

Corpus Assembly as Text Data Integration from Digital Libraries and the Web



Udo Hahn & Tinghui Duan

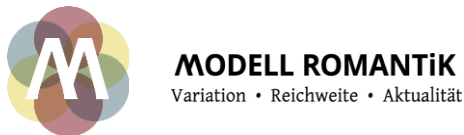
Jena University Language & Information Engineering (JULIE) Lab

<https://julielab.de/>

DFG Graduate School „Romanticism as a Model“

<http://modellromantik.uni-jena.de>

Friedrich Schiller University Jena, Germany



Jun 3 2019 – Urbana-Champaign IL

JCDL 19' – Session 1A – *Generation and Linking*



Allgemeine Literatur-Zeitung (1785-1849)

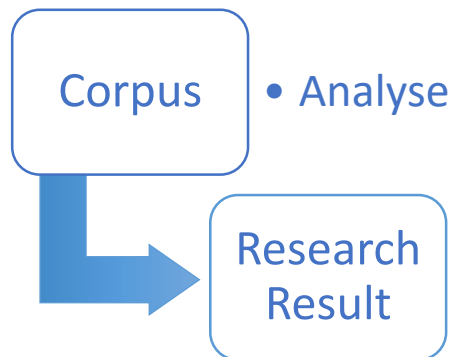
General Literature Gazette, ALZ

Jena/Halle
Germany

Very important historical text source
for literary studies
in German Romanticism (1790-1830)

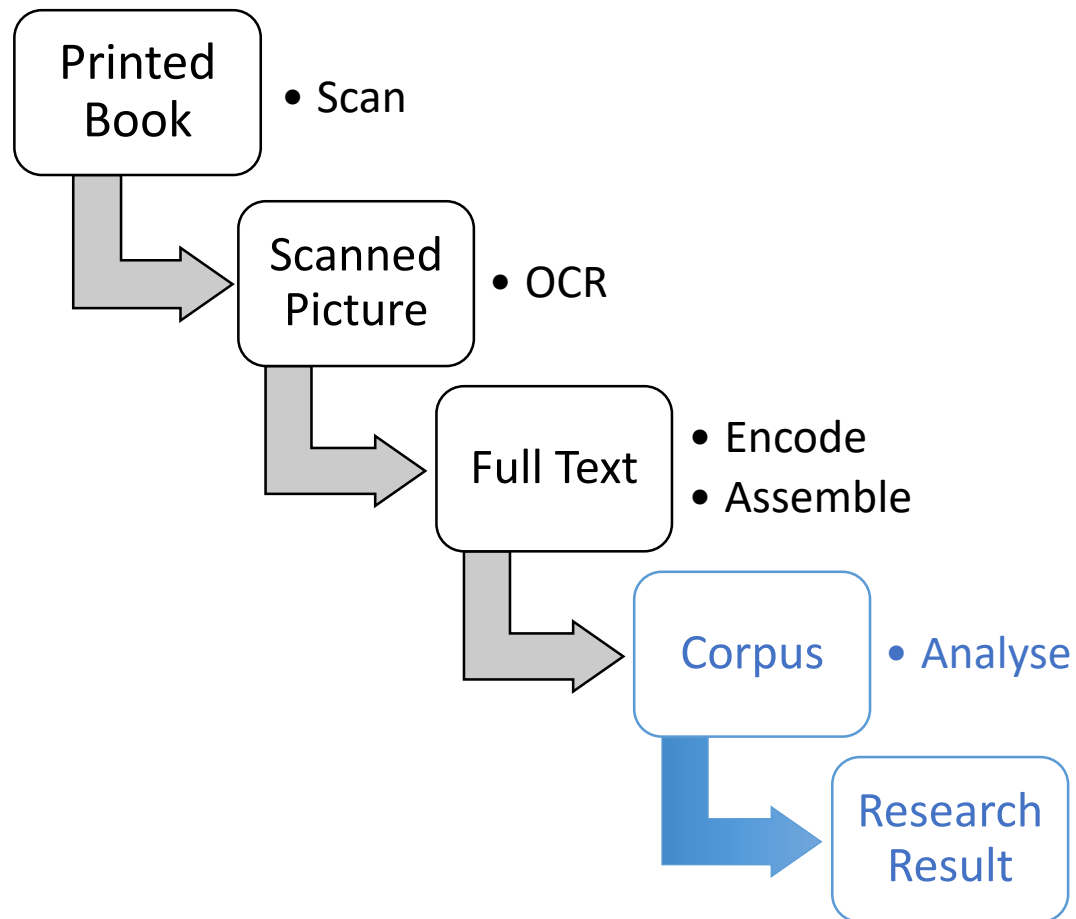


Allgemeine Literatur-Zeitung (1785-1849)



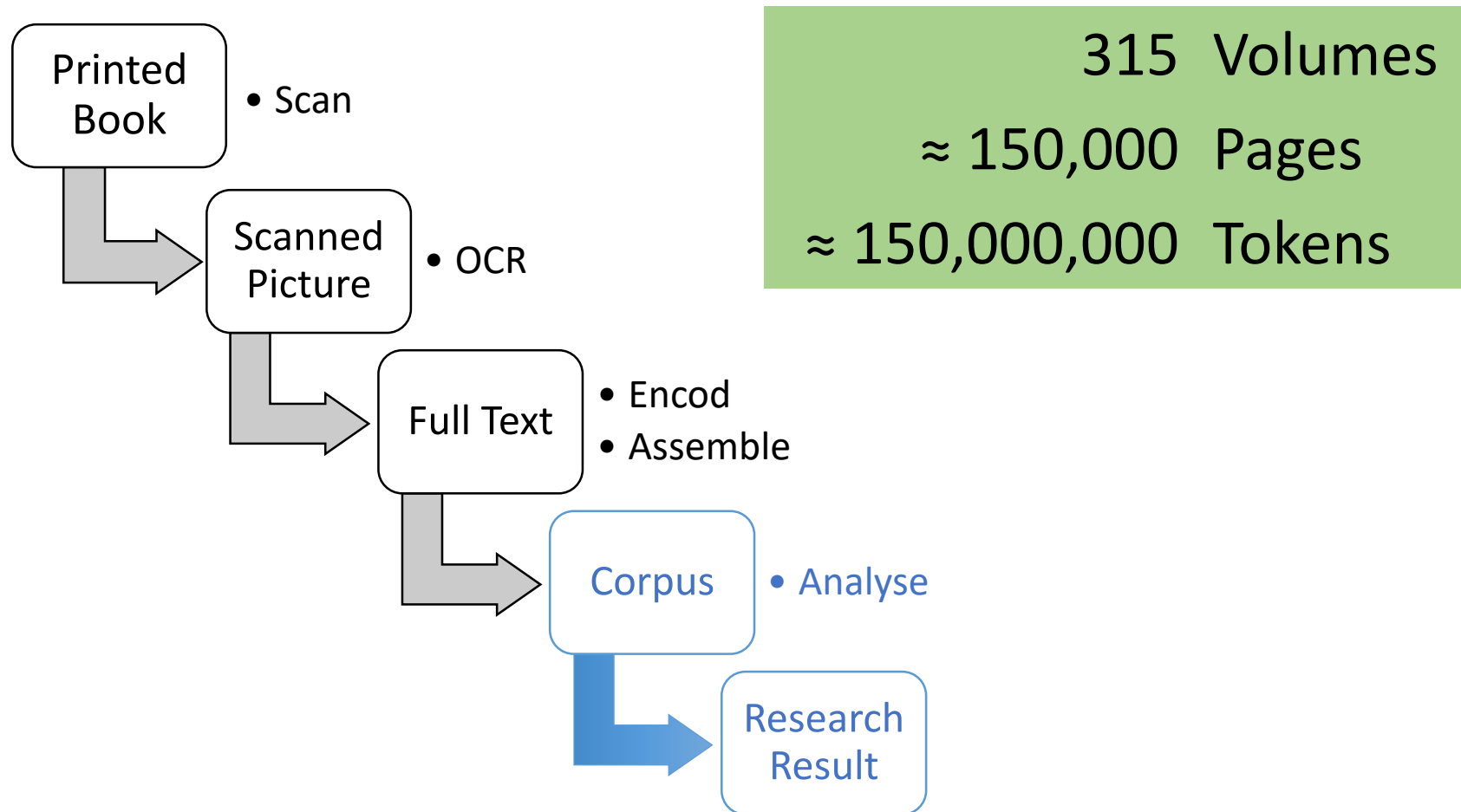
Allgemeine Literatur-Zeitung (1785-1849)

Traditional Workflow



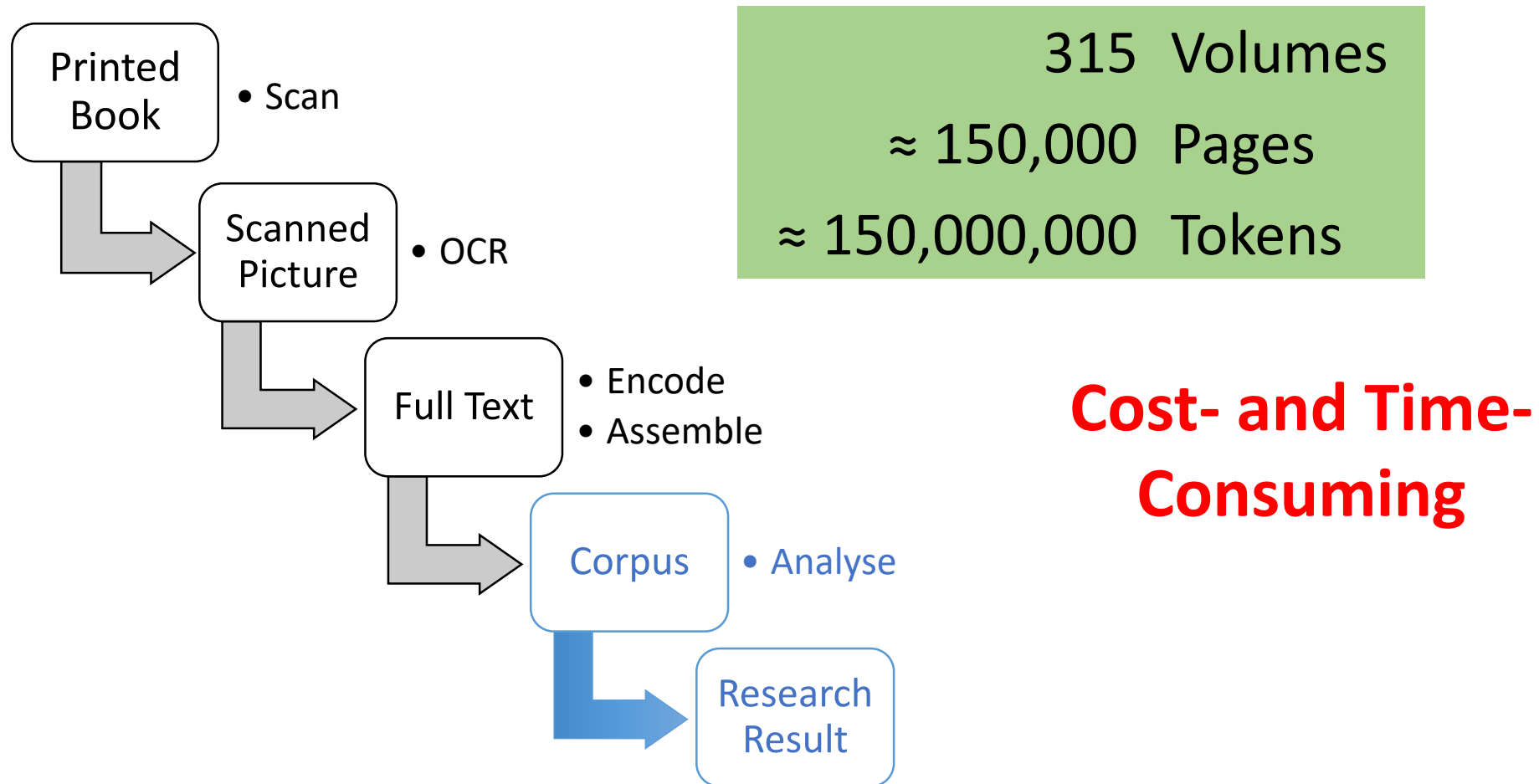
Allgemeine Literatur-Zeitung (1785-1849)

Traditional Workflow



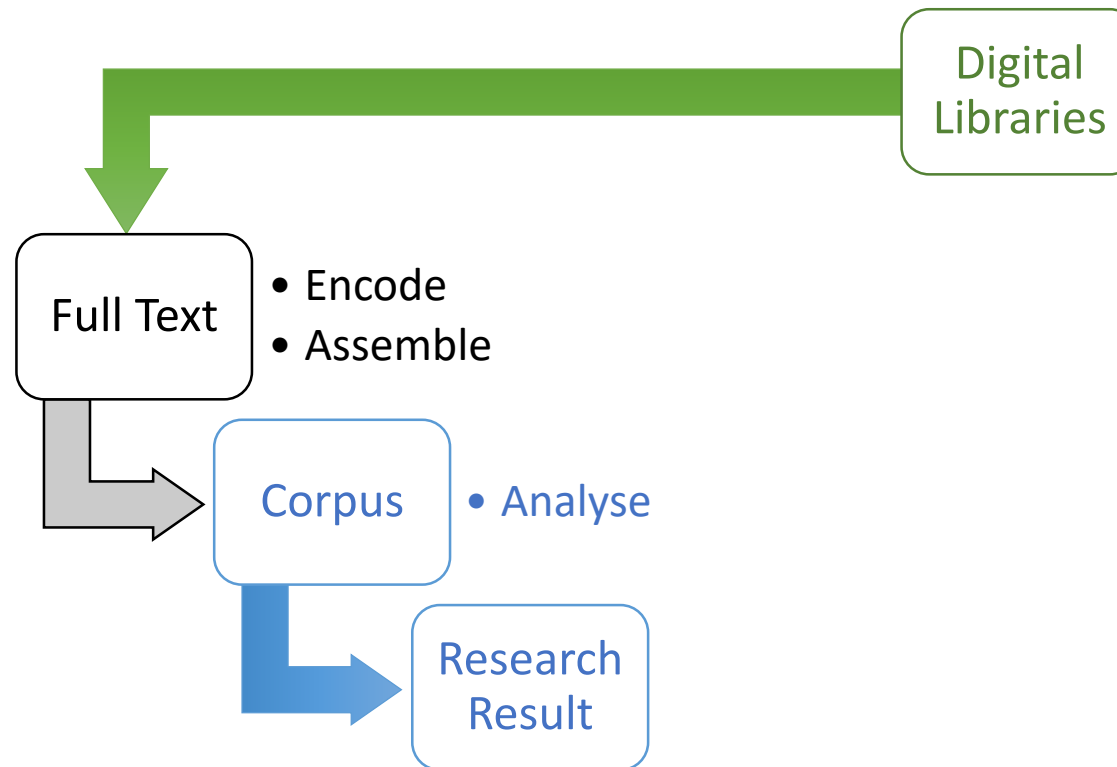
Allgemeine Literatur-Zeitung (1785-1849)

Traditional Workflow

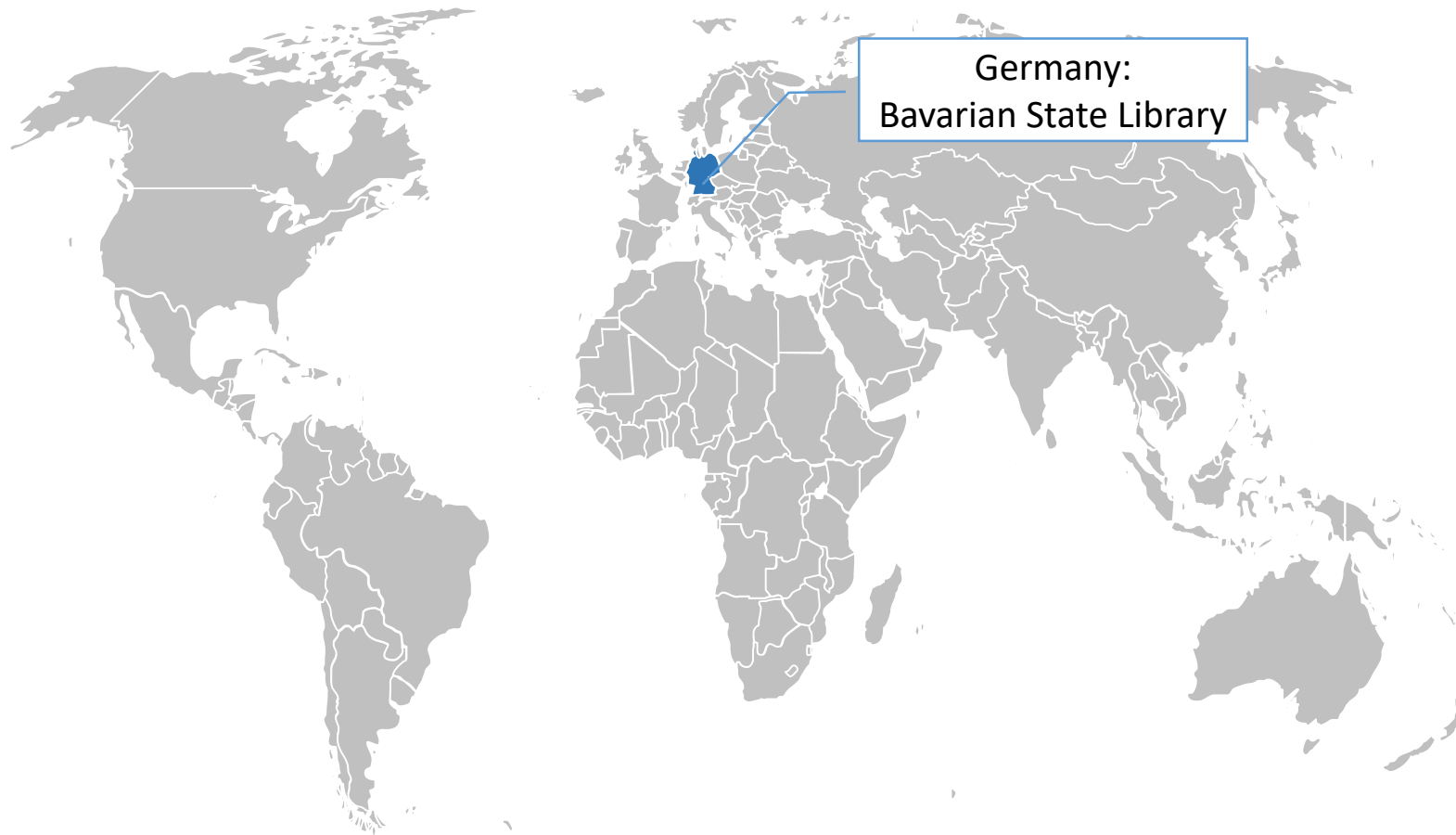


Allgemeine Literatur-Zeitung (1785-1849)

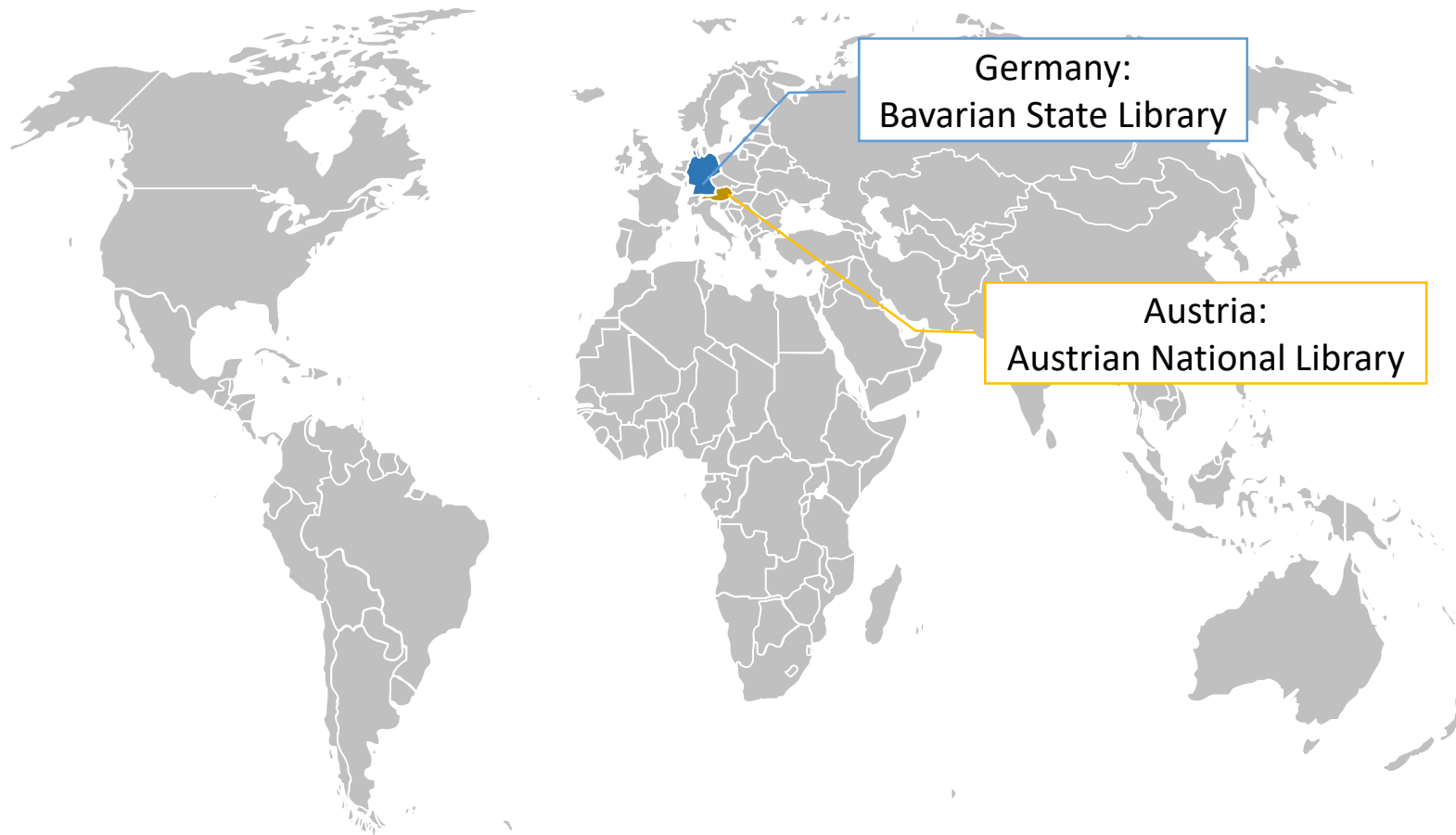
Alternative Workflow



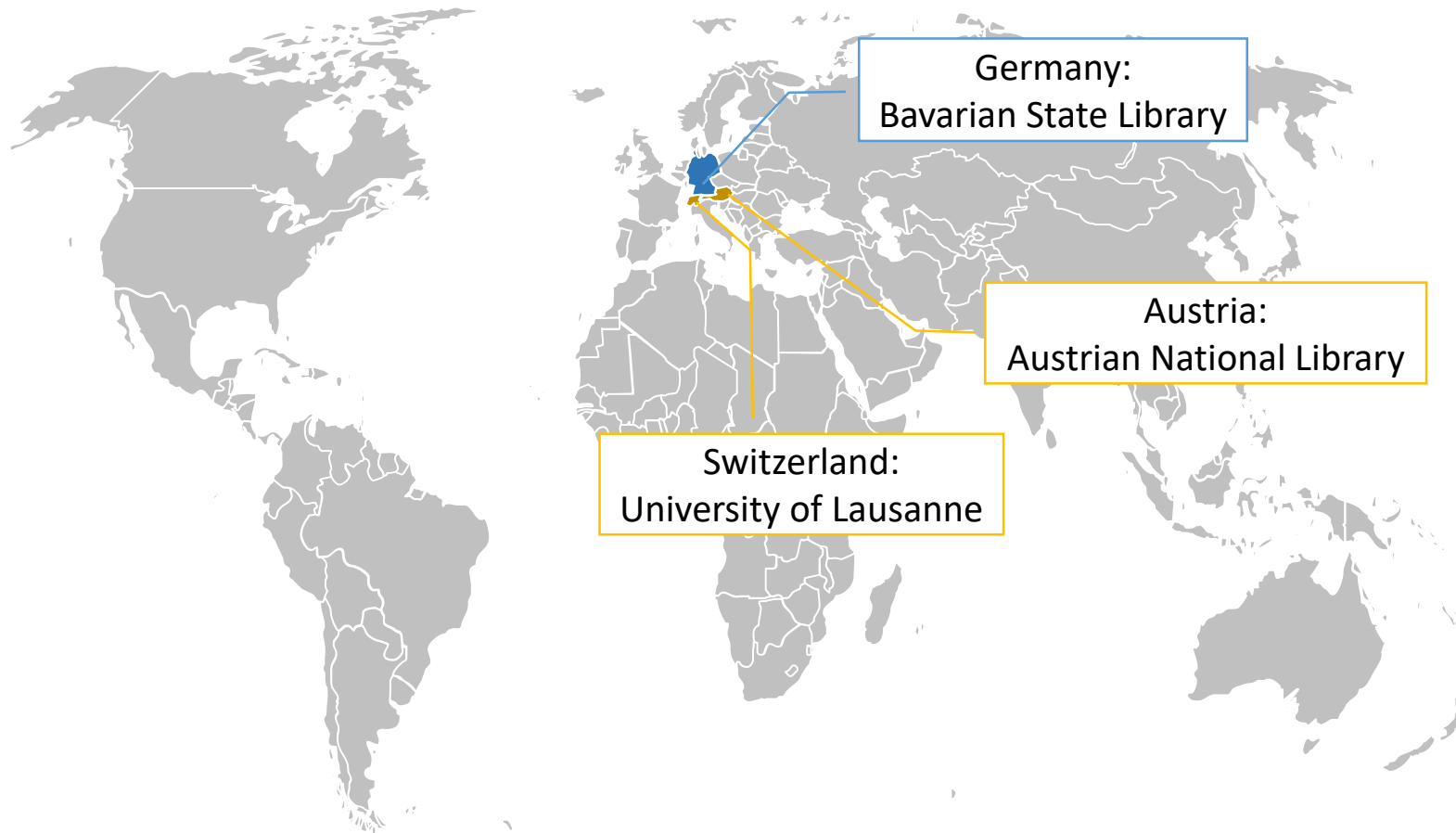
Scattered Digital Resources of ALZ



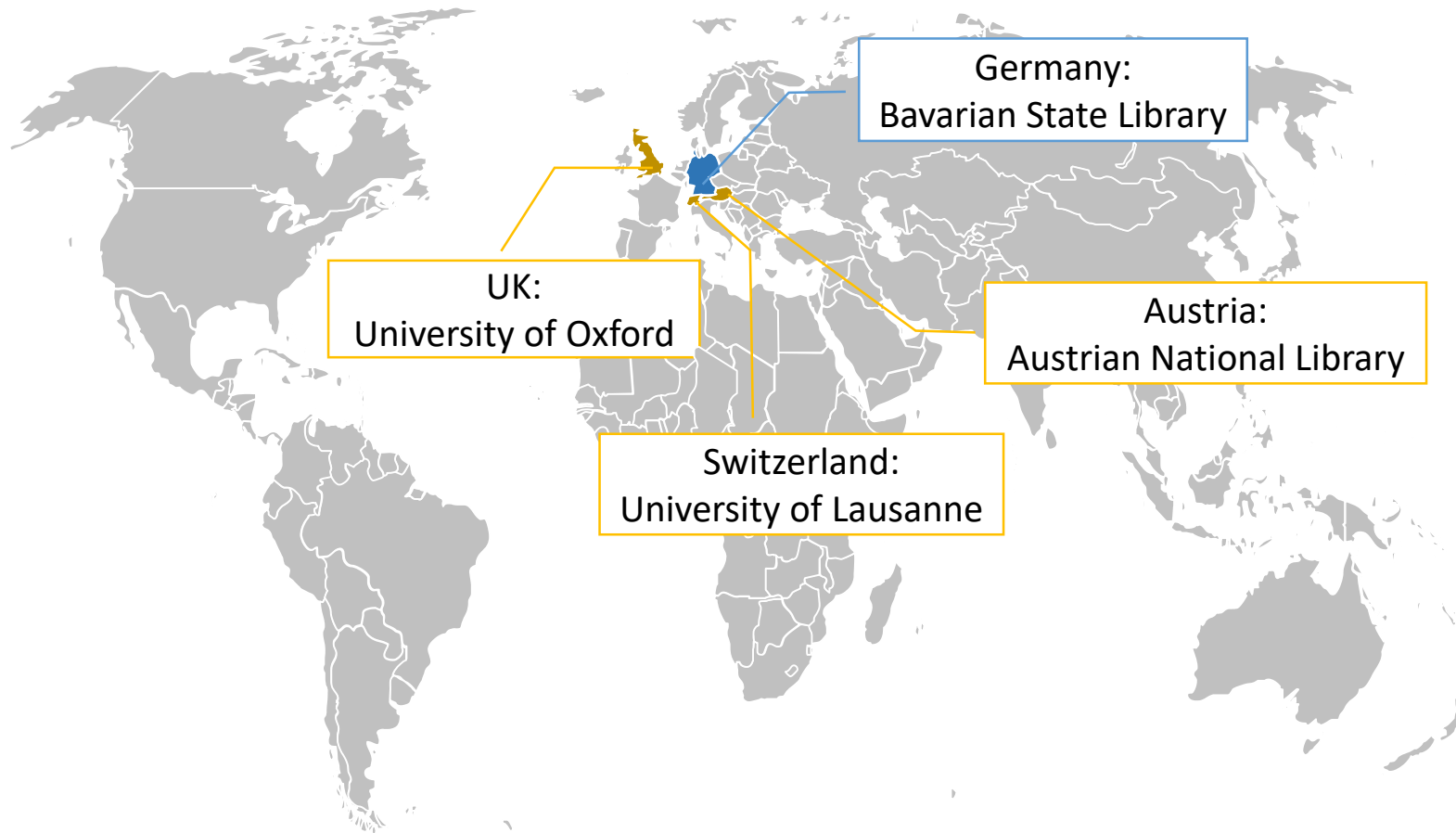
Scattered Digital Resources of ALZ



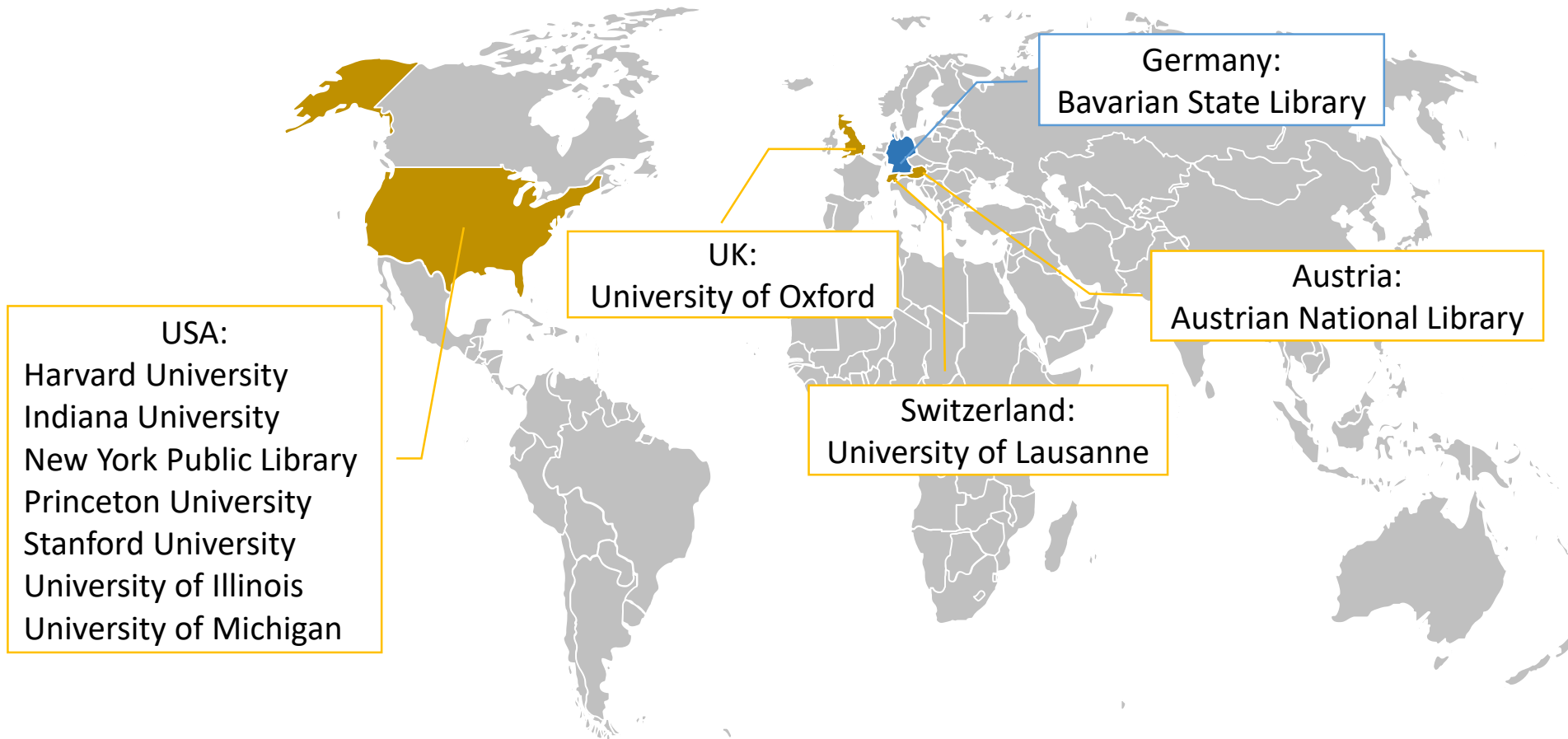
Scattered Digital Resources of ALZ



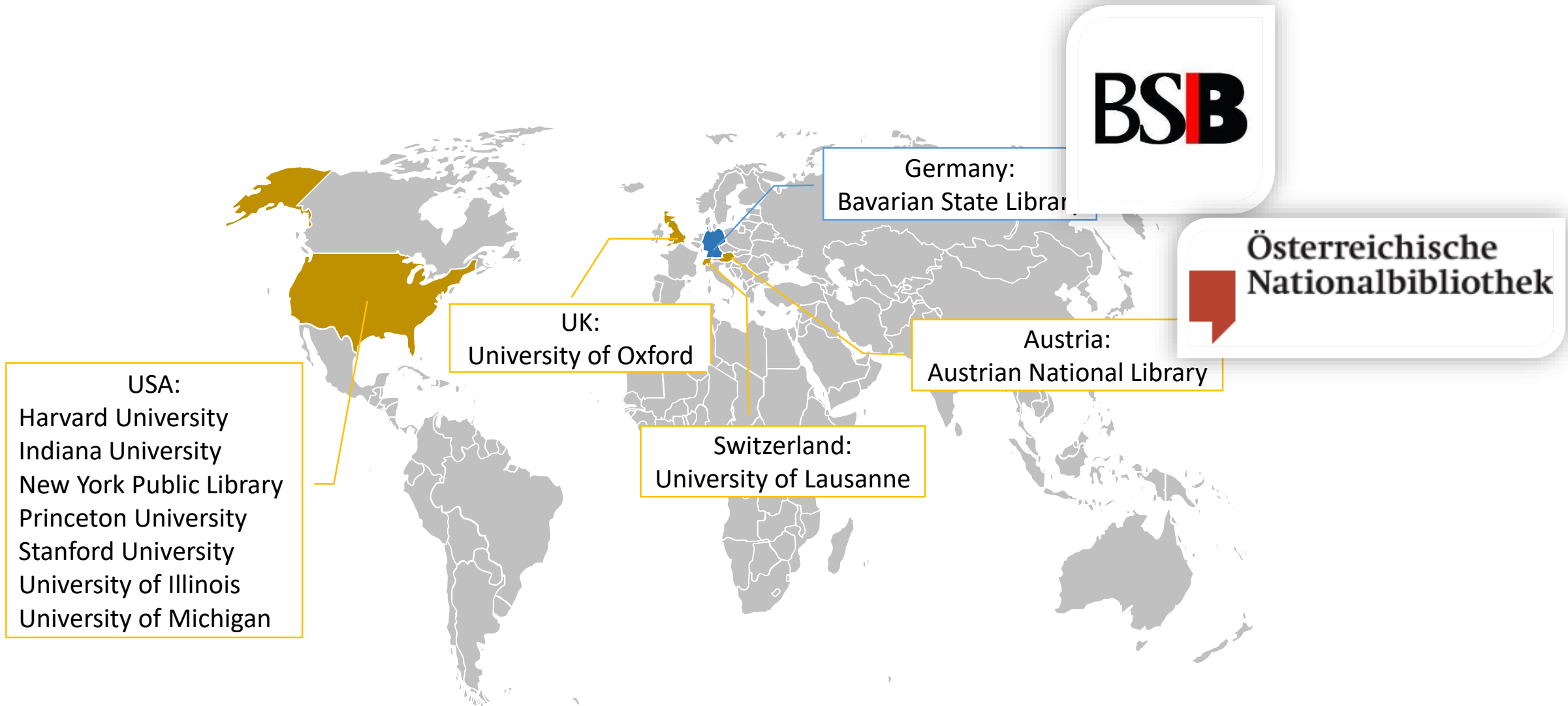
Scattered Digital Resources of ALZ



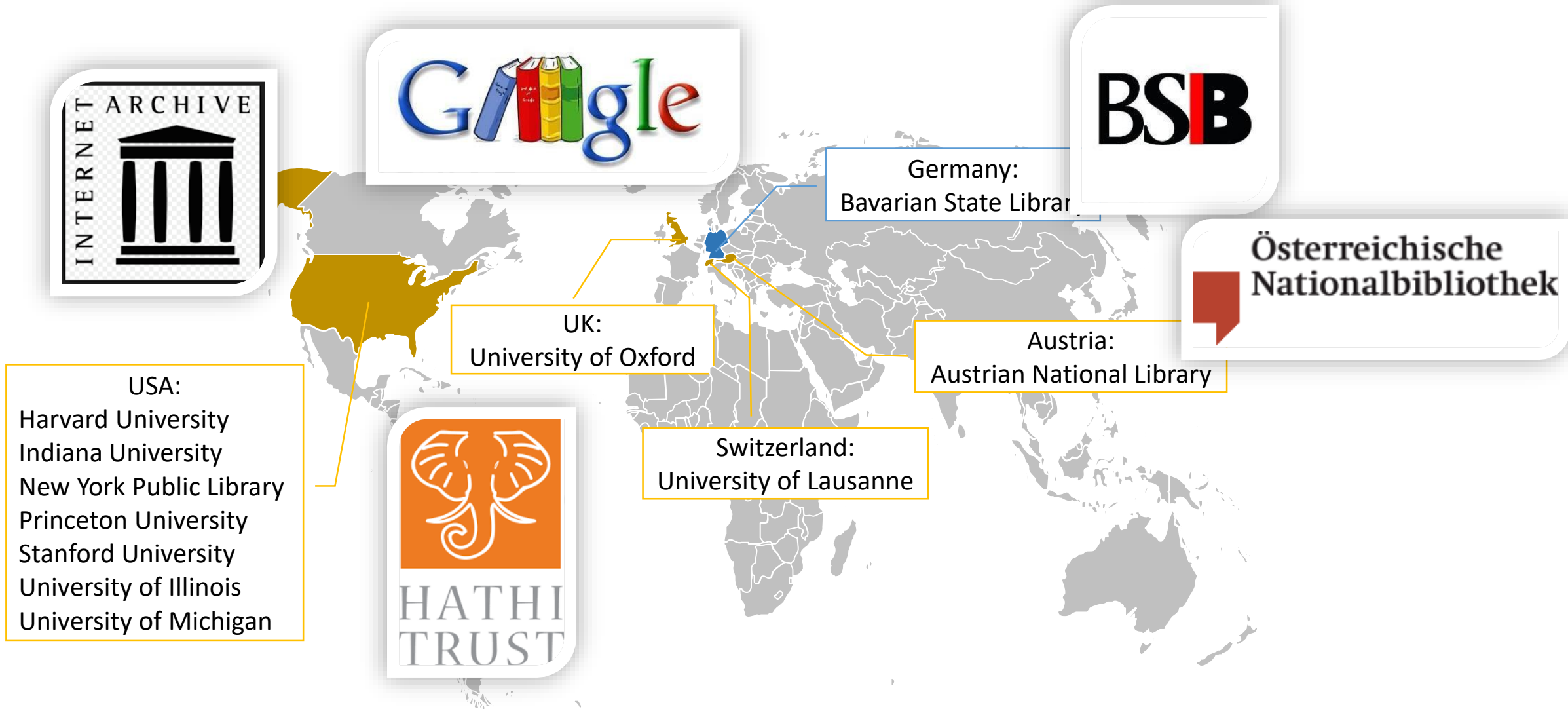
Scattered Digital Resources of ALZ



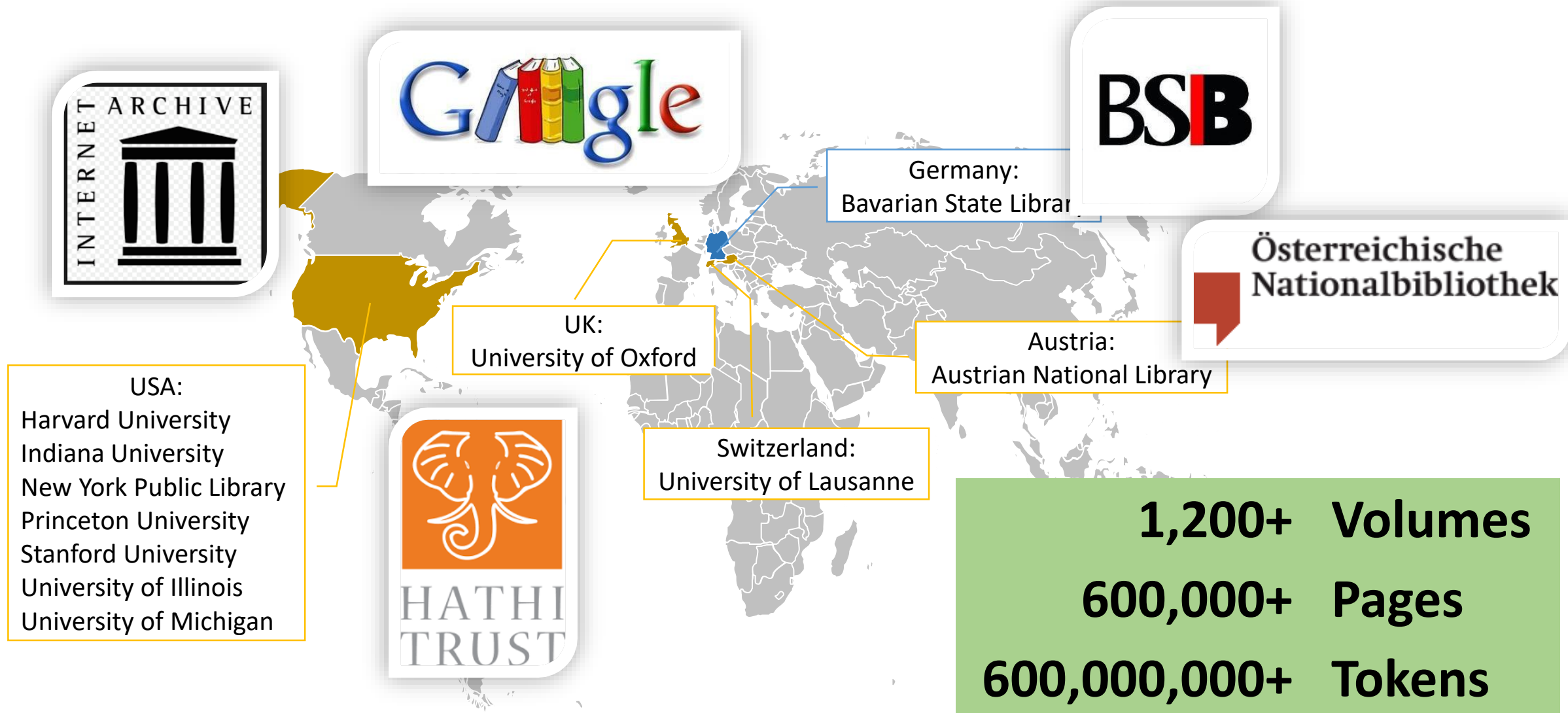
Scattered Digital Resources of ALZ



Scattered Digital Resources of ALZ



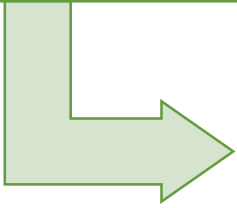
Scattered Digital Resources of ALZ



Proposed Workflow

Digital Libraries
and the Web

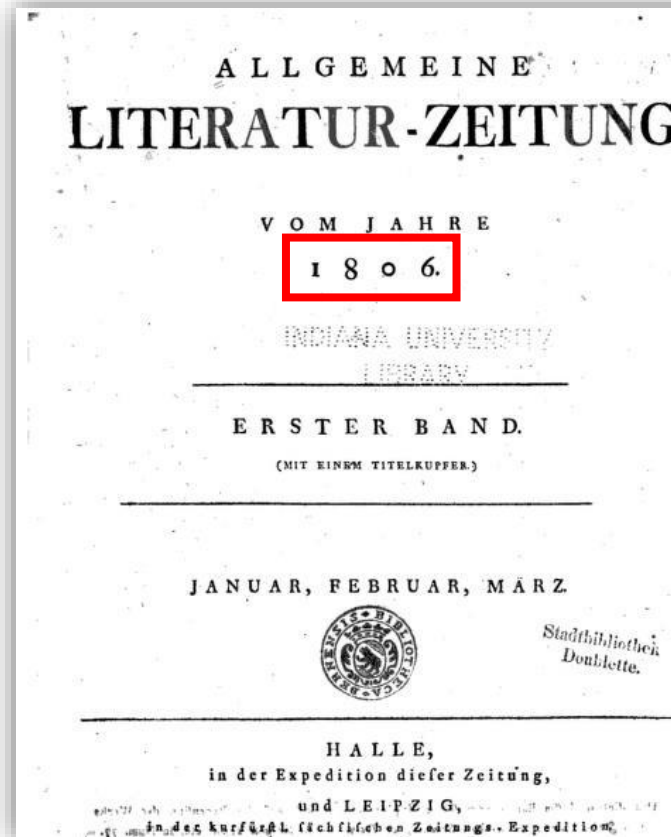
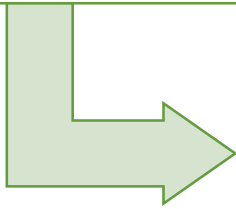
- Collect
- Correct Metadata






Proposed Workflow

Digital Libraries and the Web

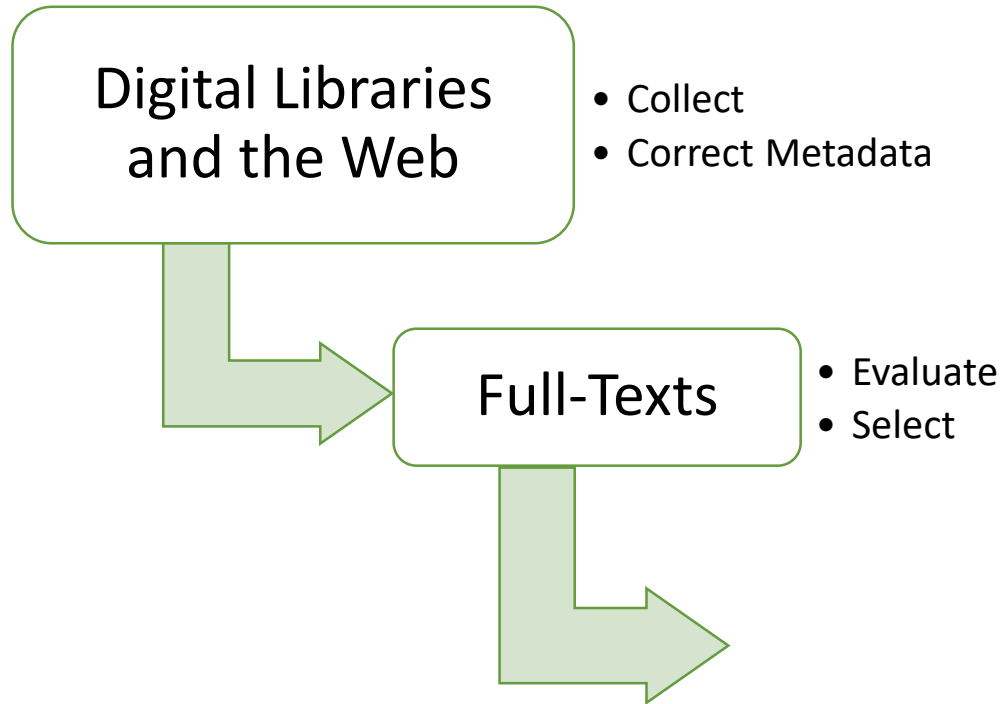
- Collect
- Correct Metadata



 Allgemeine Literatur-Zeitung	
Publication date	1785
Usage	Public Domain Mark 1.0  
Topics	bub_upload, Books
Publisher	Jena [etc.]
Collection	americana
Digitizing sponsor	Google
Book from the collections of	Indiana University
Language	German
Has supplement: Intelligenzblatt, 1787-1807, 1829-1849	
Mode of access: Internet	
Google-id	udTjAAAAAMAAJ
Identifier	bub_gb_udTjAAAAAMAAJ
Identifier-ark	ark:/13960/t34217g7g
Ocr	ABBYY FineReader 11.0
Olsearch	post
Pages	535
Scanner	google
Source	http://books.google.com/books?id=source=gbs_api
Worldcat (source edition)	1479154
Year	1806

https://archive.org/details/bub_gb_udTjAAAAAMAAJ/

Proposed Workflow



Proposed Workflow

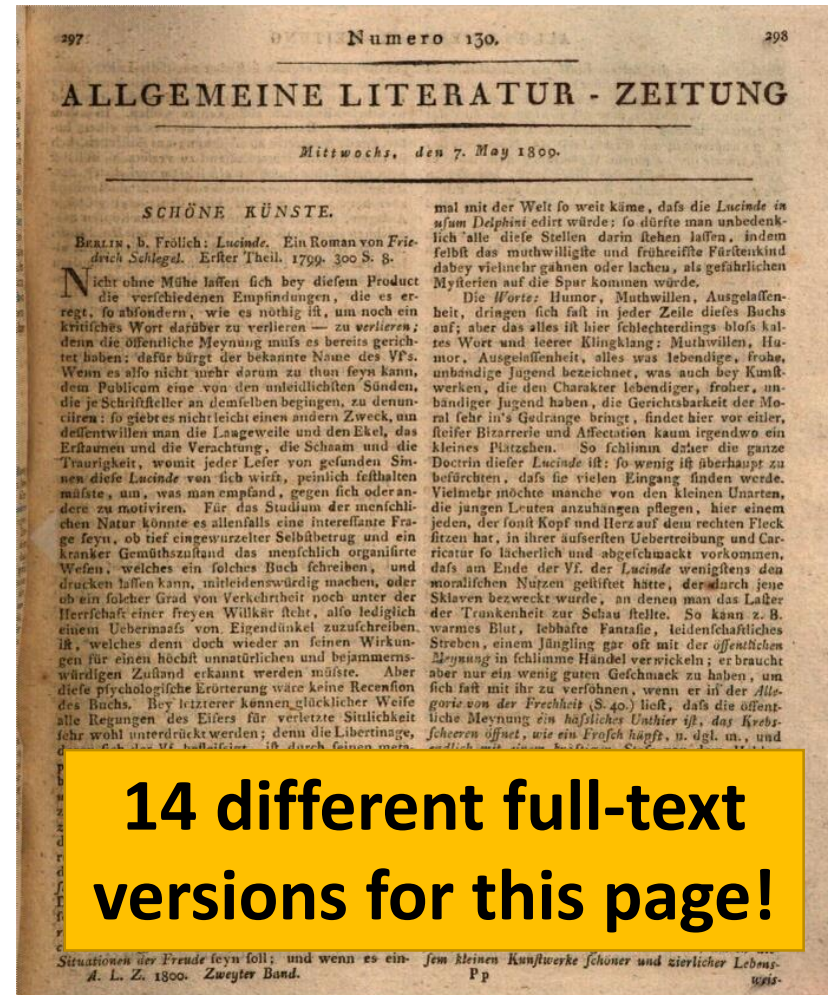
Digital Libraries
and the Web

- Collect
- Correct Metadata

Full-Texts

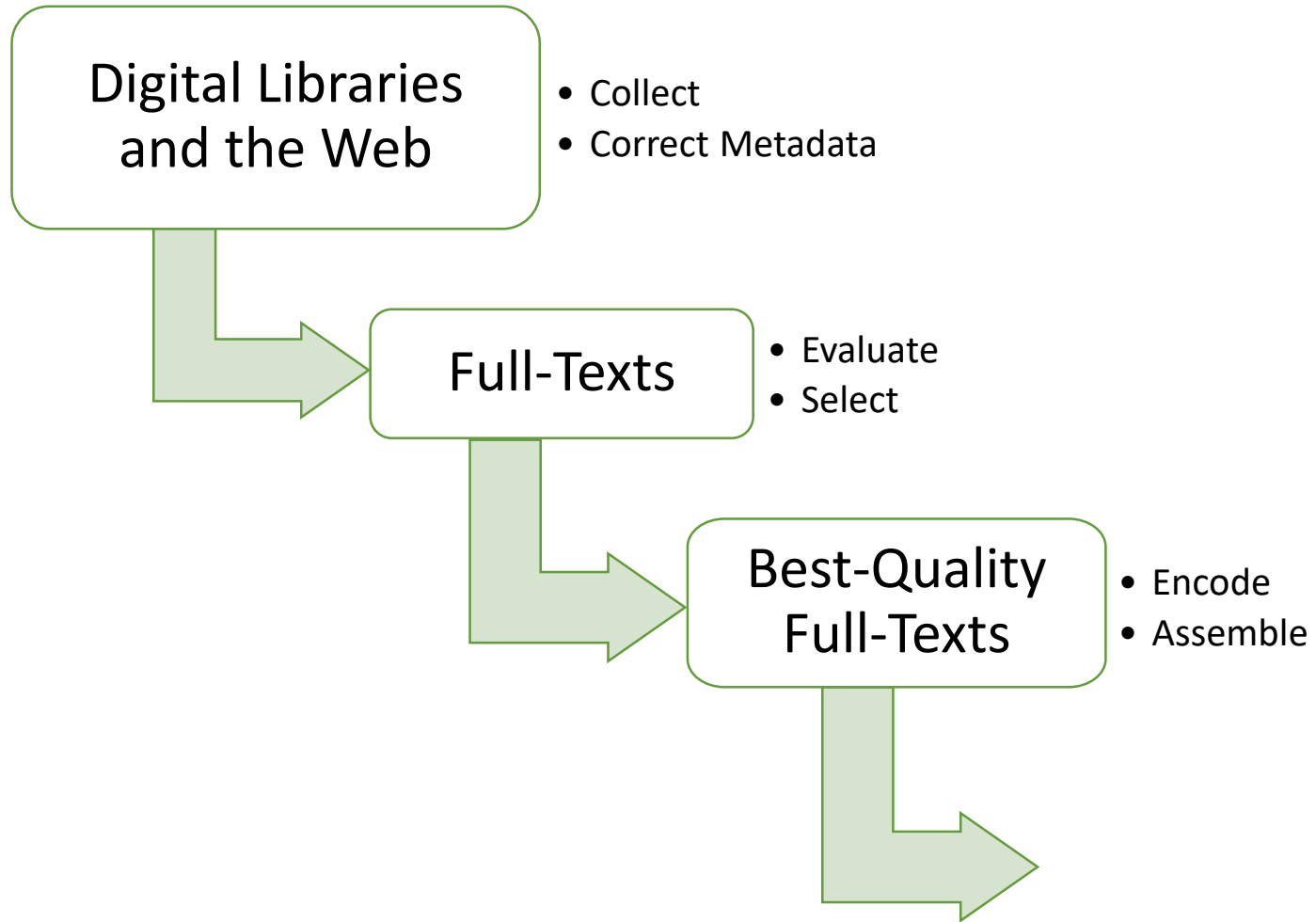
- Evaluate
- Select

14 different full-text
versions for this page!

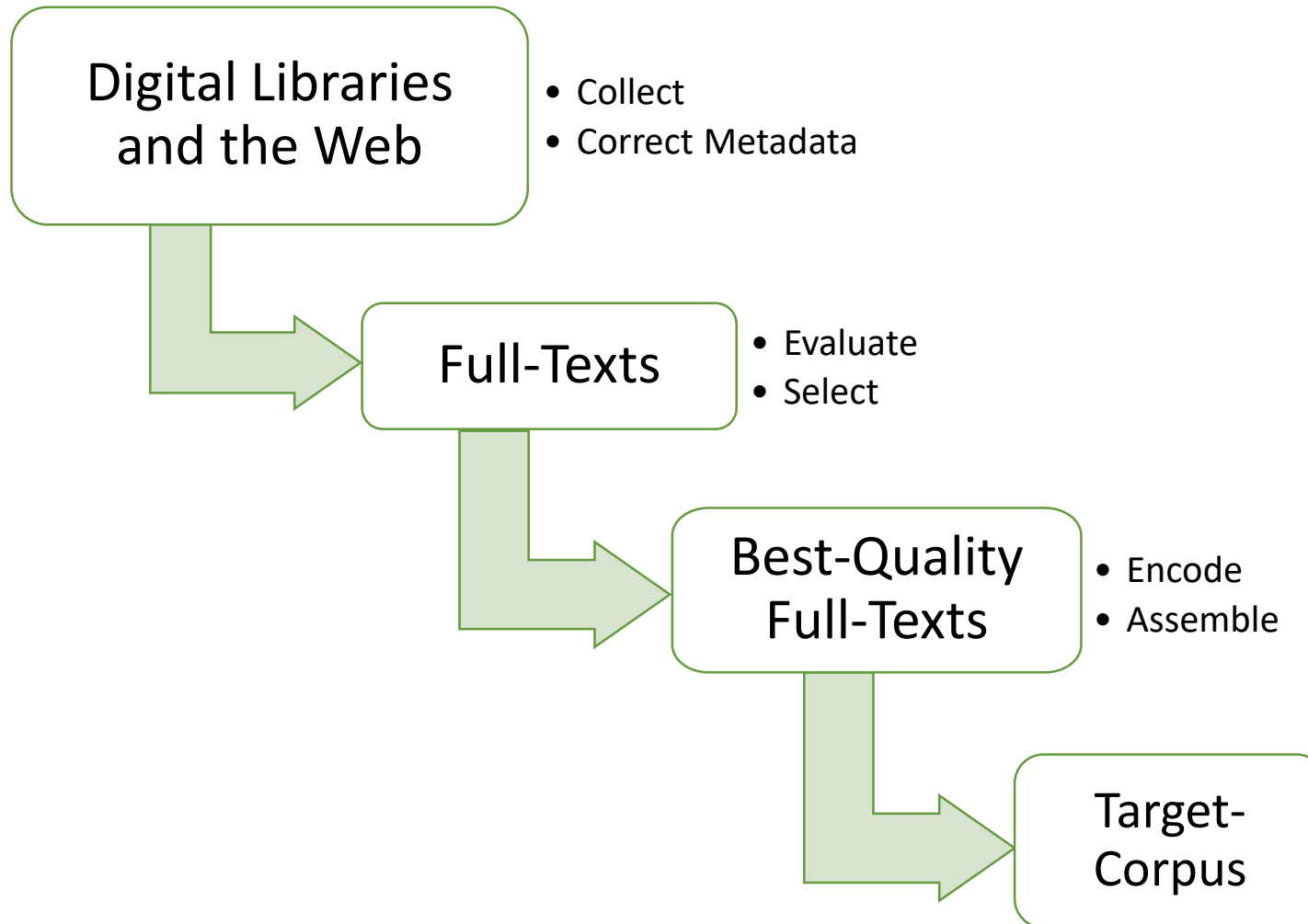


297
130, 298
N u m e r o
ALLGEMEINE LITERATUR - ZEITUNG
Mittwochs,
den 7. May 1809.
S C H Ö N E K Ü N S T E .
BERLIN, b. Frölich: Lucinde. Ein Roman von Friedrich Schlegel. Erster Theil. 1799. 300 S. 8.
Nie ohne Mühe lassen sich bey diesem Product die verschiedenen Empfindungen, die es erregt, so absondern, wie es nöthig ist, um noch ein kritisches Wort darüber zu verlieren – zu verlieren; denn die öffentliche Meynung muß es bereits gerichtet haben: dafür bürgt der bekannte Name des Vf's. Wenn es also nicht mehr carum zu thun seyR kann, dem Publicum eine von den unelidlichen Sünden, die je Schriftsteller an demselben begingen, zu denunciren: so giebt es nicht leicht einen andern Zweck, um dessentwillen man die Langeweile und den Ekel, das Traurigkeit, womit jeder Leser von gefunden Sinnen diese Lucinde von sich wirft, peinlich festhalten müßte, um, was man empfand, gegen sich oder andere zu motiviren. Für das Studium der menschlichen Natur könnte es allenfalls eine interessante Frage seyn, ob tief eingewurzelter Selbstbetrug und ein kranker Gemüthszustand das menschlich organisirte Wesen, welches ein solches Buch schreiben, und drucken lassen kann, mittheilenswürdig machen, oder ob ein solcher Grad von Verkehrtheit noch unter der Herrschaft einer freyen Willkür steht, also lediglich einem Uebermaafs von Eigendünkel zuzuschreiben ist, welches denn doch wieder an seinen Wirkungen für einen höchst unnatürlichen und bejammernswürdigen Zustand erkannt werden müßte. Aber diese psychologische Erörterung wäre keine Recension des Buchs, Bey letzterer können glücklicher Weise alle Regungen des Eifers für verletzte Stilleckheit sehr wohl unterdrückt werden; denn die Libertinage, deren sich der Vf. befließigt, ist durch seinen metaphysisch-poetischen Unflinn unschädlich gemacht, und bey dem tölpelhaften Enthusiasmus (S. 30) der ihm ursprünglich und wesentlich in der Natur des Mannes zu liegen scheint, der, wie er hinzusetzt, leicht bis zur Grobheit göttlich ist, heben sich, zum Glück für die Jugend, die an plumpen oder feurigen Schilderungen der Wollust ein schädliches Wohlgefallen finden könnte, die wirkliche Tölpelheit und der soifalsam Enthusiasmus so ziemlich gegen einander auf. Diefes gilt selbst ven den schändlichsten, aretinisch feynfaltenden Stellen, z. B. von der faubern dithyranbischen Fantasie über die schönste Situation, welche die witzigste und schönste unter den Gestalten und Situationen der Freude seyn soll; und wenn es ein A. L. Z. 1800. Zuvveyter Band.

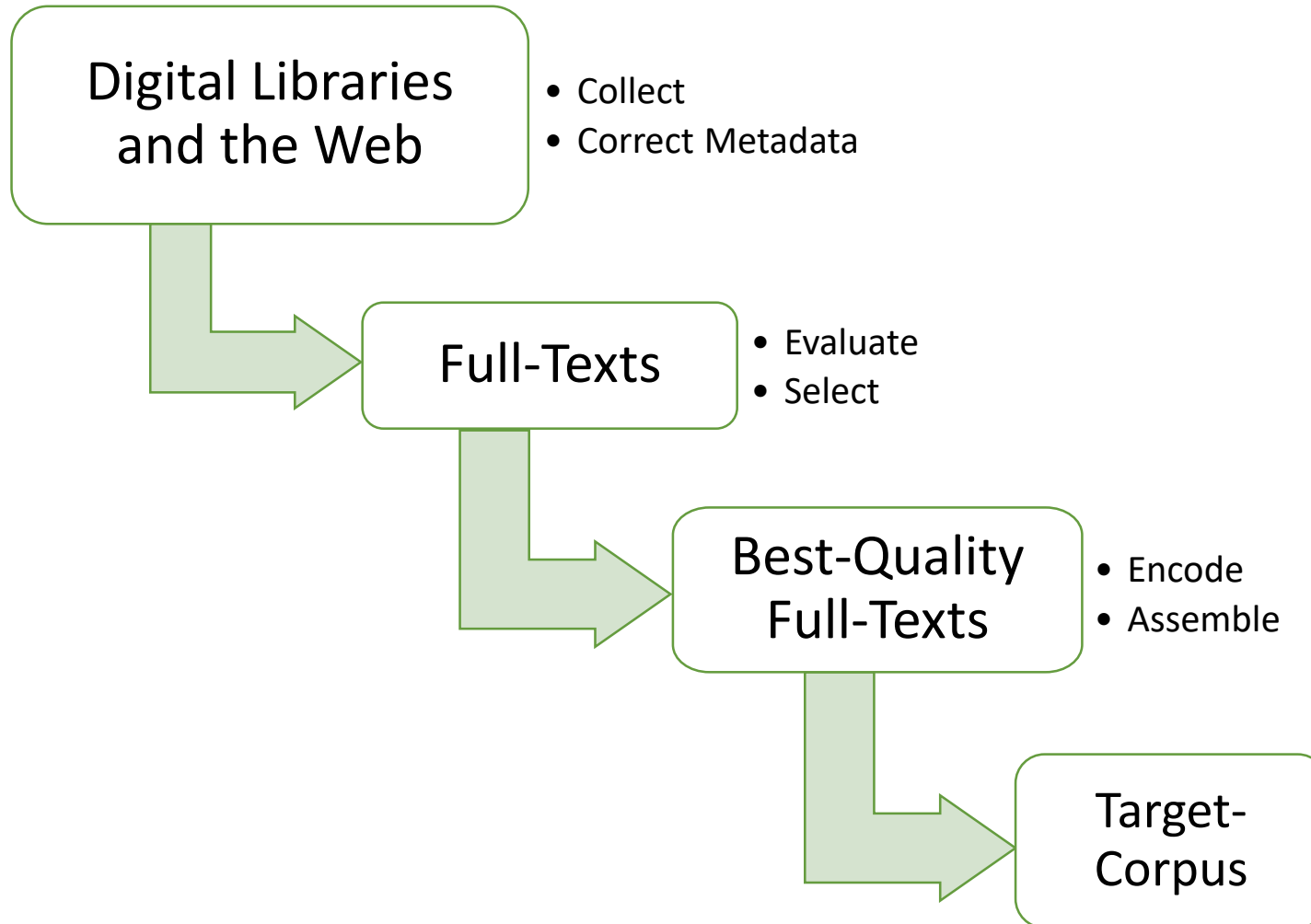
Proposed Workflow



Proposed Workflow

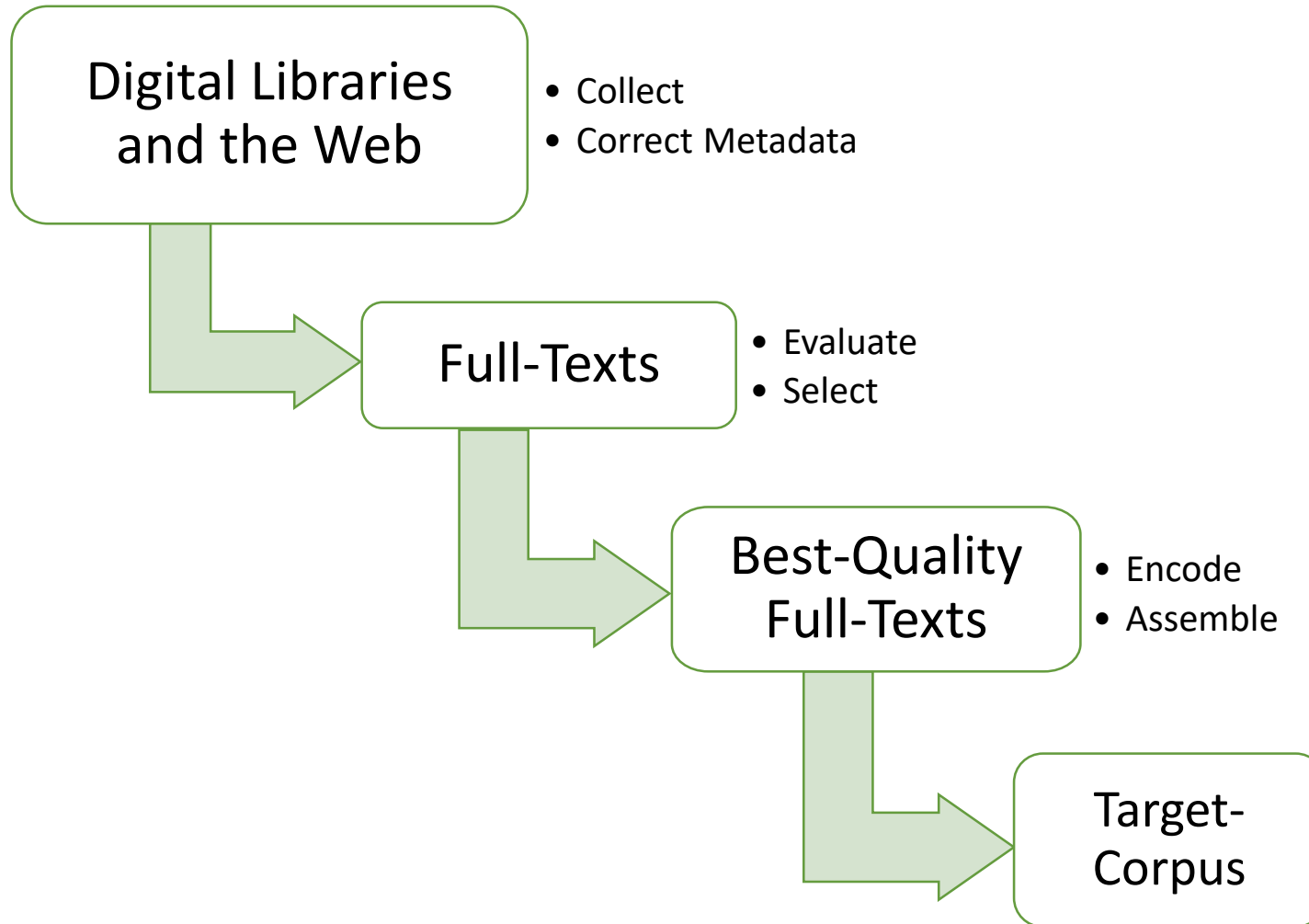


Result



261	Volumes
126,612	Pages
120,369,005	Tokens

Result



261 Volumes
126,612 Pages
120,369,005 Tokens

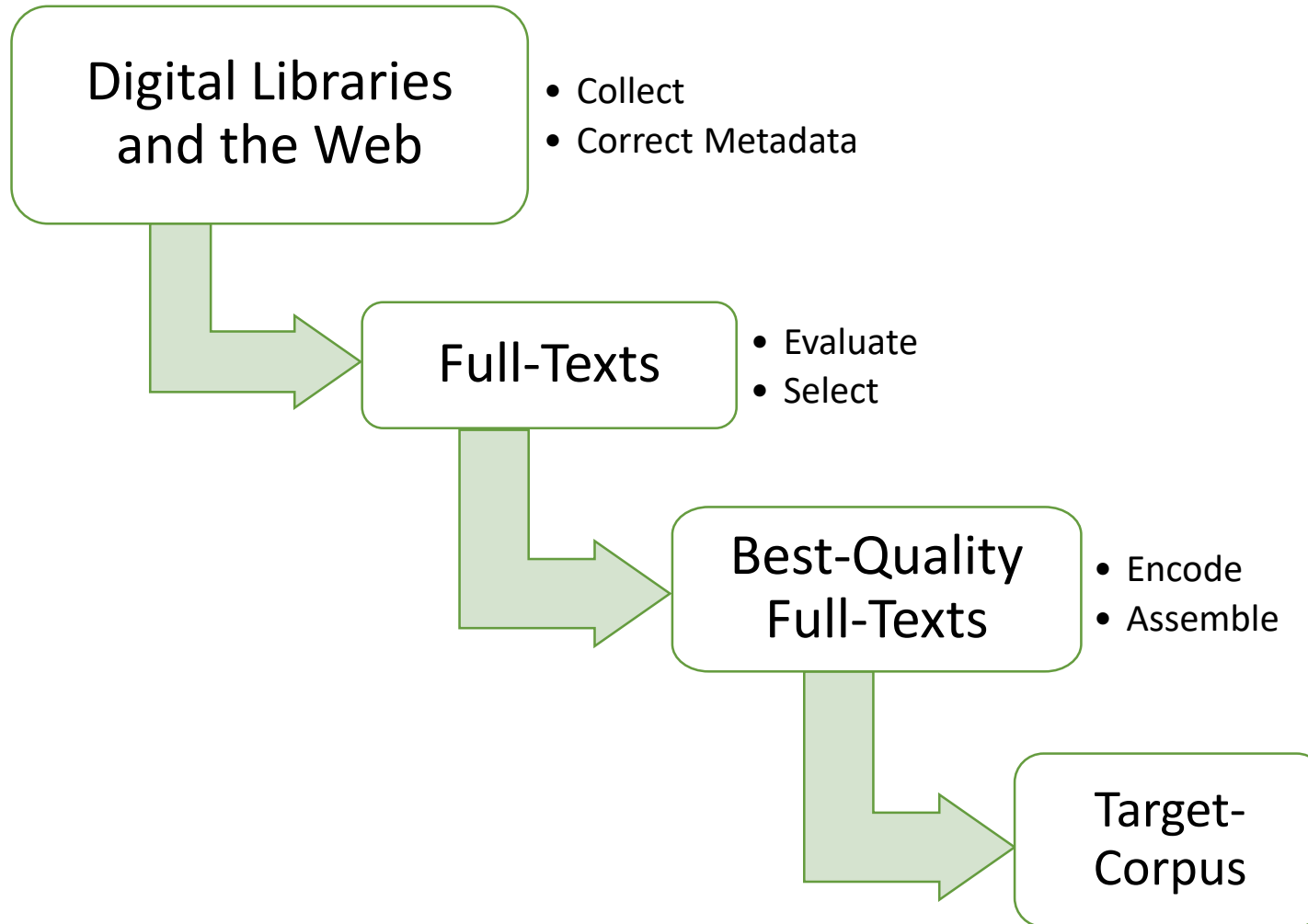
≈ 82% coverage

315 Volumes
≈ 150,000 Pages
≈ 150,000,000 Tokens

Result

The Largest Corpus for German Romanticism

<https://github.com/JULIELab/ALZ>



261 Volumes

126,612 Pages

120,369,005 Tokens

≈ 82% coverage

315 Volumes

≈ 150,000 Pages

≈ 150,000,000 Tokens

Problems

- Restricted Accessibility
- Heterogeneous Digitizing Conditions and OCR-Qualities

Conclusion

- The Largest Corpus for German Romanticism
- Big Potential of DLs for Computational Literary Studies
- More Cooperation Between DLs Desirable
- Better Metadata and OCR-Quality are Desirable

Corpus Assembly as Text Data Integration from Digital Libraries and the Web

Thank you!



Udo Hahn & Tinghui Duan

Jena University Language & Information Engineering (JULIE) Lab

<https://julielab.de/>

DFG Graduate School „Romanticism as a Model“

<http://modellromantik.uni-jena.de>

Friedrich Schiller University Jena, Germany

