Erstellung einer XML-Datei als Ressource für berufliche Kompetenzen

Johanna Binnewitt, Institut für Digital Humanities

Die folgende Dokumentation soll einen Überblick verschaffen wie mithilfe des AMS-Kompetenzenkatalogs[1] eine TEI-konforme Ressource mit kategorisierten Kompetenzen erzeugt wurde. Der benannte Katalog hierarchisiert Kompetenzen, die im Bereich der Stellenbesetzung als relevant gelten, auf verschiedenen Ebenen. Beispielsweise werden alle Kompetenzen auf oberster Ebene in fachliche berufliche und überfachliche berufliche Kompetenzen sowie Zertifikate und Ausbildungsabschlüsse unterteilt. Die Kategorisierung bekannter Kompetenzen kann bei der Extraktion dieser Kompetenzen aus Stellenausschreibungen von Vorteil sein, da die extrahierte Informationseinheit so mit weiterem Wissen wie beispielsweise die zugehörige Branche angereichert werden kann.

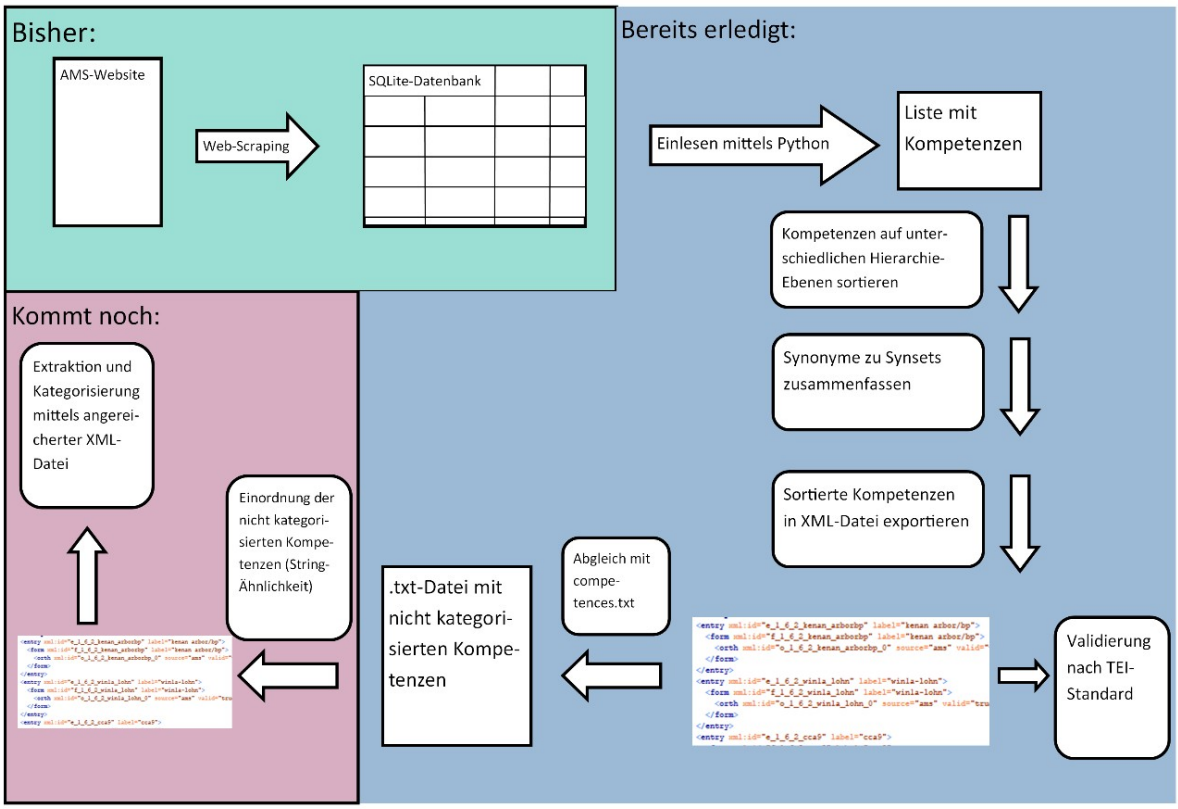

Abbildung 1Workflow XML-Datei-Erstellung

Im Verlauf des Projekts wurden die Kompetenzen des AMS-Katalogs bereits mittels Web Scraping aus der Website extrahiert und in einer SQLite-Datenbank aufbereitet. Diese konnte beispielsweise dafür genutzt werden, um unbekannte Kompetenzen nach der Extraktion mithilfe von String-Ähnlichkeits-Vergleichen zu kategorisieren. Um bekannte Kompetenzen in neuen Stellenausschreibungen

1 https://www.ams.at/bis/bis/KompetenzstrukturBaum.php (aufgerufen: 08.03.19)

aufzufinden, wurde jedoch weiterhin eine Wortliste zum Abgleich verwendet. Somit ist mit dem Auffinden einer Kompetenz im Text noch keine weitere Information über die ihre semantische Einordnung vorhanden. Deshalb wurde nun im Rahmen einer Projektarbeit eine XML-Datei erzeugt, die Wortliste und SQLite-Datenbank vereinen soll. So können in Zukunft Extraktion bekannter Kompetenzen sowie Kategorisierung dieser Kompetenzen in einem Schritt erfolgen. Die Kompetenz kann dabei auf einer der verschiedenen Hierarchieebenen kategorisiert werden, je nachdem wie generisch die gefundenen Kompetenzen zusammengefasst werden sollen. In dem unten eingefügten Auszug aus der XML-Datei repräsentiert der Tag <*div3*> den Knoten „SAP-Kenntnisse“ im Kompetenzbaum[2] des AMS. Diesem Knoten sind weitere Kompetenzen untergeordnet, die in Synsets(<*entry*>) gegliedert sind. Jedes Synonym(<*form*>) innerhalb eines Synsets kann wiederum verschiedene Orthographien(<*orth*>) speichern. Jeder Orthographie-Knoten enthält hierbei eine lemmatisierte Kompetenz äquivalent zur Wortliste in der .txt-Datei.

Publikationen

```
<div3 xml:id="se_1_3_2" label="sap-kenntnisse">
  <entry xml:id="e_1_3_2_kenntnis_von_sap_lösungen" label="kenntnis von sap-loesungen">
    <form xml:id="f_1_3_2_kenntnis_von_sap_lösungen" label="kenntnis von sap-loesungen">
      <orth xml:id="o_1_3_2_kenntnis_von_sap_lösungen_0" source="ams" valid="true" label="kenntnis von sap-loesungen">kenntnis von sap-lösung</orth>
    </form>
    <form xml:id="f_1_3_2_sap_kenntnisse" label="sap-kenntnisse">
      <orth xml:id="o_1_3_2_sap_kenntnisse_0" source="ams" valid="true" label="sap-kenntnisse">sap-kenntnis</orth>
    </form>
    <form xml:id="f_1_3_2_sap_datenbanken" label="sap-datenbanken">
      <orth xml:id="o_1_3_2_sap_datenbanken_0" source="ams" valid="true" label="sap-datenbanken">sap-datenbank</orth>
    </form>
    <form xml:id="f_1_3_2_sap_db" label="sap db">
      <orth xml:id="o_1_3_2_sap_db_0" source="ams" valid="true" label="sap db">sap db</orth>
    </form>
    <form xml:id="f_1_3_2_sap_ausbildung" label="sap-ausbildung">
      <orth xml:id="o_1_3_2_sap_ausbildung_0" source="ams" valid="true" label="sap-ausbildung">sap-ausbildung</orth>
    </form>
  </entry>
  <entry xml:id="e_1_3_2_sap_jco" label="sap jco">
    <form xml:id="f_1_3_2_sap_jco" label="sap jco">
      <orth xml:id="o_1_3_2_sap_jco_0" source="ams" valid="true" label="sap jco">sap jco</orth>
    </form>
    <form xml:id="f_1_3_2_jco" label="jco">
      <orth xml:id="o_1_3_2_jco_0" source="ams" valid="true" label="jco">jco</orth>
    </form>
    <form xml:id="f_1_3_2_sap_java_connector" label="sap java connector">
      <orth xml:id="o_1_3_2_sap_java_connector_0" source="ams" valid="true" label="sap java connector">sap java connector</orth>
    </form>
  </entry>
```

Abbildung 2 Auszug aus der XML-Datei (Beispiel SAP-Kenntnisse)

Um nun bekannte Kompetenzen aus Stellenausschreibungen zu extrahieren und zu kategorisieren, besitzen die Kompetenz-Konzepte in der XML-Datei auf jeder Ebene ein Label, das beispielsweise eine Synset beschreibt. Bei der Extraktion bekannter Kompetenzen kann nun, wie unten zu sehen, die Fundstelle mit diesem Label – in diesem Fall auf Ebene *div3* annotiert werden. So werden im vorliegenden Beispiel Derivate und ihre Wortstämme unter einem Label zusammengefasst.

2 https://www.ams.at/bis/bis/Kompetenzen.php?noteid=4628 (aufgerufen: 08.03.19)

Sentence	Label	Comp
Filtern	Filtern	Filtern
Zuverlässigkeit .	zuverlaessigkeit	zuverlässigkeit
Belastbarkeit , Zuverlässigk...	zuverlaessigkeit	zuverlässigkeit
Als zukünftiger Mitarbeiter ...	zuverlaessigkeit	zuverlässigkeit
Sie sind teamfähig , flexibel...	zuverlaessigkeit	zuverlässig
motivierende , zuverlässige...	zuverlaessigkeit	zuverlässig
Arbeiten nach Zeichnung .	kuenstlerische fachkenntnisse	zeichnung
Als zukünftiger Mitarbeiter ...	kuenstlerische fachkenntnisse	zeichnen

Abbildung 3 Auszug aus den extrahierten und kategorisierten Kompetenzen

Bei einem Abgleich der Wortliste mit der XML-Datei wurden ca. 4000 Kompetenzen identifiziert, die sich noch nicht in der XML-Datei befinden. In einem nächsten Schritt können diese mithilfe von Kookkurrenz-Analysen und String-Ähnlichkeits-Vergleichen in die abgebildeten Konzepte eingeordnet werden.