XQuery for Data Integration

Joseph Wicentowski
joewiz@gmail.com
Office of the Historian, U.S. Department of State
United States of America

Clifford Anderson
clifford.anderson@vanderbilt.edu
author.email@domain.com
Vanderbilt University, United States of America

Introduction

This half-day tutorial shows how XQuery integrates digital humanities data from multiple sources and formats. Drawing on the latest features of XQuery 3.1, the instructors demonstrate how to draw together information from the most common structured data formats, namely, JSON, CSV, RDF, and XML. We will teach some of the latest features of the XQuery language, including how to work with maps, arrays, and new functions like json-doc(), parse-json() and json-to-xml().

Specifically, we will explore the following data sources: (1) dictionary data (in JSON) from the Oxford English Dictionary; (2) an Open Publication Distribution System (OPDS)-based ebook catalog that makes publications at the U.S. Department of State searchable, browse-able and downloadable via OPDS-compliant ebook reader apps like Shubook, Hyphen, etc.; (3) an OpenRefine reconciliation endpoint API built to let people run their own lists of people against biographical databases; and (4) interacting with IIIF APIs.

Using a free and easy to install XQuery learning environment, participants (who must bring their own laptops) will gain hands-on experience writing queries against open datasets in CSV, JSON, and RDF and integrating this data with XML. Participants will gain exposure to the latest features of the XQuery language as well as best practices for connecting data across systems and formats. We presuppose that participants will have come with a basic understanding of XQuery.

Brief outline

I. Requesting remote data and storing it into an XML database
II. Querying CSV with XQuery
III. Querying JSON with XQuery
IV. Querying RDF with XQuery
V. Enriching TEI with data from other sources and formats

Each section will include hands-on exercises.

Target audience

Students, scholars, and practitioners who use or are interested in using digital methods in their humanities work in academic departments, libraries, "alt-ac" fields, or their private capacity; no previous programming experience required; some experience with XML or an XML-based format (TEI, EAD, MODS, METS, Atom) useful but not required. Participants will work with a common dataset provided by the tutorial leaders, but they may bring their own datasets for practice during the lab and consultation period.

Tutorial leaders

Clifford B. Anderson is Associate University Librarian for Research and Learning at Vanderbilt University in Nashville, Tennessee. He has a M.Div. from Harvard Divinity School and a Th.M. and Ph.D. from Princeton Theological Seminary. He also holds a M.S. in Library and Information Science from the Pratt Institute in New York City. Cliff started working with XQuery in 2006 before the first official version of the language was released. In 2014, he served as the project leader of the NEH-funded XQuery Summer Institute at Vanderbilt University. He has also taught sessions on XQuery for iterations of Laura Mandel's Programming for Humanists course at Texas A&M and leads the weekly XQuery working group at Vanderbilt University for digital humanists

Joseph C. Wicentowski is the Digital History Advisor in the Office of the Historian at the U.S. Department of State. He received his Ph.D. from Harvard University in modern East Asian history. He started using XQuery in 2007 to analyze and publish the Office of the Historian's TEI-encoded publications and datasets. For more on the project, see Wicentowski (2011). All code and data from the project are freely available on GitHub. He recognized XQuery's potential to empower students, scholars, and practitioners to take control of their own data and build their own applications. But he knew that without resources geared toward people with a humanities background, others would struggle as he first did. He began writing about XQuery in various digital humanities forums, contributing to the XQuery Wikibook online textbook, and giving work-

shops at the TEI@Oxford and Digital Humanities@Oxford Summer School programs. Joe regularly speaks and writes in the fields of history, documentary editing, and open government. He also actively participates in TEI, XQuery, and digital humanities communities, and fosters discussion about XQuery on Twitter at @XQuery.

Bibliography

Wicentowski, J. (2011) "history.state.gov: A case study of Digital Humanities in Government," *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, vol. 1 no. 3 , https://letterpress.uchicago.edu/index.php/jdhcs/article/view/80.