Tracing Johnson’s Circle: Practical Experiments with EAC-CPF

Ellen Doon, Susan Pyzynski, Michael Rush, and Melanie Wisner

Introduction

In the autumn of 2011, staff members from the Beinecke Rare Book and Manuscript Library at Yale University and the Houghton Library at Harvard University mapped out a year-long, collaborative, experimental project to test-drive the newly minted archival standard Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC-CPF). Intrigued with the potential of EAC-CPF to supplement archival description and facilitate discovery of related but disparate collections, the team devised a project centered on lexicographer Samuel Johnson (1709-1784) and his circle of fellow writers, artists, political thinkers, and friends, as well as modern collectors thereof.

The project initially posed these questions: How might separate institutions collaborate in creating and maintaining EAC-CPF records? How might they derive and preserve best practices for this new standard? What makes a content-rich EAC-CPF record? What contextual information and linking will be most beneficial to our researchers? Project deliverables were to be a modest 25 to 50 EAC-CPF records, shared documentation of best practices, and the intangibles of collaboration and promotion of a new standard. This paper will look briefly at project details and perceived benefits; we will then concentrate on challenges and lessons learned as we tried to answer these questions.
**Project staffing and funding**

Beinecke and Houghton share similarities in researchers and staffing structures, and the two libraries care for strongly related collections. The Houghton Library's Early Modern Department holds the Donald & Mary Hyde Collection of Dr. Samuel Johnson, overseen by Curator John Overholt. The Beinecke Library's James Marshall and Marie-Louise Osborn Collection, led by Curator Kathryn James, holds the papers of James Boswell, Johnson's biographer. The two curators donated time early on in the project, collaborating on selection of a cast of characters to define the Johnsonian circle for project purposes. The project's title, "Connecting the Dots: Using EAC-CPF to Reunite Samuel Johnson with His Circle", emphasized the linking ability of EAC-CPF as well as this joining of cross-institutional collections. Archivists comprising the initial project team were Ellen Doon and Michael Rush from Beinecke; and Krista Ferrante, Susan Pyzynski, and Melanie Wisner from Houghton.

Harvard's Office of Scholarly Communication had recently instituted Library Lab, an internal grant program supported by the Arcadia Fund to foster library innovation and collaboration. A Library Lab grant allowed us to hire two temporary project employee-interns who would share the work of researching and hand-crafting the records: Brooke McManus at Houghton and Mary Yearl at Beinecke. They would also compile our collaborative best-practice decisions in a Harvard-hosted wiki eventually to be made public. The two interns were chosen particularly for their academic and practical experience with Encoded Archival Description (EAD) and other metadata standards, the software used to code these standards, and their experience in writing documentation. The resulting team of seven invited Kathy Wisser, Assistant Professor in the
Graduate School of Library and Information Science at Simmons College, to serve as outside consultant for the project; previously chair of the Working Group that developed EAC-CPF, she currently co-chairs SAA’s Technical Subcommittee for Encoded Archival Context (TS-EAC)\textsuperscript{ii} which maintains the standard. She contributed her notable expertise to an all-hands project kick-off meeting in the spring of 2012, during which we reviewed the EAC-CPF tag library and considered project goals. It is notable that team member Mike Rush currently serves as co-chair of the Technical Subcommittee on Encoded Archival Description (TS-EAD)\textsuperscript{iii}, and ex-officio member of the TS-EAC, bringing to the project his own extensive experience with the inner workings of these closely related standards.

**Project description**

The funding for the project covered an 11-month period, so we set out a timeline and project plan to fill 8 months, knowing that some slippage was bound to occur. We spent the first few months setting up the project: installing software; setting up the documentation wiki; creating job descriptions and preparing to hire the two temporary project employee-interns; and finalizing the list of names/persons for the EAC-CPF records. The middle of the project period was taken up with creation of the EAC-CPF records which included evaluating software used, and writing, compiling, and reviewing documentation as the record creation went along.

Finally, the last month or so of the project was spent distilling our best practices document; and working with a developer to install a test instance of the SNAC (Social Networks and Archival Context Project)\textsuperscript{iv} software, an opportunity to view and share the records in a graphical interface (this installation has since been taken down).
We planned for the two project employee-interns to create the EAC-CPF records and to write and compile the documentation. They worked independently creating records, each intern selecting secondary sources from the online and analog resources of their respective research library. Beinecke-Houghton project staff reviewed and edited the records, reviewed the documentation, and tried to answer questions as they arose, elevating some questions to whole team consideration. Given our geographic distance, we relied on email and conference calls to stay current.

Records were created using the XML editor oXygen. The plan also included Harvard's hosting of a local instance of the open source Social Networks and Archival Context Project (SNAC) prototype interface. This local instance could have supported vetting of the records by targeted researchers if time had allowed for publicity and data gathering; with time and technological constraints, we were not able to build this step into the project but consider it important for future work.

We decided initially that a minimum EAC-CPF record would require name authority information, at least one link to another record for a corporate body, person or family and, where possible, at least one link to an archival resource, though our sights were set on much richer encoding. The small number of extant systems able to display and search EAC-CPF records meant we had little to compare to, but we conceived of a "rich" record as accurate and detailed, built upon a wealth of authoritative secondary sources; it also came to mean that wherever the standard offered both a data-centric and a narrative way to encode information, we would use both. To aid in consistency of record creation, Michael Rush developed a
Schematron schema to enforce certain project encoding decisions, to be used in validating records alongside the EAC-CPF schema. This became a valuable part of our project documentation.

[Insert Figure 1]

Project successes

To summarize the successes of the project during our year-long grant period:

- 78 EAC-CPF records--28 more than the maximum expected due in part to faster record creation than anticipated
- Thorough familiarity with the intent of each element of the standard, encouraged by the hand-crafting approach
- A temporary local test instance of the SNAC prototype installed and hosted by the Berkman Center for Internet and Society at Harvard as part of the grant, which enabled us to preview our 78 records within a searchable user interface
- A productive and smooth interstate collaboration, with the satisfaction of learning a new standard from the ground up in good company
- A publicly accessible project documentation wiki, including 78 records, the Schematron schema, and our best practices, which can be found here:

  https://wiki.harvard.edu/confluence/display/connectingdots/
The collaboration was of value not just in and of itself but also in ensuring that one institution's descriptive habits did not prevail, small though the differences may have been. As we hope our records will become part of a greater aggregation, this kind of cross-check was helpful. The collaboration was smooth in part because we knew each other from regional and other gatherings but more importantly because we are all practicing archivists. We brought shared knowledge of basic archival principles, DACS\textsuperscript{vi}, EAD, and authoring finding aids; we did not parse our roles within the project further than distinguishing between the work of the interns and that of the rest of the team, in fact sharing most tasks evenly across the team.

**Challenges and lessons learned**

In spite of the evident productivity of our two interns, one of the chief difficulties was managing this relational data in a flat document-based environment. Early experimentation with ICA-AtoM\textsuperscript{vii} (Web-based archival description software that can both import and export EAC-CPF records – rare functionality for this early stage of the standard) showed that it was not going to serve as a shared record creation and display system as had been hoped, as it lacked functionality for the full set of EAC-CPF elements and attributes. Without such a shared test bed, our project wiki became the de facto exchange point for both records and documentation, resulting in a more cumbersome workflow than we had wished for. Without a shared relational database environment, the growing number of entity-to-entity (cpfRelations) and entity-to-resource (resourceRelations) relationships to track became unwieldy for one intern, let alone two in different locations. While we had hoped the interns could share work on the same records, their terms of employment overlapped only briefly; they instead worked on separate
halves of the master list of names. Ensuring that all relationships were properly coded for symmetry where applicable (e.g., two entities who were associated with each other, one record coded at Harvard and one at Yale) was therefore challenging, involved maintaining Excel spreadsheets, and fell largely to the intern whose term was later in the project. We were therefore unable to share equally the editing of records; one of our goals had been to share in record creation, illustrating that two repositories with related collections could co-own record creation and maintenance, distributing effort and cost.

We set out to explore what a rich record might be. How rich is rich: what are the natural or reasonable limits in terms of time spent, information included, and links created? We knew that one aspect of “rich” was accuracy of detail: all the information would be grounded in the scholarly resources concerning the people and period. Each entity was therefore carefully researched, and we often found ourselves wondering where to stop. For example, when two established sources such as the *Dictionary of National Biography* and the definitive biography of an entity disagreed on a detail that we planned to encode, and the name authorities in our hierarchy did not resolve the discrepancy, we had to make a decision about which source to cite, whether or not to do further research into the matter, and whether or not to include the information at all if its accuracy was not certain—all of which took time.

Once the accurate data was ascertained, there were many ways within each record to encode it, both structured and unstructured. There are chronological lists, roles that can be assigned to places, and notes explaining relations, as well as the narrative biogHist note. It takes time to
make decisions about how and where to include what information, as well as when and how
redundancy of information within a record might be useful. Without a particular access system
in mind or known user expectations to guide our choices, we tended to choose redundancy and
more detail rather than less.

Our engagement with the multipleIdentities element, for example, involved nuanced choices
about where information belonged in a record. George III clearly required two entityType
values ("person" and "corporateBody"), but when encoding separate descriptions of man and
king we found it difficult to disentangle these identities: ought his existDates as a corporate
body to commence with his coronation rather than his birth? Is his marriage a personal or a
corporate event, or both? Also murky was whether to use multipleIdentities to distinguish
between personal and corporate activities of bookseller-publishers; in this era a person and his
or her business were not always thought to be different or separable from one another. Our
final decision was not to use multipleIdentities for publishers, encoding individuals only as
persons with separate records for firms when there was a discernibly separate corporate body.

[Insert Figure 2.]

Our best practices in the end mandated a minimum of one cpfRelation per record, and we
aimed for but did not require at least one resourceRelation; in practice most records had many
more of both. (We excluded functionRelation as not germane to the project.) We linked to all
related resources at Houghton and Beinecke, routinely searched for and included links to
material at the British Library, and concentrated chiefly on manuscript and archival collections. One factor that held us back in this area was the lack of permanent links for most resources outside our own repositories. We did not exhaustively research all possible related resources; these connections, in a Linked Data environment, will come in time.

In the end, how much time we chose to spend on a record, and thus its richness, depended on the degree of the entity’s centrality in the Johnson circle. The time study we undertook for the records showed that on average we spent slightly more than four hours per record. This included initial research, description, establishment of relationships, and final revisions. Where the time was spent varied, depending on type of person or corporate body being described. Lesser known people and those furthest out in Johnson’s circle took more initial research time to establish basic facts; but these individuals took very little time in creating the relations since, for many, the only link was to Johnson himself. We put the most time and effort into the records for Johnson, Boswell, and their closest associates, and in this we were consciously subjective. Much detail was left out of records for more distantly connected entities, detail that might have seemed crucial had the focus of the project been elsewhere or had the records been created outside the context of any project. But even with that acknowledged slant, each of the many small decisions made within each record, no matter where it fell on the spectrum of time and effort dedicated, added up to a general subjectivity that seems inherent to the process of creating EAC-CPF records by hand.

[Insert Figure 3.]
Another nuance to relations that we hoped to elucidate was the nature of each relation between our chosen CPF entities and their related CPF entities and resources. We hoped to provide structured, machine-actionable categories for each relation. EAC-CPF includes type attributes for specifying the nature of each relation. For the cpfRelation element, the cpfRelationType attribute may have one of the following values: "identity," "hierarchical," "hierarchical-parent," "hierarchical-child," "temporal," "temporal-earlier," "temporal-later," "family," or "associative." For the resourceRelation element, the resourceRelationType attribute may have one of the following values: “creatorOf,” “subjectOf,” or “other.” These options are, of necessity for a standard such as EAC-CPF, highly generic. For cpfRelations, we assigned cpfRelationType values of “associated” or “family.” The relationships we were describing did not follow any before/after or inferior/superior pattern that would recommend using the temporal or hierarchical options. For resourceRelations, we assigned resourceRelationType values of “creatorOf” or “subjectOf”. The distinction between “creatorOf” and “subjectOf” was not always clear when linking to archival materials. For instances of a relation between a CPF entity and a group of archival materials for which that entity is not the primary archival creator but is the creator of some portion thereof (e.g. a letter in a larger collection) we opted to use “subjectOf” unless the entity described created a significant portion of the linked archive.

The lack of specificity within EAC-CPF’s internal vocabulary for classifying relations led us to explore assigning our own terms to explain the nature of each relation. Within EAC-CPF, which
includes the XLink attribute set, the xlink:arcrole attribute is the appropriate place to encode a term to characterize the nature of a link (as opposed to xlink:role, which is intended to describe the nature of the destination of the link). As they created our EAC-CPF records, our interns began assigning values to xlink:arcrole on cpfRelation and resourceRelation, using terms of their choosing. However, as the project progressed, we realized that assigning terms at random as xlink:arcrole values satisfied our intellectual desire to categorize, but, from the perspective of machine processing, added little functionality beyond what could be provided in a descriptiveNote. Implementing xlink:arcrole in a way that added value to our EAC-CPF instances would have required us either to research and identify an appropriate existing Linked Data vocabulary or to create our own. Due to time constraints we decided to do neither and removed all xlink:arcrole attributes from the cpfRelation and resourceRelation elements in our data set.

In the absence of xlink:arcrole attributes, we required a descriptiveNote to provide some narrative context for each cpfRelation, and as necessary for a given resourceRelation. Initially we were inconsistent in composing these notes, interchanging the entity being described and the linked entity as sentence subjects. After some reflection we tried to standardize our sentences so that they followed the directionality of the link, with the entity being described as the subject and the linked entity as the direct object.

[Insert Figure 4.]
In our pursuit of precise description, we found certain elements difficult to employ. We valued the place element as a useful characteristic of the Johnsonian circle; but we sometimes had difficulty attributing definite dates or even a definite span of time to an activity by an entity at a given location. Richard Savage is said to have moved to Richmond when his fortunes began to improve, but how long did he remain there? Was it until he was charged for murder? William Adams as a clergyman was associated with several places of appointment, but it is not clear from his biography how much time he spent in each (for instance, he was archdeacon of Llandaff while also master of Pembroke College). We also felt the lack of a controlled vocabulary for placeRole element.

In one case, we found the lack of an element limiting: for this intertwined literary crowd, a bibliography element, as found in EAD, would have been very useful. We resorted to marking up titles using the span formatting element where they occurred within the narrative portion of the bioghist element, not a very satisfying or easily machine-actionable encoding.

**Next Steps**

After completing the Harvard Library Lab-funded project and successfully creating 78 EAC-CPF records, a Schematron against which they can be validated, and detailed documentation, the members of the project team agreed that enough remained to be learned about EAC-CPF and its potential impact on archival description to warrant another phase to the project. Having 78 of the richest hand-crafted records that we could imagine, we were no longer interested in creating EAC-CPF. We agreed, however, that our records could be improved by revisiting the
xlink:arcrole attribute on cpfRelation and resourceRelation. With the support of the Houghton and Beinecke Libraries, we began a year-long follow-up project in the fall of 2013 to research existing Linked Data vocabularies, adapt something extant or create our own, and update our records. We all saw this as a crucial way to further our understanding of the Linked Data component that EAC-CPF introduces to archival description. We hope that whatever vocabulary we create will be useful to others within the archival community. We also hope that our continued collaboration on the project will provide opportunities for us to advocate locally for the implementation of EAC-CPF within the finding aid databases at Harvard and Yale and thereby to contribute to broader community awareness of the standard's possibilities.

Wrap-up

Was attempting to connect the dots by hand worth it? Boswell, Johnson’s biographer, wondered about this as well, and this was his conclusion:

““The circle of his friends [...] was extensive and various, far beyond what has been generally imagined. To trace his acquaintance with each particular person, if it could be done, would be a task, of which the labour would not be repaid by the advantage.””

If Boswell had been able to imagine Linked Data, he might have come to a different conclusion. The EAC-CPF work undertaken to date lacks an accompanying body of evaluation by users, and Linked Data is also, within archival circles, in early days. The uses of machine-harvested versus hand-coded EAC-CPF records will be revealed as further projects and systems come forward. Certainly we have to draw a line. We can’t practice careful, handcrafted description for every
entity, nor should we. But when we do, the advantages may eventually repay the original effort as the rich, structured data is repurposed in multiple ways and the circle of relations is extended by archivists and researchers in the future. These records are not so much an end product in themselves as a number of potential beginnings.

---

i http://www.loc.gov/ead/index.html
ii TS-EAC on the SAA site: http://www2.archivists.org/governance/handbook/section7/groups/Standards/TS-EAC-CPF
iii TS-EAD on the SAA site: http://www2.archivists.org/governance/handbook/section7/groups/Standards/TS-EAD
v oXygen XML Editor. http://www.oxygenxml.com/
vii https://www.ica-atom.org/
ix Boswell, James. The life of Samuel Johnson, LL.D. : comprehending an account of his studies and numerous works in chronological order ... : the whole exhibiting a view of literature and literary men in Great-Britain for near half a century during which he flourished (London: Printed by Henry Baldwin for Charles Dilly, 1791). Volume I, page 32.