Session B

1. Linking multi-disciplinary data sources for a historical research platform

Kalliopi Zervanou[1], Wouter Klein[2], Peter van den Hooff [2], Frans Wiering[1] and Toine Pieters[2]
[1] Information & Computing Sciences Department, Utrecht University
[2] Freudenthal Institute, History & Philosophy of Science, Utrecht University

The problem of *information access* is a challenge in making digitised data sources available. Historians need to identify information in digital material pools, scattered across collections and often lacking semantic links to a topic of interest. This problem is addressed by the development of various collection-specific metadata schemas, such as MARC 21 (Library of Congress, 2010), and generic metadata schemas, such as the Dublin Core Metadata Initiative (DCMI, 2011). Moreover, diverse metadata schemas are mapped to each other (Bountouri and Gergatsoulis, 2009), or to custom (Liao et al., 2010) or standard ontologies (Lourdi et al., 2009), such as the CIDOC Conceptual Reference Model (CIDOC, 2006). A dominant trend in recent approaches is the *linked-data* approach (Berners-Lee, 2006; Bizer et al. 2009). Besides information access, the amount and the complexity of information accessible gives rise to an *information presentation* challenge, whereby data overviews should highlight interesting data aspects requiring detailed inspection. Additionally, for digital methods to support collaborative research, the problem of *information validation and sharing* must be addressed. This issue calls for transparency of data and methods and reproducibility of results, or verification of the arguments made. It also entails validation of computational results and algorithmic processes determining information access, in such a way that the eventual data or system limitations and biases are known and the processes are trustworthy and verifiable.

In our work, we address these challenges of information access, presentation, validation and sharing from a twofold research perspective:

I. Integration and semantic linking of existing, multidisciplinary data sources;
II. Development of a research platform that supports access, presentation, validation and sharing of complex, interlinked data.

Our particular domain of application relates to the history of botanical drug components from the New World in the early modern period (17-18th century). More specifically, it concerns highlighting phenomena denoting developmental processes of remedies or *drug trajectories*, such as the evolution of economic importance, ethical attitudes, scientific interests, trade and knowledge circulation (Gijswijt-Hofstra et al., 2002; Pieters, 2004; Friedrich and Müller-Jahncke, 2009; Klein & Pieters, 2016).

For this purpose, we integrate sources comprising of pharmaceutical data, such as the *Pharmaceutical Historical Thesaurus* (Klein & van den Hooff, 2013), archaeobotanical data, such as *RADAR* (van Haaster & Brinkkemper, 1995; RCE, 2013), botanical data, such as the *National Herbarium of the Netherlands* (Creuwels, 2014), the *Economic Botany database* (Hoffman, 2011) and the *Snippendaal Catalogue database* (van Reenen, 2007), colonial trade data, such as the database of the accounting books (*Boekhouder-Generaal*) of the Dutch East India Company (Schooneveld-Oosterling et al., 2013) and linguistic dictionaries, such as the Chronological Dictionary of Dutch (van der Sijs, 2001).

A notable recent approach to the issue of digital formats integration is the one adopted in the Timbuctoo infrastucture (Andersen, 2013). Most approaches opt for conversion to a recommended metadata schema, such as SKOS (Miles & Bechhofer, 2009), or a common data model such as the Europeana Data Model (EDM, 2016). However, apart from the diversity of digital formats an

important aspect in integration lies in the reuse and re-purposing of resources originally built for a different audience and purpose.

In our approach, integration entails concept mapping, not only across disciplines, but also in time. Thus, data source integration calls for support for the evolution of science from the 16th century onwards to re-classify and re-define concepts. Additionally, it entails dealing with phenomena of historical term variation and ambiguity which gradually give way to spelling standardisation and current nomenclature conventions in e.g. botany and biology. Furthermore, we account for under-specificity and ambiguity of information found in historical sources while maintaining associations with potentially related concepts and context. Most importantly, we provide references for information provenance tracing and validation. For these purposes, we resort to designing our own ontology, where e.g. ambiguous terms are connected to multiple concepts, temporal periods and reference sources, and where mappings are provided across essential historical and current taxonomies. Our data sources are semi-automatically enriched with additional information, such as geographical coordinates and named entities. Moreover, inconsistencies within and across data sets are semi-automatically identified and normalised. Finally, data sources are integrated following a linked data approach allowing for extensions to other linked open data and eventually capitalising on techniques such as reasoning, which may extend explicit information in our data sets with implicitly inferred information.

Our *Time Capsule* research platform[9] implements our solutions to information access, presentation and validation challenges. It is a scalable working platform currently querying more than 55 million RDF triples. It is often difficult for a non-expert user to perform queries, either because they are unfamiliar with the required terminology, or because they are unfamiliar with the underlying data model. Our solution to this issue lies in providing two querying strategies, one that supports a faceted, exploratory, guided search and browsing of information by means of links, photos, and keyword auto-completion suggestions and one that supports the creation of ad hoc queries. Our exploratory search mode is intended to engage a wider audience and reveal to both experts and non-expert users the underlying data content and structure. Ad hoc queries are in essence ad hoc RDF SPARQL queries (Prud'hommeaux & Seaborne, 2008) to our data. However, given that most users are neither familiar with SPARQL, nor with the content and structure of our datastore, a query wizard is provided that assists users in forming natural language queries, such as "*Which drug component(s) are made out of the plant Acorus calamus L. and which parts of the plant were used?*"

Search results are presented as an overview of all available information on the query topic and users may "zoom-in" on specific information by following links that provide more detailed geographical, temporal and concept relation visualisations. Such visualisations are mainly intended to provide overviews in the evolution of phenomena related to drug trajectories, such as for instance change in a plant part used as medical ingredient, trade routes of botanical products, or geographical distributions in time of known concepts in Latin/scientific terms vs. lay terms, the latter indicating public knowledge and familiarity with a given plant or drug.

References

Andersen, J. A., Filarski, G. J., Haentjens Dekker, R., Maas, M. & Ravenek, W. (2013). Timbuctoo data repository infrastructure (version 1.0), Huygens ING – ICT, Amsterdam, The Netherlands.

Berners-Lee ,T. (2006). Linked Data. Document version: June 2009. In: Design Issues, W3C. Available online at: https://www.w3.org/DesignIssues/LinkedData .html

[9] Time Capsule system: http://timecapsule.science.uu.nl/timecapsule/#/login Logging in as a *Guest* allows full access to the system functionality except saving your search results.

Bizer, C., Heath T. and Berners-Lee T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, vol. 5(3), pp. 1-22. DOI: 10.4018/jswis.2009081901

Bountouri L. and Gergatsoulis M. (2009). Interoperability between archival and bibliographic metadata: An EAD to MODS crosswalk. Journal of Library Metadata, 9(1-2):98–133.

CIDOC (2006). The CIDOC Conceptual Reference Model. CIDOC Documentation Standards Working Group, International Documentation Committee, International Council of Museums. Available online at: http://www. cidoc-crm.org/.

Creuwels, J. (2014). The National Herbarium of the Netherlands. Naturalis Biodiversity Center, Leiden. Available online at: http://herbarium.naturalis.nl/

DCMI (2011). The Dublin Core Metadata Initiative. Available online at: http:// dublincore.org/.

EDM (2016). Europeana Data Model – Mapping Guidelines v2.3, 18 November 2016, Europeana Network Association. Available online at: http://pro.europeana .eu/page/edm-documentation

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.

van Haaster H. and Brinkkemper O. (1995). RADAR, a Relational Archaeobotanical Database for Advanced Research. Vegetation History and Archaeobotany, vol. 4(2), pp. 117-125, Springer.

Hoffman, B. (2011). The Naturalis Economic Botany database. Naturalis Biodiversity Center, Leiden.

Klein, W. and Pieters, T. (2016). The Hidden History of a Famous Drug: Tracing the Medical and Public Acculturation of Peruvian Bark in Early Modern Western Europe (c. 1650–1720). Journal of the History of Medicine and Allied Sciences, Vol. 71(4), pp. 400–421. DOI: 10.1093/jhmas/jrw004

Klein, W. and van den Hooff, P. C. (2013). Farmaceutische Historische Thesaurus. National Museum for the History of Pharmacy, Utrecht.

Liao, S.-H., Huang, H.-C., and Chen, Y.-N. (2010). A semantic web approach to heterogeneous metadata integration. In: Proceedings of ICCCI ’10, LNCS vol. 6421, pp. 205–214, Kaohsiung, Taiwan. Springer.

Library of Congress (2010). MARC standards. Network Development and MARC Standards Office, Library of Congress, USA. Available online at: http://www.loc. gov/marc/index.html.

Lourdi, I., Papatheodorou C., and Doerr M. (2009). Semantic integration of collection description: Combining CIDOC/CRM and Dublin Core collections application profile. D-Lib Magazine, 15(7/8).

Miles, A. and Bechhofer S. (eds) (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. Available online at: http://www.w3.org/TR/skos-reference

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

Prud'hommeaux, E. and Seaborne A. (eds.) (2008). SPARQL Query Language for RDF. W3C Recommendation, 15 January 2008. Available online at: https:// www.w3.org/TR/rdf-sparql-query/

RCE (2013). RADAR, a Relational Archaeobotanical Database for Advanced Research. Rijksdienst voor het Cultureel Erfgoed, Ministerie van Onderwijs, Cultuur en Wetenschap. Available online at: https://archeologieinnederland.nl/ bronnen-en-kaarten/radar

van Reenen, G. (2007). Snippendaalcatalogus database. Hortus Botanicus Amsterdam. Available online at: http://dehortus.nl/en/Snippendaal-Catalogue

Schooneveld-Oosterling, J., Knaap, G., Karskens, N., Smit-Maarschalkerweerd, D., Tetteroo, S., van den Tol, J., Nijhuis, H., van Wijk, K., Kunst, A., Buijs, J., Jongma, M., Boer, R. (2013). Boekhouder-Generaal Batavia. Huygens ING. Available online at: http://resources.huygens.knaw.nl/ boekhoudergeneraalbatavia

van der Sijs, N. (2001). Chronologisch Woordenboek. Available online at: http://dbnl.org/tekst/sijs002chro01_01/

2. A Linked Data Approach to Disclose Handwritten Biodiversity Heritage Collections

Lise Stork, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
l.stork@liacs.leidenuniv.nl

Andreas Weber, Department of Science, Technology and Policy Studies (STePS), University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
a.weber@utwente.nl

Over the last decade, natural history museums in and beyond the Netherlands have heavily invested in digitizing and extracting biodiversity information from manuscript and specimen collections (Heerlien et al. 2015; Pethers and Huertas, 2015; Svensson, 2015). In particular handwritten fieldnotes describing occurrences of species in nature (see illustration) form an important but often neglected starting point for researchers interested in long-term habitat developments of a specific area and the history of scientific ordering, writing and collecting practices (Blair 2010; Bourget 2010; Eddy 2016). In order to disclose handwritten descriptions of flora and fauna and related specimen and drawings collections, natural history museums usually resort to manual enrichment methods such as full text transcription or keyword tagging (Ridge 2014; Franzoni et al. 2014). Often these methods rely on crowdsourcing, where online volunteers annotate pages with unstructured textual labels (Field Book Project 2016). More recently, curators of archives, data scientists and historians have started to experiment with semi-automatic annotation systems for historical manuscript collections such as the MONK system (Schomaker et al. 2016). Since MONK is a supervised learning system, a large amount of properly recognized textual labels is necessary to safeguard the system's recognition abilities.

Thus, although such practices have the potential to yield high quality data, merely annotating pages with unstructured textual labels raises two problems: First, without suggestions driven by semantic