

Data Processing for Aquifer Records

Data Processing

Initially, remove out of scope records.

- Records with zero location/url or identifier/uri elements. **UM processing:** taken care of through the re-exposed records.
- Records with a subject/hierarchicalGeographic element, that:
 - Has a country subelement but none with the value United States (or reasonable other values, such as U.S., America, etc.), or;
 - Has a state subelement but none with a value that matches any of the 50 states, written out or in abbreviation

Basic search implies Google-like functionality, so when basic is noted in the following table, it means the element[s] are part of the basic search index. This index should also be an option in the advanced search page, as "keyword".

Levels of adoption, according to [SWG's page](#):

- Level A. A user is able to identify a resource (to reference, for future re-discovery).
- Level B. A user is able to find resources through a process (search and/or browse) that offers a modest amount of precision.
- Level C. Everything else. These fields allow users to perform searches with a high degree of precision, browse winnowing, and disambiguation between related resources.

| Item being Processed | Processing Notes | XPath Query | Level of Adoption | Brief /Full Display | Record Display Label | Basic /Advanced Search Index | In Advanced Search? | Browse Facet |
|----------------------|--|---|-------------------|---------------------|--|------------------------------|--|--------------|
| title | <p>For brief display, clicking on the title hotlinks to the URL.</p> <p>In the brief display, only show one title, and prefer one with no type attribute. If more than one title lacks a type attribute, display the first title without a type. If all titles include a type attribute, display the first.</p> <p>In the full display, show all titles with the label as described elsewhere on this page, none hyperlinked.</p> <p>nonSort and the title should be separated by a space. The title and the subTitle, partName or partNumber should be separated by a space colon space.</p> <p>The non-sort attribute should not be in the title for sorting purposes.</p> | <p>indexing: mods/titleInfo/ [title and subTitle]partName partNumber]</p> <p>brief display: mods/titleInfo/ [nonSort title[0] and subTitle]partName partNumber]</p> <p>full display:</p> <ul style="list-style-type: none"> mods /titleInfo/ [nonSort ti title and subTitle p artName p artNumbe r] mods /titleInfo/ [nonSort ti title [abbrevi ated alter native tran slated unif orm] | level A | brief, full | Title For type attributes: Abbreviated Title Alternative Title Translated Title Uniform Title | basic, advanced | yes, selectable from drop-down | no |
| date ¹ | <p>First, use keyDate, if it exists. Should be one and only one keyDate. @w3cdtf strongly preferred.</p> <p>Second, dateIssued and dateCreated are the priority dates for indexing and display. One or the other of these sub-elements should be available in the record. If neither is, copyrightDate or dateOther should be used.</p> <p>Exclude dateCaptured, dateModified, dateValid.</p> <p>When normalized dates are available, these should be used for sorting and searching purposes only, not for display.</p> | <p>(in order) 1. mods /originInfo/date* [/@keyDate] [/@="w3cdtf"] 2. mods /originInfo/ [dateIssued or dateCreated] 3. mods /originInfo /copyrightDate d ateOther</p> <p>Further processing rules below.⁴</p> | level A | brief, full | Date Specifically, <ul style="list-style-type: none"> Issue Date Creation Date Copyright Date Date (for dateOther) | basic, advanced | yes, choice of single date entry, range, and era /decade | yes |
| language | <p>UM processing: re-exposed records contain exploded language codes. If there is a @type="code", another sub-element is added with @type="text" that includes the exploded code. If @type="text" already exists, it is left alone.</p> | mods/language /languageTerm [/@type="text"] | level C | full | Language | advanced | yes, selectable from drop-down | yes |
| URL | <p>Two fields can contain clickable URLs: location/url and identifier@uri. For display, only the primary URL in location/url should be used, if available. For brief display, clicking on the title hotlinks to the primary URL. For full display, the URL displays as-is.</p> <p>In the event there is no location/url, identifier@uri may be used. UM processing: both location/url and identifier@uri are used to filter digital object records, in the event the latter may be useful.</p> <p>Exclude any that lead to a 404.</p> | mods/location/url [/@usage="primar y display"] or mods/identifier [/@type="uri"] | level B | [brief], full | URL | neither | no | no |
| creator | <p>Separate name and namePart, affiliation, role or description with a space comma.</p> <p>Explode the role/roleTerm@type="code" attributes. Add a new sub-element that contains the exploded code in a @type="text" attribute.</p> | mods/name /namePart affiliati on description and role /roleTerm [/@type="text"] | level B | full | Related Names | basic, advanced | yes, selectable from drop-down | no |

| | | | | | | | | |
|-----------------------------|---|---|---------|-------------|-----------------------------|---|------------------------------------|---|
| subject ² | Record display and browse facets are driven by the subject indexes. They should be generated from all subelements of subject, regardless of whether they appeared within a single subject container. Therefore, split pre-coordinated headings (e.g., United States - Social conditions - 1980- - Juvenile literature - Bibliography) into their component parts for indexing and browse display, but not for record display. As noted, geographicCode should be exploded, as language codes are at UM and as roleTerm is recommended to be done. Indexes should: <ul style="list-style-type: none"> Combine geographic, hierarchicalGeographic, geographicCode (exploded) into one "geographic" index Combine topic, occupation, titleInfo into one "topic" index No index for cartographic All other subject subelements (temporal, name, genre) should be their own indexes <ul style="list-style-type: none"> Genre facet should include data from both mods/genre and mods/subject/genre | mods/subject/geographic hierarchicalGeographic geographicCode mods/subject/topic occupation titleInfo mods/subject/[temporal or name or /genre] | level B | full | Subject | basic, advanced | yes, limiter by subject index type | yes Specifically, <ul style="list-style-type: none"> Subject: Geographic Subject: Topical Subject: Temporal Subject: Genre Subject: Related Names |
| physical description | Sub-elements should be separated by a space semicolon space. Ignore element and note sub-element attributes. | mods/physicalDescription/* | level C | full | Physical Description | basic | no | no |
| publisher and place | placeTerm@code should be exploded as described above for subject /geographicCode, roleTerm and language. placeTerm and publisher should be separated by a space semicolon space. | mods/originInfo/place/placeTerm[@type="text"] and publisher | level B | full | Publisher | basic (publisher), advanced (publisher and place) | yes, selectable from drop-down | no |
| origin aspects | Sub-elements should be separated by a space semicolon space. | mods/originInfo/edition issuance frequency and mods/part/* | level C | full | Publication Specifics | basic, advanced | no | no |
| resource type | Ignore attributes. | mods/typeOfResource | level B | full | Resource Type | basic, advanced | yes, limiter by value | no |
| genre | Ignore attributes. | mods/genre | level B | full | Genre | basic, advanced | yes, selectable from drop-down | yes |
| location | Separate multiple instances of physicalLocation by a comma space. | mods/location/physicalLocation | level C | full | Physical Location | advanced | no | no |
| identifiers | If @type="uri" is used for URL, exclude it here. Separate multiple instances of identifier by a comma space. | mods/identifier | level C | full | Identifier | neither | no | no |
| classification ³ | Ignore attributes, for now. Separate multiple instances of classification by a comma space. | mods/classification | level C | full | Classification | neither | no | no |
| table of contents | Ignore attributes. | mods/tableOfContents | level C | full | Table of Contents | basic | no | no |
| abstract | Ignore attributes. | mods/abstract | level C | full | Abstract | basic, advanced | yes, selectable from drop-down | no |
| note | Ignore attributes. Separate multiple instances of note by a comma space. | mods/note | level C | full | Note | basic | no | no |
| audience | Ignore attributes. | mods/targetAudience | level B | full | Audience | basic, advanced | no | yes |
| rights | Ignore attributes. | mods/accessCondition | level B | full | Terms and Conditions of Use | neither | yes, limiter by value | no |
| related item | Exclude the "difaqcoll" attribute here, because used for collection. Separate multiple instances of relatedItem by a semicolon space. If possible, use the processing logic enumerated above to handle subelements of relatedItem. | mods/relatedItem/* | level C | full | Related Item | basic | no | no |
| preview | UM processing: re-exposed records contain a thumbnail image in the @access="preview" attribute. If the re-exposed records do not contain a preview image, the Thumbgrabber can be used to gather them. | mods/location/url[@access="preview"] | level A | brief, full | n/a | n/a | n/a | n/a |
| collection | UM processing: re-exposed records contain the "difaqcoll" attribute that concatenates repository name and OAI setName into a readable collection phrase. | mods/relatedItem/titleInfo[@authority="difaqcoll"]title | level A | brief, full | Collection | basic, advanced | yes, limiter by collection | yes |

¹ We would recommend including the following in this methodology:

- Some set-level analysis to determine which date to use (only feasible for relatively small harvesters)
- If more than one date appears, throw out any dates after about 1996 or so as they're likely digitization dates
- Use the one that's machine readable if some are not

² Investigate supplementing the time browse facet that contains mods/subject/temporal with data from date elements. Also, investigate using @authority to determine if certain controlled vocabularies (e.g., LCSH) can help us create more consistent subject indexes. If clustering is a possibility, this will also aid this effort.

³ Look into whether classification can supplement genre or subject. For instance, [High Level Browse](#) at UM can be used to map classification numbers to a set of topics.

⁴ Date processing rules per the MWG and the SWG:

for both indexing and sorting:

- choose keyDate="yes" and w3cdtf="yes", if exists
- if those two attributes don't exist, choose keyDate="yes"
- if no keyDate, choose one of these:
(in order) dateCreated, dateIssued, copyrightDate, dateOther
- if none of those dates exist, choose one of these:
(in order) dateCaptured, dateValid, dateModified
- assumption is there is only one sort date and only one indexing field
(which may have multiple values)

other indexing rules:

- all chosen dates are normalized to a year value
- for a single date, e.g., 1986, index only that date
- for a range of known dates, e.g., 1944-1950, index each of those dates inclusive
- for uncertain dates, e.g., 198?, 1908s, 198-, each date is indexed inclusive, e.g., 1980-1989
- for circa dates, e.g., ca. 1945, each date is expanded for indexing +/- 5 years, e.g., 1940-1950
- for expanded indexing, these dates will be searchable across decades, e.g., ca. 1945 will be searchable in the 1940s and the 1950s
- non-dates, e.g., n.d., don't get indexed or normalized
- centuries are indexed as such, e.g., 17th century/cent. as 1601-1700
- for date elements with start and end attributes, use first start/end pair and treat these as a known date, e.g., start=1900, end=1920, index 1900-1920

sorting:

- for a range of known or circa dates, choose the mid-point date, e.g., for 1944-1950, choose 1947; for ca. 1945, choose 1945 (because indexing expanded to 1940-1950)
- for a range of uncertain dates, choose the beginning date, e.g., for 1940?, choose 1940
- for a single date, choose that date
- non-dates, e.g., n.d., [no date], should sort at the end, no matter whether sort is chronological or reverse chronological

display:

- display all dates in the original encoding
- do not display the normalized value for the indexed date field
- records with no date should not appear if a date or date range is searched
- display copyright, circa and uncertain dates as is, e.g., c1945, 1845?, ca. 1944

MODS fields not used for data processing, although they may be used for other things, are:

- mods
- modsCollection
- recordInfo
- dateCaptured
- dateModified
- dateValid
- extension -- being used to contain asset action information, but not correctly; on hold for now

Remaining questions:

- Should all elements be displayed in full display?
- How does one choose the best URL to use for display? Is /mods/location/url[@usage='primary display'] sufficient?
- Should we ignore location/url@displayLabel?
- Is typeOfResource beneficial as a browse facet?

[original page](#)