Extracted Features Schema Version 3.0:

Documentation – 5-December-2019

The following documentation is the result of collaboration between the HathiTrust Research Center, JSTOR, and Portico intended to result in a common vocabulary for the exchange of extracted features datasets derived from a number of text object types.

Editors: Boris Capitanu, Timothy W. Cole, Jacob Jett, Deren Kudeki, Amy Kirchoff, Ted Lawless, Ron Synder, Dongqing Xie

The simplified data model (Figure 1 below) is designed to seamlessly fold JSTOR and Portico's fine-grained article database metadata with the HathiTrust's coarse-grained MARC records.

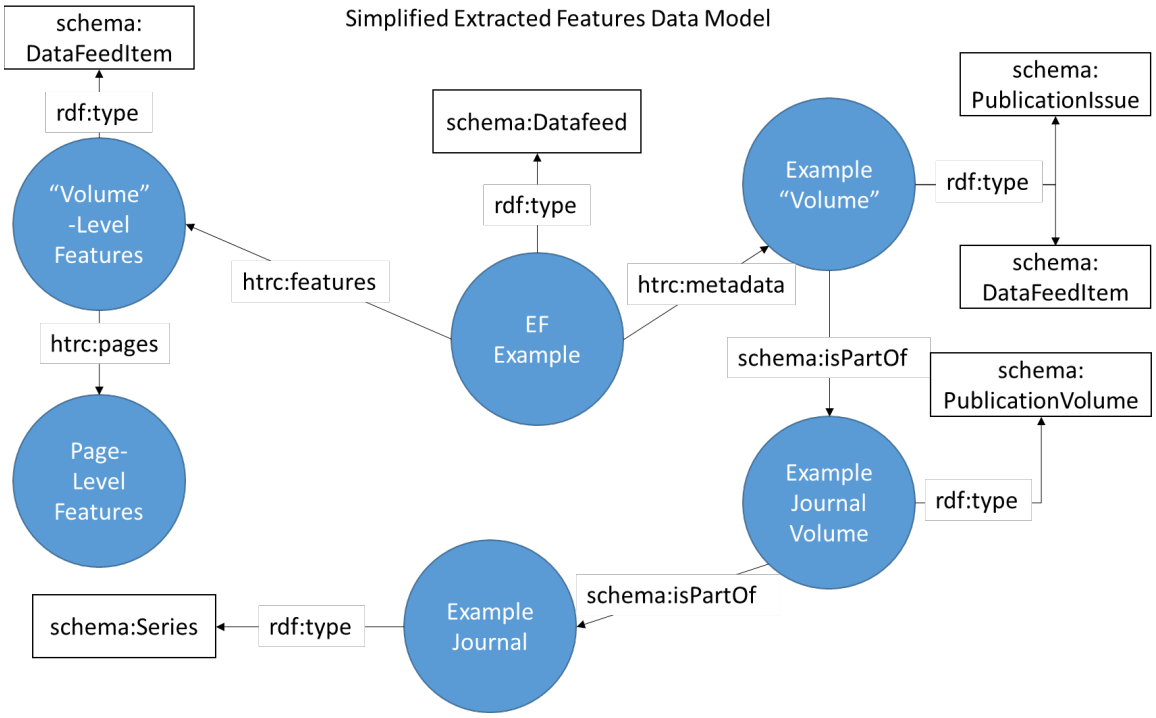

Figure 1: Simplified Extracted Features Data Model

The primary expectation is that the resulting Extracted Features (EF) datasets will be serialized through JSON-LD that conforms to the accompanying JSON-LD context document. Figures 2-6 (below) graphically showcase the overall document model for the EF JSON files.

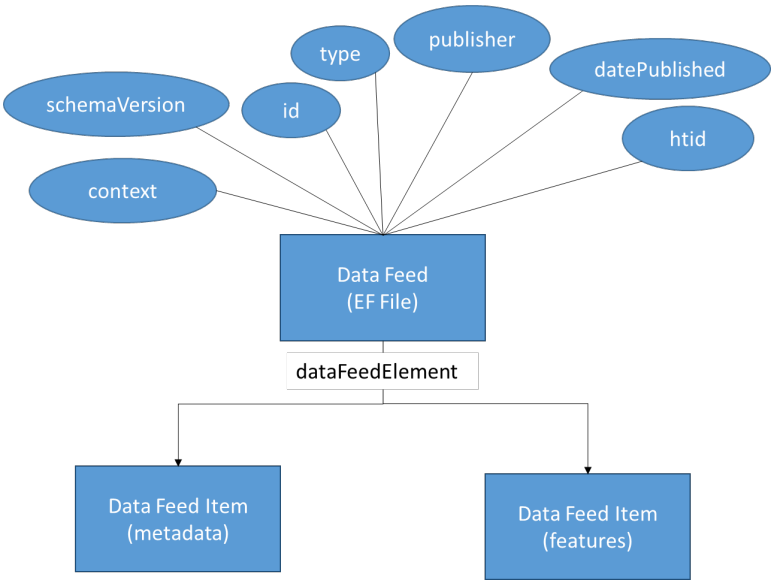

Figure 2: JSON-LD Document Model Root/Header Section

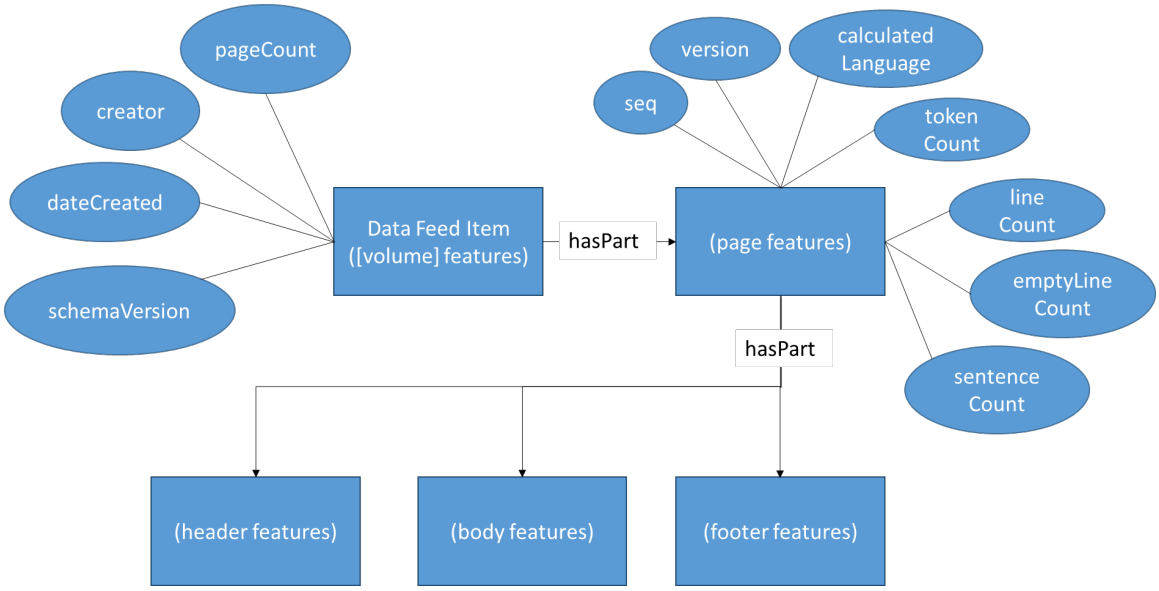

Figure 3: JSON-LD Volume/Page Features Section

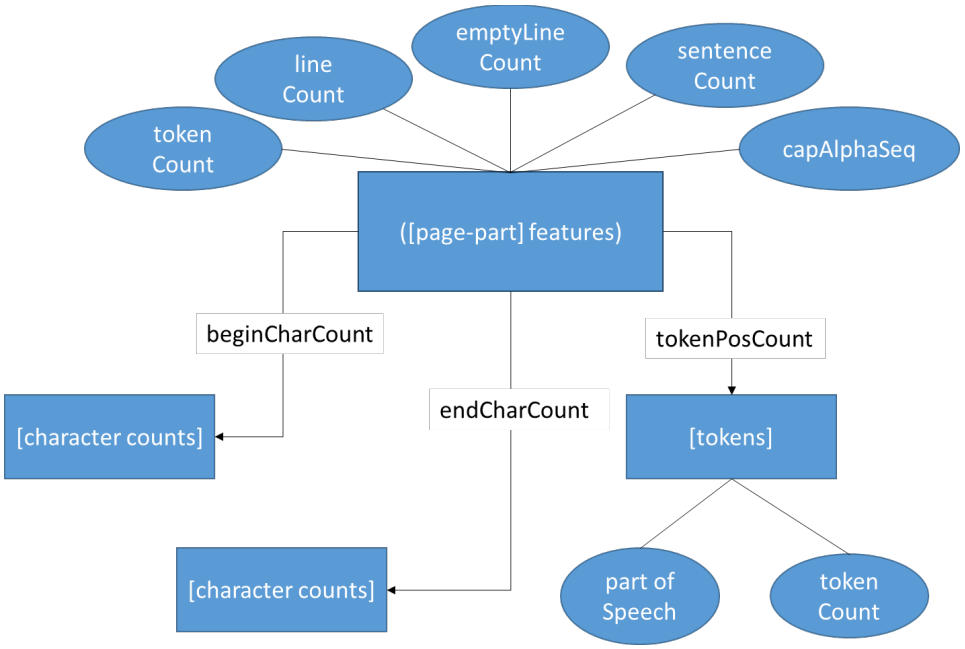

Figure 4: JSON-LD Header/Body/Footer Features Sections

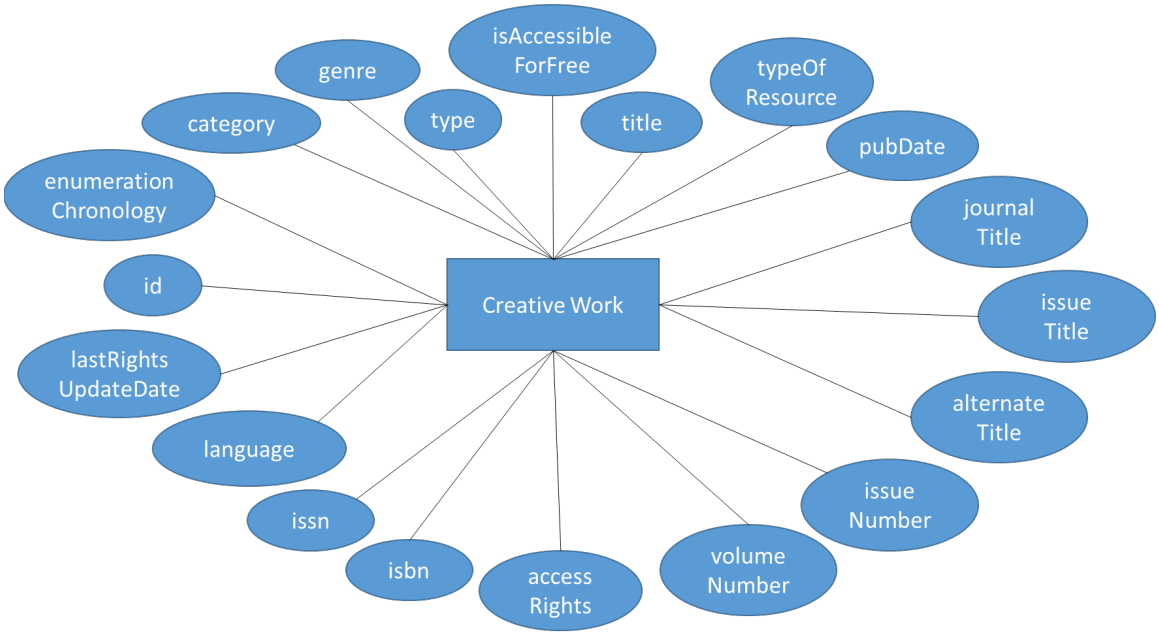

Figure 5: JSON-LD Metadata Section (Attributes)

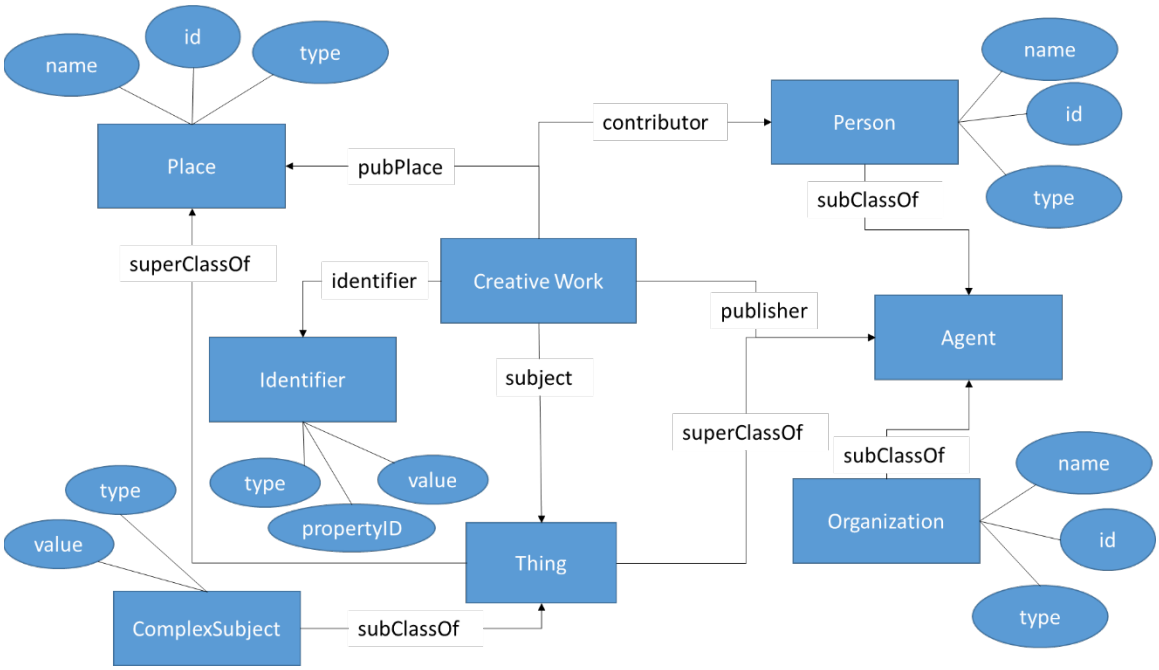

Figure 6: JSON-LD Metadata Section (Entity Relationships)

The following tables define the keys to be used in the JSON-LD EF serialization. Users should note that additional keys conforming to additional schema.org entity properties may also appear in some EF data files, especially those produced by JSTOR/Portico which have finer-grained metadata from which to draw information. Among these additional keys (not defined here) users may see: pageEnd, pageStart, and pagination.

Note: The primary purpose of this documentation is to explicate the specific semantics and intended use of each key in the schema. As such, no information is provided regarding the provenance of the data that appears as values for those keys.

Table 1: EF Common JSON Object Keys

Key Name	Definition
id	This key is used to identify node objects within the JSON document. Note that when present, its value *MUST* be null, an absolute URL, a relative URL, or a compact URL. Note also in some cases this id will be synonymous with previously used ids (in older schema versions) such as the value used by the deprecated "volumeIdenitifier" key. See JSON-LD spec (https://json-ld.org/spec/latest/json-ld/#node-objects) for additional information.
type	This key used to assert that the node is a particular kind of entity. If the node object contains the @type key, its value MUST be either an absolute URL, a relative URL, a compact URL, a term defined in the active context expanding into an absolute URL, or an array of any of these. See section 3.4, Specifying the Type, for further discussion on @type values. See JSON-LD spec (https://json-ld.org/spec/latest/json-ld/#node-objects) for additional information.

	The following types are typically used in EF files for bibliographic entities (including the EF file and its features and metadata parts): • https://schema.org/DataFeed • https://schema.org/DataFeedItem • https://schema.org/CreativeWork • https://schema.org/Book • https://schema.org/PublicationIssue • https://schema.org/PublicationVolume • https://schema.org/Article • https://schema.org/CreativeWorkSeries Agents will typically have one of the following types: • https://schema.org/Organization • https://schema.org/Person

Table 2: Extracted Features File Header and Primary Objects

context	This key is used to provide the context document that aliases all of the document's various keys to their correct namespace references. When present, its value *MUST* be null, an absolute URL, a relative URL, or a compact URL.
schemaVersion	This key is used to convey version information for the EF JSON-LD document's overall schema as well as those portions particular to the metadata and features objects. When present, its value *MUST* be null, an absolute URL, a relative URL, or a compact URL.
publisher	This key is used to convey information about the agent who published a bibliographic entity. It is semantically synonymous with schema.org's **publisher** property (https://schema.org/publisher). Note that here at the top level the publisher is the institution publishing the Extracted Features dataset (e.g., HathiTrust Research Center).
name	This key records a name label for an entity. It is semantically synonymous with schema.org's **name** property (https://schema.org/name). ***Scope Note:*** This key is intended to be used for node objects representing agents (e.g., publisher above). For node objects that represent bibliographic entities, see title below.
datePublished	This key records the date information regarding when the JSON-LD file was published. Note that this can vary from the dates the data is created at. It is semantically synonymous with schema.org's **datePublished** property (see: https://schema.org/datePublished).
htid	This key records an ambiguous name for the family of objects related to a particular "volume" in the HathiTrust context. It is reused in the identifiers that denote various versions of the "volume" (e.g., the OCR text files that compose it, a book on a shelf, etc.), metadata records describing the "volume", and even this extracted features JSON-LD file. When present, this key's value *MUST* be null or a string.

metadata	This key is used to demarcate the section of the EF file that contains metadata describing the entity from which the EF data was generated. When present, the value of this key *MUST* be null or a node object. Semantically, the metadata predicate is a sub-property of schema.org's **dataFeedElement** property (https://schema.org/dataFeedElement). ***Scope Note:*** This key is used specifically for denoting the part of the data feed that communicates metadata about the bibliographic object that was analyzed to produce the features.
features	This key is used to indicate the section of the JSON-LD serialization that contains the actual EF data. When present, the value of this key *MUST* be null or a node object. Semantically, the features predicate is defined as a sub-property of schema.org's **dataFeedElment** property (https://schema.org/dataFeedElement). ***Scope Note:*** This key is used specifically for denoting the part of the dataset that communicates the data that was generated by analysis of the bibliographic object described by the metadata block.

Table 3: Features Object Keys

schemaVersion	This key is used to convey version information for the metadata-specific portion of the EF schema. When present, its value *MUST* be null, an absolute URL, a relative URL, or a compact URL.
creator	This key records which institution generated the EF file. It is semantically synonymous with schema.org's **creator** property (https://schema.org/creator).
dateCreated	This key records the date information regarding when the features data in this section was created. It is semantically synonymous with schema.org's **dateCreated** property (see: https://schema.org/dateCreated).
pageCount	This key is conveys the number of pages in the "volume". When present, the value of this key *MUST* either be null or an integer.
pages	This key is used to indicate the section of the JSON-LD serialization that contains the page-level EF data. When present, the value of this key *MUST* be null or an array. Semantically, the pages predicate is defined as a sub-property of schema.org's **hasPart** property (https://schema.org/hasPart). ***Scope Note:*** This key is used specifically for denoting the part of the dataset that communicates the data. For other meronymic relations see *isPartOf* and *hasPart* above.
seq	This key conveys the relative position of a page within a "volume". When present, the value of this key *MUST* be null or an integer.
version	This key provides an MD-5 Hash that identifies the version of the analyzed page. When present, the value of this key *MUST* be either a string or null.

calculatedLanguage	This key provides the algorithmically estimated language (the one with the highest probability value) of the text on the page. When present, the value of this key *MUST* be either a string or null.
tokenCount	The total number of tokens in the section (e.g., page, header, body, or footer) indicated by the node object. When present, the value of this key *MUST* be null or an integer.
lineCount	The total number of non-empty lines in the section (e.g., page, header, body, or footer) indicated by the node object. When present, the value of this key *MUST* be null or an integer.
emptyLineCount	The total number of empty lines in the section (e.g., page, header, body, or footer) indicated by the node object. When present, the value of this key *MUST* be null or an integer.
sentenceCount	The total number of sentences in the section (e.g., page, header, body, or footer) indicated by the node object. When present, the value of this key *MUST* be null or an integer.
header	This key is used to indicate the section of the JSON-LD serialization that contains the actual EF data. When present, the value of this key *MUST* be null or a node object. Semantically, the header predicate is defined as a sub-property of schema.org's **hasPart** property (https://schema.org/hasPart). ***Scope Note:*** This key is used specifically for denoting the part of the dataset that communicates the data. For other meronymic relations see *isPartOf* and *hasPart* above.
body	This key is used to indicate the section of the JSON-LD serialization that contains the actual EF data. When present, the value of this key *MUST* be null or a node object. Semantically, the body predicate is defined as a sub-property of schema.org's **hasPart** property (https://schema.org/hasPart). ***Scope Note:*** This key is used specifically for denoting the part of the dataset that communicates the data. For other meronymic relations see *isPartOf* and *hasPart* above.
footer	This key is used to indicate the section of the JSON-LD serialization that contains the actual EF data. When present, the value of this key *MUST* be null or a node object. Semantically, the footer predicate is defined as a sub-property of schema.org's **hasPart** property (https://schema.org/hasPart). ***Scope Note:*** This key is used specifically for denoting the part of the dataset that communicates the data. For other meronymic relations see *isPartOf* and *hasPart* above.
tokenPosCount	This key conveys the tokens that appear in a section. When present, the value of this key must either be null or an array. ***Scope Note:*** An unordered list of all tokens (characterized by part of speech using OpenNLP), and their corresponding frequency counts, in this page section. Tokens are case-sensitive, so a capitalized "Rose" is shown as a separate token. There will be separate counts, for instance, for "rose" (noun) and "rose" (verb). Words separated by a hyphen

	across a line break are rejoined. No other data cleaning or OCR correction was performed.
beginCharCount	This key conveys the aggregated frequency counts of the first non-whitespace character on each line. When present, this key's value *MUST* either be null or an array.
endCharCount	This key conveys the aggregated frequency counts of the last non-whitespace character on each line. When present, this key's value *MUST* either be null or an array.
capAlphaSeq	This key conveys the longest length of the alphabetical sequence of capital characters starting a line. When present, this key's value *MUST* either be null or an integer. ***Scope Note:*** This key only appears in node objects that are the value of *body* keys.

Table 4: Metadata Object Keys

schemaVersion	This key is used to convey version information for the metadata-specific portion of the EF schema. When present, its value *MUST* be null, an absolute URL, a relative URL, or a compact URL.
dateCreated	This key records the date information regarding when metadata in this section was created. Note that this date is not reflective of the metadata's original creation date, just the most recent date a new version (the version used here) of the original metadata was created. It is semantically synonymous with schema.org's **dateCreated** property (see: https://schema.org/dateCreated).
title	This key is used to convey the title of the entity the node object represents. Semantically, the title predicate is defined as a sub-property of schema.org's **name** property (https://schema.org/name). ***Scope Note:*** This key is intended to be used for node objects representing bibliographic entities. For node objects that represent agents, see name above.
journalTitle	This key is used to convey the title the entity that a Journal node object represents. This predicate is defined as a sub-property of the title predicate (above). ***Scope Note***: This is key is used for node objects representing entities that possess the schema:Journal class.
issueTitle	This key is used to convey the title the entity that a Publication Issue node object represents. This predicate is defined as a sub-property of the title predicate (above). ***Scope Note***: This is key is used for node objects representing entities that possess the schema:PublicationIssue class.
alternateTitle	This key is used to convey an additional title of the entity the node object represents. This predicate is defined as a sub-property of the title predicate (above). ***Scope Note:*** This key is intended to be used for node objects representing bibliographic entities. For node objects that represent agents, see name above. It may be used in conjunction with the journalTitle and issueTitle keys in addition to the title key. This key may have an array of strings as its value.

enumerationChronology	This key is used to convey information regarding which volume, issue, or anum a creative work is. When present, the value of this key must be null or a string. Typically, the string value is derived from a part of the title string.
issueNumber	This key is used to convey enumeration information regarding a publication issue. When present, the value of this key must be null, an integer, or a string. It is semantically synonymous with schema.org's **issueNumber** property (https://schema.org/issueNumber).
volumeNumber	This key is used to convey enumeration information regarding a publication volume. When present, the value of this key must be null, an integer, or a string. It is semantically synonymous with schema.org's **volumeNumber** property (https://schema.org/volumeNumber).
publisher	This key is used to convey information about the agent who published a bibliographic entity. It is semantically synonymous with schema.org's **publisher** property (https://schema.org/publisher).
pubPlace	This key is used to convey information regarding where a bibliographic entity was first published. It is semantically synonymous with schema.org's **location** property (https://schema.org/location).
pubDate	This key is used to convey information regarding when a bibliographic entity was first published. It is semantically synonymous with schema.org's **datePublished** property (https://schema.org/datePublished).
genre	This key is used to convey information regarding the genre of a bibliographic entity. It is semantically synonymous with schema.org's **genre** property (https://schema.org/genre).
category	This key is used to convey information regarding the overall topic of a bibliographic entity. When present its value must be NULL, a string value, or an array of string values. It is semantically synonymous with schema.org's **about** property (https://schema.org/about). ***Scope Note:*** Note that the strings in the category conform to an established standard mapping from Library of Congress Classification numbers (LCC) to a taxonomy [using string data] describing broad topical categories maintained by the University of Michigan (https://www.lib.umich.edu/browse/categories/).
subjects	This key is used to convey additional topical information about a bibliographic entity. This predicate is defined as a sub-property of the category predicate (above). When present its value must be NULL or an array which may contain strings or objects. ***Scope Note:*** Note that this key is intended to record all other kinds of topical information including subject headings from a variety of standards (e.g., LCSH, etc.) and mappings from classification systems (e.g., DDC, UDC, etc.).

language	This key is used to convey information regarding the primary language of a bibliographic entity. It is semantically synonymous with schema.org's **inLanguage** property (https://schema.org/inLanguage). ***Scope Note:*** Note that this key is also used in the features section to annotate the primary language each analyzed page possessed.
accessRights	This key is used to convey the copyright status of the bibliographic entity from which the EF were created. When present this key's value must either be null or a string. ***Scope Note:*** When the EF file's source is the HTRC, string values for this key will conform to the attributes described by the HathiTrust's rights database (https://www.hathitrust.org/rights_database#DatabaseLayout). EF files generated by JSTOR/Portico also conform to these attributes but typically only use the "pd", "ic", and "und" values.
isAccessibleForFree	This key is used as a coarser method than the previous key for indicating when something is publicly available. It is semantically synonymous with schema.org's **isAccessibleForFree** property (https://schema.org/isAccessibleForFree). ***Scope Note:*** End users should note that the value of this key may not match the actual accessibility state as copyright reassessments are occurring for the works in the dataset all of the time. See related *lastRightsUpdateDate* key below.
lastRightsUpdateDate	This key is used to convey the most recent date that the access rights for a bibliographic entity were updated on. It is semantically synonymous with schema.org's **dateModified** property (https://schema.org/dateModified).
contributor	This key replaces the older *names* key (see list of deprecated keys below). It is semantically synonymous with schema.org's **contributor** property (https://schema.org/contributor), except that it may additionally have as a value, an array. ***Scope Note:*** As JSTOR/Portico have richer datasets from which to draw information from, narrower keys indicating richer role information, including *author*, *editor*, etc., *MAY* appear in some EF files. These keys are semantically synonymous with analogous properties defined by the schema.org standard. See: • https://schema.org/author • https://schema.org/character • https://schema.org/editor • https://schema.org/producer • https://schema.org/actor • https://schema.org/director • https://bib.schema.org/artist • https://bib.schema.org/colorist • https://bib.schema.org/inker • https://bib.schema.org/letterer • https://bib.schema.org/penciler

	https://schema.org/illustrator
typeOfResource	This key is used to indicate the coarse-grained format of a bibliographic entity, e.g., text, image, etc. When present, this key's value must either be null or a string.
sourceInstitution	This key is used to indicate which institution provided a bibliographic entity. It is semantically synonymous with schema.org's **provider** property (https://schema.org/provider).
isPartOf	This key is used to convey information regarding an entity's meronymic relation to another entity. When present, the value of this key *MUST* be null or a node object. It is semantically synonymous with schema.org's **isPartOf** property (https://schema.org/isPartOf). ***Scope Note:*** This key is expected to be used in JSTOR and Portico EF files where article-issue-volume-series chains are better supported by the data. In this manner JSTOR and Portico will be able to provide additional relevant metadata at each level of work granularity.
hasPart	This key is used to convey information regarding an entity's meronymic relation to another entity. When present, the value of this key *MUST* be null or a node object. It is semantically synonymous with schema.org's **hasPart** property (https://schema.org/hasPart). ***Scope Note:*** This key is expected to be used in JSTOR and Portico EF files where article-issue-volume-series chains are better supported by the data. In this manner JSTOR and Portico will be able to provide additional relevant metadata at each level of work granularity. Note that the *features* key (described below) also maps to schema's hasPart relation.
mainEntityOfPage	This key is used to indicate additional sources of metadata that describe a bibliographic entity, e.g., such as through links to catalog records. It is semantically synonymous with schema.org's **mainEntityOfPage** property (https://schema.org/mainEntityOfPage) except that it may additionally have as a value, an array.
identifier	This key is used to indicate additional IRIs or labels that identify a bibliographic, e.g., such as through the source institution's handle for the entity. It is semantically synonymous with schema.org's **identifier** property (https://schema.org/identifier) except that it may additionally have as a value, an array.
issn	This key is used to convey a creative work series' ISSN number. It is semantically synonymous with schema.org's **issn** property (https://schema.org/issn).
isbn	This key is used to convey a book's ISBN number. It is semantically synonymous with schema.org's **isbn** property (https://schema.org/isbn).

Table 5: List of Deprecated Keys

Key	Reason for Deprecation
names	Superseded by more informing keys (e.g., contributor, author, etc.).
imprint	Superseded by publisher, pubPlace, and pubDate
hathiTrustRecordNumber	Integrated into identifier array.
htBibUrl	Integrated into mainEntityOfPage array.
handleURL	Used as node ID for *metadata* node object
sourceInstitutionRecordNumber	Integrated into identifier array.
oclc	Integrated into identifier array.
lccn	Integrated into identifier array.
classification	Integrated into identifier array.
lastUpdateDate	Renamed to *lastRightsUpdateDate* to better reflect purpose.
languages	Renamed to *calculatedLanguage* to better reflect purpose.
governmentDocument	Superseded by more accurate information appearing in genre.