

# WEB OF SCIENCE™ CORE COLLECTION

## XML USER GUIDE

February, 2022

# Table of Contents

---

Overview.....	3
Delivery Frequency and Files.....	4
Weekly Files and Corrections.....	5
Deletions.....	6
Times Cited File.....	7
Selection Criteria/Products.....	8
Counting Citations.....	9
Support and Questions.....	10
XML Schema.....	11
Schema Diagram.....	12
Source Record Identifiers.....	13
Document and Source Titles.....	15
Source Author Names.....	16
Authors and Addresses.....	19
Organizations.....	22
Contributors.....	23
Categories.....	24
Funding Information.....	25
Cited References.....	27
Citation Context.....	30
Appendix – Helpful Links.....	31

Your contract for raw data entitles you to get timely updates, which you may store and process according to the terms of your agreement. The associated XML schemas describe the record structure of the data and the individual elements that define the fields. You should familiarize yourself with these schemas as you configure your repository to manage this data.

The starting point for every new customer is a delivery of the full WOS re-extraction. This extraction takes place every year at the start of the year, and is generally delivered in February, allowing time for generation, QA, delivery, etc.

Some customers may only receive these annual extractions year over year, however you also have the option to purchase weekly updates or a mid-year update. The details of these options are outlined below.

# Delivery Frequency and Files

---

**Annual:** Annual files are produced at the start of each year to coincide roughly with weekly #1 of the year. These files are generally delivered by the end of February, given the time it takes to generate the files, perform QA on them, deliver them, etc. These files represent a total re-extraction of the WOS database as it is at that time, separated into zip files containing XML for papers from each year. These file names are of the form *YYYY\_PROD.zip* where YYYY is year and PROD is the product code. For instance 2012\_CORE.zip would contain the CORE records from year 2012.

**Weekly:** Weekly updates include several files that are delivered on a set schedule. Primarily this includes a weekly file with both newly indexed records and updated/corrected records. This also includes daily delete files which include the IDs of papers deleted on a given day when applicable, and a times cited file. More information on these files can be found in the subsequent few section.

In addition, a load order file will also be produced and provided with annuals which outlines where the annual extraction files fit in with the weekly files, and which delete files should be processed after which weeklies.

**Mid-Year:** Starting in 2021, a mid year extraction file has also been produced. These files look identical to the annual files, but contain only a combination of new or corrected records from the first half of the year. Effectively they can be processed as a single update on top of the annual files to bring your database up to date as of end of June of that year.

# Weekly Files/Corrections

---

Weekly update files for WOS Core Collection are generally produced over the weekend and delivered on Monday mornings US time. They contain data from the previous Saturday through Friday. These files are numbered in accordance with the week of the year in the form `WOS_RAW_YYYYWW_PROD.tar.gz`, where YYYY again is year, WW is week number and PROD is the product code. So for instance `WOS_RAW_202010_CORE.tar.gz` would be the 10<sup>th</sup> weekly from the year 2020 for CORE records.

These files contain the full record data of any papers either added to WOS or updated within the past week. Note that these files are not year-restricted, so if you have only purchased a limited number of more recent years, you would still get occasional older content in these weekly files. That content can be skipped over or ignored.

## Corrections -

Any XML file you receive may contain corrections (in some cases a more apt word is “update”). There is no data element or indicator that flags a record as an update vs a new record. An updated record will always be a *complete* record. Consequently, if a record in a newly delivered file has a UID that matches the UID of a record in your repository, it should replace the old record. Note also that certain updates to records in WOS occur outside of the normal correction workflow. This pertains primarily to backfills, where we have added a new field or done some particular initiative that involves mass data-updating. In this case those changes would ONLY come through in the next year’s re-extraction.

## Gap Records -

You cannot simply use the publication year of a record to determine which are new and updates, since it is possible for some new content to be loaded from other years. We have traditionally called these ‘gap records.’ Gap record are new records of articles from journals published before the current year. Generally these occur for one of two reasons – when onboarding a new source, we tend to index at least two more years of back content for JCR purposes, and to fill gaps where we may have missed an issue or supplement.

Note that a gap record may contain indexing or data enhancements that were not in practice in that record’s year of publication, causing it to appear to contain more or different data than other records published in the same year, indexed closer to their publication date. Current indexing and data entry policies are applied whenever new records are added to the database, regardless of the year of the source publication. For example, Web of Science Core Collection began including author email addresses for authors in 1997. If a gap record for a 1995 article is created in 2013, and if the article includes author email addresses, then the gap record will include the email addresses.

This is worth noting particularly for occasional backfill project where large data may be processed from older dates, adhering to the current policy. This is rare, but two recent examples would be the ESCI backfill, started in 2018, and the CPCI backfill started in 2017.

Customers entitled to weekly files also receive daily delete files. These files are produced every day, seven days a week, when there are papers deleted from WOS. There are not deletes every day and thus there will not be files every day. The file names will be in the form *WOSYYYYDDD.del* where YYYY is year and DDD is the day of the year numbered from 1 to 365 (or 366 in a leap year). So for instance *WOS2021021.del* would be the deletes from the 21<sup>st</sup> day of 2021. These files contain simply a list of record IDs to be deleted on that day. They are not restricted to any products or year, so if a particular paper is already missing from your database that is not indicative of an issue.

The format for the file itself is very simple, containing simply a CSV files with two fields: a UID of the paper to be deleted followed by a confirmation field that should always be 'Y.'

*Sample List of Records Deleted from Web of Science Core Collection:*

```
WOS,0002085180000001,Y
WOS,0002085180000002,Y
WOS,0002085180000003,Y
WOS,0002085180000004,Y
WOS,0002085180000005,Y
```

# Times Cited File

---

A file containing Times Cited numbers from Clarivate Analytics is also available. This file, updated weekly, provides up to-date Times Cited data in a tab-delimited format. The file provides the UID of the cited source item and current Times Cited values for that source item broken down by various combinations of WOS editions (as well as a total).

Each Times Cited file has a timestamp. You should process the files as they are received or if updating less frequently, simply use the latest version.

*Example:*

COLLECTION:UID	WOS_TOTAL	SCIE_SSCI_AHCI	SCIE	SSCI	AHCI	CPCI-S_CPCI-S:
WOS:000003907500001	4 4	3 1 0 0	0 0	0 0	0 0	
WOS:000003907500002	61 57	54 7 2 6	5 1	2 2	0 0	
WOS:000003907500003	23 23	23 0 0 0	0 0	0 0	0 0	
WOS:000003907500004	26 25	25 0 0 2	2 0	1 1	0 0	
WOS:000003907500005	108 108	108 0 0 1	1 0	0 0	0 0	
WOS:000003907500006	29 28	28 0 0 3	3 0	1 1	0 0	
WOS:000003907500007	34 34	34 0 0 1	1 0	0 0	0 0	
WOS:000003907500008	673 643	643 4 0 32	32 0	13 13	0 7	
WOS:000003907500009	2 2	2 0 0 0	0 0	0 0	0 0	

# Selection Criteria/Products

---

You may specify criteria in a variety of ways for ad hoc XML deliveries, including querying on essentially all fields searchable in WOS product including things like publication years, journal title/ISSN, subject categories, addresses/institutions, and so on.

Annual/Weekly XML data can be purchased based on publication year ranges and edition combinations. See the tables below for a mapping of available products to editions included.

Product Name	WOS Editions
DU	SCIE
DSSHU	SCIE, SSCI, AHCI
DSSHPSHU	SCIE, SSCI, AHCI, CPCI-S, CPCI-SSH
CORE	SCIE, SSCI, AHCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH
ESCI	ESCI



# Counting Citations

---

It is possible to count the number of times a source item has been cited from reference data in the XML files. Each source record has a primary key, the UID. A cited reference also contains a UID, and when that reference resolves to a WOS Core record, those IDs will match (all WOS Core UIDs will be prefixed with 'WOS:'). The XPaths for these elements respectively are –

- /records/REC/UID
- /records/REC/static\_data/fullrecord\_metadata/references/reference/uid

The number of times a UID is found in a reference list is the number of times the paper was cited.

We recommend maintaining a two-field table of the source UID to all cited UIDs for every ingested record. This can then serve as the linking between both citing and cited items, and allows for easy analysis of citation via group-by queries. You could also supplement this table with date information, citation context information, etc as you like.

Note that while a UID should never change, the actual reference resolution could change and thus a reference could resolve to a record with a different UID at some point in time due to corrections.

For citation counts alone you can also reference the times cited file outlined above.

# Support and Questions

---

If you have questions about the raw XML format or data presentation, send an email to [Thomson-RawDataProductionandSupport@clarivate.com](mailto:Thomson-RawDataProductionandSupport@clarivate.com), or feel free to follow up directly to [joseph.brightbill@clarivate.com](mailto:joseph.brightbill@clarivate.com).

Clarivate URL Schema, new xmlns

```
<?xml version="1.0" encoding="UTF-8"?> <!-- Copyright (c) 2022 Clarivate Analytics Web of Science -->
<records xmlns="http://clarivate.com/schema/wok5.30/public/FullRecord">
```

This is the core schema. It defines the basic XML framework for a record of a source document. Each record enclosed by the REC element consists of:

- **UID** - Unique item identifier
- **static\_data** - Static bibliographic elements derived from source publications or from database-specific, value-added indexing
- **dynamic\_data** - Bibliographic elements and metadata generated by database processing and integration

EWUID.rawxml.xsd

Elements in this schema define the identifiers that uniquely identify a database record and that supply additional processing capabilities.

summary.rawxml.xsd

Elements in this schema define the core bibliographic fields that make up a summary record in Web of Science.

common\_types.rawxml.xsd

Elements in this schema extend the core list of elements in summary.xsd. Not every element defined in this schema is found in all editions. Conversely, some elements in this schema may occur in only one or two databases outside of WOS Core.

fullrecord\_metadata.rawxml.xsd

Elements in this schema describe bibliographic fields and record metadata not displayed in summary records.

item\_wos.rawxml.xsd

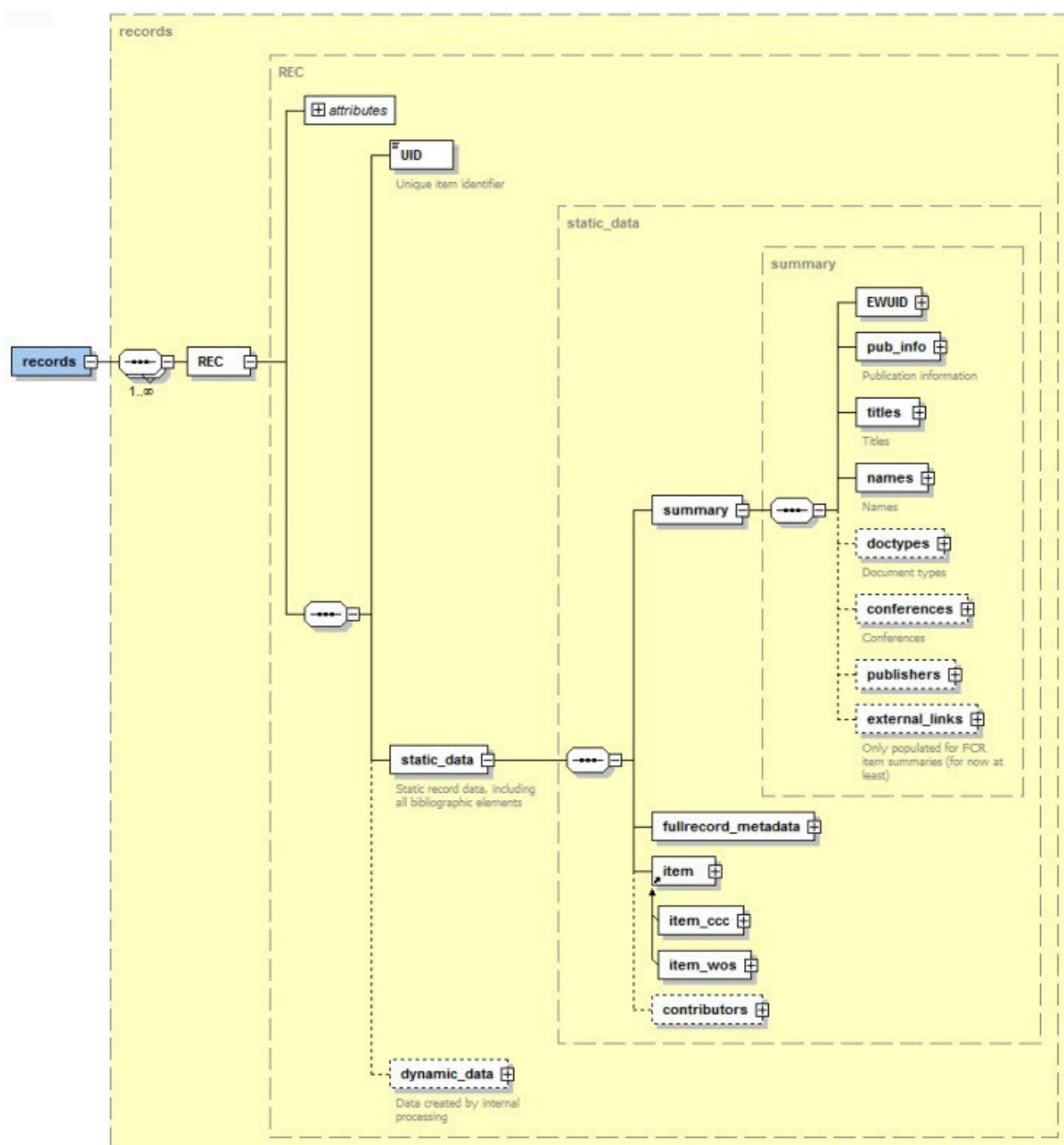
Elements in this schema describe bibliographic fields and record metadata unique to Web of Science Core Collection.

Additional “item\_\*.xsd” files that you may see in the schema contain fields that are specific to those given collections as well.

# Schema Diagram

This graphic illustrates the basic hierarchy of the schema documents that make up the schema for Web of Science Core Collection. The starting point is the <records> element in the core document, scientific.thomsonreuters.com.schema.wok5.X.rawxml.xsd

This diagram does not reveal the relationship of the document common\_types.xsd to the other schemas. All schema documents except the core schema document and EWUID.xsd include common\_types.xsd.



# Source Record Identifiers

Each source record in Web of Science Core Collection has a unique identifier, the UID. The UID is prefaced by an abbreviation of the collection (database) from which the record is retrieved (WOS for Web of Science Core Collection). The UID is always the first child element of the <REC> element:

```
<REC r_id_disclaimer="ResearcherID data provided by Clarivate Analytics">
  <UID>WOS:000306312500009</UID>
```

In Web of Science and over time, the UID has also been variously labeled **Accession Number**, **UT** and **ISI\_LOC**.

## Other Identifiers

The WUID (**W**eb of Science **U**nique **I**dentifier) identifies the collection and edition where the record is stored. The WUID is a child of EWUID (**E**dition **W**UID). A mapping table to the possible edition values and their full name counterparts in product is below.

## Web of Science™ Core Collection Database to Edition Mapping

Database in Product	Edition
Science Citation Index Expanded	WOS.SCI
Social Sciences Citation Index	WOS.SSCI
Arts & Humanities Citation Index	WOS.AHCI
Conference Proceedings Citation Index- Science	WOS.ISTP
Conference Proceedings Citation Index- Social Sciences & Humanities	WOS.ISSHP
Book Citation Index– Science	WOS.BSCI
Book Citation Index– Social Sciences & Humanities	WOS.BHCI
Emerging Science Citation Index	WOS.ESCI

## Pubmed Identifier (PMID)

Either when supplied by the source publication, or when a link is established between a WOS Core item and an item in Medline/Pubmed, an article's PMID is included in the source record in <identifier type="pmid">

```
<identifiers>
  <identifier type="accession_no" value="UM7RR"/>
  <identifier type="issn" value="0944-1344"/>
  <identifier type="eissn" value="1614-7499"/>
  <identifier type="doi" value="10.1007/s11356-021-16239-6"/>
  <identifier type="pmid" value="34480705"/>
</identifiers>
```

## Digital Object Identifier (DOI)

When supplied by the source publication, an article's DOI is included in the source record in `<identifier type="doi">`

```
<identifiers>
<identifier type="accession_no" value="0740J"/>
<identifier type="issn" value="1936-6582"/>
<identifier type="doi" value="10.1007/s10696-011-9117-0"/>
</identifiers>
```

If a DOI is not supplied in the source item, but we can find a match in Crossref, `<identifier type="xref_doi">` will be added.

```
<identifiers>
<identifier type="accession_no" value="241EK"/>
<identifier type="issn" value="0021-4922"/>
<identifier type="xref_doi" value="10.1143/JJAP.38.L872"/>
</identifiers>
```

The DOI is a persistent identifier for a document, regardless of where the document appears. Note that not all records in Web of Science Core Collection have DOI's or PMIDs. DOI's were captured from source publications starting in 2002.

# Document and Source Titles

Document and source titles are given in the <title> element and categorized by the **type** attribute. Typically, the **item** type identifies the article title, and the **source** type identifies the publication title (journal or book). Note that for books in series, the **source** type identifies the book title, and the **book series** type identifies the series title.

## Journal Article:

```
<pub_info issue="2" pubtype="Journal" sortdate="2007-04-01" has_abstract="Y" coverdate="APR 2007"
pubmonth="APR" vol="12" pubyear="2007">
  <page end="157" page_count="5" begin="153">153-157</page>
</pub_info>
<titles count="6">
  <title type="source">ANNALS OF NONINVASIVE ELECTROCARDIOLOGY</title>
  <title type="source_abbrev">ANN NONINVAS ELECTRO</title>
  <title type="abbrev_iso">Ann. Noninvasive Electrocardiol.</title>
  <title type="abbrev_11">ANN NONINVA</title>
  <title type="abbrev_29">ANN NONINVASIVE ELECTROCARDIO</title>
  <title type="item">Preliminary observations on the effect of amitriptyline treatment in
preventing syncope recurrence
in patients with vasovagal syncope</title>
</titles>
```

## Book:

```
<pub_info has_abstract="N" coverdate="2011" pubtype="Book" pubyear="2011"
sortdate="2011-01-01">
  <page end="244" page_count="65" begin="1">1-244</page>
</pub_info>
<titles count="2">
  <title type="source">OPTICAL FLUORESCENCE MICROSCOPY: FROM THE SPECTRAL TO
THE NANO DIMENSION</ title>
  <title type="item">Optical Fluorescence Microscopy: From the Spectral to the
Nano Dimension</ title>
</titles>
```

## Book in Series:

```
<pub_info pubtype="Book in series" sortdate="2011-01-01" has_abstract="Y"
coverdate="2011" vol="1239" pubyear="2011"> <page end="70" page_count="12"
begin="59">59-70</page>
</pub_info>
<titles count="8">
<title type="source">CRITICAL CONTRIBUTIONS OF THE ORBITOFRONTAL CORTEX TO BEHAVIOR</title>
<title type="series">Annals of the New York Academy of Sciences</title>
<title type="source_abbrev">ANN NY ACAD SCI</title>
<title type="abbrev_iso">Ann.NY Acad.Sci.</title>
<title type="abbrev_11">ANN NY ACAD</title>
<title type="abbrev_29">ANN N Y ACAD SCI</title> <title type="item">Representations of
appetitive and aversive information in the primate orbitofrontal cortex</title> <title
type="book_series" translated="N">Annals of the New York Academy of Sciences</title>
</titles>
```

# Source Author Names

The names of all authors of source publications are captured in Web of Science Core Collection. The names are listed in database records in the same order in which they are listed in the source publications. Child elements of the name element contain author name data:

Element	Description
name	Parent element for one author name.
display_name	Full name. If no full name is given, then the display_name is the wos_standard name.
full_name	Full name as given by the source publication
wos_standard	Surname followed by a comma and up to five initials.
first_name	First (given) name
last_name	Surname or family name
suffix	Generational suffix from a given name (JR, III, etc)
email_addr	Email address

In addition, the name element itself has the following attributes:

Attribute	Description
seq_no	Position of author in author list
addr_no	Indicates which address in the address field is associated with this author. An author can be associated with multiple addresses.
role	Role. Possible values include author, editor and inventor. The full list of roles can be found in the schema document <i>common_types.rawxml.public.xsd</i> .
reprint	Reprint flag. A value of Y indicates that the author is the reprint author.

Here is an example of an author element with many of the fields noted above.

```
<name seq_no="1" addr_no="1 2" role="author" reprint="Y">
  <display_name>Aaltonen, Jonna</display_name>
  <full_name>Aaltonen, Jonna</full_name>
  <wos_standard>Aaltonen, J</wos_standard>
  <first_name>Jonna</first_name>
  <last_name>Aaltonen</last_name>
  <email_addr>jmaalt@utu.fi</email_addr>
</name>
```



## Full Names and Abbreviations

Starting in May 2006, full names were captured from source journals. Before that, only full surnames were captured. First and middle names were abbreviated, and a name could have a maximum of five initials. Note that a paper would be subject to the policy at the time of capture, not the time of publication, and so issues processed later (either for backfill or correction) would be subject to whatever the policy was at that time. In addition, some backfill projects have taken place over the years to “fill out” some older names, so only some content from these older years will be subject to these older forms.

*Before May 2006:*

Published Name	Processed Name
Albrecht-Schmitt, Theodore Ernest	Albrecht-Schmitt, TE
Brea, Rachel J.	Brea, RJ
Fournier, Jean-Baptiste	Fournier, JB
Sheng, D.	Sheng, D

*May 2006 and Later:*

Full names are captured and presented in the database. The <wos\_standard> element contains the Web of Science abbreviation.

Published Name	Processed Name <full_name>	Processed Name wos_standard
Albrecht-Schmitt, Theodore Ernest	Albrecht-Schmitt, Theodore Ernest	Albrecht-Schmitt, TE
Brea, Rachel J.	Brea, Rachel J.	Brea, RJ
Fournier, Jean-Baptiste	Fournier, Jean-Baptiste	Fournier, JB
Sheng, D.	Sheng, D.	Sheng, D

*Author Names 1964-1975:*

During data years 1964 to 1975, source author names were captured with a maximum of 11 characters: 8-character last names, followed by a space or a period (if truncated), and up to two initials. If the length of the last name permitted, more than 2 initials were captured.

For example, the majority of source authors were captured during 1964-1975 like this:

A. Johnston was captured as Johnston A

D.E. Hofstadter was captured as Hofstadt.De

A. Rodriguez was captured as Rodrigue.A

A. Rodrigues was captured as Rodrigue.A

G.E.P. Box was captured as Box GEP

## Chinese Author Names

If the journal is a Chinese publication, our approach is that the author name is in original Chinese name order: surname followed by first and middle names.

If the journal is not a Chinese publication, we assume that the Chinese names are in the same order as the other names in the journal (that is, not in original Chinese name order).

*Hyphenated Names* - The hyphenated portion of the name is presented as an initial, and the unhyphenated portion of the name is presented as the surname.

Published Name	Processed Name (full_name)	Processed Name (wos_standard)
Chang Hui-Lan	Chang, Hui-Lan	Chang, HL

*Three-Part Hyphenated Names* - If all three parts of a Chinese name are hyphenated, the name is processed as if there are no hyphens. The last name element becomes the last name; the other two parts become initials.

The name is processed following the normal rules for American names. For example:

Published Name	Processed Name (full_name)	Processed Name (wos_standard)
Lian-Tien-Sun	Sun, Lian-Tien	Sun, LT

*Four-Part Names* - Some Chinese names are presented in four parts. If some of the parts are hyphenated and some are not, the unhyphenated portion is processed as the last name; the other parts as initials. For example:

Published Name	Processed Name (full_name)	Processed Name (wos_standard)
W. Chia-Mo Wan	Wan, W. Chia-Mo	Wan, WCM

*Unhyphenated Names* - If no hyphens are present in the name, the first part of the name is processed as the surname. If the second part has only one syllable, only one initial is processed. For example:

Published Name	Processed Name (full_name)	Processed Name (wos_standard)
Ju Rui	Ju, Rui	Ju, R
Sun Shu	Sun, Shu	Sun, S
Hu Chau	Hu, Chau	Hu, C

If the second part of the name has two syllables, the first letter of each syllable is presented as initials. For example:

Published Name	Processed Name (full_name)	Processed Name (wos_standard)
Hong Longsheng	Hong, Longsheng	Hong, LS
Zhang Wanhua	Zhang, Wanhua	Zhang, WH
Shi Youngshan	Shi, Youngshan	Shi, YS
Chang Cheng-hseuh	Chang, Cheng-hseuh	Chang, CH

Chinese names that present a last name, first/middle name and an initial are processed following our policy for unhyphenated Chinese Names with two syllables, plus an initial. For example:

Published Name	Processed Name (full_name)	Processed Name (wos_standard)
Yu Seungju M	Yu, Seungui M.	Yu, SGM

# Authors and Addresses

This excerpt from a Web of Science® Core Collection record shows that five names are associated with the source document and that one of them is Alvaro Rodriguez-Prieto.

```
<names count="5">
<name seq_no="1" role="author" reprint="Y" addr_no="1 2">
  <display_name>Rodriguez-Prieto, Alvaro</display_name>
  <full_name>Rodriguez-Prieto, Alvaro</full_name>
  <wos_standard>Rodriguez-Prieto, A</wos_standard>
  <first_name>Alvaro</first_name>
  <last_name>Rodriguez-Prieto</last_name>
  <email_addr>alvaro.rodriguez@invi.uned.es</email_addr>
</name>
```

Attributes of the name element reveal that Alvaro Rodriguez-Prieto is:

- The first of five names (seq\_no="1").
- An author (role="author" - Most, but not all, names are author names).
- The reprint author (reprint="Y").

In addition, starting at the beginning of 2008, Web of Science data capture policy was changed to index the links between authors and addresses. This linking is done via the sequence numbers associated with the addresses. The addr\_no attribute identifies the sequence number of the addresses linked to this author (separated by a space if there are more than one). So in this case we also know that Alvaro Rodriguez-Prieto is linked to the first and second addresses.

This excerpt from the addresses section shows the article has two addresses. Three authors are associated with the first address. The first of these is Alvaro Rodriguez-Prieto.

```
<addresses count="2">
<address_name>
<address_spec addr_no="1">
  <full_address>Argonne Natl Lab, Appl Mat Div, Lemont, IL 60539 USA</full_address>
  <organizations count="4">
    <organization>Argonne Natl Lab</organization>
    <organization pref="Y">Argonne National Laboratory</organization>
    <organization pref="Y">University of Chicago</organization>
    <organization pref="Y">United States Department of Energy (DOE)</organization>
  </organizations>
  <suborganizations count="1">
    <suborganization>Appl Mat Div</suborganization>
  </suborganizations>
  <city>Lemont</city>
  <state>IL</state>
  <country>USA</country>
  <zip location="AP">60539</zip>
</address_spec>
<names count="3">
  <name seq_no="1" role="author" reprint="Y" addr_no="1" r_id="">
    <display_name>Rodriguez-Prieto, Alvaro</display_name>
    <full_name>Rodriguez-Prieto, Alvaro</full_name>
    <wos_standard>Rodriguez-Prieto, A</wos_standard>
    <first_name>Alvaro</first_name>
    <last_name>Rodriguez-Prieto</last_name>
    <email_addr>alvaro.rodriguez@invi.uned.es</email_addr>
  </name>
```

```

<name seq_no="3" role="author" addr_no="1" r_id="">
  <display_name>Aragon, Ana M.</display_name>
  <full_name>Aragon, Ana M.</full_name>
  <wos_standard>Aragon, AM</wos_standard>
  <first_name>Ana M.</first_name>
  <last_name>Aragon</last_name>
</name>
<name seq_no="5" role="author" addr_no="1">
  <display_name>Yanguas-Gil, Angel</display_name>
  <full_name>Yanguas-Gil, Angel</full_name>
  <wos_standard>Yanguas-Gil, A</wos_standard>
  <first_name>Angel</first_name>
  <last_name>Yanguas-Gil</last_name>
</name>
</names>
</address_name>
<address_name>
<address_spec addr_no="2">
  <full_address>Univ Nacl Educ Distancia, Dept Mfg Engr, E-28040 Madrid, Spain</full_address>
  <organizations count="2">
    <organization>Univ Nacl Educ Distancia</organization>
    <organization pref="Y">Universidad Nacional de Educacion a Distancia (UNED)</organization>
  </organizations>
  <suborganizations count="1">
    <suborganization>Dept Mfg Engr</suborganization>
  </suborganizations>
  <city>Madrid</city>
  <country>Spain</country>
  <zip location="BC">E-28040</zip>
</address_spec>
<names count="3">
  <name seq_no="1" role="author" reprint="Y" addr_no="2" r_id="">
    <display_name>Rodriguez-Prieto, Alvaro</display_name>
    <full_name>Rodriguez-Prieto, Alvaro</full_name>
    <wos_standard>Rodriguez-Prieto, A</wos_standard>
    <first_name>Alvaro</first_name>
    <last_name>Rodriguez-Prieto</last_name>
    <email_addr>alvaro.rodriguez@invi.uned.es</email_addr>
  </name>
  <name seq_no="2" role="author" addr_no="2" r_id="M-1685-2014">
    <display_name>Camacho, Ana M.</display_name>
    <full_name>Camacho, Ana M.</full_name>
    <wos_standard>Camacho, AM</wos_standard>
    <first_name>Ana M.</first_name>
    <last_name>Camacho</last_name>
  </name>
  <name seq_no="4" role="author" addr_no="2" r_id="">
    <display_name>Sebastian, Miguel A.</display_name>
    <full_name>Sebastian, Miguel A.</full_name>
    <wos_standard>Sebastian, MA</wos_standard>
    <first_name>Miguel A.</first_name>
    <last_name>Sebastian</last_name>
  </name>
</names>
</address_name>
</addresses>

```

Finally, the article has 2 reprint addresses. The same author Alvaro Rodriguez-Prieto is associated with two different reprint addresses.

```

<reprint_addresses count="2">
<address_name>
<address_spec addr_no="1">
<full_address>Argonne Natl Lab, Appl Mat Div, Lemont, IL 60539 USA</full_address>
<organizations count="4">
  <organization>Argonne Natl Lab</organization>
  <organization pref="Y">University of Chicago</organization>
  <organization pref="Y">United States Department of Energy (DOE)</organization>
  <organization pref="Y">Argonne National Laboratory</organization>
</organizations>
<suborganizations count="1">
  <suborganization>Appl Mat Div</suborganization>
</suborganizations>
<city>Lemont</city>
<state>IL</state>
<country>USA</country>
<zip location="AP">60539</zip>
</address_spec>
<names count="1">
<name seq_no="1" role="author" reprint="Y" addr_no="1">
  <display_name>Rodriguez-Prieto, Alvaro</display_name>
  <full_name>Rodriguez-Prieto, Alvaro</full_name>
  <wos_standard>Rodriguez-Prieto, A</wos_standard>
  <first_name>Alvaro</first_name>
  <last_name>Rodriguez-Prieto</last_name>
  <email_addr>alvaro.rodriquez@invi.uned.es</email_addr>
</name>
</names>
</address_name>
<address_name>
<address_spec addr_no="2">
<full_address>Univ Nacl Educ Distancia, Dept Mfg Engn, E-28040 Madrid, Spain</full_address>
<organizations count="2">
  <organization>Univ Nacl Educ Distancia</organization>
  <organization pref="Y">Universidad Nacional de Educacion a Distancia
(UNED)</organization>
</organizations>
<suborganizations count="1">
  <suborganization>Dept Mfg Engn</suborganization>
</suborganizations>
<city>Madrid</city>
<country>Spain</country>
<zip location="BC">E-28040</zip>
</address_spec>
<names count="1">
<name seq_no="1" role="author" reprint="Y" addr_no="2">
  <display_name>Rodriguez-Prieto, Alvaro</display_name>
  <full_name>Rodriguez-Prieto, Alvaro</full_name>
  <wos_standard>Rodriguez-Prieto, A</wos_standard>
  <first_name>Alvaro</first_name>
  <last_name>Rodriguez-Prieto</last_name>

```

Prior to 1998, a research address that matches a reprint address is not included in the list of research addresses. Beginning in 1998, we do not remove a duplicate address if it appears as both a research and a reprint address. Prior to 2016, one reprint/corresponding author/address was indexed per paper. Beginning in 2016, we index all reprint/corresponding authors and addresses per paper.

No addresses were processed for the following editions and years (except in the case where a gap issue is processed):

- Science Citation Index Expanded 1945-1964
- Social Sciences Citation Index 1956-1965

# Organizations

The names of organizations are extracted from the author address and identified by the *organization* element as follows:

```
<address_spec addr_no="5"> <full_address>Univ Penn, Dept Pathol, Div Transfus Med, Stem Cell Lab, Philadelphia, PA 19104 USA</ full_address> <organizations count="2">
  <organization>Univ Penn</organization>
  <organization pref="Y">University of Pennsylvania</organization>
</organizations>
<suborganizations count="3">
  <suborganization>Dept Pathol</suborganization>
  <suborganization>Div Transfus Med</suborganization>
  <suborganization>Stem Cell Lab</suborganization>
</suborganizations>
<city>Philadelphia</city>
<state>PA</state>
<country>USA</country>
<zip location="AP">19104</zip>
</address_spec>
```

Organization names undergo a certain amount of normalization at data capture, but can still present differently across various publications. These names can refer to constituent organizations, and often contain abbreviations.

Any address captured in WOS is subject to unification, and as such many addresses contain a preferred organization name (sometimes referred to as “Org Enhanced”), signifying that that address was indeed unified. The unification of addresses is a constantly improving process. Any unified address will have a separate “organization” element with attribute pref=Y to denote it is the unified form. There may be multiple preferred names per one address, most often when an address is unified to a single school and its parent org as part of a larger system. For example:

```
<address_spec addr_no="3">
  <full_address>Texas A&M Univ Syst, Hlth Sci Ctr, Scott & White Healthcare, Donor Serv, Temple, TX 76508 USA</full_address>
  <organizations count="3">
    <organization>Texas A&M Univ Syst</organization>
    <organization pref="Y">Texas A&M Health Science Center</organization>
    <organization pref="Y">Texas A&M University System</organization>
  </organizations>
  <suborganizations count="3">
    <suborganization>Hlth Sci Ctr</suborganization>
    <suborganization>Scott & White Healthcare</suborganization>
    <suborganization>Donor Serv</suborganization>
  </suborganizations>
  <city>Temple</city>
  <state>TX</state>
  <country>USA</country> <zip location="AP">76508</zip>
</address_spec>
```

The contributors element contains the names of authors for whom a Publons Profile/ResearcherID or an ORCID identifier is provided. Some authors have both IDs.

Element	Description
contributors	Parent element for the list of contributor information coming from RID/ORCID. The count attribute shows the number of contributors in the contributor list
contributor	The information for a single contributor
name	Attributes of the name element contain the ResearcherID or ORCID identifier.
display_name	Name as given in the RID or ORCID account
full_name	Full name, same as display_name
first_name	First (given) name, as parsed from the full_name
last_name	Surname or family name, as parsed from the full_name

Attributes of <name>

Attribute	Description
orcid_id	ORCID identifier
rid_id	PublonsID/ResearcherID. This attribute is always accompanied by the role attribute whose value is researcher_id.
seq_no	The value of this attribute is the sequence number in the list of contributors.

## Example

```
<contributors count="4">
  <contributor>
    <name orcid_id="0000-0003-1069-212X" r_id="A-7779-2008" role="researcher_id"
seq_no="1">
      <display_name>Calbet, Albert</display_name>
      <full_name>Calbet, Albert</full_name>
      <first_name>Albert</first_name>
      <last_name>Calbet</last_name>
    </name>
  </contributor>
  <contributor>
    <name orcid_id="0000-0003-2611-0067" r_id="K-4263-2014" role="researcher_id"
seq_no="2">
      <display_name>Saiz, Enric</display_name>
      <full_name>Saiz, Enric</full_name>
      <first_name>Enric</first_name>
      <last_name>Saiz</last_name>
    </name>
  </contributor>
</contributors>
```

# Categories

The subject element contains the subject category of a journal, and every record from a journal in a Web of Science Core Collection database should have this element. A mapping of journals to these categories is maintained in our journal system, and when a new item is indexed in a given journal, it picks up whatever that journal's The term *ascatype*, which is an attribute of subject, is a system term for *subject category*.

XML Tag	Example
<subject ascatype="traditional">	<subject ascatype="traditional">Engineering, Manufacturing</ subject>

A "traditional" ascatype (tASCA) indicates that the subject category comes from what we consider the traditional ~250 Web of Science categories. Every journal indexed in Web of Science Core Collection should be assigned to at least one tASCA type. It is also not unusual for a journal to be assigned more than one.

An "extended" ascatype (eASCA type) indicates that the subject category comes from the list on page 34. This is referred to as a "research area" in WOS product. The eASCA types provide a small level of aggregation on top of the tASCA types, with the aim of providing a single subject category scheme across all Web of Science databases. They are added by applying a mapping to the tASCA values. As such, there are fewer unique eASCA values than tASCA values. eASCA types themselves also map to "heading" and "subheading" values (which are also elements within the "category\_info" node). These are even broader fields which are not really used in the product, other than for display purposes.

XML Tag	Example
<subject ascatype="extended">	<subject ascatype="extended">Engineering</subject>

You can find some more information on specific categories in the appendices of this document.



# Funding Information

The “fund\_ack” element contains all of the funding data associated with a record, including the funding text, funding organization, and grant number. This data has been either captured and processed by us, or indexed from third part sources. For the data captured by us, that capture of funding acknowledgements began in 2008. The English-language statement containing funding information (“funding statement”) is captured, typically contained in the paper’s acknowledgements section. Grant agency names and grant numbers are captured and indexed by extracting them from that funding text. The following table outlines when funding was captured in which editions / for which document types:

Start Year	Edition
2005	ESCI
2008	SCIE (Article and Reviews)
2015	SSCI
2017	SCIE (All Doc Types)
2017	AHCI
2017	CPCI
2019	BKCI

The fund\_ack element itself contains two main sub-elements. The first is the “fund\_text” element which contains the full funding text as indexed from the record (the text itself is indexed in paragraph elements at fund\_text/p). The second is the “grants” element which can contain one or more “grant” elements which contain the actual funding organizations (grant/grant\_agency) and grant IDs (grant/grant\_ids/grant\_id). Note that each grant element should have only one funding org, but can have multiple grant IDs. A simple example follows –

```
<fund_ack>
<fund_text>
<p>This is example funding acknowledgement text as it appears on the full record</p>
</fund_text>
<grants count="1">
<grant>
<grant_agency>German Research Foundation (Deutsche Forschungsgemeinschaft)</grant_agency>
<grant_agency pref="Y">German Research Foundation (DFG)</grant_agency>
<grant_ids count="1">
<grant_id>ABC-123</grant_id>
</grant_ids>
</grant>
</grants>
</fund_ack>
```

Note that in late 2019 unified funder was added to WOS platform. This was added to the XML starting in January 2021. The format follows similar to the format of unified organizations in addresses, where multiple funding agency values are present and the unified version is distinguishable by the presence of a “pref” attribute that equals “Y”. In the example above you can see that the unified name of the grant agency captured from the paper is “German Research Foundation (DFG)”.

In addition to the funding data that we capture ourselves, funding data from third party sources is also indexed in WOS. This data looks the same as the captured funding data, but will contain a “grant\_source” attribute within the “grant” element (for example – ‘<grant grant\_source=“NIH RePORTER”>’). Because these data are also coming from a third party source and not directly from the funding text, these records need not necessarily have any funding text captured. As of 2021 there are six possible grant source values, and those are: Researchfish, Medline, NIH RePORTER, Federal RePORTER, KAKEN, and Custom. Custom does not come directly from a

third party source but rather has been carried from a small bit of custom funder mapping done as parts of older InCites 1.x projects. New source will be onboarded into 2022 and beyond, so the possible values here will continue to grow.

# Cited References

All references cited by the source document are included in the source record in Web of Science™ Core Collection. Cited references may be classified into two broad categories: 1) references to source items in Web of Science Core Collection and 2) references that do not have matching source items in Web of Science Core Collection.

Element	Description
<uid>	<p>Cited reference identifier. There are two types of uid values: 1) the UID of a matching source item in Web of Science and 2) the UID of the parent (citing) document, followed by an increment. A value of the second type indicates a reference for which there is no matching source item.</p> <p>Note that because of data corrections and deletions, the uid of a cited reference can change. In addition, a uid can be added to a cited reference that previously had none.</p>
<citedAuthor>	First author of the cited document.
<year>	Publication year of the cited document.
<page>	<p>Starting page number of the cited document.</p> <p>Be aware that the value of the &lt;page&gt; element may be an identifier such as ARTN (article number). The identifier may appear twice in a cited reference, once in the &lt;page&gt; element and once in the &lt;art_no&gt; element.</p>
<citedTitle>	<p>Title of the cited document.</p> <p>For references processed from 2012 forward, cited references are captured with full titles when those titles are supplied by the citing article, regardless of whether the cited reference matches a source item.</p> <p>For references processed prior to 2012, it is likely that a citing title will not be included. However, some earlier cover dates may have been updated in 2012 or later. In this case there may be a full citing title presented if the title is covered as a source, or the author included the full cited title in the reference.</p>
<citedWork>	<p>Title of the cited publication.</p> <p>The value of this element may be a full work title or an abbreviated work title.</p> <p>The full work title is shown if the reference is from an article processed in 2012 or later <i>and</i> the cited publication is also a source publication <i>or</i> the author included the full title in the reference. An abbreviated work is shown if the reference refers to a publication that is not covered as a source and the author did not provide the full work title or the cited reference is from an article processed before 2012.</p>
<doi>	<p>Digital Object Identifier.</p> <p>From 2002 forward, the doi of a cited reference is captured when supplied by the citing article.</p>
<art_no>	Article number.



The article number is a unique item identifier assigned by the journal in which the citing article is published and not by the authors of the citing article. This identifier is prefaced by ARTN (for article number), PII (for publisher item identifier), or UNSP (for unspecified).

Not all cited references have this element.

**Sample Cited Reference to a Source Item** - The value of the uid is the UID of a matching source item in Web of Science Core Collection

```
<reference>
<uid>WOS:000253911800008</uid>
<citedAuthor>Gouw, AA</citedAuthor>
<year>2008</year>
<page>247</page>
<volume>25</volume> <citedTitle>Reliability and sensitivity of visual scales versus volumetry for
evaluating white matter hyperintensity progression</citedTitle> <citedWork>CEREBROVASCULAR
DISEASES</citedWork>
<doi>10.1159/000113863</doi>
</reference>
```

**Sample Cited Reference to a Non-Source Item** - Here the value of the uid is the UID of the parent (citing) document, followed by a sequence number pertaining to that item's location in the paper's bibliography.

```
<reference>
<uid>000313229500012.8</uid>
<citedAuthor>Clark, L.</citedAuthor>
<year>2008</year>
<page>349</page>
<citedWork>Heart Failure</citedWork>
</reference>
```

## Citations to Articles from Journal Supplements

When both a volume number *and* a supplement number are provided in the cited reference, the volume number is keyed in the volume field, and an S is appended to the cited work, along with the supplement number.

Example:

Johnson, L.A., Albers, J.G., Willems, C.M.T. and Sybesman, W. Effectiveness of fresh and frozen boar semen under practical conditions. J Anim. Sci. 49: Suppl. 1, 306, (1979).

```
<reference>
<citedAuthor>JOHNSON LA</citedAuthor>
<year>1979</year>
<page>306</page>
<volume>49</volume>
<citedWork>J ANIM SCI S1</citedWork>
</reference>
```

When only one number is present, the number is keyed in the volume field and an S is appended to the cited work

### Example:

Bojensen, E. *A method for determination of insulin in plasma and urine, Acta med. scand. Suppl. No. 266, p. 275, 1952.*

```
<reference>
<citedAuthor>BOJENSEN E</citedAuthor>
<year>1952</year>
<page>275</page>
<volume>266</volume>
<citedWork>ACTA MED SCAND S</citedWork>
</reference>
```

There will only be an 'S' in the citation data when the citation itself indicates a Supplement. Sometimes an 'S' precedes the page number, to indicate a supplement. In that case, we will include this S with the page number.

### Issue Information in the Volume Field

Following is an example of a cited reference presentation that is different from the usual. In this case we process the issue number in the volume field. Here is a reference from a source article - C. Poriel, Y. Ferrand, P. Le Maux, G. Simonneaux, Synlett 1 (2002)

```
<reference>
<citedAuthor>Poriel C</citedAuthor>
<year>2002</year>
<page>71</page>
<volume>1</volume>
<citedWork>Synlett</citedWork>
</reference>
```

In WOS, Synlett does not have volume numbers. 1 is the issue number.

### Cited Authors in Reference to Proceedings and Patents

The cited author name in a reference to a proceedings paper has a limit of 38 characters before the name is truncated while the patent assignee field has a limit of 20 characters.

# Citation Context Data

In late 2021, Web of Science started indexing contextual data for cited references. This includes further information about a particular citation in relation to the rest of the paper. This information manifests in the “physicalSection” element within each reference. This element includes a “physicalLocation” attribute as well as a value representing the label of the section within the paper in which this item was referenced. For instance:

```
<reference occurrenceOrder="3">
  <physicalSection physicalLocation="0.030996844">Introduction</physicalSection>
  <physicalSection physicalLocation="0.075745187">Review of Previous Experiments</physicalSection>
  <physicalSection physicalLocation="0.118709582">Review of Previous Experiments</physicalSection>
  <physicalSection physicalLocation="0.202730494">Review of Previous Experiments</physicalSection>
  <physicalSection physicalLocation="0.219529721">Review of Previous Experiments</physicalSection>
  <physicalSection physicalLocation="0.265739984">Configurations of Test Beams</physicalSection>
  <physicalSection physicalLocation="0.44292475">Experimental Results</physicalSection>
  <physicalSection physicalLocation="0.45053148">LoadDeflection Behavior</physicalSection>
  <physicalSection physicalLocation="0.484006046">LoadDeflection Behavior</physicalSection>
  <physicalSection physicalLocation="0.66163681">Strain in CFRP Composites</physicalSection>
  <physicalSection physicalLocation="0.713471592">Strain in CFRP Composites</physicalSection>
  <uid>WOS:000672278500001</uid>
  <citedAuthor>Al-Ghrery, K</citedAuthor>
  <year>2021</year>
  <page>ARTN 04021009</page>
  <volume>26</volume>
  <citedTitle>Experimental Investigation of Curved-Soffit RC Bridge Girders Strengthened in Flexure Using CFRP Composites</citedTitle>
  <citedWork>JOURNAL OF BRIDGE ENGINEERING</citedWork>
  <doi>10.1061/(ASCE)BE.1943-5592.0001691</doi>
</reference>
```

In the example above, the item is cited 11 times total in this paper. You can see that it is cited multiple times in certain sections, which can be differentiated by the different physicalLocation values. For instance a physicalLocation value of .333 would indicate that a reference was made roughly one third of the way into a paper.

# Appendix - Helpful Links

---

Older versions of this document contained full lists of various things like categories, document types and address abbreviations. In the interest of having the most up to date data I have included here links to a few helpful WOS help pages that should be kept up to date and are less likely to grow stale than a document.

Primary WOS help information can be found here:

<http://webofscience.help.clarivate.com/en-us/Content/home.htm>

In particular here are some useful specific pages in relation to what is outlined above:

- WOS Categories - <http://webofscience.help.clarivate.com/en-us/Content/wos-core-collection/wos-core-collection.htm> (the Subject Categories section at the end)
- WOS Research Areas – <http://webofscience.help.clarivate.com/en-us/Content/research-areas.html>
- Document Types - <http://webofscience.help.clarivate.com/en-us/Content/document-types.html>
- Open Access Status Information - <http://webofscience.help.clarivate.com/en-us/Content/open-access.html>
- Address Abbreviations – <http://webofscience.help.clarivate.com/en-us/Content/address-abbreviations.html>

If you have access to Web of Science platform, you can search unified org/org enhanced values by selecting the “Affiliations” hyperlink beside the OG search tag in advanced search.

In addition, you can find more detailed information on journals, journal coverage, categories etc by access the MJL page located here: <https://mjl.clarivate.com/home>.