CREATING TIME CAPSULES FOR COLONIAL BOTANICAL DRUGS IN THE EARLY MODERN LOW COUNTRIES

Zervanou, Kalliopi
K.A.Zervanou@uu.nl
Utrecht University, The Netherlands

Klein, Wouter
W.Klein@uu.nl
Utrecht University, The Netherlands

Van den Hooff, Peter
P.C.vandenHooff@uu.nl
Utrecht University, The Netherlands

Bron, Marc
mbron@yahoo-inc.com
Yahoo! Labs, London, UK

Wiering, Frans
F.Wiering@uu.nl
Utrecht University, The Netherlands

Pieters, Toine
T.Pieters@uu.nl
Utrecht University, The Netherlands

Category: Short paper

1 Introduction

The digitisation efforts undertaken by cultural heritage institutions have been gradually transforming conventional historical research methods, if anything, by making historical data sources available in electronic form. However, although data sources are increasingly available in digital form, they are typically not easy to access and comprehend. Historians face the problem of identifying those sources and the material relevant to their research in pools of information scattered across various archives, libraries and collections, often lacking metadata annotations relevant to historical research and the means to associate all relevant sources relating to a research topic of interest.

In terms of metadata, different metadata standards currently exist both within the same institution and across different institutions. For example, bibliographic data is generally described using the MARC21 standard (Library of Congress, 2010), analogue archival data is described by EAD (Library of Congress, 2002), while multimedia material may be described using a range of formats, such as RAI, INA, and IMMIX (Bauer et al., 2005). Recent years have seen increasing efforts in standardisation and integration of data formats and data sources through standard data models, such as CIDOC-CRM (Doerr, 2003) and TEI (TEI consortium, 2014), or other, linguistically motivated annotation metadata, such as FoLiA (Van Gompel and Reynaert, 2013), as well as the use of thesauri and controlled vocabularies.

However, metadata descriptions have been typically both created by and addressed to librarian and archivist experts who have been expected to assist visitors in their search. For this reason, they primarily refer to bibliographic descriptions (e.g. author/creator, title, etc.), or physical descriptions (e.g. size, shape, material, etc.), and location. Thus, they fail to address the needs of researchers working with cultural heritage data and they constrain the exploration of available information based on more intuitive semantic and content criteria.

In this paper, we present a methodology for linking and associating existing historical resources so as to, not only aggregate the respective digital information available, but also to contextualise existing resources about a given topic of interest by semantically enriching and linking these resources and by automatically detecting topical, temporal and spatial relations in structured (existing databases, dictionaries and thesauri) and unstructured (free-text) data.

Our work employs the concept of *'time capsule'* as envisaged by Andy Warhol[1] to provide integrated exploration of cultural heritage data. Our *time capsules* are user-generated collections of data and information of interest. They are representations of not only aggregated data for a given query, but also visualisations of topically and spatio-temporally contextualised digital heritage items in a way that supports innovative forms of data overviews, sharing and manipulation. Our notion of *time capsule* aims at both academic scholars and other groups of professionals or amateur researchers who can "create and open the capsules" via our platform and use the data of interest. They are thus provided with the digitised background material to re-contextualize knowledge in a specific field of research. The demand-driven approach guarantees the central position of the individual scholar's intuition: Time Capsules challenge a scholar or an amateur researcher to critically select, interpret and interrelate large amounts of heritage data.

The novelty of this work, which is still in progress, lies in the integration of existing structured data, such as databases, dictionaries and thesauri, originally created within the context of different research disciplines (linguistic, pharmaceutical, medical, botanical and historical research), with free-text sources from the early modern period into a semantic interoperable ontology and respective knowledge base. This interdisciplinary resource building method which exploits semantic web, linked data, reasoning and text mining technologies presents a novel research paradigm that will enable numerous digital humanities researchers to examine and contextualise heritage data in innovative ways.

2 The History of Medicinal Plants from the New World Case Study

As proof of concept for semantic interoperability, our work focuses on cultural heritage data relating to the history of medicinal plants in the Low Countries roughly from the moment natural drug components from the New World started penetrating Europe until the introduction of chemical and synthetic drugs in the nineteenth century (i.e. 1550-1850). The resulting system aims at enabling a historian to investigate the trajectories of colonial drug components in the Low Countries, and the patterns of correspondence, trade and knowledge exchange that lie underneath it. Our case study is meant to provide the aggregated data required for new information to surface, which an individual researcher is unable of producing by means of traditional historical research alone.

The factors leading to the adoption of a given drug component are revealing of the dynamics of the medical market at that historical period of time. Recent research showed that drug adoption has only partly to do with therapeutic qualities. Equally important are non-medical factors such as public acceptance, marketing and availability (Pieters, 2004; Gijswijt-Hofstra et al., 2002; Friedrich and Müller-Jahncke, 2009). A special focus on the trajectories of individual drugs has proven itself a valuable research method in this respect. For colonial botanical drugs in the early modern period, several publications have paved the way for this novel approach in the history of tropical

[1] From the early '70s until his death in 1987, Warhol selected items from correspondence, newspapers, gifts, photographs, and other material and preserved them in sealed boxes, which he marked with a date or title. These so-called time capsules provide a unique view into Warhol's private world, as well as an enlightening window on the interrelatedness of culture, media, politics, economics and science in the '70s and '80s.

medicine (e.g. Vöttiner-Pletz, 1990 on guaiacum, used to treat syphilis; Foust, 1992 on rhubarb; Jarcho, 1993 on Peruvian bark).

Another interesting aspect of our case study relates to the circulation of knowledge and goods in the early modern period with particular attention to exotic botany and-pharmacy in the Netherlands (as in e.g. Wilson, 2000; Schiebinger and Swan, 2005; Cook, 2007; Dauser, et al., 2008; Dupré and Lüthy, 2011; Snelders, 2012; Van Gelder, 2012). Knowledge about the vicissitudes of individual drugs across time and place may enrich current debates about the circulation of knowledge and goods. Moreover, sequenced data about colonial drugs may reveal trajectory patterns and the mechanisms of trade and exchange in the early modern period.

These aspects of our case study make it thus an excellent ground for digital research methods experimentation and a resource for other humanities areas, such as pharmacy, ethno-botany, philosophy and science of the enlightenment period, and socio-economic studies of the respective colonial trade period.

3 Creating Time Capsules[2]

Our data sources are of two main categories: structured database data, referring to linguistic, pharmaceutical and botanical data, such as the Dutch Chronological Dictionary and the Economic Botany Database, and unstructured free text corpora collections, such as pharmacopoeias, medical books, and other Early Dutch text corpora. The complete list of our current data sources is illustrated in Table 1.

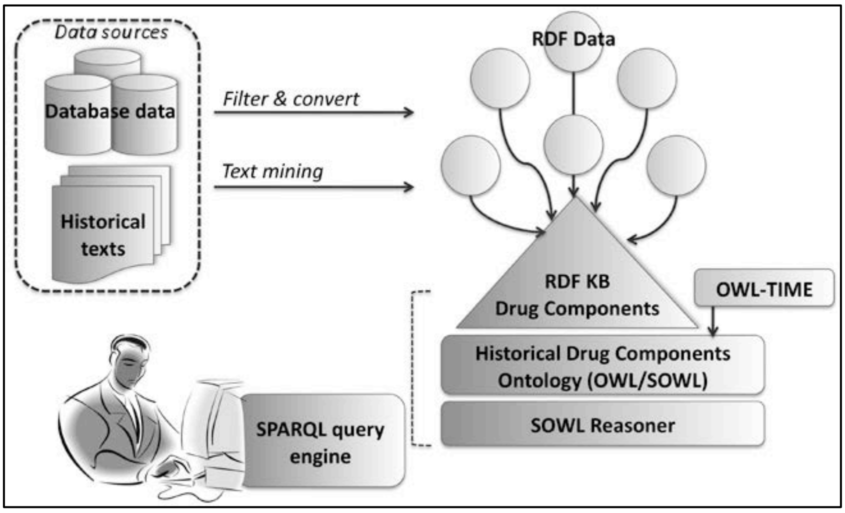

Figure 1 Time Capsule architecture overview

An important consideration and challenge in our database data, lies in their various formats (relational database, excel tables, etc.) and in our text data (OCR errors and transcribed data in Latin or Early Dutch). Most importantly, a serious issue lies in the spelling and semantic variation of both drug component as well as plant names and in the ambiguity of their taxonomic classification, which introduces fuzziness into our data classification. For example, a drug ingredient instance, may or may not originate from a certain plant, which may or may not belong to a given plant taxonomy class and may or may not correspond to a given contemporary name.

For the semantic integration of our data, we adopt a linked-data approach whereby our data sources are first converted into RDF and then linked to our Historical Drug Components Ontology, so as to enrich existing information of the respective KB, as illustrated in Fig. 1. In order to address the issues associated to variation, we incorporate historical lexicon resources and we further enrich these by application of

[2] http://www.timecapsule.nu/

automatic spelling variation detection (Reynaert, 2014). The issue of uncertainty in our data classification is currently addressed by adopting a fuzzy classification approach, whereby an ambiguous instance may be classified into more categories, e.g. a name can be an instance of both a plant concept and a drug component concept. Finally, our structured data sources are used as a resource in detecting relevant information about drug components and their potential use in text documents. The resulting semantic annotations in these historical text corpora are in turn also converted into RDF and automatically linked to our structured sources, so that original historical written evidence is provided to the researchers.

Table 1: Time Capsule Structured and Free-text Data

Structured data	
Botanical, pharmaceutical, historical and image resources	
Ontology of Historical Drug Components *historical Drug Components information found in historical pharmaceutical sources*	*National Museum for the History of Pharmacy*
Economic Botany Database *metadata of objects in the Economic Botany Collection*	*Naturalis Biodiversity Center*
BRAHMS *metadata for 1.2 million records of plant collections*	*Naturalis Biodiversity Center*
Snippendaalcatalogus *inventory of plants of the Snippendaal's 1646 Amsterdam botanic garden*	*Hortus Botanicus Amsterdam*
IrisBG *current information about the plants of the Amsterdam botanic garden, including pictures*	*Hortus Botanicus Amsterdam* *http://dehortus.gardenexplorer.org/*
Dutch Nature Images Collection *images of flora & fauna in the Netherlands & the Dutch Antilles*	*Netherlands Institute for Sound and Vision-Stichting Natuurbeelden* *http://www.natuurbeelden.nl/*
RADAR *geographical & research data about botanical macro remains collected during archaeological excavations on Dutch territory*	*Cultural Heritage Agency (RCE)*
Lexical resources	
PLAND *plant names in various Dutch dialects, including sources, dates, locations and name distributions*	*Meertens Institute* *http://www.meertens.knaw.nl/pland/*
Chronological Dictionary *historical/etymological Dutch dictionary including first observed sources and dates*	*Meertens Institute* *http://dbnl.org/tekst/sijs002chro01_01/*
GiGaNT *diachronic Dutch lexicon (6th century-now), including spelling variations, proper names and morpho-syntactic information*	*Institute for Dutch Lexicology (INL)* *http://www.inl.nl/onderzoek-a-onderwijs/projecten/gigant*
Free-text data	
Looted Letters *transcriptions & metadata of c. 3,500 letters, taken as loot from Dutch ships during the Anglo-Dutch wars (1652-1805)*	*Meertens Institute* *http://www.gekaaptebrieven.nl/*
Letters as Loot *linguistically analysed subset of c. 1,000 letters from the Looted Letters collection*	*Institute for Dutch Lexicology (INL)* *http://brievenalsbuit.inl.nl/*
CKCC corpus *transcriptions & metadata of c. 20,000 letters from the correspondence of 17th century scholars in the Netherlands*	*Huygens ING (and partners)* *http://ckcc.huygens.knaw.nl/*
DBNL corpus *subset of DBNL 16th-19th century texts, including literary, medical, biographical and other texts*	*DBNL – National Library of the Netherlands (KB)* *http://dbnl.org/index.php*
Pharmacopoeias *collection of scanned images of apothecary data*	*Google Books*

Apart from data aggregation, our *Time Capsules* should allow for information re-contextualisation. This issue is partially addressed by our semantic data integration, thus providing the researcher with different but related information sources and aspects (botanical, linguistic, historical). Another important aspect in contextualisation lies in making explicit the relation of data to space and time, thus allowing the virtual spatio-temporal “reconstruction” of the knowledge transfer trajectories. For this purpose, we exploit a combination of spatio-temporal information in our data sources with spatio-temporal reasoning using OWL-Time (W3C, 2006) and SOWL (Batsakis and Petrakis, 2011), an ontology framework for representing and reasoning over spatio-temporal information in OWL.

In the current version of our system, our structured and free-text sources are being processed and integrated, while a demo interface is gradually developed for querying the aggregated data. Our next steps include the development of an exploratory search interface and visualisations of data overviews. A particular challenge lies in the development of a user-friendly web interface for querying our RDF data using SPARQL (W3C, 2013). Future work finally includes the application of our system to a different case study and different types of data (text and images), so as to test its portability and its functionality in a different domain.

References

Batsakis, S. and Petrakis, E. G.M. (2011) SOWL: A Framework for Handling Spatio-temporal Information in OWL 2.0. In: Bassiliades, N., Governatori, G. and Paschke, A. (eds.), 5th International Symposium on Rules: Research Based and Industry Focused (RuleML' 2011), LNCS vol. 6826:242–249. Springer.

Cook, H.J. (2007). Matters of exchange: commerce, medicine, and science in the Dutch Golden Age. New Haven: Yale University Press.

Dauser, R., Hächler, S., Kempe, M., Mauelshagen, F., Stuber, M. (2008). Wissen im Netz: Botanik und Pflanzentransfer in europäischen Korrespondenznetzen des 18. Jahrhunderts. Berlin: Akademie Verlag.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine, 24(3):75.

Dupré, S. and Lüthy C. (2011). Silent messengers: the circulation of material objects of knowledge in the early modern Low Countries. Berlin: LIT Verlag.

Foust, C. (1992). Rhubarb: the wondrous drug. Princeton: Princeton University Press.

Friedrich, C. and Müller-Jahncke, W.-D. (eds.) (2009). Arzneimittelkarrieren: zur wechselvollen Geschichte ausgewählter Medikamente: die Vorträge der Pharmaziehistorischen Biennale in Husum vom 25-28. April 2008, Stuttgart: Wissenschaftliche Verlagsgesellschaft.

Gijswijt-Hofstra, M., Van Heteren, G.M. and Tansey, E.M. (eds.) (2002). Biographies of remedies: drugs, medicines and contraceptives in Dutch and Anglo-American healing cultures. Clio medica 66, Amsterdam: Rodopi.

Jarcho, S. (1993). Quinine's predecessor: Francesco Torti and the early history of cinchona. Baltimore and London: Johns Hopkins University Press.

Library of Congress (2002). Encoded archival description (EAD), version 2002. Encoded Archival Description Working Group: Society of American Archivists, Network Development and MARC Standards Office, Library of Congress. http://www.loc.gov/ead/ (accessed 17 February 2015).

Library of Congress (2010). MARC standards. Network Development and MARC Standards Office, Library of Congress, USA. http://www.loc.gov/marc/index.html (accessed 17 February 2015).

Pieters, T. (2004). Historische trajecten in de farmacie: medicijnen tussen confectie en maatwerk. Inaugural lecture – Hilversum.

Reynaert, M. (2014). TICCLops: Text-Induced Corpus Clean-up as online processing system. In: Proceedings of COLING 2014 System Demonstrations, Dublin, Ireland, pp. 52–56.

Schiebinger, L. and Swan, C. (2005). Colonial botany: science, commerce, and politics in the early modern world. Philadelphia: University of Pennsylvania Press.

Snelders, S. (2012). Vrijbuiters van de heelkunde: op zoek naar medische kennis in de tropen 1600-1800. Amsterdam: Atlas.

TEI Consortium, eds. (2014) TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.7.0.:16th September 2014. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (accessed 17 February 2015).

Van Gelder, E. (2012). Bloeiende kennis: groene ontdekkingen in de Gouden Eeuw. Hilversum: Verloren.

Van Gompel, M. and Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study; Computational Linguistics in the Netherlands Journal; 3:63-81.

Vöttiner-Pletz, P. (1990). Lignum sanctum: zur therapeutischen Verwendung des Guajak vom 16. bis zum 20. Jahrhundert, Frankfurt am Main: Govi-Verlag.

W3C (2013). SPARQL 1.1 Overview. W3C Recommendation 21 March 2013. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/ (accessed 17 February 2015).

W3C (2006). Time Ontology in OWL. W3C Working Draft 27 September 2006. http://www.w3.org/TR/owl-time/ (accessed 17 February 2015).

Wilson, R. (2000). Pious traders in medicine: a German pharmaceutical network in eighteenth-century North America. University Park: Pennsylvania State University Press.