Digital Edition Publishing Cooperative for Historical Accounts and the Bookkeeping Ontology

Christopher Pollin[1]

Abstract: The Project "*Digital Edition Publishing Cooperative for Historical Accounts*", a Andrew W. Mellon funded cooperation of five US partners and the Centre for Information Modelling at Graz University, aims to link the knowledge domain of economic activities to historical accounting records. For this purpose the so-called *Bookkeeping Ontology* is developed. DEPCHA creates a publication hub for digital editions on the web. It converts multiple formats into RDF and publishes these in combination with the associated transcriptions. DEPCHA also allows the usage of retrieval and visualization functionalities, as well as interoperability and reuse of information in the sense of *Linked Open Data*.

Keywords: Web of Data; GAMS; Historical Financial Records; Bookkeeping Ontology; Knowledge Domain; History; Digital Humanities; Semantic Web; Linked Open Data

1 Introduction

On the first of August 1808 *James Haley* purchased ¼ lb of powder, 1 lb of shot and 1 lb of sugar for the price of 2 shilling and 6 pence each from *Stagville Plantation* in North Carolina (USA). We can find information like that, and numerous similar ones, in historical financial records. In the 1980s, two groups emerged which applied different approaches to tackle such historical sources: the "traditionalists" and the "quantifiers". While the "traditionalists" used a hermeneutic approach to historical sources, the "quantifiers" tried to formally describe and evaluate the historical dimensions, in order to support an intuitive process of understanding with empirically identifiable facts [JAT85]. This divide tells us that, when using formal methods on historical data, research should distinguish between the representation of the original source and its interpretation. The latter is the core knowledge domain of historical research. It is advised to share the basic assumptions and definitions in a knowledge domain in a formal way [Th17]. In this context, the *Web of Data* (aka *Semantic Web*) and *Linked Open Data* are central concepts that offer technologies going hand in hand with that new understanding of historical research. The hermeneutic method and the method of transforming historical phenomena into formal models, as well as its connections to other domains and its reuse by the scientific community, makes the work of historians more dynamic and comprehensible.

[1] University of Graz, Centre for Information Modelling - ACDH, Elisabethstraße 59/III, 8010 Graz, Austria
christopher.pollin@uni-graz.at

The Project "*Digital Edition Publishing Cooperative for Historical Accounts*" (DEPCHA), a Mellon funded cooperation of five US partners and the *Centre for Information Modelling* at Graz University, aims to link the knowledge domain of economic activities to historical accounting records. After the discussion about common entities in historical financial records in the beginning of this paper, the second part focuses on the formalization of these entities within the *Bookkeeping Ontology*. The third section defines a workflow to publish RDF data, as part of the digital editions of historical financial records, as *Linked Open Data*. In conclusion, future challenges and results concerning ontology engineering, retrieval and visualization functionalities of the web prototype[2] are discussed briefly.

2 Historical Financial Records and Relevance

Historical financial records provide rich and highly structured data sets over long periods, containing substantial amounts of individual information. This individual information is often not in the core of research interest. Instead, the records acquire their significance in aggregation of the single entries. Pure transcription does not cover the full range of dimensions of such a source: the linguistic/textual, the quantifiable and the semantic dimension [Vo15], [Vo16]. For research purpose, historical sources are subject to a transformation process towards (linked) information sources that can be used in various research scenarios. In order to illustrate this we will discuss three case studies of project partners and their respective research interests, which go far beyond economic and administrative aspects.

The **George Washington Financial Papers (1748-1799)**[3] gives insight into the life of George Washington and other topics such as the material culture, social history, manufacturing and agriculture. The financial papers exist as digital edition, created and published via a *Drupal*[4] based editorial platform, and aim to make Washington's records freely accessible. The platform allows editing and publishing financial documents and gives the users the possibility to perform simple analytical functionalities. Samples of research questions that could be of interest to historians are: How much money did Washington spent annually and for which specific commodities? Which role slave trade plays in his business? How did the price of certain commodities fluctuate? What did the network of partners look like and who did business with him? How was the value of tobacco calculated through different currencies [St14]?

The **Wheaton Accounts (1828-1859)** contain a daybook[5] of a store selling commodities of daily life. The digital edition follows a TEI/XML approach. It extends the range of questions to historical narratives and geographical information. It is interesting to follow an individual or a family as they appear in the daybook over time and reconstruct their social background

[2] DEPCHA Prototype, `https://gams.uni-graz.at/depcha`

[3] The George Washington Financial Papers Project, `www.financial.gwpapers.org`

[4] Open-source content management framework, `www.drupal.org`

[5] Daybook of L.M. Wheaton's Store, expenses of building houses and barns, and expenses of constructing Wheaton Female Seminary buildings, `http://hdl.handle.net/11040/17982`

for a historical narrative. The same applies to geographic information allowing to track geographic relationships of people or the origin of commodities [TB13].

The digitization project of the **Stagville Financial Papers (1767-1892)**, including daybooks and ledgers from the Stagville plantation store in North Carolina, follows an open science and crowdsourcing approach using *From the Page*[6] to transcribe and encode the material. Research questions in this context include the numbers and connections of customers, as well as commodities, which "go together". Furthermore, economic dependencies (who is in debt of whom) as well as the social status of customers (e.g. free/enslaved) are of interest [BA15].

3 Methods

Common structures can be drawn from the research questions mentioned above. To do so, data must be prepared and structured according to a formal set of rules. The ***Bookkeeping Ontology***[7], a conceptual data model based on the *REA* [Mc82] model and CIDOC CRM[8], is developed in an ontology engineering process, involving historians, software developers and digital humanists. The ontology is published in a stable version in *GAMS*[9] [SS18] and in *OntoMe*[10] and can be further discussed by the scientific community.

The *Bookkeeping Ontology* formalizes the interpretation of a transaction (*bk:Transaction*) as combination of transfers (*bk:Transfer*) of measurable objects (*bk:Measurable*) from one accounting object *(bk:Between)* to another. *bk:Between* defines an abstract class, which unites bookkeeping categories (*bk:Accounts*, e.g. a cash account) and actors (individual *bk:Party* e.g. Washington or an unknown group of individuals *bk:Group* e.g. four farmers). Its physical representation in a historical source is an entry in a written accounting record (*bk:Entry*). The *bk:Entry* is an information fragment of a *bk:Transaction* often naming only one party, while the other party is implicit in the textual context of the entry. Further information on the temporal (*bk:when*), spatial (*bk:where*) dimension of a *bk:Transaction* as, well as the status (*bk:status*) of it ("*partly paid*"), can be expressed optionally. In regard to one of our research questions named above a *bk:Transaction* can be assigned to a specific context. Every transaction consists (*bk:consistsOf*) at least one transfer (*bk:Transfer*). A single *bk:Transfer* describes the action of transferring a *bk:Measurable* in one direction (*bk:from* or *bk:to*). *bk:Measurable* is defined as everything that can be quantified. It has subclasses for economic goods (*bk:EconomicGood*, as labor: *bk:Service* or as physical things: *bk:Commodity*) and money (*bk:MonetaryValues*). *bk:Measurable* is describe by its quantity (*bk:quantity*) and the unit of calculation (*bk:unit*). The *bk:Entry* is described by the transcription fragment of the original source (*bk:text*). *bk:EconomicGoods* can be

[6] Crowdsourcing manuscript transcription plattform, `www.fromthepage.com`
[7] Bookkeeping Ontology in DEPCHA, `https://gams.uni-graz.at/o:depcha.bookkeeping`
[8] Semantic framework for mapping cultural heritage information, `www.cidoc-crm.org`
[9] Humanties Asset Management System at Graz University, `https://gams.uni-graz.at`
[10] Ontology Management Environment, `ontologies.dataforhistory.org`

categorized (what is measured) and can be assigned a price. A *bk:Transfer* can be carried out by (*bk:by*) someone who conducts the transfer process in place of the business partner (*bk:Agent*). When writing it down into the ledger, accounting categories (*bk:debit* and *bk:credit*) are coded optionally.

The DEPCHA web prototype is realized in the GAMS infrastructure[11], an open source, FEDORA[12] based digital repository for storing and publishing data in the humanities. Digital objects, containing multiple data streams and methods, allow disseminating data via HTML, as archival data in XML, and via various APIs. Furthermore, GAMS implements a disseminator for RDF data via the triplestore Blazegraph.[13] An encapsulated query object stores predefined SPARQL queries including fulltext search over the RDF data sets. This provides data for retrieval, analysis and visualization [St18].

4 From Digital Edition to Linked Open Data

TEI/XML serves as an interchange format for the data from different systems (*Drupal*, *From the Page*, or generic CSV). Historical financial records are transcribed and annotated in these respective formats or directly in TEI/XML. TEI allows structuring the textual dimension of a source, marks up text-specific phenomena and normalizes places, dates or persons. The semantic relations covered by the Bookkeeping Ontology are inserted into the XML/TEI via attribute @*ana*. It allows a global, multiple and outgoing annotation of any structure in the TEI markup. During ingest into GAMS, XSLT extracts the annotated structures in the XML/TEI and transforms it into RDF. The repository stores this data in the triple store. This approach has already been successfully applied in other projects regarding historical sources, like the *Municipal accounts of the city of Basel 1535-1611*[14], the *Urfehdebücher of the city of Basel - digital edition*[15] [PV17]. The following example illustrates the workflow from the historical source to RDF data. Following figure represents the origin of an entry in the Stagville Accounts.

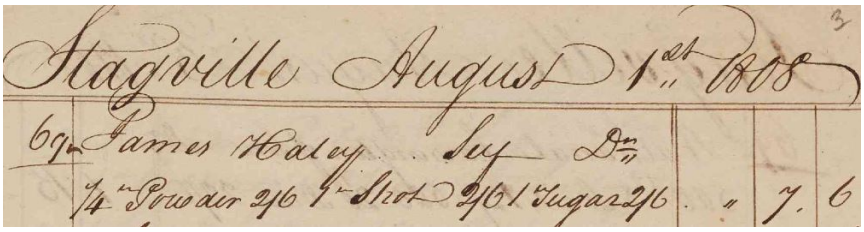

Fig. 1: Entry in the Stagville Accounts

[11] GAMS and Cirilo Client: Policies, documentation and tutorial, https://gams.uni-graz.at/o:gams.doku
[12] Flexible Extensible Digital Object and Repository Architecture, https://duraspace.org/fedora
[13] Triplestore and graph database, https://www.blazegraph.com
[14] Municipal accounts of the city of Basel 1535-1611, https://gams.uni-graz.at/srbas
[15] Urfehdebücher of the city of Basel - digital edition, https://gams.uni-graz.at/ufbas

In addition to the table structure of this example, historical financial records can also be represented in other textual structures, such as continuous text or lists. To make the connection to the *Bookkeeping Ontology* comprehensible, referencing concepts are identified by the bk-prefix in the XML annotation. The entry in the sample above can be interpreted as follows: The person James Haley (*bk:Party*) buys ¼ lb of powder, 1 lb of shot and 1 lb of sugar (*bk:EconomicGood*) from Stagville (*bk:Party*) and transfers in return the monetary value of 7 shilling and 6 pence to Stagville (*bk:MonetaryValue*). Some information is not explicitly mentioned in a *bk:Entry* and could be found in the header or is known by the editor of the source. As semantics are defined through the attribute @*ana*, the textual structure of the TEI document is not relevant for further processing and the full expressiveness of TEI can be used.

The starting point is to define a container for a *bk:Transaction* by annotating the textual representation of an *bk:Entry* by *ana="bk:entry"*. The following simplified XML/TEI snippet illustrates this very entry and its conceptual counterpart *bk:Transaction*, which consists of two *bk:Transfers*. One of them transfers powder, sugar and shot, the other *bk:Transfers* transfers 7 shilling and 6 pence.

List. 1: Simplified XML/TEI-Snippet

```
<head>
  <name ana=''bk:to'' ref=''depcha:stagville''>Stagville</name>
  <date ana=''bk:when'' when=''1808−08−01''> August 1st 1808</date>
</head>
<table>
  <head>
   69 <name ana=''bk:from'' ref=''depcha:pers.2''>James Haley</name>
   <span ana=''bk:status''>Self Dr.</span>
  </head>
  <row ana=''bk:entry'' xml:id=''Transaction−0''>
    <cell>
     <measure ana=''bk:commodity'' commodity=''wd:Q12861'' quantity=''0.25''
       unit=''wd:Q100995''>1/4 lb Powder <measure ana=''bk:price''
         quantity=''0.33'' unit=''wd:Q213142''>2/6</measure>
     </measure>
    </cell>
    <cell>
     <measure ana=''bk:money'' quantity=''7'' unit=''wd:Q213142''>7</measure>
    </cell>
    <cell>
     <measure ana=''bk:money'' quantity=''6'' unit=''wd:Q234129''>6</measure>
    </cell>
  </row>
</table>
```

To elaborate the given code further, I am going to explain the entries in detail. Besides trivial markup, e.g. the date of the entry using *<date ana="bk:when">*, the flow of money is expressed through the values *ana="bk:from"* and *ana="bk:to"*. Since in most sources of this kind a purchase of a commodity is documented, this is an assumption. The flow of economic goods - a commodity or service - can then be deduced from this conclusion. The single entry usually names only one party in the transaction, while the shopkeeper Stagville is already identified in the heading of the account or as external archival knowledge. Furthermore, the text "*Self Dr.*" in the header of the entry can be interpreted as transaction and booking status and expresses that the entry was made by that person on the debit side of a double entry style ledger.

Another important aspect is the normalization of entities such as persons, places, currencies, commodities or services, which can become very complex in the context of such a source. For example, certain weight units or currencies are affected by regional and temporal differences. Due to the large variety of context sensitive entities, normalization of the data is achieved by linking to *Wikidata.org* or by extending workflows for other LOD vocabularies, which seems useful. The use of Wikidata, however, allows to create independent new concepts, which are not represented in any vocabulary yet. The following example illustrates this: another dataset ingested in DEPCHA contains lists of goods from the city of Regensburg in the 13th century.[16] In this data set the term "*Scheffen Hafer*" is used and refers to a quite specific temporal and regional historical unit, a bushel of oat. The workflow described above allows to directly connect such terms to *Wikidata.org* concepts by adding the Q-number directly to an attribute. This pragmatically approach is supported by tools like OpenRefine[17] for semi-automatic identification. As a result of this workflow an XSLT extracts the annotated structures in the TEI to RDF - the following RDF/Turtle snippet shows the processed XML/TEI:

List. 2: Simplified RDF/Turtle

```
@prefix bk: <https://gams.uni-graz.at/o:depcha.bookkeeping#> .
@prefix depcha: <https://gams.uni-graz.at/depcha#> .
@prefix wd: <https://www.wikidata.org/wiki/> .

<depcha:Transaction-0> a bk:Transaction ;
  bk:consistsOf <depcha:Transfer-1>, <depcha:Transfer-2> ;
  bk:when "1808-08-01" ;
  bk:text "1/4 lb Powder 2/6 1 lb Shot 2/6 7 6" .

<depcha:Transfer-1> a bk:Transfer ;
  bk:transfers <depcha:Measurable-1> ;
  bk:from <depcha:stagville> ;
  bk:to <depcha:pers.2> .
```

[16] StadtA Regensburg, Cameralia 3

[17] Open source desktop application for data cleanup and transformation, `https://openrefine.org`

```
<depcha:Transfer-2> a bk:Transfer ;
  bk:transfers <depcha:Measurable-2-1> ;
  bk:from <depcha:pers.2> ;
  bk:to <depcha:stagville> .

<depcha:Measurable-1> a bk:Commodity ;
  bk:unit <wd:Q100995> ;
  bk:quantity "0.25" ;
  bk:commodity <wd:Q2908004> ;
  bk:price <depcha:Price-1> ;
  bk:text "1/4 lb Powder" .

<depcha:Measurable-2-1> a bk:Money ;
  bk:unit <wd:Q213142> ;
  bk:quantity "7" .

 <depcha:pers.2> a bk:Between;
  bk:name "James Haley".
```

5 Conclusion

This project work has shown that workflows can be defined transferring historical financial records, and its annotations according the *Bookkeeping Ontology*, from different data sources, via a standardized XML/TEI, to RDF. The outcome of this process enables us to implement basic functionalities when dealing with entities in the sources and make digital editions on a shared platform - the DEPCHA prototype - available. In addition to very general structures within transaction processes in historical financial documents, other source-specific phenomena have been identified, which the ontology has not yet been able to deal with. An example for this is the annotation of unspecified groups of participants, such as four farmers and a baker, who equally pay a certain tax. In an iterative ontology engineering process, comparable to a hermeneutic circle, the project team discussed such phenomena during so-called "*deep-dives*", trying to understand the sources together to further improve our knowledge domain. In this particular case *bk:Group* and *bk:Tax* were added to the ontology.

The advantages in summary: a common vocabulary makes communication in the project team easier and it can be reused by other projects. Through the reuse of the model the development process is promoted, in which existing concepts can be criticized and missing concepts can be added. This leads to a more widely accepted and comprehensible conceptual model. At the data level in the sense of LOD, references to *Wikidata.org* support interoperability and further use of the data in different scientific disciplines. Synergies arise

in particular when this data can be used for the implementation of generic information retrieval, analyzation and visualization functionalities, facing project-specific use cases. Such functionalities allow us to use larger amounts of data more efficiently. The prototype has shown that the JavaScript library *d3.js* can be used to implement such functionalities based on a SPARQL result for the above-mentioned use cases. A future challenge in this endeavor comes with the difficulty to normalize classifications of currencies, goods and services. The assignment of these classifications to superordinate unified platforms, such as *Wikidata.org*, or independent taxonomies in the *teiHeader* may allow a higher level of normalization, even though this may still be work in the future.

References

[BA15] Brumfield, Ben; Anna, Agbe-Davies: , Encoding Account Books Relating to Slavery in the U.S. South. https://medea.hypotheses.org/182, 2015. Accessed: 2019-03-20.

[JAT85] Jarausch, Konrad Hugo; Arminger, Gerhard; Thaller, Manfred: Quantitative Methoden in der Geschichtswissenschaft. Wissenschaftliche Buchgesellschaft, 1985.

[Mc82] Mccarth, W.: The REA accounting model: A generalized framework for accounting systems in a shared data environment. Accounting Review, pp. 554–578, 1982.

[PV17] Pollin, Christopher; Vogeler, Georg: Semantically Enriched Historical Data. Drawing on the Example of the Digital Edition of theÜrfehdebucher der Stadt Basel". In: WHiSe@ ISWC. pp. 27–32, 2017.

[SS18] Stigler, Johannes; Steiner, Elisabeth: GAMS–Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 71(1):207–216, 2018.

[St14] Sterzer, Jennifer: Working with the Financial Records of George Washington: Document vs. Data. Digital Studies/Le champ numérique, pp. 554–578, 2014.

[TB13] Tomasek, Kathryn; Bauman, Syd: Encoding financial records for historical research. Journal of the Text Encoding Initiative, (6), 2013.

[Th17] Thaller, Manfred: Historical Information Science: Is there such a Thing? New Comments on an old Idea [1993]. Historical Social Research/Historische Sozialforschung. Supplement, pp. 260–286, 2017.

[Vo15] Vogeler, Georg: Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert? In (Baum, Constanze; Stäcker, Thomas, eds): Grenzen und Möglichkeiten der Digital Humanities. 2015.

[Vo16] Vogeler, Georg; Jahnke, Carsten; Gleba, Gudrun; Cordes, Albrecht; Franzke, Cordula A; Laczny, Joachim; Würz, Simone: The Content of Accounts and Registers in their Digital Edition. XML/TEI, Spreadsheets, and Semantic Web Technologies. In (Sarnowsky, Jürgen, ed.): Konzeptionelle Überlegungen zur Edition von Rechnungen und Amtsbüchern des späten Mittelalters, pp. 13–41. Göttingen, 2016.