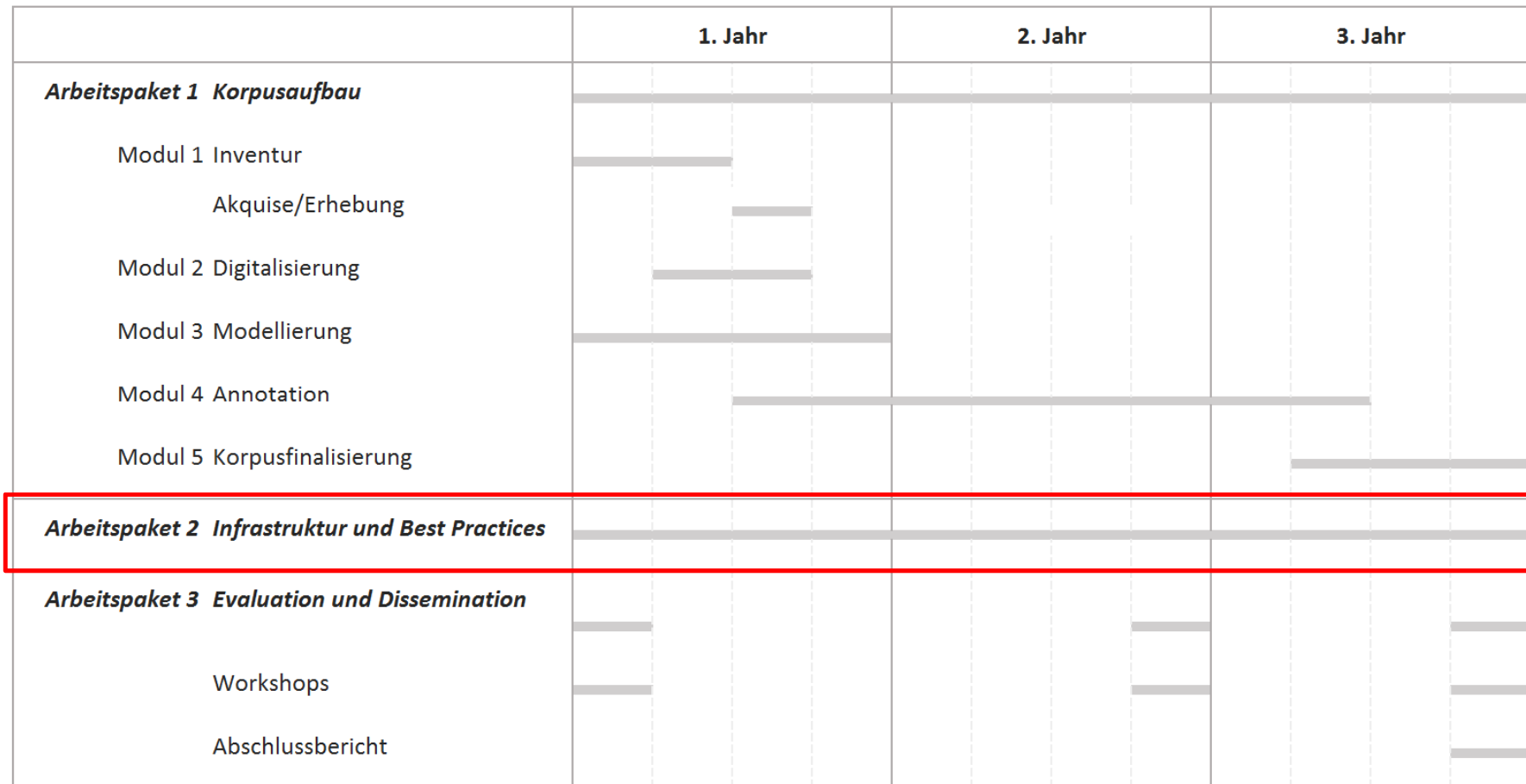
Arbeitsprogramm – AP1 Korpusaufbau

	1. Jahr	2. Jahr	3. Jahr
Arbeitspaket 1 Korpusaufbau	[Redacted]		
Modul 1 Inventur	[Redacted]		
Akquise/Erhebung	[Redacted]		
Modul 2 Digitalisierung	[Redacted]		
Modul 3 Modellierung	[Redacted]		
Modul 4 Annotation	[Redacted]		
Modul 5 Korpusfinalisierung	[Redacted]		
Arbeitspaket 2 Infrastruktur und Best Practices	[Redacted]		
Arbeitspaket 3 Evaluation und Dissemination	[Redacted]		
Workshops	[Redacted]		
Abschlussbericht	[Redacted]		

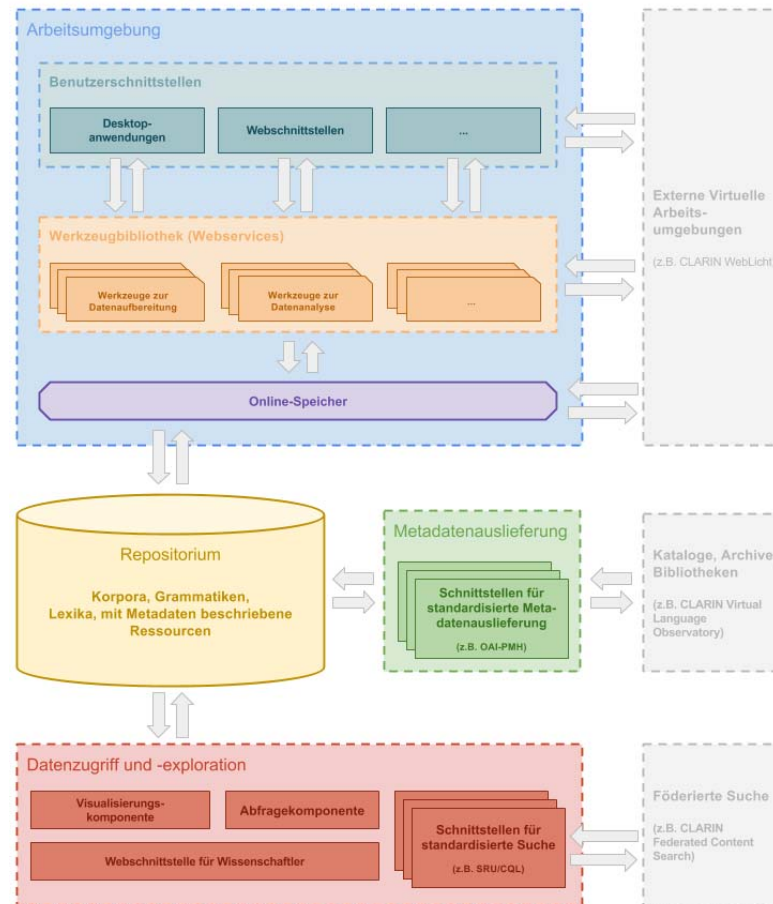
Arbeitsprogramm – AP1 Korpusaufbau



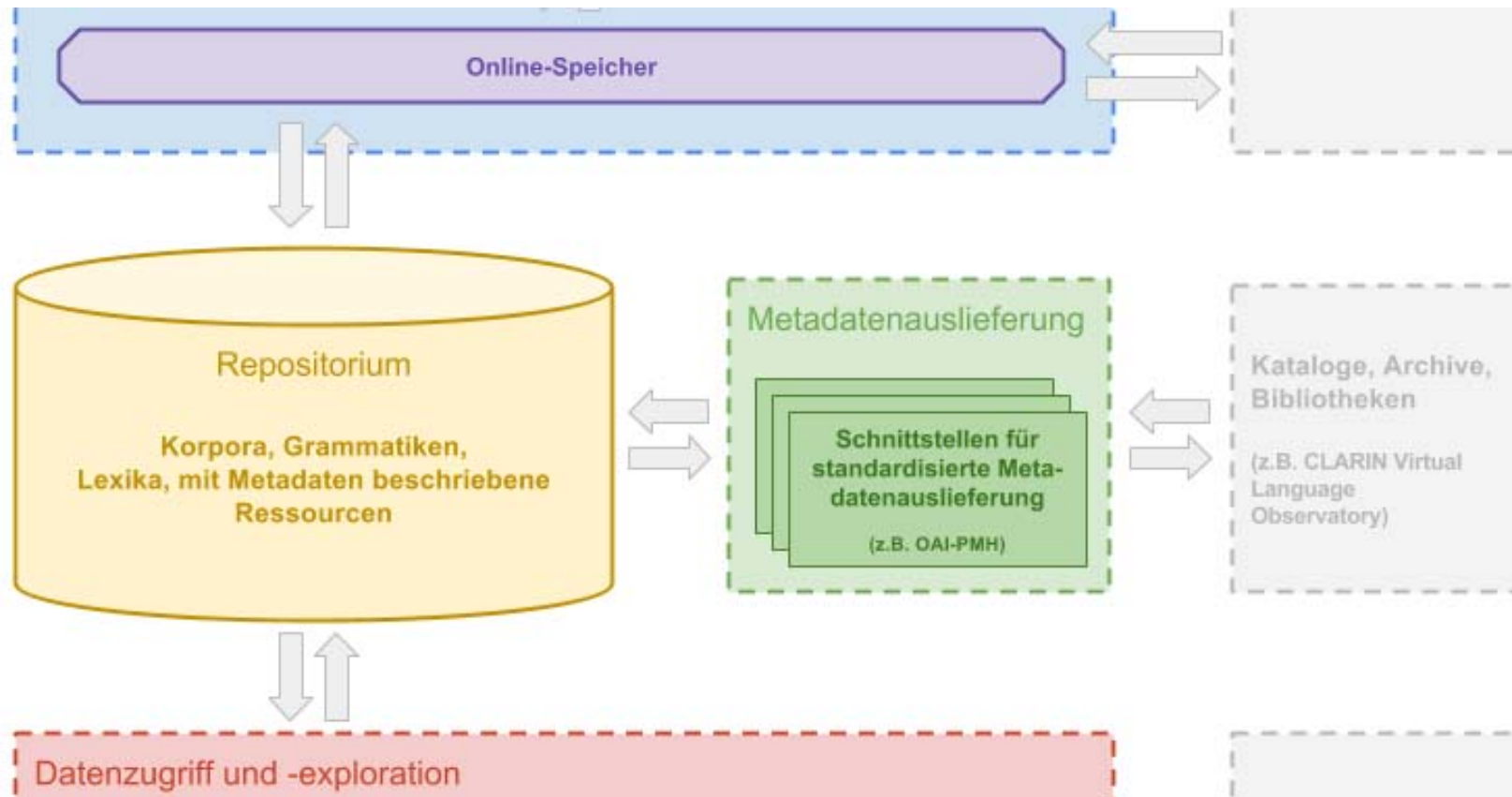
Arbeitsprogramm – AP1 Korpusaufbau



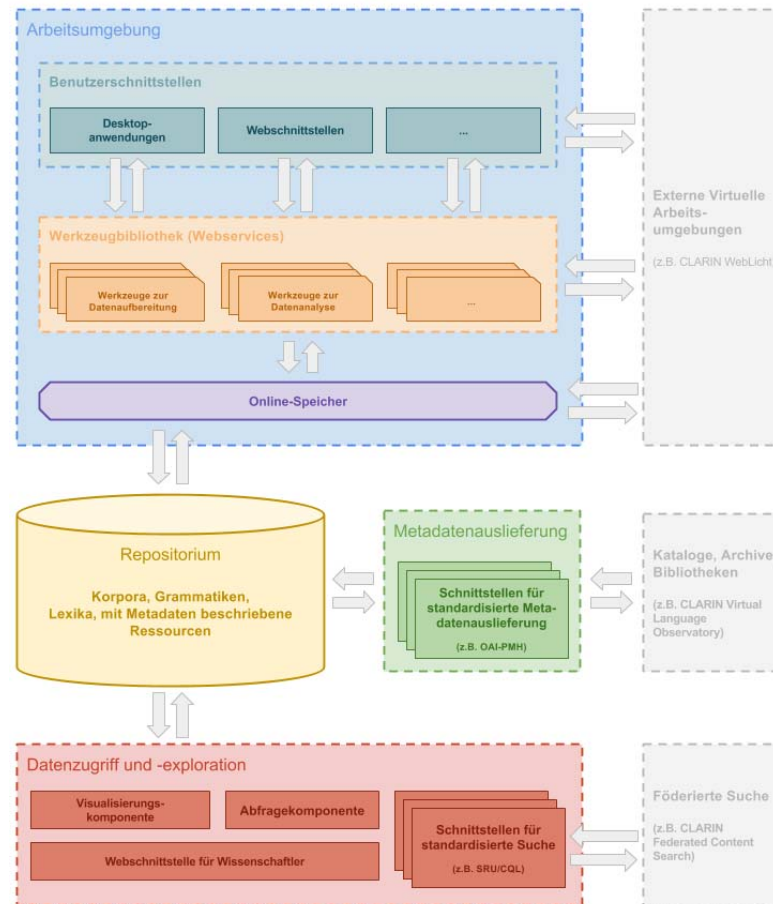
Forschungsumgebung



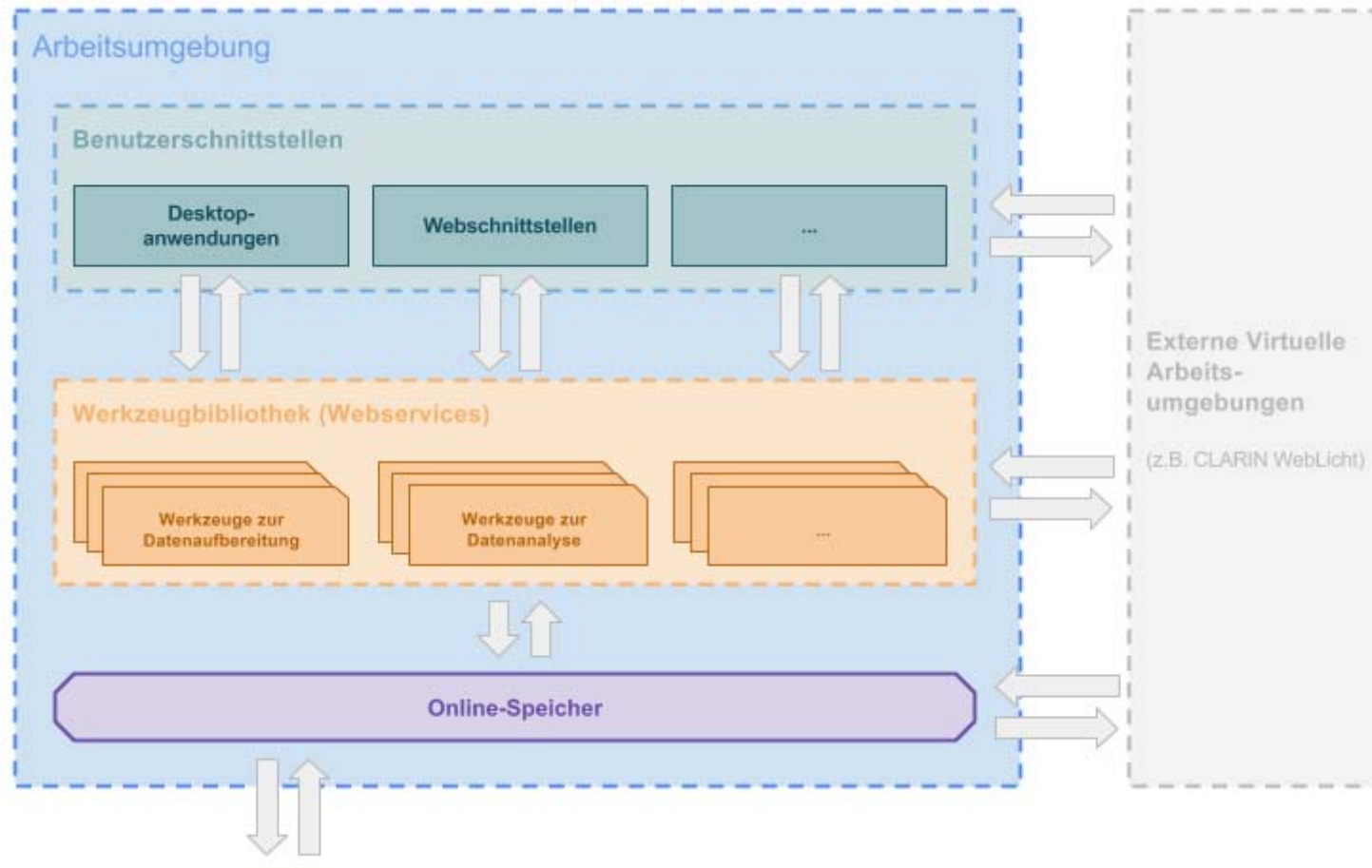
Forschungsumgebung



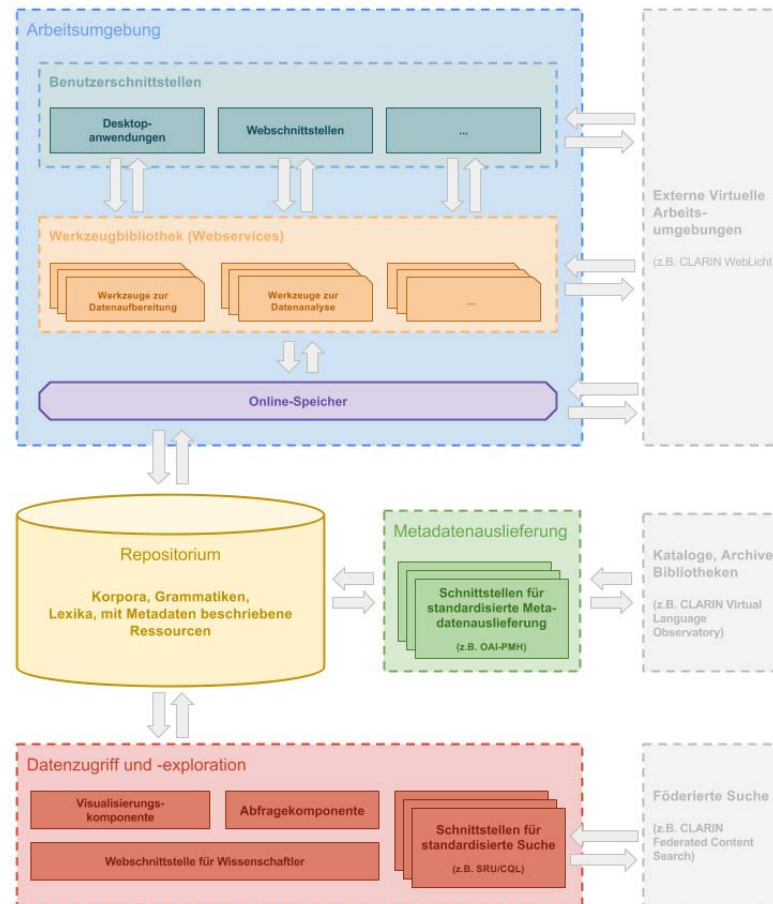
Forschungsumgebung



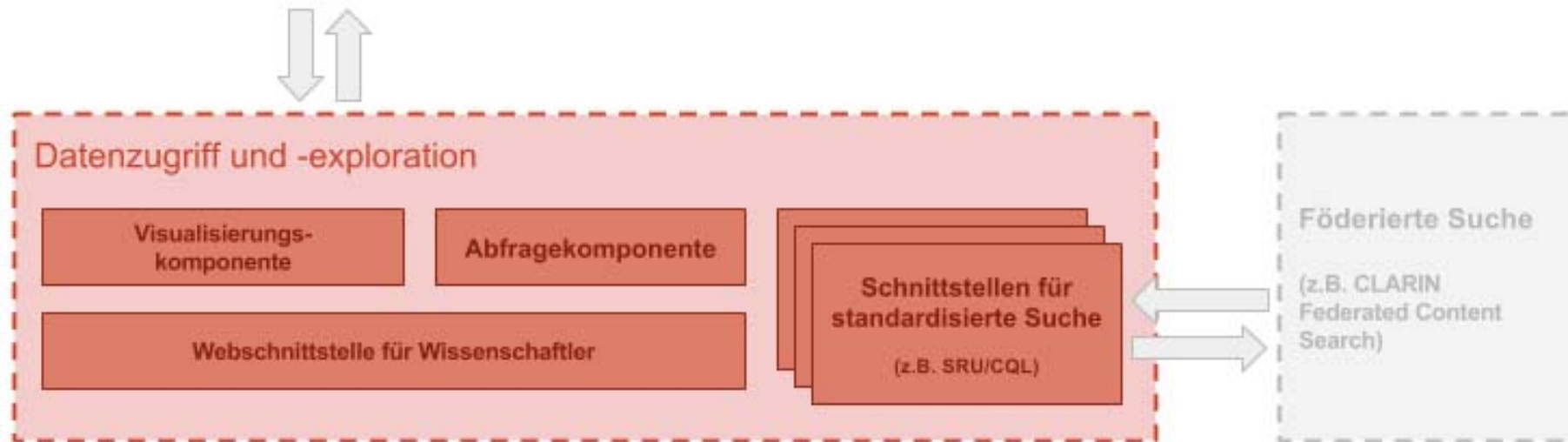
Forschungsumgebung



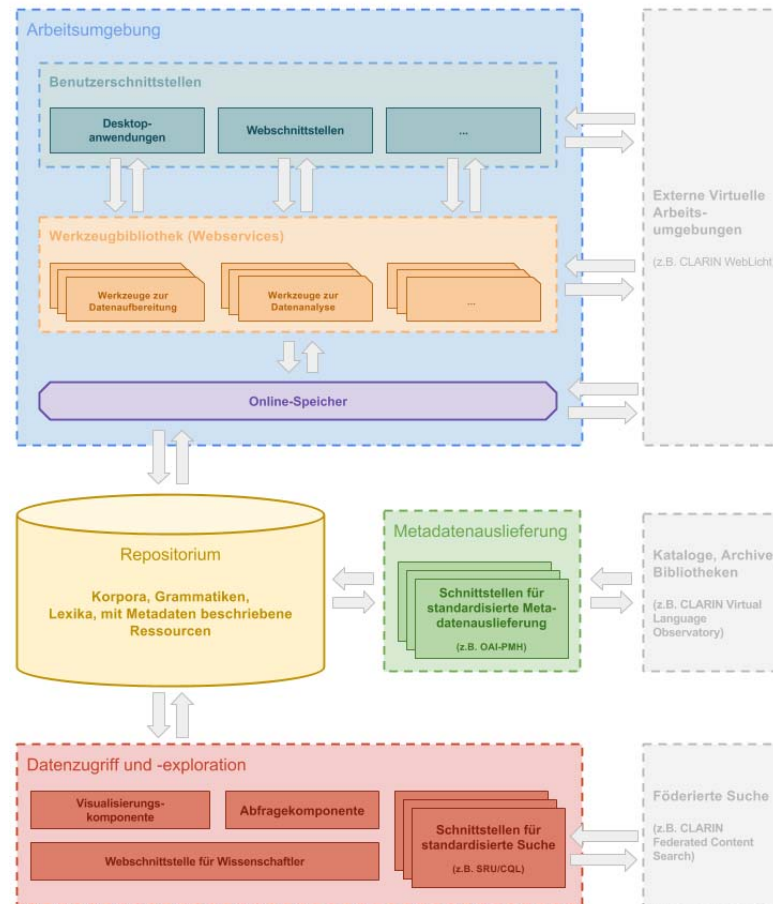
Forschungsumgebung



Forschungsumgebung



Forschungsumgebung



Arbeitsprogramm – AP1 Korpusaufbau

	1. Jahr	2. Jahr	3. Jahr
Arbeitspaket 1 Korpusaufbau	[Timeline bar spanning all three years]		
Modul 1 Inventur	[Timeline bar spanning all three years]		
Akquise/Erhebung	[Timeline bar in Year 1]		
Modul 2 Digitalisierung	[Timeline bar in Year 1]		
Modul 3 Modellierung	[Timeline bar in Year 1]		
Modul 4 Annotation	[Timeline bar in Year 1]	[Timeline bar in Year 2]	[Timeline bar in Year 3]
Modul 5 Korpusfinalisierung			[Timeline bar in Year 3]
Arbeitspaket 2 Infrastruktur und Best Practices	[Timeline bar spanning all three years]		
Arbeitspaket 3 Evaluation und Dissemination	[Timeline bar in Year 1]		[Timeline bar in Year 3]
Workshops	[Timeline bar in Year 1]	[Timeline bar in Year 2]	[Timeline bar in Year 3]
Abschlussbericht			[Timeline bar in Year 3]

Alles wird gut!

- Usecase für digitale Forschungsinfrastrukturen
- Sprachbeschreibung(en) auf der Grundlage adäquater empirischer Ressourcen
- Integration der uralistischen Community in bestehende Forschungsinfrastrukturen

