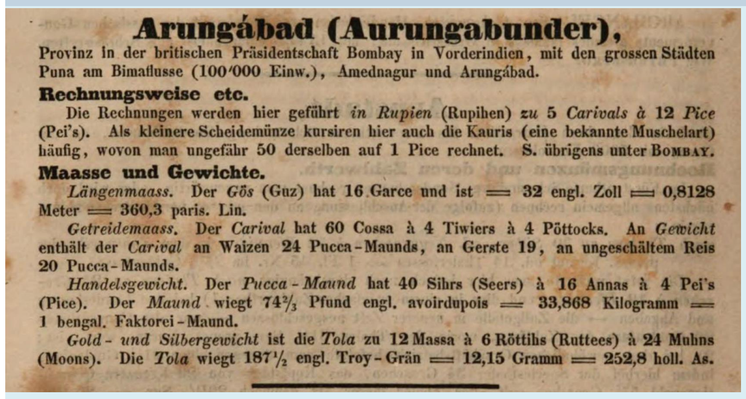
Universität Bamberg

Scenario-based planning for the semantic digitization of historical reference works

Werner Scheltjens 0000-0002-5209-9052 Christoph Schlieder 0000-0002-7226-8204



Arungábad (Aurungabunder),

Provinz in der britischen Präsidentschaft Bombay in Vorderindien, mit den grossen Städten Puna am Bimaflusse (100'000 Einw.), Amednagur und Arungábad.

Rechnungsweise etc.

Die Rechnungen werden hier geführt *in Rupien* (Rupihen) *zu 5 Carivals à 12 Pice* (Pei's). Als kleinere Scheidemünze kursiren hier auch die Kauris (eine bekannte Muschelart) häufig, wovon man ungefähr 50 derselben auf 1 Pice rechnet. S. übrigens unter BOMBAY.

Maasse und Gewichte.

Längenmaass. Der *Gös* (Guz) hat 16 Garce und ist = 32 engl. Zoll = 0,8128 Meter = 360,3 paris. Lin.

Getreidemaass. Der *Carival* hat 60 Cossa à 4 Tiwiers à 4 Pöttocks. An *Gewicht* enthält der *Carival* an Waizen 24 Pucca-Maunds, an Gerste 19, an ungeschältem Reis 20 Pucca-Maunds.

Handelsgewicht. Der *Pucca-Maund* hat 40 Sihrs (Seers) à 16 Annas à 4 Pei's (Pice). Der *Maund* wiegt 74⅔ Pfund engl. avoirdupois = 33,868 Kilogramm = 1 bengal. Faktorei-Maund.

Gold- und Silbergewicht ist die *Tola* zu 12 Massa à 6 Röttihs (Ruttees) à 24 Muhns (Moons). Die *Tola* wiegt 187½ engl. Troy-Grän = 12,15 Gramm = 252,8 holl. As.

(a) Digitized page in Noback's Handbook

Arungábad (Aurungabunder) ,
Provinz in der britischen Präsidentschaft Bombay in Vorderindien , mit den grossen Städten Puna am Bimaflusse (100'000 Einw .) , Amednagur und Arungábad .
Rechnungsweise etc.
Die Rechnungen werden hier geführt in Rupien (Rupihen) zu 5 Carivals à 12 Pice (Pei's) . Als kleinere Scheidemünze kursiren hier auch die Kauris (eine bekannte Muschelart) häufig , wovon man ungefähr 50 derselben auf 1 Pice rechnet . S. übrigens unter BOMBAY .
Maasse und Gewichte .
Längenmaass . Der Gös (Guz) hat 16 Garce und ist == 32 engl. Zoll == 0,8128 Meter == 360,3 paris. Lin.
Getreidemaass . Der Carival hat 60 Cossa à 4 Tiwiers à 4 Pöttocks . An Gewicht enthält der Carival an Waizen 24 Pucca - Maunds , an Gerste 19 , an ungeschältem Reis 20 Pucca - Maunds .
Handelsgewicht . Der Pucca - Maund bat 40 Sihrs (Seers) à 16 Annas à 4 Pei's (Pice) . Der Maund wiegt 74 **2/3** Pfund engl. avoirdupois == 33,868 Kilogramm == 1 bengal. Faktorei - Maund .
Gold - und Silbergewicht ist die Tola zu 12 Massa à 6 Röttihs (Ruttees) à 24 Muhns (Moons) . Die Tola wiegt 187 **1/2** engl. Troy - Grän == 12,15 Gramm == 252,8 holl. As

(b) Survey of OCR quality

```
<article>
    <lemma>Arungábad (Aurungabunder),</lemma>
    <search-area-0>
        <search-area-1>Provinz in der britischen Präsidentschaft Bombay in Vorderindien, mit den grossen Städten
Puna am Bimaflusse (100'000 Einw.), Amednagur und Arungábad.</search-area-1>
        <indicator-1>Rechnungsweise etc.</indicator-1>
        <search-area-1>Die Rechnungen werden hier geführt in Rupien (Rupihen) zu 5 Carivals à 12 Pice
(Pei's). Als kleinere Scheidemünze kursiren hier auch die Kauris (eine bekannte Muschelart)
häufig, wovon man ungefähr 50 derselben auf 1 Pice rechnet. S. übrigens unter BOMBAY.</search-area-1>
        <indicator-1>Maasse und Gewichte.</indicator-1>
        <search-area-1>
            <indicator-2>Längenmaass. </indicator-2>
            <search-area-2>Der Gös (Guz) hat 16 Garce und ist = 32 engl. Zoll = 0,8128
Meter 360,3 paris. Lin.</search-area-2>
            <indicator-2>Getreidemaass. </indicator-2>
            <search-area-2>Der Carival hat 60 Cossa à 4 Tiwiers à 4 Pöttocks. An Gewicht
enthält der Carival an Waizen 24 Pucca - Maunds, an Gerste 19, an ungeschältem Reis
20 Pucca - Maunds.</search-area-2>
            <indicator-2>Handelsgewicht. </indicator-2>
            <search-area-2>Der Pucca - Maund bat 40 Sihrs (Seers) à 16 Annas à 4 Pei's
(Pice). Der Maund wiegt 7493 Pfund engl. avoirdupois 33,868 Kilogramm
1 bengal. Faktorei - Maund.</search-area-2>
            <indicator-2>Gold - und Silbergewicht </indicator-2>
            <search-area-2>ist die Tola zu 12 Massa à 6 Röttihs (Ruttees) à 24 Mubns
(Moons). Die Tola wiegt 187'/2 engl. Troy - Grän = 12,15 Gramm = 252,8 boll. As.</search-area-2>
        </search-area-1>
    </search-area-0>
</article>
```

(c) Test annotation with lexicographical ontology

Semantic Digitization

The retrodigitization of library collections that exempt from copyright restrictions has made lexica, handbooks and encyclopaedias available as digital sources. The result of the **first digitization** is mostly a scan (a) and an OCR'ed fulltext, often with many mistakes (b).

For specific historical research questions, a **second, semantic digitization** is necessary (c). The second digitization aims to extract and explicitly model the semantic structure of encyclopaedic knowledge.

Planning semantic modelling

In principle, we can rely on the general methods of ontological modelling to plan the second digitization and make semantic relations in our reference work explicit.

Most importantly, we refer to **scenario-based methods** that relate the planning process to so-called **competency questions**, i.e. questions that specialist users of the source would like to examine and answer with the help of a data model (Kendall, McGuinness, 2019; Lodi et al., 2017, Carriero et al. 2021).

Examples of competency questions asked during the planning phase

Modelling level	**1**	**2**	**3**
Question	What kind of measures does Noback (1850) provide for the lemma Arungabad?	Which units of measure for length are defined under Arungabad?	In which trading places in "Vorderindien" (part of India, denoted as such by the Nobacks) were local units of measure assimilated with the measurement system of a colonial power?
Answer	"Längenmaass", "Getreidemaass", ...	"EN Guz, DE Gös", "EN DE Garce"	"Benares", "Bombay", ... , "Ajinga"
Algorithmisation	Can be answerered by clicking through the digitized source. No semantic modelling needed.	NER should not identify the German transcription of the unit of measure "Gös" as different from "Guz".	Users should be able to define what assimilation of a unit of measure means. Inference based on the semantic modelling is applied for answering the question.

Results of planning

15 historical-metrological competency questions

3 levels of modelling of different complexity

- Even at the first level, only few questions can be answered relatively easily by hand using the digitized source.
- The OCR quality is sufficient for answering only few competency questions and modelling level one.
- For answering the research questions, a domain ontology for historical metrology as well as for trade and historical geography are necessary.
- Competency questions of modelling level 3 require a lexicographical ontology that can display the lexicographical choices and referencing structure of the hanhdbook.

Digitized source

Noback, Christian, Noback, Friedrich (1850): Vollständiges Taschenbuch der Münz-, Maass-, und Gewichtsverhältnisse, der Staatspapiere, des Wechsels- und Bankwesens, und der Usanzen aller Länder und Handelsplätze. Leipzig: F.A. Brockhaus.

Bayerische Staatsbibliothek, Münchner DigitalisierungsZentrum, Digitale Bibliothek

Literature

Carriero, V. et al. (2021): "Pattern-based Design Applied to Cultural Heritage Knowledge Graphs", in: Semantic Web 12: 313 – 357.

Kendall, E., McGuinness, D.(2019): Ontology engineering. (= Synthesis Lectures on The Semantic Web: Theory and Technology, Lecture 18). Morgan and Claypool.

Lodi, G. et al. (2017): "Semantic Web for Cultural Heritage Valorisation", in: Hai-Jew, Shalin (ed.): Data Analytics in Digital Humanities. Multimedia Systems and Applications. Springer: Cham 3-37.

Contact

Prof. Dr. Werner Scheltjens, Professur für Digitale Geschichtswissenschaft, Otto-Friedrich-Universität Bamberg, werner.scheltjens@uni-bamberg.de