# Corpus Building: WorldCat, Part 1

Betti, A.

**Publication date**
2020

**Document Version**
Final published version

Link to publication

**Citation for published version (APA):**
Betti, A. (Author). (2020). Corpus Building: WorldCat, Part 1. Web publication or website, quine1960. https://quine1960.wordpress.com/2020/05/28/corpus-building-worldcat-part-1/

quine1960

*A teaching blog on Quine's W&O (1960) | 4-6 §§/week sep-dec every year | taught by @ariannabetti to 2nd year students at @UvA_Amsterdam | 1960 pop hits & art | illustration by Floris Solleveld*

# Corpus Building: WorldCat, Part 1

May 28, 2020 Library stuff

**Author: Arianna Betti**

*Last changes: February 18, 2022 (minor changes)*

Next: Corpus Building: World Cat, Part 2

Suppose you want to put together a corpus of 16th century writings, in particular textbooks, on physics, in Latin. Here's one method I will call *Pseudo-Random Digital Corpus First*. The Google Books variant of the method is to go to Google Books and search their database by using keywords that you know appear in the title of the type of books you are interested in. This method will probably give you some good hits after some filtering of irrelevant results, but your corpus will be skewed towards works that happen to have been digitized by Google, plus you don't know what you are missing by typing stuff that just happens to come to your (expert) mind. The method is 'pseudo-random' because the choice of texts is not random – it just seems so. And it is 'corpus first' because you don't start from metadata, you immediately gather full-text material. This is not necessarily a problem if all you want is to use some material digitized by Google without following any principle of well-grounded corpus building. If however you are interested in proper corpus building, here's a second method. I will call that *Universal Bibliography First.*

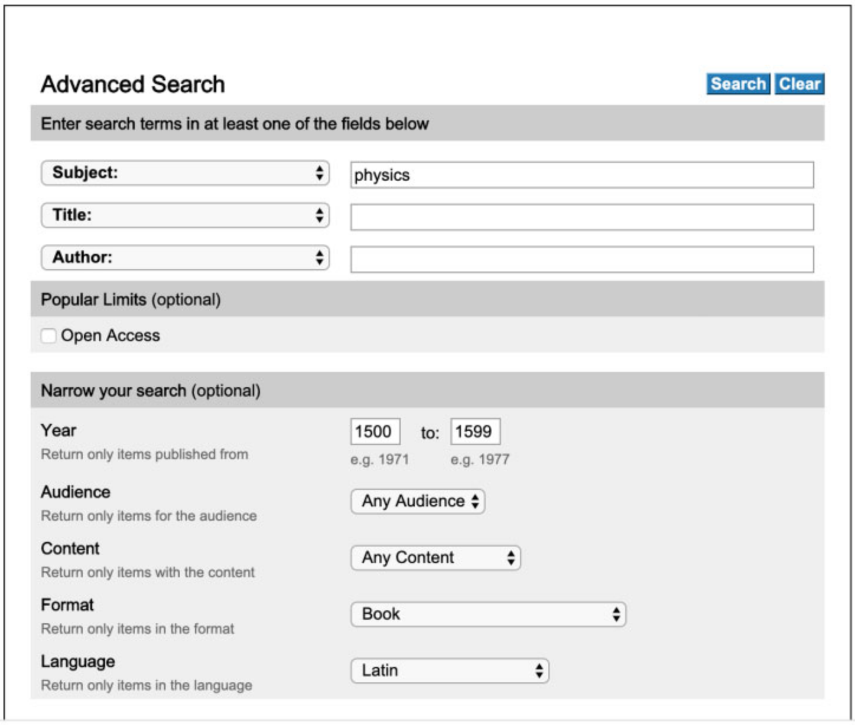
## Universal Bibliography First

WorldCat is the biggest bibliographic catalogue on earth. It has plenty of mistakes, for a number of understandable reasons, but it is still the best resource you can turn to if you are after comprehensiveness (or *recall*): getting all possibly relevant data, also data that might be noisy and difficult to gather.

Why not using the Universal Short Title Catalogue (USTC), at least for early prints, you might ask? The short answer is that USTC's subject search is, at a first glance, way too rough: try to search for *Logic* as a subject. I might do a comparison in a later post.

The data we are after is *book metadata*: bibliographic records of 16th century textbooks on physics written in Latin. WorldCat has an advanced search option that you can use to take the first step:

WorldCat®



Advanced Search

Enter search terms in at least one of the fields below

Subject: physics

Title:

Author:

Popular Limits (optional)

□ Open Access

Narrow your search (optional)

Year: 1500 to: 1599
Return only items published from
e.g. 1971 e.g. 1977

Audience: Any Audience
Return only items for the audience

Content: Any Content
Return only items with the content

Format: Book
Return only items in the format

Language: Latin
Return only items in the language

We hit enter on this search, and we get 755 results (*added on Feb 18, 2022: it is now 833*). Note that we have restricted the search to 'physics' by using the subject field of records (su:physics), *not* by using free keywords. Note that WorldCat inherits the subject classification from original library records, so here's two warnings:

**Warning 1**: If the original library record does not have 'physics' in the subject field, or it does not link in any way to a 'physics' subject heading, the record won't show up among WorldCat results for su:physics.

**Warning 2**: If the original library record has 'physics' in the subject field, but in a different language than English, the record might not show up among WorldCat results for su:physics.

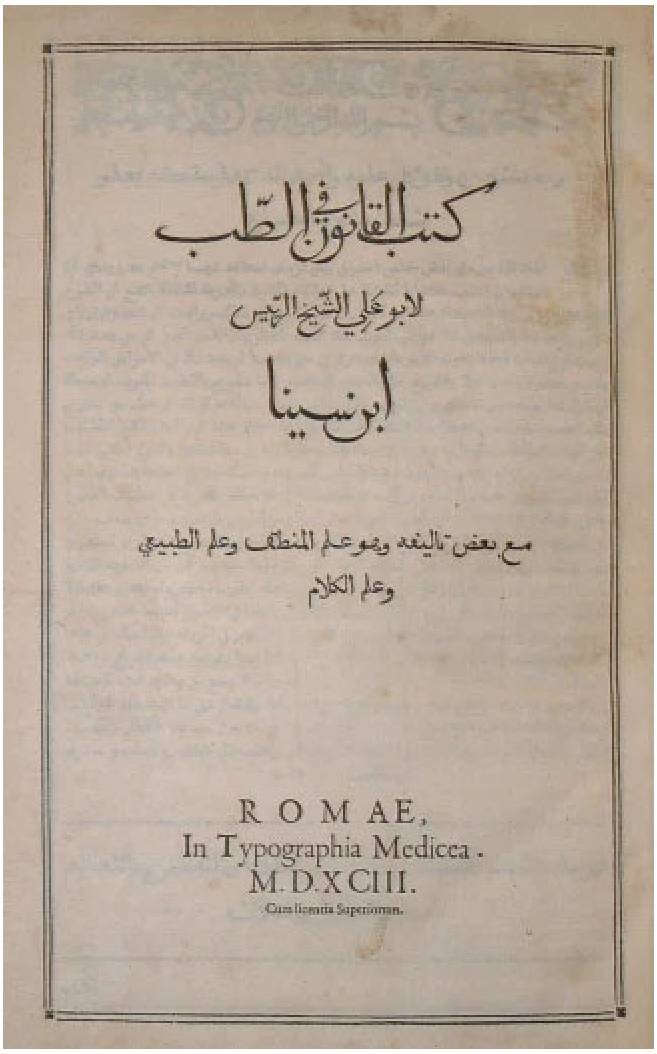
Despite having gotten our 755 results by searching only for Latin, the language facet on the landing page indicates 741 records in Latin, but also in four other languages: Italian (1), Arabic (1), Ancient Greek (11) and English (1). (Note: the same search repeated a week later, June 3, 2020, gives 750 Latin hits, as ingestion of new records in WorldCat proceeds rather fast.) Here below I go over the details of three records – Arabic, Greek and Italian. Let's first however see one problem, and how to pragmatically proceed to corpus building without solving that problem and all other problems first.

**Problem 1**: You cannot save all of these 755 results in one go. You can only make an account, create a list, and save 10 records at a time. You can fortunately export your list in csv.

**Quick and dirty corpus building procedure**: The results of the search we did has hits from about 100 different authors. Given **Problem 1**, your best bet is, in a query like this one, to go author by author using the author facet: the hits contain actually quite a number of duplicates, plus, as you will see below, the author list contains editors and printers, so by hundred clicks, and saving each time the selection of your interest from the results to your list, you'll be able to have a collection of metadata for your corpus in a reasonably manageable manner. When you are happy with your list, you can export it in csv (comma separated value) format and proceed further.

**1. Arabic**

The record in Arabic has #670051662 as its unique OCLC number and it comes from Toronto University libraries. It describes a truly remarkable volume: the 1593 edition of the Canon of Medicine (*al-Qānūn fī aṭ-Ṭibb*), a classic from the history of medicine written by the Arabic medieval philosopher Ibn Sina (Avicenna), and elegantly printed in Rome. This was, according to Britannica the first Arabic edition in the West: why do we get this record by searching for Latin texts?



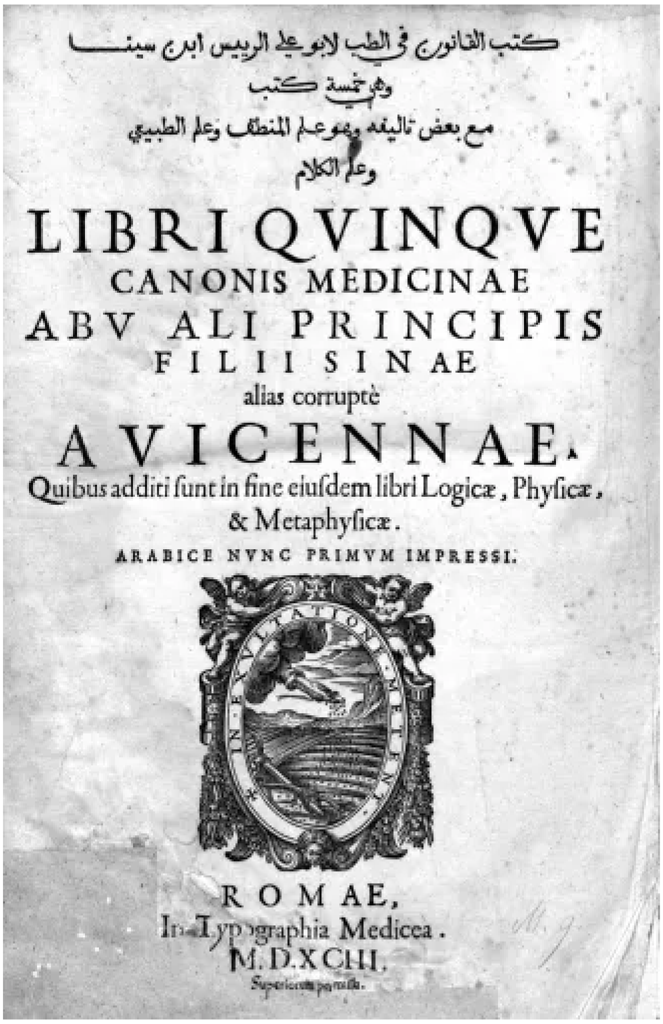
كتب القانون في الطب
لابو علي الشيخ الرئيس
ابن سينا
مع بعض تاليفه وهو علم المنطق وعلم الطبيعي
وعلم الكلام

ROMAE,
In Typographia Medicea.
M.D.XCIII.
Cum licentia Superiorum.

WorldCat does not link a digital full-text for this record. So we find a digital copy at the American University in Beirut for inspection outside WorldCat (by googling ‘Kutub al-qānūn fī ăṭ-ṭibb’, landing on the Wikipedia page of the publishing house, Typographia Medicea, inspecting the Wikipedia page, and following, from there, a link to the digital copy). Page by page inspection of the over thousand pages of the book reveals that the text is only in Arabic. So where does the Latin classification comes from? In WorldCat we do find another record (#257644251) with a longer title, part in Arabic, part in Latin, partially overlapping with the title of #670051662:

*Kutub al-qānūn fi ‘ṭ-ṭibb li Abū ‘Alī ar-ra’īs Ibn Sīnā wa-hiya ẖamsat kutub : ma‘a ba‘ḍ ta’līfihī wa-huwa ‘ilm al-manṭiq wa-‘ilm aṭ-ṭabī‘ī wa-‘ilm al-kalām = Libri Qvinqve Canonis Medicinae Abv Ali Principis Filii Sinae alias corruptè Avicennae : Quibus additi sunt in fine eiusdem libri Logicæ, Physicæ & Metaphysicæ*

This record indicates the item’s language is Latin. After some – again, independent – research we find that #257644251 is a record coming from Erfurt University Library. No digital copy linked. After some more surfing, we find this image:



كتب القانون في الطب لابو علي الرييس ابن سينا
وهو خمسة كتب
مع بعض تاليفه وهو علم المنطق وعلم الطبيعي
وعلم الكلام

LIBRI QVINQVE
CANONIS MEDICINAE
ABV ALI PRINCIPIS
FILII SINAE
alias corruptè
AVICENNAE.
Quibus additi ſunt in fine eiuſdem libri Logicæ, Phyſicæ, & Metaphyſicæ.
ARABICE NVNC PRIMVM IMPRESSI.

ROMAE,
In Typographia Medicea.
M.D.XCIII.
Superiorum permiſſu.

This title page in Latin corresponds exactly to the title of the Erfurt record (#257644251). There is more: Google books has a scan in which this Latin title page precedes the Arabic title page: the Beirut scan lacks the Latin page title. Is the Beirut scan incomplete? Was the page cut from the Beirut exemplar for some reason?

Intricate research spanning several days involving a counterreformation Pope, Christian Arabs and the *dhimmī* communities of Eastern Christianity, a famous Sixteenth-century type designer, the *Frankfurter Buchmesse*, Roman cardinals, European doctors, medieval medicine, and a book theft, helps us fix the following. There were *two* impressions of this edition, one with the Latin title page and one without, because of different markets, Oriental and European. (For the most important sources consulted in this investigation about the typographic and commercial aspects of the 1593 *Canon* edition, see **Sources** below).

A professional librarian would have been able to get to this, I guess, a lot quicker than I did, since the Erfurt record carries also this cryptic, though (it turns out) helpful, info:

Adams, Catalogue 1501-1600, A-2322
EDIT 16 CNCE 3554

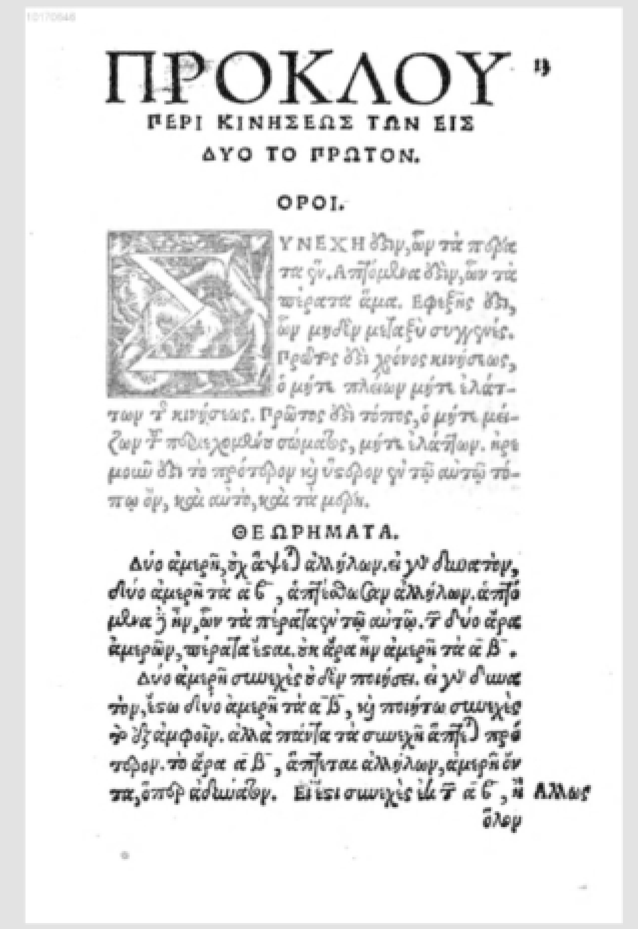
After some more research we discover that "Adams," […] is "*The Catalogue of Books Printed on the Continent of Europe, 1501-1600, in Cambridge Libraries*, compiled by H.M. Adams." (see this post); 'EDIT 16' stands for *Censimento nazionale delle edizioni italiane del XVI secolo*, and CNCE 3554 is the record of Avicenna's 1593 *Canon* in the EDIT 16 database: some more info (titles images visualization) show two variants for the title, A (with the Latin title) and B (with Arabic title). (There is another such reference work for the 16th century: *Index Aureliensis: catalogus librorum sedecimo saeculo impressorum,* normally noted in library records as 'Ind. Aur.' – the *Canon* is in vol. II, entry 453).

The Erfurt Latin record is obviously describing an item of the variant A, and it is either mistaken (because after all, the book is fully in Arabic), or it is incomplete. To say it's merely *incomplete*, and not mistaken, we need two more assumptions: (i) the assumption that (some?) libraries, when a text is in a certain language contains even a small portion of another language (even if the first language takes up more than thousand pages and second language takes up one sentence) record both languages; (ii) the assumption that the language field of the Erfurt record indicates both languages, despite both WorldCat and Erfurt library interfaces showing only one. It's a bit hard to check these assumptions, especially getting to know – by surfing – how the interface of the Erfurt library actually works.

**Corollary 1 to Worldcat Tenet 1** If, in a search, a record with two languages in the relevant field is among those returned, then WorldCat displays the additional languages via the language facets.

**2. Greek**

Among the records of books indicated as being in Greek we find this one: *Procli insignis philosophi Compendiaria de motu disputatio,: posteriores quinque Aristotelis De auscultatione naturali libros, mira brevitate complectens.* Per Io. Bebelium, et Mich. Ysingrinium, 1531 (#79189104). Where's the Greek? We check the record page of the record for more info in the hope to get, ideally a link to a digital copy. The links to the English libraries appear broken, while the French and German libraries do not link to any scan. So, again, we google the title, and we find a scan at the Bayern State Library, peruse it, and we establish clearly what the work actually is: a 1531 edition of Proclus' *Elements of Physics*, indeed in Greek:



ΠΡΟΚΛΟΥ

ΠΕΡΙ ΚΙΝΗΣΕΩΣ ΤΩΝ ΕΙΣ
ΔΥΟ ΤΟ ΠΡΩΤΟΝ.

ΟΡΟΙ.

ΣΥΝΕΧΗ ἐστιν, ὧν τὰ πέρατα ἕν. Ἁπτόμενα ἐστιν, ὧν τὰ πέρατα ἅμα. Ἐφεξῆς ἐστιν, ὧν μηδὲν μεταξὺ συγγενές. Πρῶτος ἐστὶ χρόνος κινήσεως, ὁ μήτε πλείων μήτε ἐλάττων τῆς κινήσεως. Πρῶτος ἐστὶ τόπος, ὁ μήτε μείζων τοῦ περιεχομένου σώματος, μήτε ἐλάττων. ἠρεμοῦν ἐστι τὸ πρότερον καὶ ὕστερον ἐν τῷ αὐτῷ τόπῳ ὄν, καὶ αὐτὸ, καὶ τὰ μέρη.

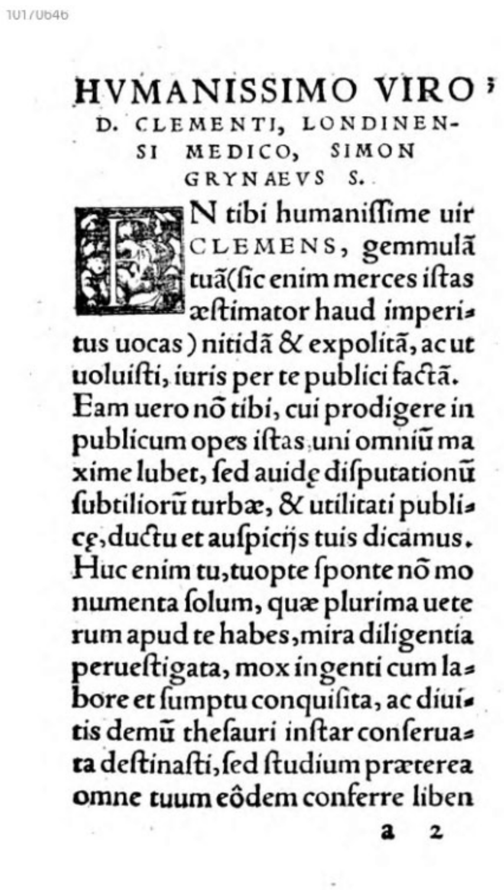
ΘΕΩΡΗΜΑΤΑ.

Δύο ἀμερῆ, οὐχ ἅψεται ἀλλήλων. εἰ γὰρ δυνατὸν, δύο ἀμερῆ τὰ α β, ἁπτέσθωσαν ἀλλήλων. ἁπτόμενα δὲ ἦν, ὧν τὰ πέρατα ἐν τῷ αὐτῷ. τῶν δύο ἄρα ἀμερῶν, πέρατα ἔσται. οὐκ ἄρα ἦν ἀμερῆ τὰ α β.

Δύο ἀμερῆ συνεχὲς οὐδὲν ποιήσει. εἰ γὰρ δυνατὸν, ἔστω δύο ἀμερῆ τὰ α β, καὶ ποιείτω συνεχὲς τὸ ἐξ ἀμφοῖν. ἀλλὰ πάντα τὰ συνεχῆ ἅπτεται πρότερον. τὰ ἄρα α β, ἅπτεται ἀλλήλων, ἀμερῆ ὄντα, ὅπερ ἀδύνατον. Εἰ ἐστι συνεχὲς τὸ τῶν α β, ἢ Ἄλλως

ὅλον

But with a preface by Simon Grynaeus in Latin:

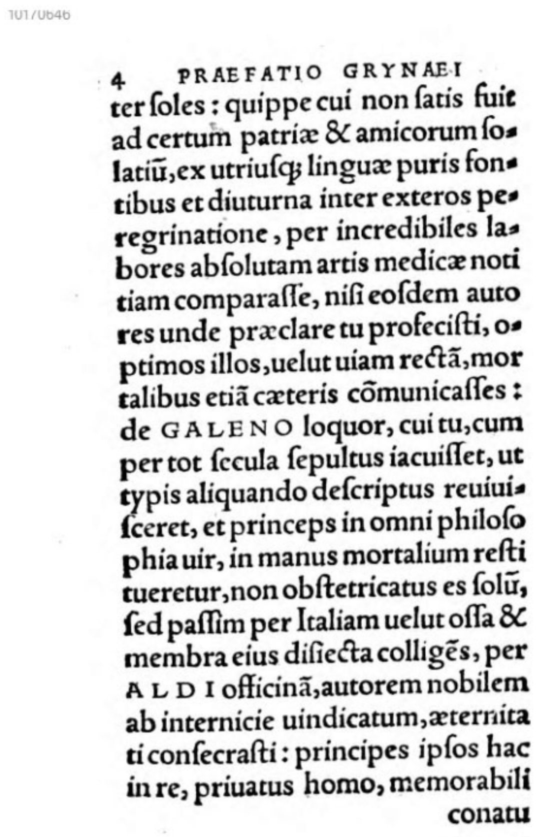


101/0646

HVMANISSIMO VIRO
D. CLEMENTI, LONDINENSI MEDICO, SIMON GRYNAEVS S.

EN tibi humaniſſime uir CLEMENS, gemmulã tuã(ſic enim merces iſtas æſtimator haud imperitus uocas) nitidã & expolitã, ac ut uoluiſti, iuris per te publici factã. Eam uero nõ tibi, cui prodigere in publicum opes iſtas uni omniũ maxime lubet, ſed auidę diſputationũ ſubtiliorũ turbæ, & utilitati publicę, ductu et auſpicijs tuis dicamus. Huc enim tu, tuopte ſponte nõ monumenta ſolum, quæ plurima ueterum apud te habes, mira diligentia perueſtigata, mox ingenti cum labore et ſumptu conquiſita, ac diuitis demũ theſauri inſtar conſeruata deſtinaſti, ſed ſtudium præterea omne tuum eôdem conferre liben

a 2



101/0646

4 PRAEFATIO GRYNAEI

ter ſoles: quippe cui non ſatis fuit ad certum patriæ & amicorum ſolatiũ, ex utriuſq; linguæ puris fontibus et diuturna inter exteros peregrinatione, per incredibiles labores abſolutam artis medicæ notitiam comparaſſe, niſi eoſdem autores unde præclare tu profeciſti, optimos illos, uelut uiam rectã, mortalibus etiã cæteris cõmunicaſſes: de GALENO loquor, cui tu, cum per tot ſecula ſepultus iacuiſſet, ut typis aliquando deſcriptus reuiuiſceret, et princeps in omni philoſophia uir, in manus mortalium reſtitueretur, non obſtetricatus es ſolũ, ſed paſſim per Italiam uelut oſſa & membra eius diſiecta colligẽs, per ALDI officinã, autorem nobilem ab internicie uindicatum, æternitati conſecraſti: principes ipſos hac in re, priuatus homo, memorabili

conatu

So it’s correct: the record should in fact have two languages, Latin and Greek and the indication ‘Latin’ (and only ‘Latin’) is, again, incomplete rather than wrong. Note that, again, only one language appears on the OCLC record page. So by now we know: if both languages are specified in a record, this won’t be visible to users in the record itself, but WorldCat keeps the information and displays it through the language facet.

Note also that while the language facet indicates that 11 records are in Greek, clicking on the facet gets us a list of 29 records, of which 18 in Latin and 11 in Greek (the English facet returns 2, the Italian remains at 1). These discrepancies might be explained by the same type of connection that seems responsible for the connection between the Arabic and the mistaken Latin record of Avicenna’s *Canon of Medicine*: there must a record of the Grynaeus edition of Proclus’ *Elements of Physics* that indicates the language as Latin (or *also* Latin). Indeed: it’s the #600973626 from Basel UB. And something similar must be the case with the other book records that show up as ‘Greek’ among the results of the Latin search. The only difference with the Arabic case is that in the Greek case, the faulty (or better incomplete) Latin record can be found if one retrieves records via the ‘View all Editions and Languages‘ link from #79189104, while the Latin record was not retrieved in this way from any of the Arabic ones.

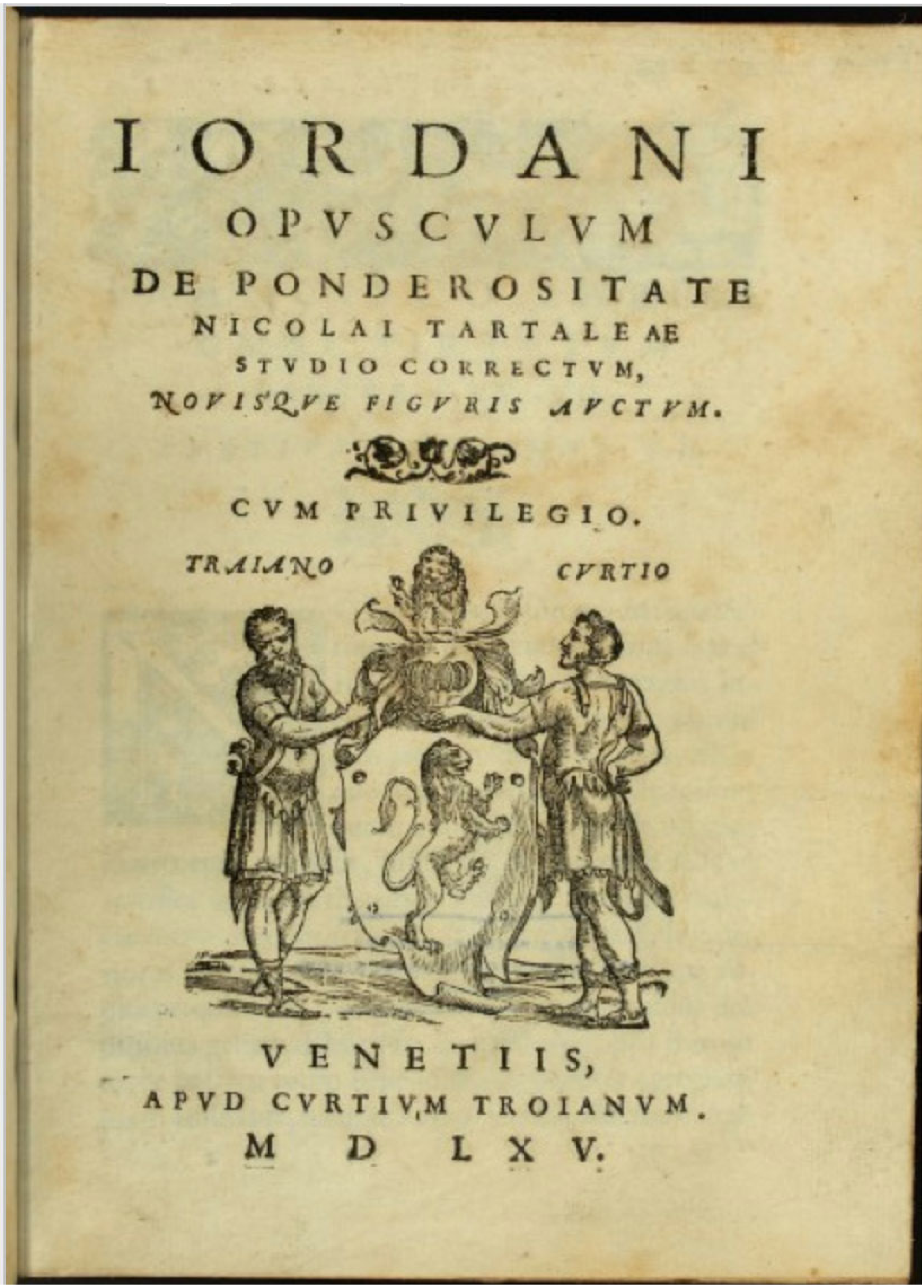
**3. Italian**

Let’s inspect the Italian record: *Iordani opusculum de ponderositate*. Apud Curtium Troianum, Venetiis, 1565, for which three authors are indicated: Nemorarius Jordanus, Niccolò Tartaglia and Curzio Troiano de Navò.

It isn’t easy to understand what this work is. Without special knowledge of Mediaeval and Renaissance history of statics, someone with sophisticated expert knowledge of general history of science and command of several languages needs to spend at least a couple of hours of advanced research on this issue. Here’s what I did. I first tracked down several sources, some online (see **Sources** below). From them I learned that Jordanus de Nemore, a 13th century mathematician, is the author of *De ratione ponderis*, the most important medieval treatise of statics (*scientia de ponderibus*); that the treatise was commented by the 16th century mathematician Niccolò Tartaglia by means on notes on his own a manuscript copy; and that in 1565, after

Tartaglia's death, Curzio Troiano de Navò, a 16th publisher and bookseller operating in Venice who had published other works by Tartaglia, put together the booklet under the name *Iordani opusculum de ponderositate* on the basis of Tartaglia's notes on the manuscript.

Is any part of the *opusculum* in Italian? A quick google search lands a full-text scan at ECHO, a digital repository at the Max Planck institute for the History of Science in Berlin:



IORDANI
OPVSCVLVM
DE PONDEROSITATE
NICOLAI TARTALEAE
STVDIO CORRECTVM,
NOVISQVE FIGVRIS AVCTVM.

CVM PRIVILEGIO.

TRAIANO CVRTIO

VENETIIS,
APVD CVRTIVM TROIANVM.
M D LXV.

The main text is in Latin, but the booklet also contains some pages in Italian reporting experiments made by Tartaglia's ( "Esperienze fatte da Nicolo Tartalea. 1541. a di XIIII. aprile", leaves 20-[23]). The record of the Oxford University Library in WorldCat, for instance, accurately described the book as having two languages: Latin and Italian. So, again, this case is similar to the Arabic and the Greek case: there is at least one record indicating the two languages even if the WorldCat interface shows only one. Again, this is not easy to see by only using the web interface at worldcat.org, as we are shown only one language in a record: we need to inspect the original records at the libraries holding the item, and hope that what the interface of these libraries show, is also the actual information recorded in their database. (You thought that a web interface at the frontend would show all the information there is in the database at the backend? Think twice.)

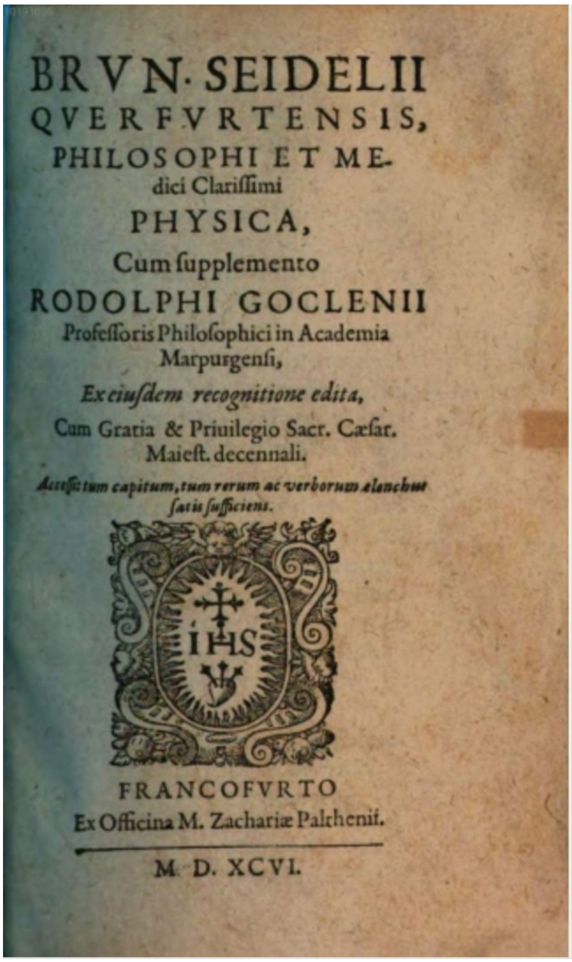
The good news is that it seems that, by extracting bibliographies, high recall of bibliographic information is substantially guaranteed, no matter the quirks, navigation inconveniences and intrasparent choices that the WorldCat interface users meet. One quick way to obviate the most pressing of these challenges on the side of WorldCat interface developers would be allowing saving search results far above the 10 at a time allowed right now, saving links to all related editions and languages for each OCLC number, and offering a more sophisticated search interface to users. Let's see how far we can go in circumventing these obstacles.

**Sources**

- Pisano & Capecchi, *Tartaglia's Science of Weights and Mechanics in the Sixteenth Century: Selections from Quesiti et inventioni diverse: Books VII–VIII,* 2015: 5;
- Iommi Echeverría, Virginia. (2010). La división del aire en los Quesiti et inventione diverse (1546) de Niccolò Tartaglia. *Dynamis*, *30*, 197-212;
- Moody, Ernest Addison, and Marshall Clagett. *The Medieval Science of Weights*. University of Wisconsin Press, 1960.

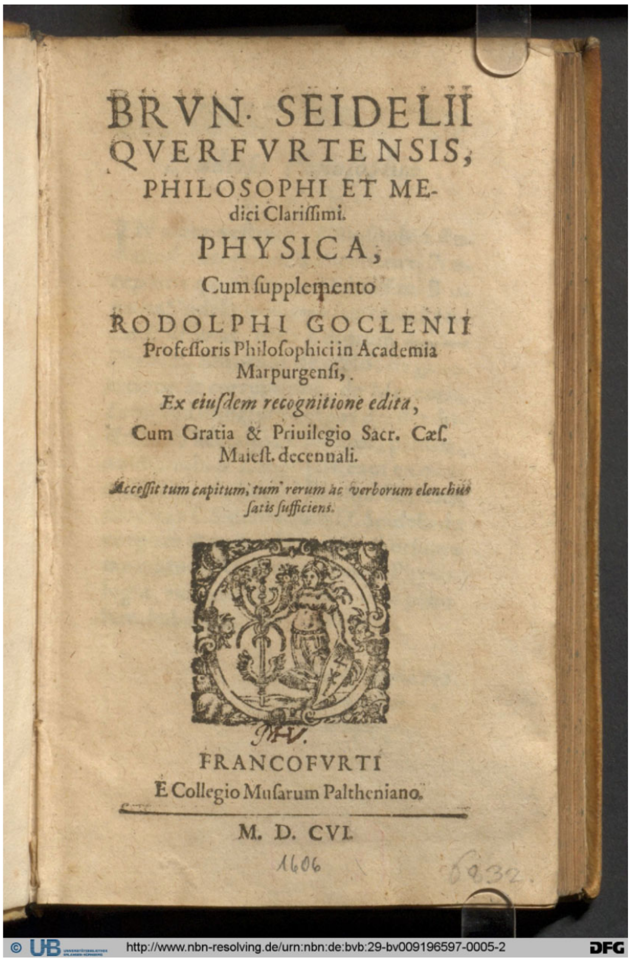
**4. Undetermined**

We saw that the language of some records is either incompletely or wrongly classified, but are there also gaps, cases of records with an empty or undefined language field? A bit by accident, something interesting emerges from looking at the clustered records of Bruno Seidel's *Physica* of which, according to the facets, 7 are in Latin and 3 are *undetermined*. Upon inspection, all the undetermined records turn out to be in Latin (*en passant*, both contain a juicy Latin basic lexicon of natural philosophy). The three records correspond to two editions, both available digitally: the 1596 edition at Bayern State Library and the 1606 edition at the Erlangen-Nürnberg University Library.



BRVN. SEIDELII
QVERFVRTENSIS,
PHILOSOPHI ET MEdici Clariſſimi
PHYSICA,
Cum ſupplemento
RODOLPHI GOCLENII
Profeſſoris Philoſophici in Academia
Marpurgenſi,
*Ex eiuſdem recognitione edita,*
Cum Gratia & Priuilegio Sacr. Cæſar.
Maieſt. decennali.
*Acceſſit tum capitum, tum rerum ac verborum elenchus*
*ſatis ſufficiens.*

FRANCOFVRTO
Ex Officina M. Zachariæ Palthenii.
M. D. XCVI.



BRVN. SEIDELII
QVERFVRTENSIS,
PHILOSOPHI ET MEdici Clariſſimi.
PHYSICA,
Cum ſupplemento
RODOLPHI GOCLENII
Profeſſoris Philoſophici in Academia
Marpurgenſi,
*Ex eiuſdem recognitione edita,*
Cum Gratia & Priuilegio Sacr. Cæſ.
Maieſt. decennali.
*Acceſſit tum capitum, tum rerum ac verborum elenchus*
*ſatis ſufficiens.*

FRANCOFVRTI
E Collegio Muſarum Paltheniano.
M. D. CVI.
1606

http://www.nbn-resolving.de/urn:nbn:de:bvb:29-bv009196597-0005-2

So: is there a way to have WorldCat get us the undetermined records? A look at the URL suggest we might do this (note the *ln:und* in place of *ln:lat*) as follows:

https://www.worldcat.org/search?q=su%3APhysics&dblist=638&fq=yr%3A1500..1599+%3E++%3E+ln%3Aund&qt=facet_ln%3A

This gives us 74 results (*18 feb 2022: 111*), but most of these are records of images of artefacts such as jewels or artworks; similar items are recorded by the 5 'computer files' (all from the University of Michigan). Only 6 are books (*18 feb 2022: 8)*. After some (brief this time!) research, again, here's the list: 4 are in Latin (at least partially), and 2 in Italian.

- Bartholomäus Arnoldi von Usingen's *Compendium naturalis philosophie* (1507), a commentary of *Parvulus philosophiae naturalis* probably written by Petrus Gerticz of Dresden (d. 1421) (see Kärkkäinen, Pekka. "Synderesis in Late Medieval Philosophy and the Wittenberg Reformers." *British Journal for the History of Philosophy* 20, no. 5 (September 1, 2012): 881–901.)
- a rare 1550 edition of Averroe's prooemium to the long commentary to Aristotle's *Physics* (see: Thomas, David, and John A. Chesworth. *Christian-Muslim Relations. A Bibliographical History: Volume 6. Western Europe (1500-1600)*. Brill, 2014; for more on Averroe's commentaries on the physics I'd turn to Glasner, Ruth. *Averroes' Physics: A Turning Point in Medieval Natural Philosophy*. OUP Oxford, 2009.)
- Simone Porzio's *De rerum naturalium principiis* (1553)
- Paolo Manuzio's *De gli elementi, e di molti loro notabili effetti.* (1557)
- Alessandro Piccolomini's *Della grandezza della terra et dell'acqua* (1558)
- the 1575 edition of Rheticus' *Canon Doctrinae Triangulorum* digitized at e-rara ETH Zürich.

Seidel's Physica is not among these 73 records: Seidel's Physica record cluster gets ascribed Latin notwithstanding being composed of 7 Latin and 3 undetermined records.

*Note added Feb 19, 2022: The list of books includes now also (ignoring a single-leaf manuscript):*

- *A translation (and commentary?) of (parts of) Aristotle's Physics by François Vatable (1495-1547) and Ioannis Argyropoulos (c.1415-1487) (1547)*

Next post.

With thanks to Proclus guru Marije Martijn, Classics scholar Noemi Lambardi, Maria Chiara Parisi, and book expert Paul Dijstelberge.