legacy4reuse

Criteria and Methods for Upcycling Data Collections in Social and Economic History

Online Workshop, 23 November 2023
9 – 18 hrs
Zoom: Register to receive the Zoomlink.

New scientific findings often reuse or repurpose existing printed, digitised or genuinely digital data collections. Disciplines such as archaeology or linguistics use the term legacy collections to indicate that such data collections share certain characteristics. They originate from the historical heritage of the respective discipline; usually, they do not meet today's scientific standards, and often they are documented poorly. Legacy collections are problematic, but hard to replace. They preserve unique data. Sometimes this data is no longer available elsewhere; often it is the result of years of research that cannot be repeated.

Social and economic history often draws on legacy collections, even though they do not usually label their historical data collections as such. One might think of printed series of historical statistics, dictionaries of ancient weights and measures, or published price statistics, customs registers and account books. The reuse of such legacy collections in social and economic history remains strongly person- and project-bound and does not comply, or only to a limited extent, with the principles of FAIR Data. Current research practice thus prevents a sustainable handling of legacy collections in social and economic history.

In order to achieve sustainability, the reuse and repurposing of historical data collections should perhaps take place within an upcycling framework. The digitisation of a source is merely the starting point. Upcycling aims at digital enhancement of the historical data collection. In the process, both the original data genesis and the technical challenges and epistemological requirements for its reuse or repurposing are scrutinized and documented. This facilitates the creation of reconfigured, upcycled data collections that comply with the FAIR data principles and can be reused in a sustainable way.

At the *legacy4reuse* workshop an international group of scholars will present current research with legacy collections in social and economic history and will discuss the methodological implications of their reuse. In so doing, the workshop aims to gain insight into the criteria, methods, and best practices for upcycling data collections in social and economic history.

Preliminary programme

Time slot	Contributors	Title
9 – 9.15	Mark Spoerer, Regensburg Werner Scheltjens, Bamberg	Introduction
	Session 1	
9.20 – 10.05	Sytze H.J. Van Herck, Utrecht Rick J. Mourits, Amsterdam	Upcycling the Dutch civil registry using Linked Data
10.10 – 10.55	Michaela Schmölz-Häberlein, Bamberg	An Ambivalent Legacy - Upcycling Genealogical Research Data Collected before 1945 for a Jewish Prosopography of the Pre-Modern Period
10.55 – 11.15	**Short Break**	
	Session 2	
11.15 – 12.00	Antonio Iodice, Genova	Ready-made legacy collections for reuse? Giuseppe Felloni's paper database of Average sources
12.05 – 12.50	Ulrich Pfister, Münster	Upcycling relational databases: the import tolls of Hamburg (1733-1798)
12.50 – 14.00	**Long break**	
	Session 3	
14.00 – 14.45	Christine Fertig, Münster	Merchant's Handbooks and Commodity Lists in the Digital Space. Transcribing and analysing digitised sources of the 18th century
14.50 – 15.35	Gabi Wüthrich, Zürich	"Tables are tricky". Testing Text Encoding Initiative (TEI) Guidelines for FAIR upcycling of digitised historical statistics
15.35 – 15.50	**Short break**	
	Session 4	
15.50 – 16.35	Lucky Ugbudian, Ebonyi	Legacy Collection and Uses in Nigeria: National Archives Experience
16.40 – 17.25	Shuai Wang, Amsterdam Angelica Maineri, Rotterdam	FAIR Implementation Profile for the SSHOC-NL Socio-Economic History Community
17.25 – 17.30	**Short break**	
17.30 – 18.00	Mark Spoerer, Regensburg Werner Scheltjens, Bamberg	Conclusions

Abstracts

Session 1

Upcycling the Dutch civil registry using Linked Data

Sytze H.J. Van Herck | UU/CLARIAH | 0000-0001-5627-042X | s.h.j.vanherck@uu.nl
Rick J. Mourits | IISG/CLARIAH | 0000-0002-2267-1679 | rick.mourits@iisg.nl

Civil certificates – registrations of birth, marriage, and death – contain a wealth of information regarding occupations, age at marriage, mortality, etcetera. In order to make this information available to scholars within the biological, historical, life, and social sciences, civil certificates have been scanned and manually indexed by employees and volunteers from regional and city archives in the Netherlands. A collaboration between the Central Bureau for Genealogy (CBG), the International Institute for Social History (IISH), and the CLARIAH project aims to upcycle the civil registry data by matching names of individuals appearing on multiple certificates to reconstitute families and life courses. The goal is to create a research-grade dataset in a FAIR format, with proper documentation and provenance information.

The project requires collaboration between socio-economic historians and computer scientists from multiple institutions. The CBG retrieves, merges, stores, and shares indexes of civil certificates from local archives. Currently, the IISH retrieves an Access to Archives (A2A) index from the CBG at irregular intervals. The civil certificate indexes are then converted into a MySQL database and cleaned before being converted into Linked Data (LD). Due to the size of the data and the complexity of the matching procedure, BurgerLinker converts the LD to a HDT format followed by a scalable and fast matching procedure based on a Levenshtein automaton. The Civil Registries Reconstitutions Cleaner (C2RC) tool is being developed next to apply a set of rules to semi-automatically validate clusters of certificates in which a single individual appears.

In our paper we will show how Linked Data's focus on formal descriptions of data structures and open licenses incentivized us to adopt a community metadata standard, develop open-source so□ware, and share results. We also discuss how we facilitate its use by writing manuals, filming tutorials, and offering courses.

An Ambivalent Legacy - Upcycling Genealogical Research Data Collected before 1945 for a Jewish Prosopography of the Pre-Modern Period

Michaela Schmölz-Häberlein, Michaela.Schmoelz-Haeberlein@uni-bamberg.de

Two relational databases were created for the two DFG projects "Fractality and the Dynamics of Jewish Lifestyles in the South of the Old Empire in the 17th and 18th Centuries" and "Religious Difference and Economic Cooperation: Christian-Jewish Business Relations in the Late Period of the Old Empire (1648-1806)". The first database is a prosopographic database that reconstructs the family relationships of the Jewish actors, records their activities and traces their involvement in the Jewish communities. In addition, migrations can be reconstructed. The second database connects the individual actors with their business partners, reconstructs the cooperation between Christian and Jewish partners in trading companies as well as in everyday life, and records individual transactions. The findings from literature and archival research were incorporated into the databases. Legacy collections were used to complete the databases. These legacy collections are "family books" and genealogies published mainly at the end of the 19th and the beginning of the 20th century, as well as unpublished works by Jewish

genealogists, which are preserved mainly in the form of collections at the Leo Baeck Institute in New York. In 1913, the Viennese Rabbi Max Grunwald founded the Archive for Jewish Genealogy, Art History and Museums, which was discontinued with the outbreak of the First World War. In 1924, Jewish genealogists joined together in the Society for Jewish Genealogical Research founded by Arthur Czellitzer, in which Jacob Jacobson, the head of the *Gesamtarchiv der deutschen Juden*, was also involved. The journal of this society appeared until 1938 and was understood as an expert platform for all questions of the ancestry of Jewish men and women. All this information was used in the Office for Genealogical Research (*Reichsstelle für Sippenforschung*), established in 1933, as well as the numerous genealogical card indexes compiled by local researchers, which were based on church register evaluations since the 17th century. The specific historical and political-ideological contexts in which these collections were created raise the question of how one can and should deal with this material today. We fed the relevant information into the relational databases and then checked, validated and correlated it. In this way, we attempted an upcycling of the genealogical data. In an introduction, the challenge of using these materials is critically questioned and documented. The aim of the data project is to make this broad database available to all researchers so that it can be used in the long term.

Session 2

Ready-made legacy collections for reuse? Giuseppe Felloni's paper database of Average sources

Antonio Iodice, ai290@exeter.ac.uk

Giuseppe Felloni was among the first historians to acknowledge the importance of studying average practices. In maritime law, the concept of general average, currently governed by the York-Antwerp Rules first drafted in 1890, refers to the damages incurred by the ship or the cargo and the extraordinary expenses incurred during the voyage as a result of a voluntary action taken by the shipmaster to safeguard the vessel and the goods from an imminent peril. Expenses and damages suffered solely for the benefit of the vessel or the cargo are classified as particular average.

In an article published in 1978, Felloni presented the findings of a sample-based study that examined 180 average practices preserved at the State Archive of Genoa, covering the period from 1599 to 1601. By cataloguing the information contained in these average practices, he did a valuable research on the history of maritime trade and the broader early modern Mediterranean economy. In the following years, he and his students catalogued thousands of additional practices spanning from the sixteenth to the eighteenth centuries. However, Felloni eventually discontinued this research and shifted his focus elsewhere. Perhaps due to the extensive workload required for the systematic cataloging of such sources, no one continued working on average procedures until the starting of the ERC project led by Maria Fusaro. This project, running from 2017 to 2022, culminated in the creation of the award-winning open access online database *AveTransRisk* (World Economic History Conference, Paris). This database contains thousands of Average procedures including those originally registered by Felloni. This contribution aims to address the technical challenges associated with reusing Felloni's work, quantify the time saved in the project as a result, and outline the different solutions we implemented over time to adapt them to modern research standards.

Upcycling relational databases: the import tolls of Hamburg (1733-1798)

Ulrich Pfister

The Hamburg Admiralty, Convoy and Revenue Books document the customs clearance of individual imports via the sea route to Hamburg. The books have been preserved for 36 years from the period 1733-1798. Schneider, Krawehl and Denzel (2001) have recorded the individual entries; there are about 180,000 records. They contain information on the dutiable, the port of origin, the value and a non-standard description of the imported goods. In the book publication, Schneider et al. report series on individual goods or places or regions of origin. For this purpose, they have developed a relational database with MS-Access, which links places of origin and imported goods via concordances with standardised spellings for place names and goods descriptions as well as descriptions for aggregates (countries, goods categories). They generated the published series with database queries, which are part of the database in an exemplary way. I myself (Pfister 2017) have cleaned up the database somewhat, added queries and, above all, supplemented the material with price information (source mainly Gerhard and Kaufhold 2001) in order to estimate import quantities for 40 goods. The project has not been completed; data cleaning has only been partially carried out. Above all, the processing for subsequent use is still pending. For relational databases, I see ensuring accessibility (A in FAIR) and broad usability (R) as challenges. Option (1) consists of a fully developed database that contains dialogues for many conceivable queries. This is convenient for users with limited requirements and knowledge, but it does not provide platform independence and comprehensive usability. Option (2) consists of keeping the data sets underlying the database separately in a platform-independent standard format (.csv) and providing a code written e.g. with R that assembles the database and already contains some queries. This variant would correspond well to FAIR, but makes certain demands on users. The paper presents the project in its current state of development and explains the dilemma addressed in the preparation of the project data for subsequent use.

Session 3

Merchant's Handbooks and Commodity Lists in the Digital Space. Transcribing and analysing digitised sources of the 18th century.

Christine Fertig (Münster)

An important element of European expansion was the growing trade in old and new luxury and consumer goods. In addition to well-known and quantitatively significant goods such as coffee, cotton or porcelain, exotic substances from the East and West reached Europe. During the long 18th century, a body of commercial knowledge about these substances developed in the German-speaking countries, which facilitated trade in exotic substances and enabled the evaluation and targeted marketing of these exotica. At the same time, merchants from Central European regions could hardly fall back on direct contacts in the regions of origin, as they were essentially dependent on imports from the Atlantic trading centres. The production of knowledge took place in specialised publications such as merchant's handbooks, commodity guides, but also pharmacopoeias as well as encyclopaedias and lexicons, which were primarily aimed at the general public. The development and systematisation of this knowledge is traced on the basis of a larger corpus of digitised source material available in various libraries and transcribed by means of OCR (optical character recognition). The evaluation with qualitative and quantitative methods is carried out in MAXQDA, an application specialised in mixed-methods approaches.

"Tables are tricky". Testing Text Encoding Initiative (TEI) Guidelines for FAIR upcycling of digitised historical statistics.

Gabi Wüthrich, University of Zurich, gabi.wuethrich@ub.uzh.ch

In the context of a digitisation project on 1918 pandemic data in Zurich, a project work on digital data management tests the implementation of XML structures based on the Text Encoding Initiative (TEI), an XML-based standardized vocabulary for text structures, on historical statistical tables. The basic idea of the data management project was to prepare tables of historical health statistics in a sustainable way to make them reusable, interoperable, and machine-readable in a platform-independent way.

After Zurich's Central Library had retro-digitized and published the serial statistical print publications from the 1910s and 1920s, the goal was to capture their content semi-automatically with OCR in Excel and to prepare them as XML documents according to the TEI guidelines including a fitting XML schema. Such clearly structured XML documents should be relatively easily convertible into formats readable by a variety of statistical tools.

Converting the tables to Excel and then to XML is not unproblematic, however. As the OCR software failed to capture the table content accurately, both steps needed to be done "by hand", which is error-prone. Ideally, XML export should already be possible in pdfs with OCR. Regarding TEI implementation, tables seem to have a shadowy existence so far – or, as TEI pioneer Lou Burnard remarked: "Tables are tricky". The main reason for this is probably the running text orientation of the existing tools and users. In principle, however, TEI data processing offers the opportunity to think conceptually about the function of data structured in tabular form, and to make changes traceable, especially in serial statistics. This is proved by a project from Basle and Graz using early-modern Basle account books upcycled according to the TEI principles. In addition, the clearly structured text preparation of TEI could provide a training basis to improve the quality of table text recognition.

Session 4

Legacy Collection and Uses in Nigeria: National Archives Experience

Lucky Igohosa Ugbudian, Alex Ekwueme Federal University Ndufu-Alike, Ebonyi, Nigeria, liugbudian@gmail.com, https://orcid.org/0000-0002-8370-7516

The paper examines legacy collection and uses in Nigeria with a focus on national archives. The National archives in Nigeria were established in the colonial era to serve as a repository of data, materials and information on Nigeria socio-cultural, political, economic, religious and the histories of the people, government, issues and events since precolonial era. Materials and reports including oral interviews were documented and copies deposited at the archives while private and government were also mandated to submit copies of their reports to the archives. In the 1940s three national archies were established and located in Enugu, Ibadan and Kaduna through the efforts of British and Nigerian scholars, particularly, Kenneth Dike. In recent year is how to address the challenges confronting access and availability of records and documents in archives. This is the aspect the paper intends to analyse. Data for the paper will be generated from archives, newspapers, journals, books and interview of 11 key informants that will be analysed qualitatively using historical approach which will reveal that the poor maintenance, high fees and destruction of the materials in the archives created access and availability challenges leading some scholars and users to make efforts to digitalised the records in some of the archives. The paper argues that the effort to digitalise the records in the archives with the

aim of enhancing access and preserved the materials from theft, destruction and high fees for access has been relatively successful in guiding users despite opposition by the management and staff of the archives.

FAIR Implementation Profile for the SSHOC-NL Socio-Economic History Community

Shuai Wang | Vrije Universiteit Amsterdam | 0000-0002-1261-9930 | shuai.wang@vu.nl
Angelica Maineri | ODISSEI & Erasmus University Rotterdam | 0000-0002-6978-5278 | angelica@odissei-data.nl

Despite the wide endorsement of the Findability, Accessibility, Interoperability, and Reusability (FAIR)-principles by research institutions and funders, the principles are implemented heterogeneously. Instead of focusing on FAIRification of individual datasets, our work is about decisions made by each community. FAIR Implementation Profiles (FIPs) have been proposed as a solution to streamline the efforts of FAIR implementation and a documentation of the FAIR effort in a community. The SSHOC-NL Socio-Economic History (SSHOC-NL-SEH) community is a sub-community of the Dutch Socio-Economic History community consisting of researchers from different institutions, especially those related to the SSHOC-NL infrastructure. More specifically, they are from the International Institute of Social History (IISH), Utrecht University (UU), and the Vrije Universiteit (VU) Amsterdam. The research output of members of the community includes publishing historical and upcycled datasets as well as their analytical results about social and economic history. The FIP from this community is a first step towards the investigation of FAIR implementation and compares the similarities and differences across communities. These FIPs are published in PDF as well as machine actionable format. Such FIPs could serve as a community standard when publishing upcycled data. More specifically, these FIPs can provide guidance on platforms to publish the data, possible metadata to use, licenses to use. Moreover, the use of FIP enables automatic comparison of decisions made by the SSHOC-NL-SEH community against other communities, which could contribute to the FAIR of upcycled datasets. Finally, based on this FIP, we would like to provide suggestions and future steps to take for improving the reusability of datasets, especially upcycled datasets.