

From TUSTEP to TEI in Baby Steps

Peter Fankhauser

Oliver Pfefferkorn, Andreas Witt

IDS-Mannheim

Mannheim Corpus of Historical Magazines and Newspapers

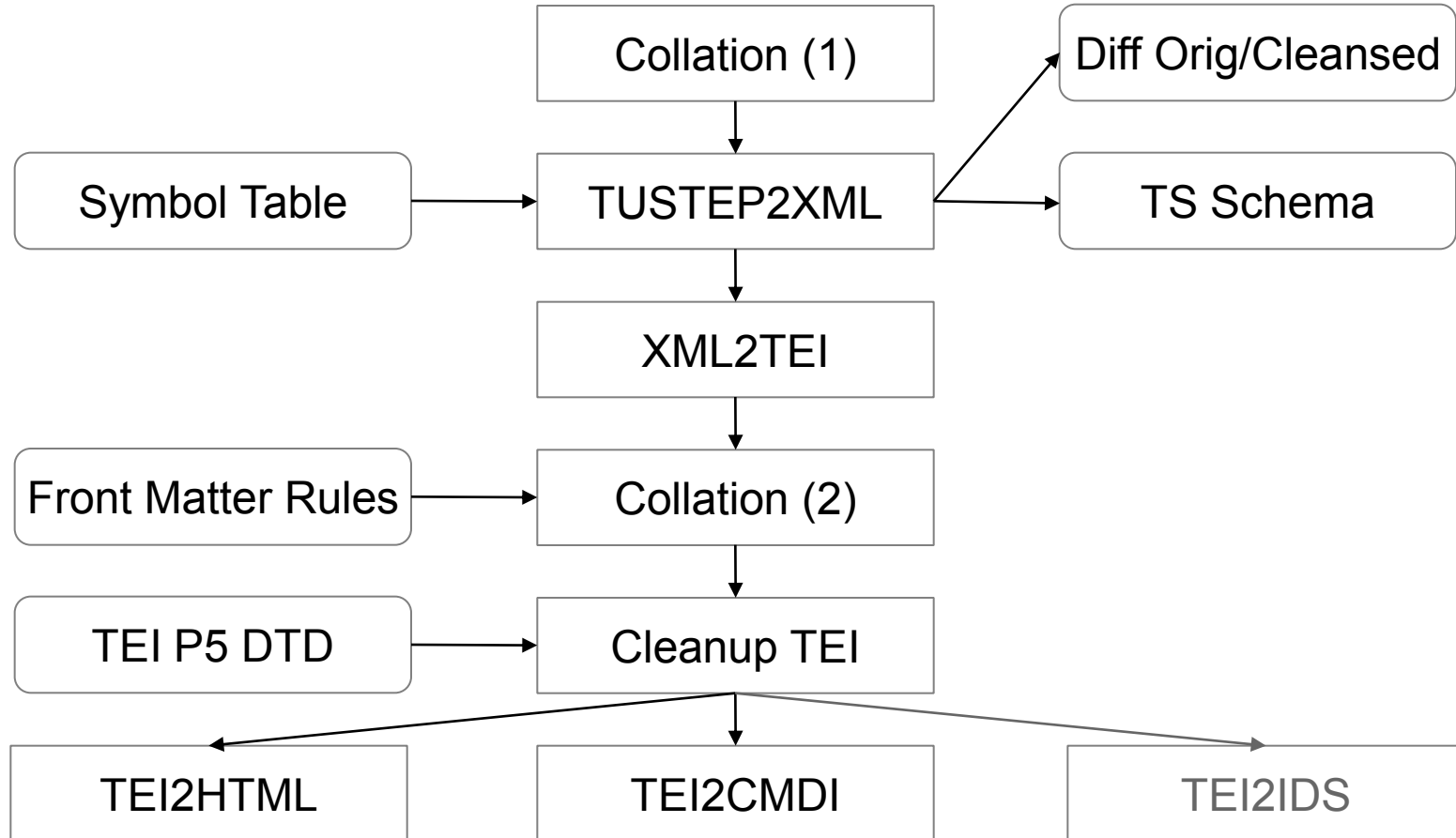
- 21 Magazines and Newspapers
- 652 Individual Volumes
- 4678 Pages (JPEG)
- 4.1 Mio Word Tokens
- Coverage
 - 18th and 19th century
 - Germany, Austria

TUSTEP

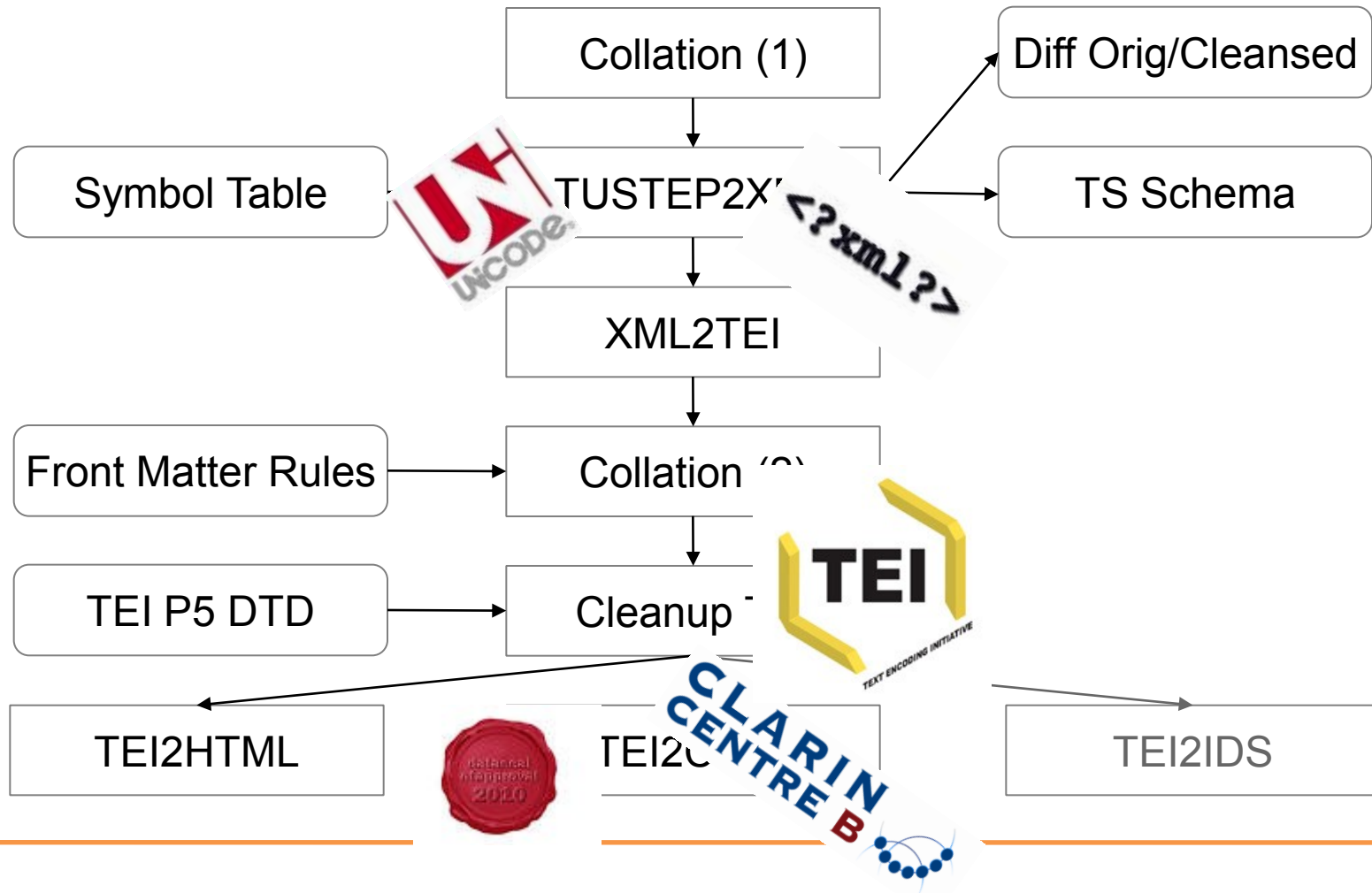
Issues

- Collation
 - Duplicates (about 1.5%)
 - M:N mapping between volumes and transcriptions
 - Missing transcriptions
- Transcription “according to TUSTEP conventions”
 - Diverse markup syntax: <P>, #S+, @@, %?i, {ri}, ...
 - Interleaved markup (typography, layout, logical structure)
 - Tag omission, errors

Overall Workflow



Collecting Seals...



TEI Encoding Choices

- Interleaved Markup
 - Columns vs. Paragraphs:
virtual joins
 - Hyphenation:
<w>...<lb break="no">
- Typographic Markup
 - <hi rend="..." style="...">
- Symbols
 - UTF-8 where available
 - Glyphs otherwise: <g>

```
<div rend="twocolumns">
  <div rend="column">
    <p rend="text-center">
      <lb n="14" />
      Der erste Artikel der Grundrechte des deutschen
      <lb n="15" />
      Volkes nach dem Entwurf der
      <w>
        Verfassungs=
        <lb n="16" break="no" />
        commission
      </w>
      <hi rend="superscript">1</hi>
    ).
  </p>
  ...
  Banus seine Stelle in die Hände des Kaisers nieder, we
  <milestone type="continuation" unit="absent" />
</div>
<div rend="column">
  <p xml:id="idm81066312" prev="#idm81069768">
    <milestone type="continuation" unit="absent" />
    <lb />
    im Fall er sich dem magyarischen Minister unterwerfen
    <lb n="80" />
    thun fest entschlossen ist, so ist der Bürgerkrieg zw
    Croatien
```

Implementation

- Workflow Orchestration, Validation
 - Perl + LibXML, LibXSLT
- Collation (1 & 2)
 - Perl + manual corrections & patterns
- TUSTEP2XML
 - Custom Event-Based Parsers (Perl)
- XML2TEI, TEI Cleanup, TEI2HTML
 - XSLT 1.0
- Schema generation for TUSTEPXML
 - Trang (James Clark)

Titel: Mainzer Journal.
Ort: Mainz
Datum: 22-06-1848
Nummer: 8
Seite: 1

Facsimile TEI Tustep

[1] Mainzer Journal.

2) Image

[3] Das Mainzer Journal erscheint täglich (mit Ausnahme der höchsten Festtage) und zwar so, daß das Hauptblatt mit den „Rheinischen Unterhaltungs-
[4] blättern“ schon am Vorabende, die ständige Beilage am Vormittage des betreffenden Tages selbst ausgegeben wird. Bestellungen nehmen alle Postämter an;
[5] für Mainz und die nächste Umgebung die Buchhandlung von Kirchheim, Schott und Thielmann am Leichhofe. Der Preis des Blattes ist hier in Mainz
[6] jährlich 8 fl. in vierteljährigen Vorausbezahlungen von 2 fl.; in dem gesammten Gebiete des Fürstlich Thurn- und Taxisschen Postbezirkes jährlich eben=
[7] falls 8 fl. Inserate aller Art werden aufgenommen und die dreispaltige Petitzeile oder deren Raum mit 3 kr. berechnet.

8) Image

9) N^{ro} 8. Donnerstag, den 22. Juni. 1848.

10) Image

[11]  Wegen des hohen Frohnleichnamstages erscheint morgen keine Zeitung. Die
[12] zu der heutigen Nummer gehörende Beilage wird Freitag Vormittag um
[13] 11 Uhr ausgegeben.

[14] Der erste Artikel der Grundrechte des deutschen
[15] Volkes nach dem Entwurf der Verfassungs=
[16] commission ¹⁾).

[17] H Vom Rheine 20. Juni. „Dem deutschen Volke ist ge=
[18] währleistet Freiheit des Bekenntnisses, vorbehaltlich der Bestra=
[19] fung von Verbrechen und Vergehen, welche bei Ausübung dieser
[20] Freiheit begangen werden, sowie vorbehaltlich aller staatsbürger=
[21] lichen Pflichten. Einer Anerkennung des Bekenntnisses durch
[22] den Staat bedarf es nicht. Für die Bekenner aller Religionen
[23] Gleichheit vor dem Gesetz. Es ist ausdrücklich die Bildung neuer

(...)leicht, daß in dem vagen Ausdruck „Freiheit des Bekenntnisses,“
[68] auch schon die religiöse und kirchliche Freiheit enthalten sey? so
[69] bedenke man, daß dasselbe Wort bisher in allen modernen Ge=
[70] setzen, wenigstens für die Bekenner der drei Hauptbekenntnisse ge=
[71] standen, aber die schreiendsten Eingriffe in die religiöse Freiheit
[72] nicht gehindert hat.
[73] Allein selbst diesem Ausdruck, „Freiheit des Bekenntnisses,“
[74] hat die Commission noch Clauseln und Bestimmungen angehängt,
[75] die, wenn sich nicht eine geheime, die volle Freiheit des Bekennt=
[76] nisses beeinträchtigende Nebenabsicht dahinter verbirgt, mindestens

```
- <TEI>
+ <teiHeader></teiHeader>
- <text>
- <front>
- <div>
  <pb n="1" facs="http://hdl.handle.net/10932/00-01B8-B175-5D03-A101-2"/>
  <milestone type="tustepsource" facs="http://hdl.handle.net/10932/00-01B8-B175-3573-A001-5" unit="absent"/>
- <p rend="text-center">
  - <hi rend="fractlarge">
    <lb n="1"/>
    Mainzer Journal.
  </hi>
  </p>
- <p>
  <lb n="2"/>
  <graphic/>
  </p>
- <p rend="posindent1">
  - <hi rend="fractsmall">
    <lb n="3"/>
    Das Mainzer Journal erscheint täglich (mit Ausnahme der
    höchsten Festtage) und zwar so, daß das Hauptblatt mit den
    „Rheinischen
  - <w>
    Unterhaltungs=
    <lb n="4" break="no"/>
    blättern
  </w>
    “ schon am Vorabende, die ständige Beilage am Vormittage des
    betreffenden Tages selbst ausgegeben wird. Bestellungen
    nehmen alle Postämter an;
```

Mannheimer Korpus Historischer Zeitungen und Zeitschriften

[de] Das Mannheimer Korpus Historischer Zeitungen und Zeitschriften besteht aus 21 deutschsprachigen Zeitungen und Zeitschriften des 18. und 19. Jahrhunderts und umfasst etwa 652 einzelne Ausgaben mit insgesamt über 4.1 Mio laufende Wortformen auf 4678 Druckseiten. Dieses Korpus wurde zwischen 2009 und 2011 zusammengestellt und digitalisiert, und 2013 von TUSTEP nach TEI P5 konvertiert.

[en] The Mannheim Corpus of Historical Newspapers and Magazines consists of 21 German newspapers and magazines from the 18th and 19th century. It comprises about 652 individual volumes with over 4.1 Mio word tokens on 4678 pages overall. This corpus has been assembled and digitized from 2009 to 2011, and been converted to TEI P5 in 2013.

Type: collection

PID: <http://hdl.handle.net/10932/00-01B8-AE41-41A4-DC01-5>

Publisher: Institut für Deutsche Sprache

- hasPart:** [Der allerneuesten Europäischen Welt- und Staats=Geschichte II. Theil \(1744\)](#)
- hasPart:** [Allgemeine Auswanderungs=Zeitung \(1846 - 1848\)](#)
- hasPart:** [Allgemeine Zeitung \(1871\)](#)
- hasPart:** [Arbeitgeber. Archiv für die gesammte Volkswirtschaft, Central-Anzeiger für Stellen- und Arbeitergesuche \(1870 - 1877\)](#)
- hasPart:** [Die Bayerische Presse \(1850\)](#)
- hasPart:** [Bayreuther Zeitungen \(1752\)](#)
- hasPart:** [Berlinische Nachrichten von Staats= und gelehrten Sachen \(1740 - 1741\)](#)
- hasPart:** [Berlinische Privilegirte Zeitung \(1737\)](#)
- hasPart:** [Conversations=Blatt zur Unterhaltung und Belehrung für alle Stände \(1836 - 1838\)](#)
- hasPart:** [Das Heller=Blatt oder Magazin zur Verbreitung gemeinnütziger Kenntnisse \(1834\)](#)
- hasPart:** [Das Pfennig=Magazin der Gesellschaft zur Verbreitung gemeinnütziger Kenntnisse \(1836 - 1855\)](#)
- hasPart:** [Das wohlfeilste Panorama des Universums zur erheiternden Belehrung für Jedermann und alle Länder \(1834 - 1836\)](#)
- hasPart:** [Deutsche Auswanderer=Zeitung \(1852\)](#)
- hasPart:** [EUROPA Wochenschrift für Kultur und Politik \(1905\)](#)

Mainzer Journal

Type: collection

PID: <http://hdl.handle.net/10932/00-01B8-B174-96E3-9C01-0>

Publisher: Institut für Deutsche Sprache

- isPartOf:** [Mannheimer Korpus Historischer Zeitschriften und Zeitungen](#)
- hasPart:** [Mainzer Journal, Nr. 8, 22-06-1848](#)
- hasPart:** [Mainzer Journal, Nr. 10, 25-06-1848](#)
- hasPart:** [Mainzer Journal, Nr. 13, 28-06-1848](#)
- hasPart:** [Mainzer Journal, Nr. 15, 30-06-1848](#)
- hasPart:** [Mainzer Journal, Nr. 16, 01-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 19, 04-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 22, 07-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 26, 11-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 29, 14-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 30, 15-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 34, 19-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 37, 22-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 39, 24-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 42, 27-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 43, 28-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 45, 30-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 46, 31-07-1848](#)
- hasPart:** [Mainzer Journal, Nr. 47, 01-08-1848](#)
- hasPart:** [Mainzer Journal, Nr. 48, 02-08-1848](#)
- hasPart:** [Mainzer Journal, Nr. 49, 03-08-1848](#)

Mainzer Journal, Nr. 10

Type: text

PID: <http://hdl.handle.net/10932/00-01B8-B176-54E3-A701-7>

Publisher: Institut für Deutsche Sprache

- isPartOf:** [Mainzer Journal](#)
- source:** [HTML](#)
- source:** [TEI](#)
- source:** [TUSTEP](#)
- source:** [TUSTEP](#)
- source:** [MainzerJournal/MainzerJournal\(1\)0007.jpg](#)
- source:** [MainzerJournal/MainzerJournal\(1\)0008.jpg](#)
- source:** [MainzerJournal/MainzerJournal\(1\)0009.jpg](#)
- source:** [MainzerJournal/MainzerJournal\(1\)0010.jpg](#)
- source:** [MainzerJournal/MainzerJournal\(1\)0011.jpg](#)
- source:** [MainzerJournal/MainzerJournal\(1\)0012.jpg](#)

Metadata Information

Creator: generateCMDI.pl
 Creaton date: 2013-06-23
 Profile: clarin.eu:c1_p_1366895758244
 Collection: Institut für Deutsche Sprache, CLARIN-D Zentrum, Mannheim

Titel: Mainzer Journal.
 Ort: Mainz
 Datum: 25-06-1848
 Nummer: 10
 Seite: 7 [Facsimile](#) [TEI](#) [Tustep](#)

[1] Mainzer Journal.

[2] [Image](#)

[3] Das Mainzer Journal erscheint täglich (mit Ausnahme der höchsten Festtage) und zwar so, daß das Hauptblatt mit den „Rheinischen Unterhaltungs=blättern“ schon am Vorabende, die ständige Beilage am Vormittage des betreffenden Tages selbst ausgegeben wird. Bestellungen nehmen alle Postämter an; für Mainz und die nächste Umgebung die Buchhandlung von Kirchheim, Schott und Thielmann am Leichhofe. Der Preis des Blattes ist hier in Mainz jährlich 8 fl. in vierteljährigen Vorausbezahlungen von 2 fl.; in dem gesammten Gebiete des Fürstlich Thurn= und Taxisschen Postbezirkes jährlich eben= falls 8 fl. Inserate aller Art werden aufgenommen und die dreispaltige Petitzelle oder deren Raum mit 3 kr. berechnet.

[6] [Image](#)

[9] N^{ro} 10.

Sonntag, den 25. Juni.

1848

[10] [Image](#)

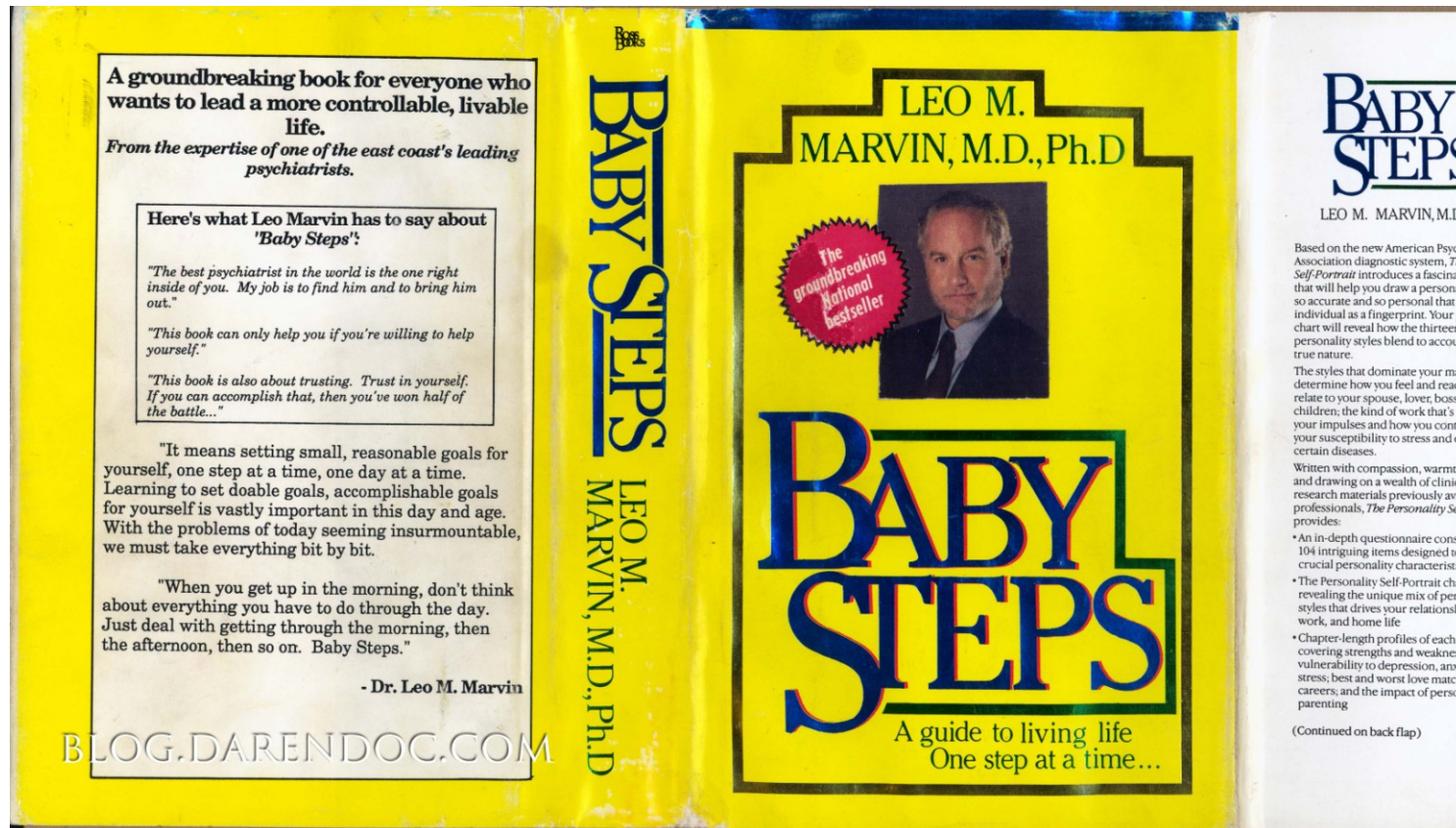
[11] Die absolute Demokratie und die Geschichte.

[12] J Mainz 23. Juni. Das, worauf eine durch Zahl ansehn=liche und durch ihre Talente und ihre Energie kräftige Partei lossteuert, ist die absolute Demokratie und eine einige und theil= bare Republik für ganz Deutschland. Hierin allein, glauben jene Männer, könne die Freiheit vollständig verwirklicht und die Größe Deutschlands nach Außen begründet werden. Was das letztere betrifft, so würde in der That vielleicht Deutschland in der gewaltigen Aufregung, in welche es dadurch sich versetzt fände, eine ähnliche Kraft nach Außen entwickeln, wie das republikanische Frankreich am Ende des vorigen und am Anfang dieses Jahrhunderts. Was aber das Eude der einigen absolut demokratischen deutschen Republik seyn würde, darüber mag die Geschichte uns belehren. Bei den Männern freilich, die man die Männer des „blinden“ Fortschritts nennen könnte, weil sie gleichsam mit verbundenen Augen voranstürmen, ohne rück=

(...)alle Gewalt in der Weise in den Händen des Volkes liegt, daß dasselbe rein nach der Kopfhalt alles entscheidet und regiert oder durch eine Versammlung, die es durch eine ebenso nach der Kopf= Zahl berechnete Majorität gewählt hat, beschließen und regieren läßt, so zwar, daß die von der Majorität gefaßten Beschlüsse das einzige und höchste Gesetz sind. Dieses ist jener majestätische, unumschränkte, allmächtige Volkswille, der Wille der 40 Millionen, von denen Hecker in seinem Manifest an das Parlament mit einer Art abgöttischer Schwärmerei spricht. Eine solche absolute Demokratie aber, sagen wir, hat nie auf die Dauer bestanden. In allen antiken Republiken war in einem Senat, mochte er heißen wie immer, in alten adeligen Geschlechtern ein höchst zähes, tief in die Geschichte und das Leben verwachsenes conservatives und aristokratisches Element vorhanden, das dem leicht beweglichen und fluthenden Willen der Menge gegenüber die Wage hielt. Außerdem waren die Magistrate, die Consuln, Prätores des alten Roms nichts weniger als bloße Beamten, son=

Lessons Learned, Recommendations

- Define useful, realistic, verifiable goals
- Standards early: Easily said
 - Tools: Edit, Verify, Process
 - Ecosystem
- Quality Assurance
 - Baby steps: Modular, verifiable workflows
 - Many eyes (beyond double keying)
 - Documentation
- Posthoc Cleansing
 - Summarize: markup, symbols, structure,...
 - Declarative: mapping tables, rules, grammars
 - Documentation: steps, diffs, mapping tables, ...
- No Free Lunch.



From: What about Bob?