Extending the Utility of the HTRC Extracted Features Dataset Through Linked Data

Jacob Jett, Boris Capitanu, Deren Kudeki, Ryan Dubnicek, Timothy W. Cole, J. Stephen Downie

HathiTrust Research Center, School of Information Sciences, University of Illinois at Urbana-Champaign

Context

- **HathiTrust Digital Library (HTDL):** A large-scale digital library comprising over 17 million volumes of digitized books and journals contributed by over 160 institutions.
- **HathiTrust Research Center (HTRC):** The academy-facing research institution charged with developing tools through which digital humanities scholars can interact with and analyze the HTDL corpus.

Prior Work

- **Non-Consumptive Research Paradigm:** The copyright preserving re-search paradigm in which computational tools are employed to analyze text in such a way that the text's intellectual property is not infringed.
- **Extracted Features (EF) Datasets:** A series of base-line datasets produced by the HTRC and designed to facilitate analysis and research of the HTDL corpus.

What's New – EF 2.0

- **EF schema mapped to Schema.org ontology.** Allows for:
  - Future interoperability (standardized semantics)
  - Easy extensions (subjects & genres)
  - Richer metadata (LCC – topic mapping)
  - Ability to leverage Linked Open Data resources (VIAF identifiers for some contributors)

EF Dataset Contents

- **Volume-Level Metadata:** Metadata gathered from the host institution's MARC records.
- **Multi-Level Features Data:** Line counts, token counts, and part-of-speech tags at the volume and page-level for each item in the corpus.
- **Wow Numbers here:**
  - **# of Volumes:** *17,123,746*
  - **# of Pages:** *6,221,631,336*
  - **# of Tokens:** *2,906,819,723,689*
- **Downloads Available From:**
  - https://doi.org/10.13012/R2TE-C227

Partial Example File


```
@context:            "https://worksets.htrc.il…ontext/ef_context.jsonld"
schemaVersion:       "https://schemas.hathitrust.org/EF_Schema_v_3.0"
id:                  "https://data.analytics.h…00204/hvd.32044072198021"
htid:                "hvd.32044072198021"
type:                "DataFeed"
publisher:           {…}
datePublished:       20200204
metadata:
  schemaVersion:     "https://schemas.hathitru…_MetadataSubSchema_v_3.0"
  id:                "http://hdl.handle.net/2027/hvd.32044072198021"
  type:              […]
  dateCreated:       20200130
  title:             "Bulletin of the United States Fish Commission."
  alternateTitle:    […]
  contributor:
    0:
      name:          "Biodiversity Heritage Library."
      type:          "http://id.loc.gov/ontologies/bibframe/Organization"
      id:            "http://www.viaf.org/viaf/161999238"
    1:
      name:          "United States Fish Commission."
      type:          "http://id.loc.gov/ontologies/bibframe/Organization"
      id:            "http://www.viaf.org/viaf/133008683"
  pubDate:           1895
  publisher:         {…}
  pubPlace:          {…}
  language:          "eng"
  accessRights:      "pd"
  sourceInstitution: {…}
  mainEntityOfPage:  […]
  lcc:               "SH11.A25"
  lccn:              "sn 98030130"
  oclc:              "1506338"
  category:          "Aquaculture. Fisheries. Angling"
  genre:             […]
  enumerationChronology: "v.15 (1895)"
  typeOfResource:    "http://id.loc.gov/ontologies/bibframe/Text"
  isAccessibleForFree: true
  lastRightsUpdateDate: 20180709
  isPartOf:          {…}
```


Volume-level information


```
features:
  schemaVersion:     "https://schemas.hathitru…_FeaturesSubSchema_v_3.0"
  id:                "http://hdl.handle.net/2027/hvd.32044072198021"
  type:              "DataFeedItem"
  dateCreated:       20200124
  pageCount:         702
  pages:
    0:               {…}
    1:               {…}
    2:               {…}
    3:               {…}
    4:               {…}
    5:               {…}
    6:               {…}
    7:               {…}
    8:               {…}
    9:               {…}
    10:
      seq:           "00000011"
      version:       "447bb20062614b90ac6aeb9dab92c5ba"
      language:      "en"
      tokenCount:    262
      lineCount:     21
      emptyLineCount: 0
      sentenceCount: 28
      header:        null
      body:
        tokenCount:  262
        lineCount:   21
        emptyLineCount: 0
        sentenceCount: 28
        capAlphaSeq: 7
        beginCharCount: {…}
        endCharCount: {…}
        tokenPosCount:
          2:         {…}
          5:         {…}
          13:        {…}
          20:        {…}
          28:        {…}
          1894:      {…}
          1896:      {…}
          Barton:    {…}
          reference:
            NN:      1
```


Page-level information

References:

- Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. Digital Humanities Quarterly (DHQ) 3(4), pp 1-30.
- Capitanu, B., Underwood, T., Organisciak, P., Cole, T., Sarol, M. J., and Downie, J. S. (2016). The HathiTrust Research Center Extracted Feature Dataset (1.0) [Dataset]. HathiTrust Research Center. DOI: 10.13012/J8X63JT3.
- Jett, J., Capitanu, B., Kudeki, D., Cole, T. W., Hu, Y., Organisciak, P., Underwood, T., Koehl, E. D., Dubnicek, R., & Downie, J. S. (2020). The HathiTrust Research Center Extracted Features Dataset (2.0). HathiTrust Research Center. DOI: 10.13012/R2TE-C227.
- Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp 423-30.
- Organisiciak, P., Capitanu, B., Underwood, T., and Downie, J. S. (2017). Access to billions of pages for large-scale text analysis. iConference 2017 Proceedings 2, pp 66-76. DOI: 10.9776/17014.