UofT - WoS database structure

`publication` table

Each row represents a publication. The `id` field (primary key) is the 15-character alphanumeric unique WoS identifier. `edition` is a code for the WoS edition the record is from (some records have multiple edition codes in the raw XML; UofT-WoS indicates only the first). `Source_id` (foreign key) is the internal ID of the publication's source (e.g. journal). `type` is one of *journal*,*book*, or *book in series*. The other fields are self-explanatory. Note that `page_begin` and `page_end` are of text type rather than numeric, to allow for page identifiers including letters; the `page_count` field however is an integer.

`source` table

A *source* represents a journal or a book. It cannot be assumed that a journal is represented by a single entry in this table. For example, *Science* is represented by six different entries that differ in their `publisher_id` (see below for the reason for the multiplicity in publisher ids).

`publisher` table

The table includes the publisher's `full_name` and `unified_name` (the latter is a more concise version of the name provided by Clarivate), `address` and city (the latter extracted from the full address for convenience). It cannot be assumed that a publisher is represented by a single entry in the table. That is due to the fact that the publisher information is not given in a consistent way in records in the WoS raw XML. For example, the *American Association for the Advancement of Science* (AAAS) is represented by six different entries that differ slightly in (full) name

("AMER ASSOC ADVAN SCIENCE" vs. "AMER ASSOC ADVANCEMENT SCIENCE") or a slightly differently-written address (or missing address in one instance). The unified name does not always help resolving this.

`author` table

Each row in this table is an author on a single publication (that is represented in the `wos_id` field). An entry does not represent a unique individual. Author names do not appear in a

consistent way in the WoS data: surnames may have only the first letter capitalized or be fully capitalized,the given name may appear in full or just the first initial (in the latter case, it may or may not be followed by a period), a middle name may appear in full, as first initial, or not at all. To make sure that a search captures all (or at least almost all) variations, one should use case insensitive wildcard search. For example, to find all papers by Albert Einstein in the database, this query could be made:

```
SELECT * FROM author WHERE full_name ILIKE 'einstein, a%';
```

This will however capture many entries associated with publications where an author had a similar name. Additional search criteria should be added to reduce false matches, but there is no recipe to guarantee that search results based on name correspond to a single individual.

The `orcid` field is uniquely associated with an individual person and could be used to track authors across publications. However, as of mid 2022, only about 1.5% of author entries have anORCID, and no publication earlier than 2007 has it. Cf. `contributor` table.

`address` table

Each row is a unique address; an author entry can have multiple addresses and they are connected through the `author_address` bridge table. The `address` field in this table is the full address(including the name of the institution, mostly employing abbreviations such as *Univ* and *Hosp*),while `street`, `city`, `admin_div` (administrative division), and `country` are extracted from the full address for convenience (some may be empty).

The information in this table has reliability issues (often due to errors in the publication). For example:

```
Univ Hlth Network, Div Cardiol, Toronto, ON, Australia
Univ Toronto, Mt Sinai Hosp, Dept Med, Toronto, CA USA
Hosp Sick Children, Toronto, Germany
McMaster Univ, Dept Med, Toronto, AL USA
```

`contributor` table

While the `author` table contains the information of authors as provided in the publication, thistable has the names of authors that are associated with the publication through some other means.For example, individuals with Publons or ORCID accounts may claim to be authors on a certain paper by adding it to their online profile. Thus it is possible that an `author` entry will not

contain an ORCID but that the corresponding ORCID could be found by examining the `contributor` entries for the same publication.

The information in this table has reliability issues. The name is often missing or malformed, and we often see individuals listed as contributors on a publication where they are clearly not an author, but one of the real authors has a similar name.

`identifier` table

Each row is a single identifier (such as DOI) associated with a publication. A single publication may have several identifiers, and some identifiers (such as ISSN) are not unique to a publication and may represent the journal or the series of the publication. To see all possible identifiers(currently there are 11), run the special command `\dT+ idtype` in psql.

`descriptor` table

Each row is a unique descriptor; a publication can have multiple descriptors and they are connected through the `publication_descriptor` bridge table. The `type` field is an enumerated data type and can have the following values: `doctype`, `language`, `heading`, `subheading`, `subject`, `subject_ext`, `keyword`, and `kw_plus`.

The WoS XML records have one or more *subjects*, each subject has an `ascatype` attribute that can be `traditional` or `extended`. Traditional subjects become a `subject`-type descriptor in UofT-WoS, while extended subjects become a `subject_ext`-type descriptor.

KeyWords Plus become `kw_plus`-type descriptors. These are generated by Clarivate:

> *The data in KeyWords Plus are words or phrases that frequently appear in the titles of an article's references, but do not appear in the title of the article itself.*

`conference` table

Each row is a unique conference; a publication entry can have multiple conferences and they are connected through the `publication_conference` bridge table. Location information is given in the `city`, `admin_div`, and `country` fields. `admin_div` in this case refers only to US states and territories (occasionally misspelled such as KT instead of KY for Kentucky).

`grant` table(s)

Currently grant data is found in three separate tables: `grant_agency`, `grant_data` and `grant_pi`. Note that as most of this data is derived from third party sources (see `funding` table below), there are often duplicates. The data has not been normalized, and the quality differs depending on the origin source - some data sources may be more robust than others. For example, basic information may be available from MedLine, with a more robust set of the same information coming from UKRI.

`grant_data` contains the bulk of the information about a particular grant, such as the `id`, `project_title`, `amount` given, `currency`, and `date_start` and `date_end` of the funding. `source` provides the third party data source that the information was derived from (for example, UKRI), and `pi_institution` gives the organization name of the first Principal Investigator listed. Granting agency information is included in a separate table, `grant_agency`, which includes `agency` and `pref`. A `pref` value of “Y” corresponds to the unified version of a funding agency, which is not present in all cases. Note that as all organizational levels represented in the XML have been preserved, one grant may have many agencies. For example, there would be two separate entries in the `grant_agency` table for a unique grant ID, if both agencies below were present in the raw XML:

```
UK Research & Innovation (UKRI)
Medical Research Council UK (MRC)
```

`grant_pi` contains the `grant_id` and `name` of a Principal Investigator (PI) on the grant. Most grants only have one or two PIs.

The `publication_grant` table is a bridging table between the grant and publication tables, and contains both a `grant_id` and corresponding `wos_id`.

`reference` table(s)

As of the 2023 update, the `reference_unindexed`,and `reference_patent` tables no longer exist. Instead there is just one `reference` table that includes data for all references. References that are indexed will have `cited_id` values starting with `"WOS:"` followed by the UID (SQL string operations could be used to strip the prefix and get the actual UID), and patents will have a non-NULL value in the `patent_no` field.

The `reference_context` table contains, for each reference in a paper, the `location` in the paper of each of its occurrences, in the form of a number between 0 and 1, and the title of the `section`.

`openaccess` table

Each row now contains a publication identifier and open access status, each publication may appear more than once with different or even the same open access status, as is the case in the rawXML data we receive from Clarivate, whether it makes sense or not…

Other tables

Other tables with self-explanatory names and fields include: `abstract`, `funding`, and `conference_sponsor`.

Note `funding` contains string data on funding data that was included as part of the original XML. In some cases, this has been parsed by Clarivate to extract basic grant information, which is included in the `grant` tables listed above. In those cases, the value of the `source` field in `grant_data` will be `null`.