XML-Workshop für Editionsprojekte

Christian Sonder M. A.

Wiss. Mitarbeiter SSRQ, Universität St. Gallen

Trilog-Verlag für Kunst, Literatur und Wissenschaft

Mail: christian.sonder@unisg.ch oder christian.sonder@trilog-verlag.de

Ziele dieses Kurses

1. Sitzung

1) Was ist XML?
2) Wofür benutzt man XML?
3) Was sind die zentralen Konzepte und Grundbegriffe?
4) XML-Syntax, Teil 1: Elemente, Attribute

Ziele dieses Kurses

2. Sitzung

1) Wie ist ein vollständiges XML-Dokument aufgebaut?
2) Wie kodiert man einfache Sachverhalte mit XML?
3) XML-Syntax, Teil 2: Escapen, XML-Deklaration, Kommentare, CDATA
4) XML und Unicode, Character References

Ziele dieses Kurses

3. Sitzung

1) Was ist die Text Encoding Initiative (TEI)
2) Wohlgeformtheit vs. Validität
3) XML-Syntax, Teil 3: Processing Instructions
4) Verknüpfung mit dem TEI-all-Schema
5) Umgang mit dem Oxygen XML Editor

Definition von XML

XML eXtensible Markup Language
erweiterbare Auszeichnungssprache

- XML ist eine künstliche Sprache mit Vokabeln, einer Syntax und einer Semantik.
- XML ist keine Programmiersprache, sondern eine Auszeichnungssprache! Sie dient dem Strukturieren, Klassifizieren, mit Metadaten Annotieren und Kommentieren von Daten.
- XML ist beliebig erweiterbar, solange die syntaktischen Regeln nicht gebrochen werden. D. h. man kann neue Vokabeln (Dialekte) einführen.

Einsatzgebiete von XML

Art der Daten	Beispiel
Text	Text Encoding Initiative (TEI)
Musik	Music Encoding Initiative (MEI)
Geodaten	Open Street Map, KML, GPX
Finanzdaten	Interbanken- und Zahlungsverkehr, SEPA
Vektorgrafiken	SVG
Webseiten	XHTML
Web-Feeds	RSS
u. v. m.	

XML als technischer Standard

- XML ist ein technischer Standard mit einer offiziellen Dokumentation:
  https://www.w3.org/TR/xml/
  Es gibt korrektes, d. h. dem Standard entsprechendes, und nicht korrektes XML.
- XML wird entwickelt und betreut vom World Wide Web Consortium (W3C),
  dem wichtigsten Gremium für die Entwicklung der Internettechnologien:
  https://www.w3.org/
- Aktuellste Version des Standards: XML Version 1.0 (Fünfte Auflage) von 2008
  (die 1. Auflage erschien 1998).
- XML Version 1.1 (2. Auflage 2006) hat sich nicht durchgesetzt und gilt als überflüssig.

XML und die X-Technologien

- Um mit XML arbeiten zu können, gibt es zahlreiche weitere, ebenfalls in Form von Standards definierte Technologien:
- **XML Schema** = Sprache zur Beschreibung von XML-Dokument-Klassen
- **XPath** = Sprache zur Navigation in XML-Dokumenten
- **XSL(T)** = Programmiersprache zur Verarbeitung von XML-Dokumenten
- **Namespaces** = Mechanismus zum Einsatz mehrerer XML-Dialekte in einem Dokument
- **XInclude** = Mechanismus zum Einbetten beliebiger Inhalte in XML-Dokumente
- **XQuery** = Abfragesprache für XML-basierte Dokumente und Datenbanken
- u. v. m.

XML als Bestandteil von Editionen

XML-Dateien bilden den Kern von digitalen Editionen:

- Sie enthalten Transkriptionen und konstituierte Editionstexte, im Bereich der Geisteswissenschaften in der Regel kodiert nach den TEI-Richtlinien.
- Sie werden mit XML-Schemata überprüft.
- Sie werden mit XSL(T) oder XQuery verarbeitet.
- Sie werden für die Präsentation im Internet, für den Satz (z. B. über TUSTEP, Latex, Indesign) als PDF oder zur weiteren Verarbeitung in anderen Programmen (z. B. Officeprogramme) in andere Formate konvertiert.
- Sie werden zur Erzeugung von Registern, Indizes, Glossaren und anderer Teile von digitalen Editionen abgefragt bzw. ausgewertet.

Wortlaut, Struktur und Aussehen/Darstellung

Auf der folgenden Folie wird ein Scan eines Textes eingeblendet.
Versuchen Sie, folgende Fragen zu beantworten:

- Was sehen Sie auf dem Scan?
- Was ist der *Wortlaut* des Textes?
- Wie ist die *logische Struktur* des Textes?
- Wie sind *Wortlaut* und *Struktur* mit dem *Aussehen* bzw. der *Darstellung* verbunden?

Gedichte.

Erste Periode.

Hektors Abschied.

Andromache.

Will sich Hektor ewig von mir wenden,
Wo Achill mit den unnahbar'n Händen
Dem Patroklus schrecklich Opfer bringt?
Wer wird künftig deinen Kleinen lehren
Speere werfen und die Götter ehren,
Wenn der finstre Orkus dich verschlingt?

Hektor.

Theures Weib, gebiete deinen Thränen!
Nach der Feldschlacht ist mein feurig Sehnen,
Diese Arme schützen Pergamus.
Kämpfend für den heil'gen Heerd der Götter
Fall' ich, und des Vaterlandes Retter
Steig' ich nieder zu dem styg'schen Fluß.

Andromache.

Nimmer lauscht ich deiner Waffen Schalle,
Müßig liegt dein Eisen in der Halle,
Priams großer Heldenstamm verdirbt.
Du wirst hingeh'n, wo kein Tag mehr scheinet,
Der Cocytus durch die Wüsten weinet,
Deine Liebe in dem Lethe stirbt.

Hektor.

All mein Sehnen will ich, all mein Denken,
In des Lethe stillen Strom versenken,
Aber meine Liebe nicht.
Horch! der Wilde tobt schon an den Mauern,
Gürte mir das Schwert um, laß das Trauern!
Hektors Liebe stirbt im Lethe nicht.

Amalia.

Schön wie Engel voll Walhallas Wonne,
Schön vor allen Jünglingen war er,
Himmlischmild sein Blick, wie Mayensonne,
Rückgestrahlt vom blauen Spiegelmeer.

Seine Küsse — Paradiesisch Fühlen!
Wie zwo Flammen sich ergreifen, wie
Harfentöne in einander spielen
Zu der himmelvollen Harmonie —

Stürzten, flogen, schmolzen Geist und Geist zusammen,
Lippen, Wangen brannten, zitterten,
Seele rann in Seele — Erd' und Himmel schwammen
Wie zerronnen um die Liebenden!

Er ist hin — vergebens, ach, vergebens
Stöhnet ihm der bange Seufzer nach!
Er ist hin, und alle Lust des Lebens
Wimmert hin in ein verlor'nes Ach!

Eine Leichenfantasie.

Mit erstorbnem Scheinen
Steht der Mond auf todtenstillen Hainen,
Seufzend streicht der Nachtgeist durch die Luft —

Wortlaut, Struktur und Aussehen/Darstellung

Wortlaut	Struktur	Aussehen/Darstellung
«Gedichte.»	Titel	grosse Schrift, Weissraum, Linien, Sperrung, Zentrierung
«Erste Periode.»	Untertitel	grosse Schrift, Sperrung etc.
«Hektors Abschied.»	Gedichttitel	etwas grössere Schrift, etc.
«Andromache.»	Strophentitel	leicht grössere Schrift, Zentrierung
«Will sich Hektor ...»	Vers	normale Schrift, Zeilenumbrüche, linksbündig, Grossschreibung am Zeilenanfang

Daraus folgt: Die logische Struktur wird durch typographische Mittel (Aussehen/Darstellung) für den Leser sichtbar und interpretierbar.
Wortlaut, Struktur und Aussehen bilden eine Einheit.

Arbeiten mit Markup statt mit Office-Programmen

- Wenn man Texte auszeichnet, werden Wortlaut, Struktur und Aussehen durch das Markup strikt voneinander getrennt!
- Der Wortlaut ist dasjenige, was mithilfe von Markup ausgezeichnet wird.
- Das Markup selbst gibt die logische Struktur wieder.
- Ein separates Stylesheet (z. B. XSL) weist jedem Strukturelement ein bestimmtes Aussehen zu.
- Für jeden Anwendungszweck kann es ein eigenes Stylesheet (z. B. für Web, Druck, Korrekturdurchgänge, Registerarbeiten etc.) geben.
- Kein «What you see is what you get» (WYSIWYG) ≠ Office-Programme!
- Stattdessen: Quelltextansicht mit Steuerzeichen.

1. Grundkonzept: Elemente

- Nahezu die gesamte Struktur von XML besteht aus Elementen.
- Elemente existieren in zwei Varianten, je nachdem, ob sie Inhalt haben oder nicht:
- Variante 1: Starttag + Inhalt + Endtag z. B. <titel>Gedichte</titel>
- Variante 2: Emptytag z. B. <seitenumbruch/>
- Starttag: < + Name des Elements + > z. B. <titel>
- Endtag: </ + Name des Elements + > z. B. </titel>
- Emptytag: < + Name des Elements + /> z. B. <seitenumbruch/>

Namen von Elementen

Korrekte Namen

- <Titel> <titel> <TiTeL>
- <_titel> <titel1> <titel_1>
- <títêl-1> <αβγδε> <ß>
- Nahezu alle Buchstaben sind denkbar. Im Zweifel ausprobieren, aber besser bei einfachen, gut lesbaren Namen bleiben!
- Achtung: Gross- und Kleinschreibung wird unterschieden: <titel> ≠ <Titel>

Nicht korrekte Namen

- <1titel> (keine Ziffern am Anfang)
- <tit el> (keine Weissräume)
- <tit/el> (kein Schrägstrich)
- <tit<el>> (kein <>)
- <.titel> (nicht alle Sonderzeichen sind am Anfang möglich)

Übung 1

Zeichnen Sie die Titel und die erste Strophe des Gedichts «Hektors Abschied» mit folgenden Elementen aus:

- <titel>
- <untertitel>
- <gedichttitel>
- <strophentitel>
- <vers>
- <seitenbeginn>

Gedichte.

Erste Periode.

Hektors Abschied.

Andromache.

Will sich Hektor ewig von mir wenden,
Wo Achill mit den unnahbar'n Händen
Dem Patroklus schrecklich Opfer bringt?
Wer wird künftig deinen Kleinen lehren
Speere werfen und die Götter ehren,
Wenn der finstre Orkus dich verschlingt?

Hektor.

Theures Weib, gebiete deinen Thränen!
Nach der Feldschlacht ist mein feurig Sehnen,
Diese Arme schützen Pergamus.
Kämpfend für den heil'gen Heerd der Götter
Fall' ich, und des Vaterlandes Retter
Steig' ich nieder zu dem styg'schen Fluß.

Andromache.

Nimmer lauscht ich deiner Waffen Schalle,
Müßig liegt dein Eisen in der Halle,
Priams großer Heldenstamm verdirbt.
Du wirst hingeh'n, wo kein Tag mehr scheinet,
Der Cocytus durch die Wüsten weinet,
Deine Liebe in dem Lethe stirbt.

Hektor.

All mein Sehnen will ich, all mein Denken,
In des Lethe stillen Strom versenken,
Aber meine Liebe nicht.
Horch! der Wilde tobt schon an den Mauern,
Gürte mir das Schwert um, laß das Trauern!
Hektors Liebe stirbt im Lethe nicht.

Amalia.

Schön wie Engel voll Walhallas Wonne,
Schön vor allen Jünglingen war er,
Himmlischmild sein Blick, wie Mayensonne,
Rückgestrahlt vom blauen Spiegelmeer.

Seine Küsse — Paradiesisch Fühlen!
Wie zwo Flammen sich ergreifen, wie
Harfentöne in einander spielen
Zu der himmelvollen Harmonie —

Stürzten, flogen, schmolzen Geist und Geist zusammen,
Lippen, Wangen brannten, zitterten,
Seele rann in Seele — Erd' und Himmel schwammen
Wie zerronnen um die Liebenden!

Er ist hin — vergebens, ach, vergebens
Stöhnet ihm der bange Seufzer nach!
Er ist hin, und alle Lust des Lebens
Wimmert hin in ein verlor'nes Ach!

Eine Leichenfantasie.

Mit erstorbnem Scheinen
Steht der Mond auf todtenstillen Hainen,
Seufzend streicht der Nachtgeist durch die Luft —

Lösung 1

```
<seitenbeginn/>
<titel>Gedichte</titel>
<untertitel>Erste Periode.</untertitel>
<gedichttitel>Hektors Abschied.</gedichttitel>
<strophentitel>Andromache.</strophentitel>
<vers>Will sich Hektor ewig von mir wenden,</vers>
<vers>Wo Achill mit den unnahbar’n Händen</vers>
<vers>Dem Patroklus schrecklich Opfer bringt?</vers>
<vers>Wer wird künftig deinen Kleinen lehren</vers>
<vers>Speere werfen und die Götter ehren,</vers>
<vers>Wenn der finstre Orkus dich verschlingt?</vers>
```

Verschachtelung

- Elemente dürfen beliebig tief ineinander verschachtelt werden.
- Die Verschachtelung ist streng hierarisch, Überlappungen sind strikt verboten!
- Beispiel:

  `<aussen><innen>foo</innen></aussen>`
- Beispiel:

  `<aussen><innen>foo</innen><innen>bar</innen></aussen>`
- Aber nicht:

  `<aussen><innen>foo</aussen></innen>`

Wurzelelement (Rootelement)

- In jedem XML-Dokument muss es ein Element geben, das alle anderen Elemente umfasst. Dieses äusserste Element wird Wurzelelement (Rootelement) genannt.
- Beispiel:

  <aussen><innen>foo</innen></aussen>

- Aber nicht:

  <aussen><innen>foo</innen></aussen><aussen><innen>bar</innen></aussen>

- XML-Dokumente können daher als Baumstruktur betrachtet werden, wobei der «Stamm» des Baums das so genannte Wurzelelement ist.

Verschachtelung und Einrückung

- Um die Verschachtelung besser lesbar zu machen, kann (und sollte man) die Tiefe der Verschachtelung durch Einrückungen ausdrücken. XML soll nicht nur für Computer, sondern auch für Menschen lesbar sein!
- Beispiel:

```
<aussen>
    <innen>foo</innen>
    <innen>
        <noch_weiter_innen>bar</noch_weiter_innen>
    </innen>
</aussen>
```

- Aber nicht:

```
<aussen><innen>foo</innen><innen><noch_weiter_innen>bar
</noch_weiter_innen></innen></aussen>
```

Übung 2

Verschachteln Sie das bisher Kodierte, um die Struktur des Textes besser wiederzugeben. Nutzen Sie dafür die folgenden Elemente:

- <text> [Gruppiert alles = Wurzelelement]
- <titelei> [Gruppiert die Überschriften]
- <gedicht> [Gruppiert den Gedichttitel und die Strophen]
- <strophe> [Gruppiert den Strophentitel und die Verse]

Vorher ohne Verschachtelung

```
<seitenbeginn/>
<titel>Gedichte</titel>
<untertitel>Erste Periode.</untertitel>
<gedichttitel>Hektors Abschied.</gedichttitel>
<strophentitel>Andromache.</strophentitel>
<vers>Will sich Hektor ewig von mir wenden,</vers>
<vers>Wo Achill mit den unnahbar’n Händen</vers>
<vers>Dem Patroklus schrecklich Opfer bringt?</vers>
<vers>Wer wird künftig deinen Kleinen lehren</vers>
<vers>Speere werfen und die Götter ehren,</vers>
<vers>Wenn der finstre Orkus dich verschlingt?</vers>
```

Lösung 2 mit Verschachtelung

```
<text>
    <seitenbeginn/>
    <titelei>
        <titel>Gedichte</titel>
        <untertitel>Erste Periode.</untertitel>
    </titelei>
    <gedicht>
        <gedichttitel>Hektors Abschied.</gedichttitel>
        <strophe>
            <strophentitel>Andromache.</strophentitel>
            <vers>Will sich Hektor ewig von mir wenden,</vers>
            <vers>Wo Achill mit den unnahbar'n Händen</vers>
            <vers>Dem Patroklus schrecklich Opfer bringt?</vers>
            <vers>Wer wird künftig deinen Kleinen lehren</vers>
            <vers>Speere werfen und die Götter ehren,</vers>
            <vers>Wenn der finstre Orkus dich verschlingt?</vers>
        </strophe>
        [... hier kommen die weiteren Strophen hin]
    </gedicht>
</text>
```

Gemischter Inhalt (Mixed Content)

Elemente mit Inhalt müssen nicht nur entweder Elemente oder Text enthalten, sie können auch beides gemischt enthalten.

- Nur Elemente als Inhalt: `<aussen><innen/><aussen>`
- Nur Text als Inhalt: `<aussen>foo</aussen>`
- Weissraum ist auch Text: `<aussen>  </aussen>`
- Gemischter Inhalt: `<aussen>foo <innen>bar</innen>.</aussen>`

Übung 3

Zeichnen Sie in dem bisher Kodierten die Namen von Personen und Orten mit folgenden Elementen aus:

- <person>
- <ort>

Vorher ohne Personen- und Ortsnamen

```
<text>
    <seitenbeginn/>
    <titelei>
        <titel>Gedichte</titel>
        <untertitel>Erste Periode.</untertitel>
    </titelei>
    <gedicht>
        <gedichttitel>Hektors Abschied.</gedichttitel>
        <strophe>
            <strophentitel>Andromache.</strophentitel>
            <vers>Will sich Hektor ewig von mir wenden,</vers>
            <vers>Wo Achill mit den unnahbar'n Händen</vers>
            <vers>Dem Patroklus schrecklich Opfer bringt?</vers>
            <vers>Wer wird künftig deinen Kleinen lehren</vers>
            <vers>Speere werfen und die Götter ehren,</vers>
            <vers>Wenn der finstre Orkus dich verschlingt?</vers>
        </strophe>
        [... hier kommen die weiteren Strophen hin]
    </gedicht>
</text>
```

Lösung 3 mit Personen- und Ortsnamen

```
<text>
    <seitenbeginn/>
    <titelei>
        <titel>Gedichte</titel>
        <untertitel>Erste Periode.</untertitel>
    </titelei>
    <gedicht>
        <gedichttitel><person>Hektors</person> Abschied.</gedichttitel>
        <strophe>
            <strophentitel><person>Andromache</person>.</strophentitel>
            <vers>Will sich <person>Hektor</person> ewig von mir wenden,</vers>
            <vers>Wo <person>Achill</person> mit den unnahbar'n Händen</vers>
            <vers>Dem <person>Patroklus</person> schrecklich Opfer bringt?</vers>
            <vers>Wer wird künftig deinen Kleinen lehren</vers>
            <vers>Speere werfen und die Götter ehren,</vers>
            <vers>Wenn der finstre <ort>Orkus</ort> dich verschlingt?</vers>
        </strophe>
        [... hier kommen die weiteren Strophen hin]
    </gedicht>
</text>
```

2. Grundkonzept: Attribute

- Mit Attributen werden Metadaten zu einem Element erfasst.
- Attribute sind immer Bestandteil des Starttags bzw. Emptytags.
- Attribute bestehen aus einen Namen und einem dazu gehörigen Wert.
- Attributname + = + "Attributwert" z. B. id="1"
- Attributname + = + 'Attributwert' z. B. id='1'
- Beispiele:

```
<person id="pers001">Hektor</person>
<seitenbeginn nummer='1'/>
```

- Achtung: Deutsche und englische Gänsefüsschen, z. B. „“, sowie französische und Schweizer Guillemets, z. B. «», sind nicht als Begrenzung von Attributwerten erlaubt, sondern nur " und '.

Regeln für Attribute

- Ein Element kann beliebig viele Attribute haben.
- Die Reihenfolge der Attribute ist beliebig.
- Zwischen den Attributen darf beliebig viel Weissraum stehen.
- Beispiele:

```
<person id="pers1" fraktion="Trojaner">Hektor</person>
<person fraktion="Trojaner" id="pers1">Hektor</person>
<person
      id="pers1"
      fraktion="Trojaner"
>Hektor</person>
```

Einschränkungen von Attributen

- Pro Element darf ein Attribut nur einmal vorkommen.
- Für Attributnamen gelten die selben Einschränkungen wie für Elementnamen.
- Negativbeispiele:

  <person id="pers1" fraktion="Trojaner" id="pers1">Hektor</person>

  [«id» kommt zweimal vor.]

  <person 1fraktion="Trojaner" id="pers1">Hektor</person>

  [«1fraktion» ist kein gültiger Name.]

Übung 4

Erweitern Sie die zuletzt kodierten Personen- und Ortselemente um jeweils ein Attribut id und nutzen Sie die unten stehenden Werte:

- Hektor pers001
- Andromache pers002
- Achill pers003
- Patroklus pers004
- Orkus ort001

Erweitern Sie ausserdem die Elemente: gedicht, strophe, vers und seitenbeginn um das Attribut nummer und fügen Sie einen passenden Wert sein.

Vorher ohne Attribute

```
<text>
    <seitenbeginn/>
    <titelei>
        <titel>Gedichte</titel>
        <untertitel>Erste Periode.</untertitel>
    </titelei>
    <gedicht>
        <gedichttitel><person>Hektors</person> Abschied.</gedichttitel>
        <strophe>
            <strophentitel><person>Andromache</person>.</strophentitel>
            <vers>Will sich <person>Hektor</person> ewig von mir wenden,</vers>
            <vers>Wo <person>Achill</person> mit den unnahbar'n Händen</vers>
            <vers>Dem <person>Patroklus</person> schrecklich Opfer bringt?</vers>
            <vers>Wer wird künftig deinen Kleinen lehren</vers>
            <vers>Speere werfen und die Götter ehren,</vers>
            <vers>Wenn der finstre <ort>Orkus</ort> dich verschlingt?</vers>
        </strophe>
        [... hier kommen die weiteren Strophen hin]
    </gedicht>
</text>
```

Lösung 4 mit Attributen

```
<text>
   <seitenbeginn nummer="1"/>
   <titelei>
      <titel>Gedichte</titel>
      <untertitel>Erste Periode.</untertitel>
   </titelei>
   <gedicht nummer="1">
      <gedichttitel><person id="pers001">Hektors</person> Abschied.</gedichttitel>
      <strophe nummer="1">
         <strophentitel><person id="pers002">Andromache</person>.</strophentitel>
         <vers nummer="1">Will sich <person id="pers001">Hektor</person> ewig von mir wenden,</vers>
         <vers nummer="2">Wo <person id="pers003">Achill</person> mit den unnahbar'n Händen</vers>
         <vers nummer="3">Dem <person id="pers004">Patroklus</person> schrecklich Opfer bringt?</vers>
         <vers nummer="4">Wer wird künftig deinen Kleinen lehren</vers>
         <vers nummer="5">Speere werfen und die Götter ehren,</vers>
         <vers nummer="6">Wenn der finstre <ort id="ort001">Orkus</ort> dich verschlingt?</vers>
      </strophe>
      [... hier kommen die weiteren Strophen hin]
   </gedicht>
</text>
```

Zusammenfassung

- XML ist eine nahezu beliebig erweiterbare Auszeichnungssprache, mit der man Daten strukturiert, klassifiziert, mit Metadaten annotiert und kommentiert.
- In XML werden Daten aller Art erfasst: Texte, Musik, Geodaten, Banküberweisungen, Webseiten u. v. m.
- Die beiden zentralen Grundkonzepte sind: Elemente und Attribute.
- Über die Namen der Elemente klassifiziert man das, was man auszeichnet.
- Über die Verschachtelung der Elemente gibt man die logische Struktur eines Textes wieder.
- Über die Attribute reichert man die Elemente mit Metadaten an.

Der komplette Kurs auf GitHub

https://github.com/ChristianSonderUniSG/xml-kurs

Dort finden Sie alle Folien, Übungsaufgaben, Hausaufgaben und weitere Materialien.