Corpus Building: WorldCat, Part 2

Betti, A.

Publication date
2020
Document Version
Final published version

Corpus Building: WorldCat, Part 2

June 6, 2020 Uncategorized

Author: Arianna Betti

This is Part 2. Go to Part 1.

Late May and before June 4, 2020. *Last changes: Feb 19, 2022 (readability improved)*

WorldCat's record identity and relatedness criteria

WorldCat clusters records of the *same edition* of the same work, and links records of *different editions* of the same work: this is evident from the fact that (a) search results do not show all WorldCat records for the same edition, that is, some are clustered; (b) some records come with a link to 'other editions and languages'; (c) individual record pages come with library holdings both for that record and for (some) other editions of the item described by the record.

The clustering criteria for (a), (b) and (c) are however mysterious: when does WorldCat cluster records (or when does it merely link them), and when not? One noticeable thing is that WorldCat tends to see difference where expert eyes see similarity and identity. WorldCat's conservatism is understandable: WorldCat is an aggregated catalogue; an unsupervised algorithm should not cluster too much when aggregating highly complex long data from different sources (even when the sources use the same data model), on pains of mistakes, and of loss of potentially relevant information. In some cases, besides, clustering and identification can only be achieved by consensus among experts of several different disciplines, as we saw in the Arabic and Italian case in Part 1.

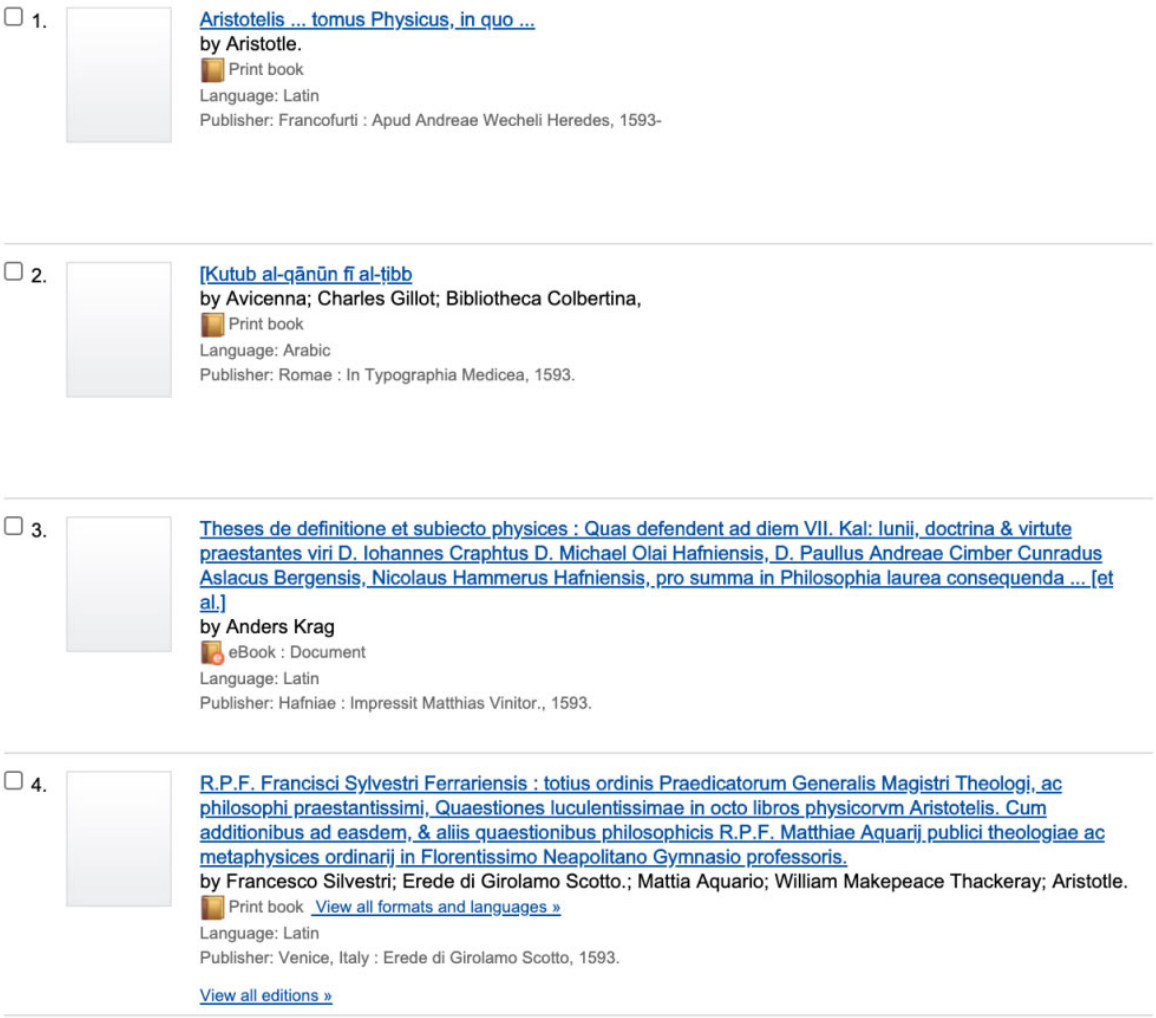
Case study: Avicenna's 1593 *Canon*

To discover some WorldCat identity and relatedness principles, we take the case of the 1593 Arabic edition of Avicenna's *Canon* described in Part 1. We take this case for two reasons (1) it comes with great, though manageable, linguistic and script variation in title, author, place of publication and publisher fields; but (2) it comes in only one edition in a specific year printed in a specific city by a specific press, albeit in two different variants.

1. **A mystery**

A search for books on physics from 1593 in Latin (done late May and before June 4, 2020)only gives us 8 items in total: surprisingly, the list does *not* include the Erfurt record indicating ln=Latin (which as we saw it's at best incomplete), but it does show the Toronto record indicating ln=Arabic (hit n. 2 in the screenshot below), exactly like it happened in the original search we did for books in Latin on physics in 1500-1599 in Part 1. Is

this result due to the presence in WorldCat of the Erfurt Latin record? How is that possible? Next to the Avicenna record below (second hit), there is no button 'View all formats and languages':



Hit n. 2 would be understandable if the Arabic and the Latin records were clustered via 'other editions and languages', or 'Other formats and editions' but they are not. The Toronto Arabic record #670051662 links as 'other editions and languages' only one other Arabic record (#604016614, from UB Basel). So far, it remains a mystery why a search for Latin books in physics in 1593 does not catch #257644251 (in Latin) but does catch #670051662 (in Arabic, correctly linked it to #604016614).

2. All the *Canons* in WorldCat

How many other records of this edition exist in WorldCat? A new search for books in physics from 1593 with **Kutub al-qānūn** as title words returns twelve records: 11 Arabic, and 1 Latin. The Latin is #257644251 Erfurt. That's just a part of the records, as we will see. The 12 records are:

Paris-Mazarine #799697155
BnF #575925056
BNE #943676706
BnF #575925040
Firenze IRIS Uffizi #908533420 +
Basel #604016614 (no library found in WorldCat for this record?)
Paris, Histoire Naturelle #30947261
Indiana #40695103
British Library #917369289 +
Erfurt #257644251 (no library found in WorldCat for this record?)

Strasbourg-BNU #492831405
Nichibunken #1020979776

Of these, only British Library #917369289 and Firenze IRIS Uffizi #908533420 link 'Other editions and languages'. But we know Toronto *is* linked to Basel – indeed, Toronto does not show up in the list, and Basel does.

Note also that all of these records describe the same edition of the work, but they aren't *all* the records in WorldCat that describe that edition (it's 69 at least). How can we get all other records? First, we save the 12 records to a list (this one). Then we observe that the search results also facet four authors, Avicenna (6 records), Avicenne (3 records), Avicena (1) and 980-1037 Ibn Sīnā (1). But these are all different names for only one author, in different language variants (Italian/Latin, French, Spanish and Arabic – but not his formal Arabic name, which was Abū ʿAlī al-Ḥusayn ibn ʿAbdillāh ibn al-Ḥasan ibn ʿAlī ibn Sīnā). This example reminds us of the Arabic/Latin puzzle we encountered in Part 1. so here's a corollary to our general tenet of record information inheritance:

Corollary 2 to Worldcat Tenet 1 If, in a search including an author field, a record with two authors in the relevant field is among those returned, then WorldCat displays the additional author via the author facets. This applies also to records that have only one name in the author field but link that name to unique identifiers in ISNI, VIAF and the like.

Corollary 2 is confirmed by inspecting records that link the following pairs via unique identifiers:

1. Avicenna | Avicenne;
2. Avicena | Avicenna;
3. Avicenna | 980-1037 Ibn Sīnā

The *first pair* is linked in two BnF records via an authority record ('Notice de personne') from field 700 of their UNIMARC record format; the UNIMARC authority record, in turn, links to ISNI, which provides unique identifiers for authors of creative work. In ISNI, Avicenna is 121430876. A similar database is VIAF (Virtual International Authority File), a database associating the same unique identifier to all the alternative names for the same author. In VIAF, Avicenna is https://viaf.org/viaf/89770781/ – a number associated with many, many alternative names. The *second* pair is linked to the BNE record via an authority record linking to VIAF, ISNI and Wikidata. The *third* pair comes from a Japanese record from Kyoto, linking to VIAF, NDK and CiNii. (This paragraph owes a lot to the Wikipedia entry on Avicenna, which records unique authority identifiers from a number of systems.)

Our hypothesis at this point is that the WorldCat interface does not directly cluster records via author fields via ISNI or VIAF: it only preserves the links to authority files contained in the incoming library records, if any such links are present. To check this, we find a Spanish record with 'Avicena' as author and with title 'Libri quinque Canonis medicinae' (Madrid Complutense #1024892954), such that we get this record as a hit if we search for 'Avicena', 1593 and 'Libri quinque Canonis medicinae', but we do *not* get it as a hit if we search for 'Avicenna', 1593 and 'Libri quinque Canonis medicinae' – which demonstrates the point that the link between pairs such as Avicena|Avicenna is not established by WorldCat, but by the records themselves: if any link is present, WorldCat somehow retains it.

At this point, the challenge of catching all the Canon records by going via the author field in WorldCat seems daunting (especially when knowing we can only save ten records at the time). But still: surely a search for 'Avicenna' and year 1593 should get us only a manageable amount of records? We should get some variation, given different titles, but that's it. Indeed, we get 42 hits for that search – which is still quite some considering

that we know it's 42 ways to described the same edition, and that this is only a search for one variant of an author's name.

To get as many relevant records as we can, we also search for

1. 1593 and kw:'Typographia Medicea';
2. 1593 and kw:'Medicea';
3. 1593 and kw:'Typographia'.

By doing so (and filtering using the author facet in the third search) and some more digging we arrive at 69 records (and we catch some mistaken records, such as #313549609, which is shared by the Württembergische Landesbibliothek and UB Freiburg, and has an error in the publisher field – 'Mediosa' instead of 'Medicea'.)

There is no reason to think that we have by now caught all the records of the *Canon*.

The 69 records (and counting)

Every record among these 69 has a different title. No search would get all of them at once – not without knowing them already. Of these 69, some are clustered, some share the same OCLC number.

What we discovered so far suggests:

Worldcat Tenet 2. If two records are clustered, then they have identical title field. They do not need to have the same author field.

Worldcat Tenet 3. If two records describe the same edition and share the same OCLC number then they have identical title field *and* are on the same material support (i.e. both books, both microfiche, etc)

Tenet 3 is *really* fine-grained: in WorldCat we also find distinct records with the same author field, same title field, same material support, same year, same publication field, *but different OCLC number*. See also **Enigma 2** below. (Note that the other 'if…then' direction of **Tenet 3** does not hold, that is, it does not hold that if two records have identical title field *and* are on the same material support then they share the same OCLC number).

Example for **Tenet 2** and **3**: Record #908533420 is clustered with other five records with the same title field (but different author fields). In this cluster, #123412029 is shared by the IRIS consortium record of a micro-form edition held by Firenze/Harvard-I Tatti-Berenson and the record at Columbia: these two holdings differ only item-wise. (I conjecture this hypothesis: they have the same reproduction ID.) The other three clustered records are of printed books: one record is from Firenze-Galleria degli Uffizi (908533420), one from Marburg (#608421254), one from Zurich (#637218590).

Enigma 1: The IRIS Consortium holding #908533420 is an aggregator of certain Firenze libraries and contains two records for two items of our work: one for the microfiche held by the Berenson Library at Villa I Tatti (#123412029), and one for the printed book at the Galleria degli Uffizi library (#908533420). But the Berenson Library is part of the Harvard library system and its record for this item are also in WorldCat, under another OCLC number (#1144485765). This creates an odd duplication since we have two different OCLC numbers for exactly the same physical item: a microfiche held at Villa I Tatti in Florence (not a bad place for microfiches and humans alike – as an aside, the Villa has the same garden designer as Villa La Foce in Val d'Orcia, Cecil Pinsent). But if the IRIS Consortium Harvard record gets the same OCLC number as Columbia, why does the

Harvard record get a new number? After all, it's the same work, same edition, same support: just two different items.

Enigma 2. The record of the item of the Galleria degli Uffizi is in WorldCat as #908533420, brought in by in IRIS, but in a search for the exact title of that record WorldCat tells us that 'no library has the item'. Odd. (This is not the only case of 'no library hit' for a record.)

The 750 Latin records: how many link a full text scan?

Back to the 750 records in Latin (as of June 6, 2020). In the small experiment made in the previous post on our three Arabic, Greek and Italian records we were able to find rather quickly a digital copy of high quality for each of our three items in academic repositories. How about then the books in Latin we are potentially interested in? How many of these link a digital copy?

That's for the next post.