GeoBib

GEOREFERENZIERTE ONLINE-BIBLIOGRAPHIE FRÜHER HOLOCAUST- UND LAGERLITERATUR

Annotationsschema für die bibliographischen Metadaten, die inhaltliche Verschlagwortung sowie die Geodaten

M4.1S

Bastian Entrup, Charlotte Kitzinger

Juli 2013

INHALT

1. Datenanalyse 2
2. Anforderungen an ein Annotationsschema 2
   2.1. (Erweiterte) bibliographische Daten 2
   2.2. Personendaten 3
   2.3. Ortsdaten 5
3. XML-Annotationsschema und Wiki-System 5
   3.1. TEI5 5
   3.2. XML und SVN 6
   3.3. MediaWiki 7
   3.4. Wiki-System und SVN 7
4. Graphische Nutzeroberfläche: Autoren-Modus 9
5. Ausblick 11
6. Anhang 12

1. DATENANALYSE

Ausgehend von den von der Arbeitsstelle Holocaustliteratur (AHL) gesammelten Daten zu einigen Beispieltexten sowie von mehreren Gesprächen und Treffen zwischen der Angewandten Sprachwissenschaft und Computerlinguistik (ASCL) und der AHL, konnten Anforderungen an ein Annotationsschema definiert werden. Neben der Erfassung von klassischen bibliographischen Daten, z.B. Titel, Autor, Verlag usw., sind hier die Besonderheiten des GeoBib-Projektes von entscheidender Bedeutung: Die Erfassung zusätzlicher biographischer und geographischer Daten sowie die Kombination von geographischen, temporalen, biographischen und bibliographischen Daten stellen besondere Anforderungen an ein Annotationsschema.

Von der ursprünglichen Annahme, man könnte all diese Daten gut in einer XML-Datei verbinden, musste auf Grund der Arbeitsweise im Projekt (viele Mitarbeiter aus verschiedenen Teilprojekten müssen mit denselben Daten arbeiten und haben evtl. verschiedene Ansprüche an diese Daten) sowie dem Umfang des Projektes (in 25 Texten konnten über 7000 verschiedene Personen- oder Ortsreferenzen festgestellt werden, die auf über 4000 verschiedene Entitäten, also Personen oder Orte, verweisen) verzichtet werden. Stattdessen wurden Wege gesucht und gefunden, die Daten eindeutig zu halten und Redundanzen zu vermeiden und dabei gleichzeitig den Workflow zu unterstützen.

Neben den klassischen XML-Annotationen umfasst das System deshalb auch ein Wiki zum Erfassen, Bearbeiten und Bereitstellen von personen- und ortsbezogenen Informationen. *Subversion* (SVN) dient der Versionskontrolle sowie dem Austausch der Daten. Eine graphische Nutzeroberfläche unterstützt das Team beim Erstellen der XML-Dateien.

Einige für das Projekt als wichtig angesehene Punkte wurden während der Phase der Datenanalyse wie folgt bewertet: Inhaltliche Schlagworte zu Texten (Keywords) wurden zwar begrüßt und gewünscht, stellten sich jedoch als durchaus nicht trivial heraus. Ein eigenes Schema zur Verschlagwortung von Texten zu finden wurde daher abgelehnt. Stattdessen soll die Möglichkeit geprüft werden, solche Informationen aus Verbundkatalogen o.ä. mit öffentlichem *Application Programming Interface* (API) zu beziehen. Durch eine Verbindung zu diesen Katalogen soll außerdem von der späteren Web-Plattform aus eine direkte Suche in Bibliotheken gestartet werden können.

2. ANFORDERUNGEN AN EIN ANNOTATIONSSCHEMA

2.1. (ERWEITERTE) BIBLIOGRAPHISCHE DATEN

Zu den klassischen bibliographischen Daten gehören die folgenden Angaben:

- Titel und Untertitel,
- Autor(en) und Herausgeber,
- der Verlag,

- der Ort der Veröffentlichung,
- das Veröffentlichungsjahr,
- die Auflage und
- der Seitenumfang.

Zusätzlich wurden die folgenden Daten ausgemacht, die entweder von besonderem Interesse sein können, oder aber den historischen Umständen geschuldet sind:

- Angaben zur Drucklizenz durch die Siegermächte (wenn nach 1945 in Deutschland veröffentlicht),
- Angaben zur Auflagenhöhe der bearbeiteten Ausgabe (meist die erste, sofern diese nicht verfügbar ist auch eine spätere), zur Gesamthöhe aller Auflagen sowie die Anzahl der Auflagen insgesamt,
- Informationen zum Verkaufspreis zur Zeit der Veröffentlichung,
- das Genre,
- ein von der AHL bzw. dem Herder Institut erstelltes Abstract,
- eine kurze Werkgeschichte mit Quellenangaben.

Zusätzlich ist ein Feld zum Eintragen einer ISBN vorhanden[1].

Außerdem wurde die Möglichkeit vorgesehen, Angaben und Informationen zu Graphiken und Abbildungen oder zur Umschlaggestaltung zu machen.

2.2. PERSONENDATEN

Eine Besonderheit des GeoBib-Projektes ist die Verknüpfung verschiedener Daten. Dazu gehören auch biographische Angaben zu Autoren und Herausgebern. Mit Blick auf die zukünftige Entwicklung des Projektes sollte es in derselben Weise auch möglich sein, weitere textrelevante Personen zu erfassen und zu beschreiben. Neben dem Namen und Vornamen werden die folgenden persönlichen Angaben gesammelt: Geschlecht, Geburtstag und -ort, Sterbetag und -ort, Namensvarianten und Pseudonyme, ein Fließtext mit biographischen Angaben für Autoren und Herausgeber, Angaben zu Quellen und Sekundärliteratur sowie ggf. verschiedene Bild- und PDF-Dateien.

[1] ISBN-Nummern wurden erst nach dem im Projekt betrachteten Zeitraum eingeführt. Die Angabe ist der Vollständigkeit halber und um das Annotationsschema auch für eine Weiterführung in einem anderen Zeitraum zu ermöglichen, dennoch nötig.

Personen werden in einem relativ ausführlichen System mit Kategorien getaggt:

Kategorie	**Unterkategorie**	**Beschreibung**
Autor		Autor einer Monographie oder eines Beitrages, Gedichtes etc., auch innerhalb eines Sammelbandes oder einer Anthologie
Herausgeber		Herausgeber einer Monographie oder eines Sammelbandes
NS-Funktionsträger		Person mit besonderen Befugnissen und Funktionen innerhalb des nationalsozialistischen Systems, etwa Reichspropagandaminister, aber auch Obersturmführer
KZ-Personal		Wachpersonal oder administratives Personal in einem KZ (keine Funktionshäftlinge)
Gefängnis-Personal		Wachpersonal oder administratives Personal in einem Gefängnis
KZ-Häftling		einfacher Häftling in einem KZ, ohne Funktion und Vergünstigungen
	Funktionshäftling	Häftling mit bestimmten Funktionen im KZ wie Blockältester oder Häftlingspfleger, ist immer auch als einfacher Häftling aufzuführen
NS-Gettofunktionär		(deutsches) Personal außerhalb der internen Gettoverwaltung
Gettobewohner		Gettobewohner: Einfacher Gettobewohner, ohne Funktionen und Befugnisse
	Mitarbeiter der jüdischen Gettoverwaltung	Personen der internen Selbstverwaltung des Gettos mit bestimmten Funktionen und Befugnissen
Angehöriger des Widerstands		Personen, die aktiv am Widerstand teilnehmen und Widerstandsgruppen angehören
Künstler		z.B. literarische, musische oder darstellende Künstler wie Johann Wolfgang von Goethe, Shakespeare etc.
literarische Figur		fiktive Figuren aus der Literatur, z.B. Sherlock Holmes, Goethes Faust etc.
mythologische/religiöse Figur		Figuren aus der Religion oder Mythologie, z.B. Gott, Adam und Eva, Zeus etc.
historische Person		bekannte und herausragende Personen der Geschichte, z.B. Otto von Bismarck, Friedrich der Große
Zeitgenosse		Personen, die zeitgleich mit den Personen oder Figuren eines Textes leben
Unbekannt		Personen, die keiner Kategorie zugeordnet werden können, aber namentlich erwähnt werden

Personendaten müssen außerdem mit Ortsreferenzen sowie temporalen Angaben versehen werden können.

2.3. ORTSDATEN

Orte werden zur eindeutigen Identifizierung neben ihrem Namen außerdem mit Angaben zum Land, zum Bezirk und ihrem Typ (Stadt, Region, Land, Lager o.ä.) versehen. Zusätzlich kann eine Zeitangabe gemacht werden, um Regionen oder andere Orte, die ihren Namen oder ihre Lage im Laufe der Zeit geändert haben, zu identifizieren. Ein Kategorisierungssystem für die Ortsdaten wird momentan erarbeitet (siehe dazu auch das Meilensteindokument M3.1R und M5.1R).

Auf Grundlage dieser Daten kann später eine (teil-)automatische Lokalisierung auf einer Karte erfolgen.

3. XML-ANNOTATIONSSCHEMA UND WIKI-SYSTEM

3.1. TEI5

Gegenüber anderen Systemen zur Erfassung von bibliographischen Daten, wie z.B. MARC21/XML oder DublinCore, bietet die Text Encoding Initiative (TEI) durch ihre Variabilität und Anpassungsfähigkeit einige Vorteile: Die TEI stellt ein XML-Schema, bzw. Module zur Erstellung eines solchen Schemas, zur Verfügung, um unter anderem (literarische) Texte genau beschreiben zu können. Viele der oben genannten Anforderungen an die Erfassung von bibliographischen Daten stellt die TEI schon in ihrer Grundversion zur Verfügung. Für einige weitere gibt es Erweiterungen oder jederzeit die Möglichkeit, das Vokabular selbst zu ergänzen.

Das Vorgehen im GeoBib-Projekt (es werden nicht ganze Texte annotiert, sondern nur Informationen über die Texte gesammelt) erfordert, dass alle Informationen im *teiHeader*-Element verfügbar sind. Das *text*-Element der TEI wird nicht benötigt. So müssen Angaben zu Personen und Orten in Form von Listen im *teiHeader* verfügbar gemacht werden. Dafür muss das TEI-Schema entsprechend erweitert werden, um Einträge wie in Abb. 1 zu ermöglichen.

```
<particDesc>
    <listPerson>
        <person xml:id="FilipFriedman" role="author">
            <ref target="http://wiki.geobib.info/index.php/Filip_Friedman">Filip
Friedman</ref>
            <note>Autor des Vorwortes </note>
        </person>
        <person xml:id="GerszonTaffet" role="author">
            <ref target="http://wiki.geobib.info/index.php/Gerszon_Taffet">Gerszon
Taffet</ref>
            <note>Autor der Einführung </note>
        </person>
        <person xml:id="WaltervonBrauchitsch" role="undef">
            <ref target="http://wiki.geobib.info/index.php/Walter_von_Brauchitsch">Walter
von Brauchitsch </ref>
            <note>Generalfeldmarschall</note>
        </person>
```

Abbildung 1: Personenliste gemäß GeoBib-TEI-Schema.

Die Verbindung von bibliographischen mit geographischen und biographischen Daten bringt jedoch einige Probleme mit sich, die mit Hilfe der TEI nicht so leicht zu lösen sind.

Personen können in verschiedenen Texten erwähnt werden. Dies ist sogar relativ häufig der Fall. Daher ist es nicht ratsam, biographische Daten in den jeweiligen XML-Dateien, die je einen Text repräsentieren, zu speichern. Die Daten würden dadurch redundant gespeichert und unterlägen der Gefahr der Inkonsistenz (Heterogenität) der Angaben. Ähnliches gilt natürlich auch für Orte.

Es muss also eine Art Repositorium oder Datenbank verwendet werden, um jeder Orts-/Personen-Entität eine ID/URI zuzuweisen und diese eindeutig identifizieren und referenzieren zu können. Eine klassische Datenbank hat dabei den Nachteil, dass sie unnötig kompliziert zu bedienen ist und daher den Workflow erheblich stören könnte. Dennoch bedarf es eines Systems, das die Daten strukturiert speichern und zur Verfügung stellen kann. Beide Anforderungen – die Bedienbarkeit sowie die geforderte Struktur, die eine automatische Verarbeitung ermöglicht– bietet das von uns gewählte MediaWiki-System.

3.2. XML UND SVN

Die Arbeit im Projekt erfordert ein hohes Maß an Kollaboration zwischen Mitarbeitern, die oftmals räumlich weit voneinander getrennt sind. Die Arbeit an den XML-Dateien ist davon in besonderem Maße betroffen. Das TEI-Schema muss erwartungsgemäß regelmäßig überprüft und gegebenenfalls angepasst werden. Gleiches wird auch für die graphische Be-

nutzeroberfläche gelten. Außerdem werden XML-Dateien nicht immer nur von einem Mitarbeiter bearbeitet. Neben Fragen der Datensicherung ergeben sich hier Probleme der Dateirechteverwaltung und der Versionskontrolle.

Diese Probleme werden im Projekt durch die Verwendung von *Subversion* (SVN) gelöst. SVN wird auf einem Server installiert und stellt dort den Nutzern ein Repositorium zur Verfügung, in dem Dateien gespeichert werden können. Änderungen werden auf den Server übertragen und mit den anderen Benutzern abgeglichen.

Ordnern und Dateien können in SVN Dateirechte zugewiesen werden, die den Kreis der Personen, die diese z.B. ändern können, einschränken. Dies ist besonders für das bereitgestellte Schema aber auch für Vorlagen u. ä. wichtig.

Bei der Aktualisierung der lokalen Arbeitskopie kann sich der Nutzer zeilenweise die Änderungen bzw. Unterschiede zwischen den verschiedenen Versionen ansehen und erkennt so, an welchen Stellen eines Dokuments welcher Nutzer Änderungen vorgenommen hat.

3.3. MediaWiki

Das MediaWiki-System bietet für die Projektmitarbeiter den Vorteil, dass das Bearbeiten der Dokumente relativ einfach ist und das Ergebnis übersichtlich in einer von z.B. Wikipedia bekannten Art und Weise dargestellt wird, wie in Abb. 2 zu sehen ist.

Die Arbeit mit sogenannten Templates erleichtert nicht nur das Eingeben der Daten, es ermöglicht auch eine einfachere maschinelle Verarbeitung. Templates können Vorlagen zur Eingabe von Daten in bestimmte Felder bereitstellen oder aber die Eingabe von Daten, z.B. Namensvarianten, erheblich vereinfachen.

Jede Seite im Wiki repräsentiert eine Entität und kann über einen eindeutigen URI angesprochen und referenziert werden.

3.4. Wiki-System und SVN

Zwischen den XML-Dateien auf der einen Seite und den Wiki-Seiten auf der anderen besteht eine viele-zu-viele Relation: Jede XML-Datei kann auf verschiedene Wiki-Seiten verweisen und jede Wiki-Seite kann in verschiedenen XML-Dateien verwendet werden.

Die Administration des Wikis erfordert daher besondere Vorsicht, sollte man Änderungen an den URI, also den Seitennamen im Wiki, vornehmen. Solche Fälle können auftreten, wenn sich nach dem Anlegen einer Seite herausstellt, dass der Seitenname mehrdeutig oder falsch ist. In Fällen, in denen Wiki-Seiten umbenannt oder verschoben werden müssen, muss zusätzlich geprüft werden, ob diese Seiten in XML-Dateien referenziert werden, um *broken links* zu vermeiden.

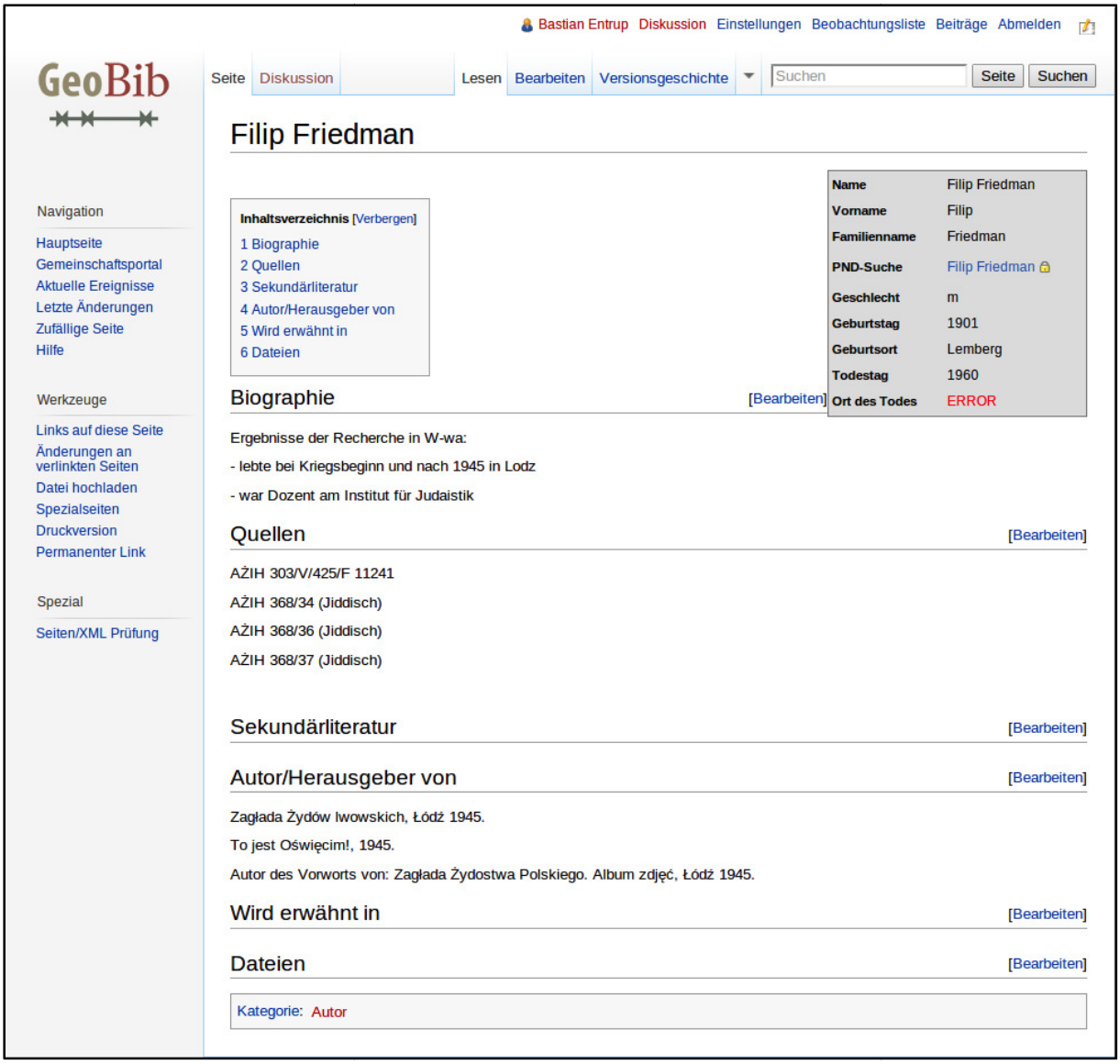

Abbildung 2: Personeneintrag im Wiki

Ebenso muss auch das einzelne XML-Dokument auf seine Konsistenz hinsichtlich der Verweise überprüft werden können. Während Ersteres durch eine Funktion im Wiki übernommen wird (s. Abb. 3.), die für das Projekt hinzugefügt wurde, wird die zweite Untersuchung durch eine Funktion in der graphischen Nutzeroberfläche übernommen.

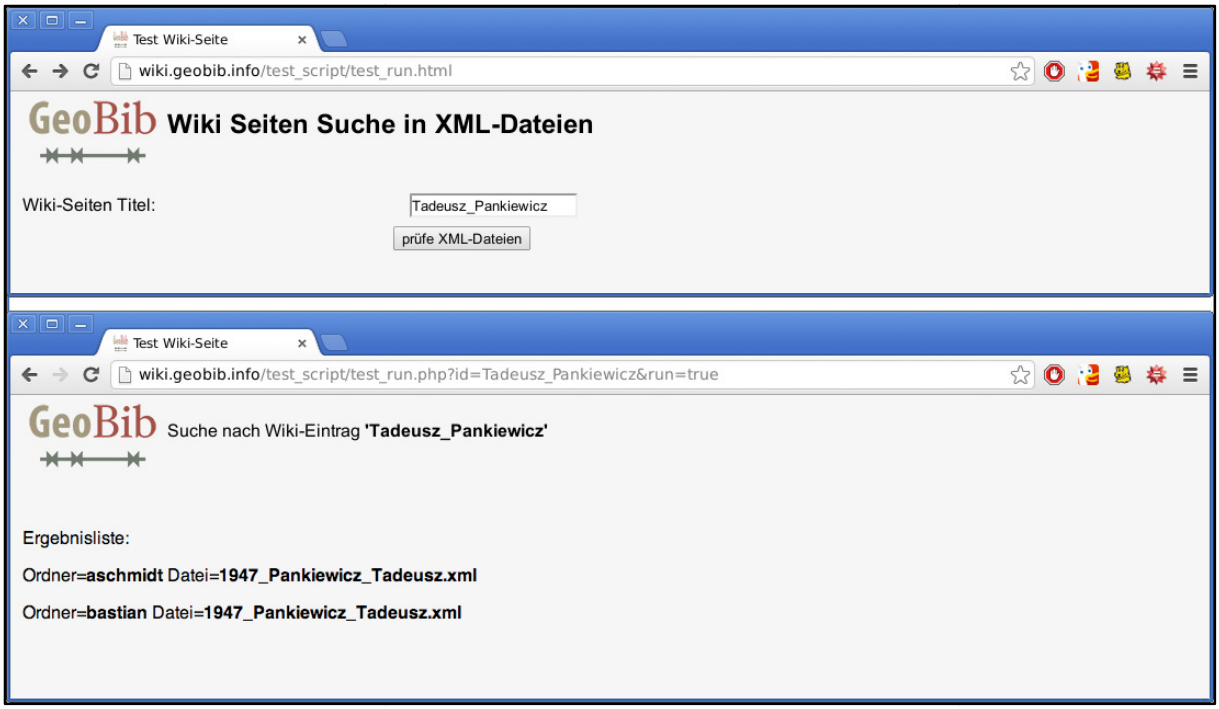

Abbildung 3: Wiki-Seiten-Suche in XML-Dateien

4. Graphische Nutzeroberfläche: Autoren-Modus

Zum Editieren der XML-Dateien wird oXygen-XML verwendet. Um die Arbeit mit diesem XML-Editor einfacher und intuitiver zu gestalten, wurde ein Autoren-Modus entworfen, der dem Mitarbeiter die XML-Datei nicht in Form ihrer XML-Syntax, sondern unterstützt durch *Cascading Style Sheets* (CSS) graphisch aufbereitet in Form von Feldern und Tabellen darstellt. Dies ist in Abb. 4 zu sehen.

Während dies die Scheu vor der XML-Syntax nehmen kann, ergibt sich der wirkliche Vorteil in der Verwendung von den von oXygen bereitgestellten Java-APIs, die es ermöglichen, komplexe Funktionen und Annotationsvorgänge zu automatisieren oder zu überwachen.

Die Eingabe von Personen-Referenzen ist z.B. ein Arbeitsschritt, der relativ komplex ist. Zuerst muss die passende Wiki-Seite gefunden werden. Dann muss die URL kopiert werden und der entsprechende Eintrag in der korrekten Syntax in XML angelegt werden. Jeder

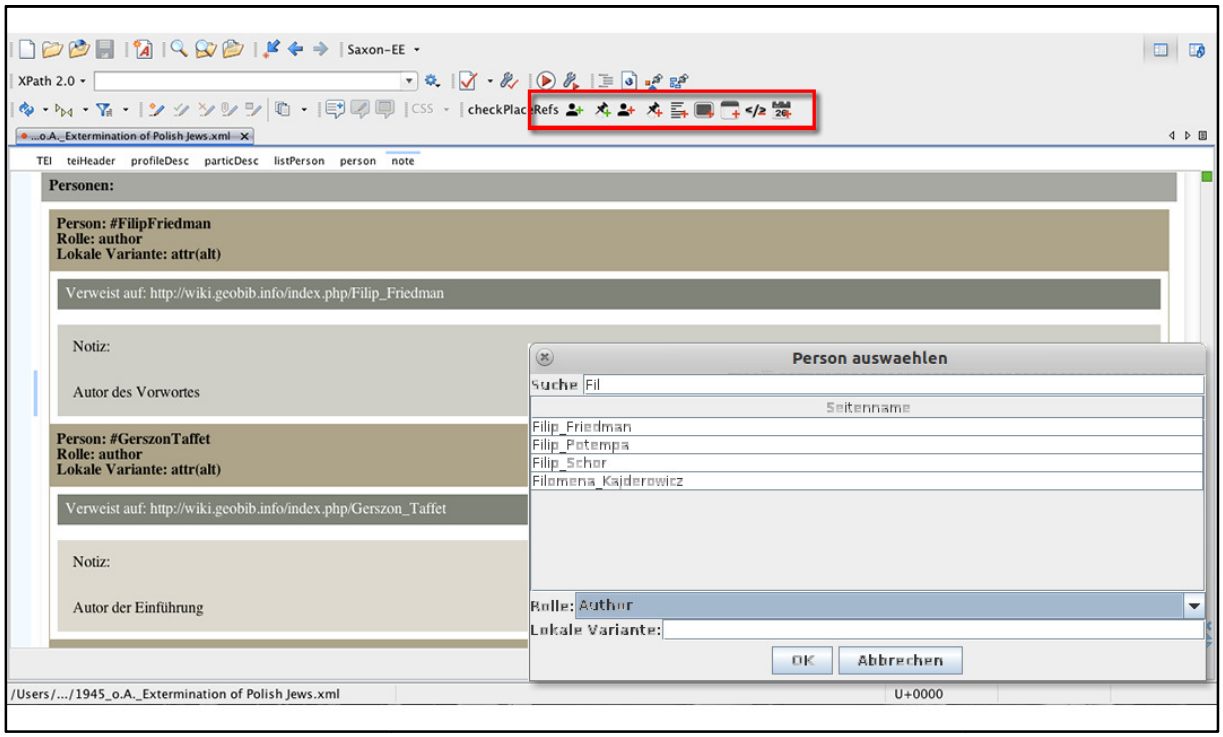

Abbildung 4: Autorenmodus

Eintrag erhält dann eine interne ID. Außerdem können weitere Attribute gesetzt werden.

Im Autorenmodus wird dieser Vorgang dahingehend vereinfacht, dass man durch Anklicken einer Schaltfläche eine Liste mit allen verfügbaren Wiki-Seiten angezeigt bekommt. Diese Liste lässt sich durch eine Nutzereingabe einschränken und sortieren. Ist der benötigte Eintrag gefunden, können im selben Arbeitsschritt die nötigen Attribute gesetzt werden. Die Vergabe der ID übernimmt die Software. Ebenso wird ein valider, mit dem Schema konformer Eintrag erzeugt und an der richtigen Stelle eingefügt. Ähnlich funktioniert auch die Eingabe der Orte.

Wie oben angedeutet, werden die Referenzen von der Software automatisch auf ihre Validität geprüft. Zusätzlich wird der Autor bei der Eingabe von Fließtexten sowie dem Editieren von Einträgen, z.B. deren Löschung usw., unterstützt.

Während die Arbeit mit dem Wiki einen Internetanschluss voraussetzt, können die hier beschriebenen Arbeitsschritte auch in einem Offline-Modus auf Reisen durchgeführt werden.

5. AUSBLICK

Das Schema betreffend sind noch folgende Einschränkungen zu machen: Das Schema stellt immer einen *status quo* dar, der sich jedoch ständig ändern kann. Gemeint sind hier weniger Änderungen im großen Stil, als vielmehr die Anpassung des Schemas an

a) neue Gegebenheiten oder
b) vorher unbekannte oder nicht bedachte Problemstellungen sowie
c) die Anpassung des Schemas an Wünsche und Bedürfnisse der Projektbeteiligten.

Durch die Arbeit mit den Primärtexten können jederzeit neue Probleme entstehen bzw. entdeckt werden, denen dann durch Änderungen im Schema begegnet werden muss. Des Weiteren sind auch zum jetzigen Zeitpunkt schon einige Probleme bekannt, die das Schema in seiner aktuellen Form noch nicht löst.

Hierbei ist beispielsweise die Erfassung von Anthologien, Gedichtbänden oder auch von Zeitungsartikeln zu nennen. Ein Gedicht oder ein Zeitungsartikel kann mit Recht als eigenständiger Text betrachtet werden. Die Verbindung zwischen einem Sammelband und den in ihm verfügbaren Texten ist ein solches bisher ungelöstes Problem. Die Anzahl möglicher Beispiele ist zu diesem Zeitpunkt noch sehr gering, so dass noch keine Anstrengungen unternommen wurden, diese Verbindung zu modellieren. Ähnliches gilt für die Verlinkung von Bildern/Scans und Texten. Eine solche Verbindung ist im Wiki schon relativ einfach gelöst worden. Für die Primärtexte steht allerdings noch nicht fest, in welchem Umfang solche Daten bereitstehen, distribuiert werden können und erfasst werden, so dass auch dieser Punkt vorerst ausgeklammert bleibt.

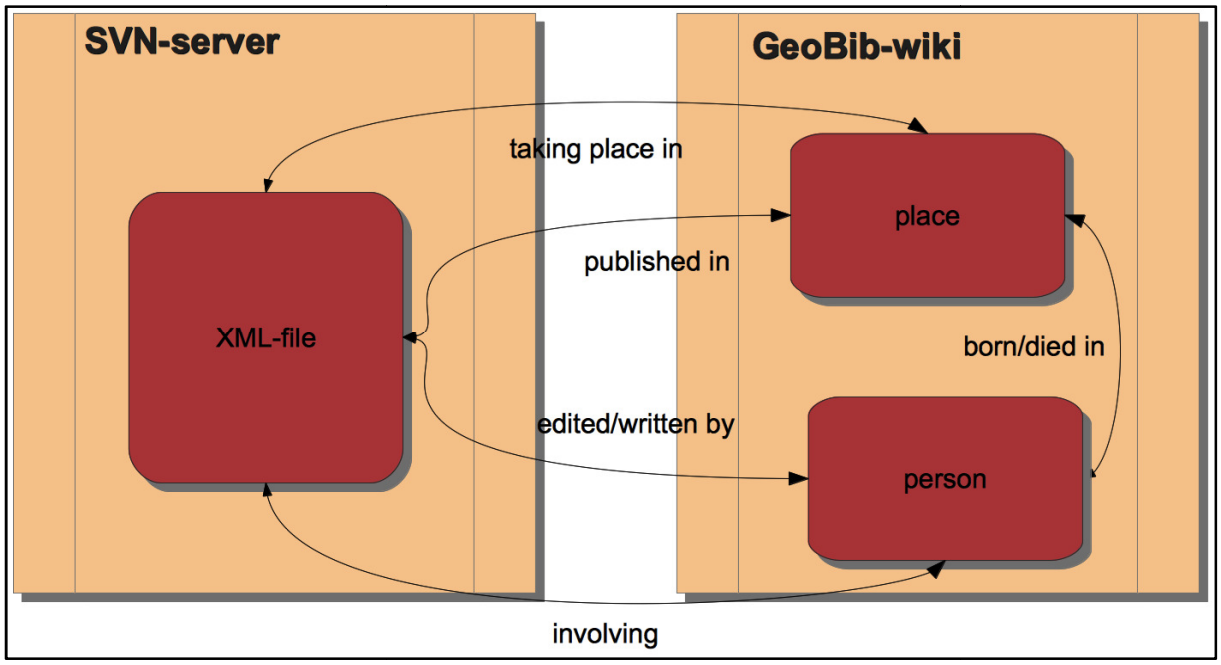

Abbildung 5: (geplanter) Workflow im GeoBib-Projekt

Aus der Verwendung der oben beschriebenen Software, die sich wiederum aus einer Analyse der zu bearbeitenden Daten ergibt, resultiert neben einer Trennung von Daten, die die Primärtexte bzw. Personen und Orte betreffen, ein besonderer Workflow (Abb. 5). Dieser sieht zum einen die Erstellung von XML-Dateien vor, wurde darüber hinaus aber um die Verwendung eines Wikis erweitert und wird durch die Verwendung von SVN zum Datenaustausch unterstützt.

Aus dem täglichen Umgang mit dem vorgeschlagenen System wird sich auch in diesem Bereich gewiss weiterer Optimierungsbedarf ergeben, der wie die oben genannten offenen Punkte im Bereich des Schemas in der noch folgenden Projektlaufzeit durchgeführt und angepasst werden soll.

6. ANHANG

Kurz_beispiel.xml:	Beispiel-Datei, die das vorliegende Schema verwendet
Vorlage_simpel.xml:	Vorlage-Dokument zur Verwendung im Projekt
TEIgeobib_strict.doc.pdf:	Ausführliches Handbuch zum Annotationsschema