Software for
Business Intelligence

BizInt Smart Charts

Not Just Duplicates – Challenges in Identifying Non-Patent Literature Articles Across Databases

John Willmore, BizInt Solutions

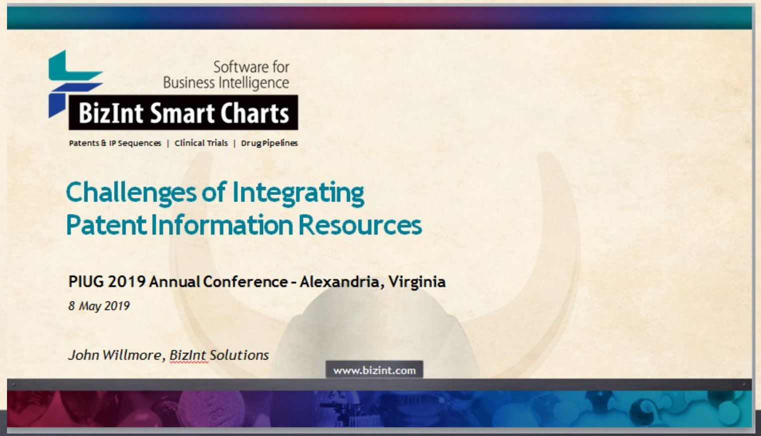
PIUG, Apritober 28, 2020

www.bizint.com

What is a Duplicate?

- Two or more records in a collection of results referring to the same underlying document
- In Patents, full-text versions of the same publication.
- In value added patent databases, we established that records on the same family are not duplicates.

Willmore, *Challenges of Integrating Patent Information Resources*, PIUG 2019 Annual Conference. bizint.com/slides



Software for
Business Intelligence

BizInt Smart Charts

Patents & IP Sequences | Clinical Trials | Drug Pipelines

Challenges of Integrating
Patent Information Resources

PIUG 2019 Annual Conference - Alexandria, Virginia

8 May 2019

John Willmore, BizInt Solutions

www.bizint.com

Duplicates are Naturally Occuring

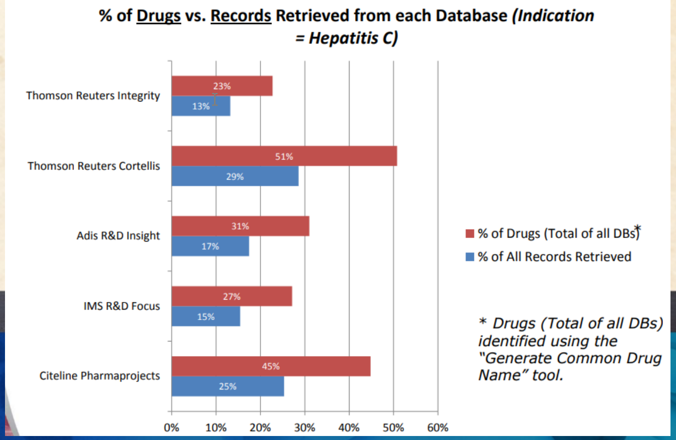
- Different databases with different journal coverage
- Differences in indexing (general vs. domain)

  SciSearch indexes over 8,600 journals
  TULSA draws on 300 journals
- Differences in timeliness
- Differences in search engine / technology
- Collections built over time

Concept of Coverage

- RECALL increases as you apply different search techniques or search different collections
- We often look at COVERAGE – the percentage of items retrieved by any one source/search/technique – as a way to evaluate sources and approaches
- Generally speaking, duplicates are a feature, not a bug (as long as you can handle them)

Surfing the Pipeline, bizint.com/slides#surfing


% of Drugs vs. Records Retrieved from each Database *(Indication = Hepatitis C)*

** Drugs (Total of all DBs) identified using the "Generate Common Drug Name" tool.*

Simple De-duplication Strategies

- Hedging – removing duplicates as part of a search strategy – can easily remove hits from consideration from the only search that will retrieve them
- Let the platform do the deduplication
- Long history of study of duplicate detection, mostly focused on detecting duplicates within collections stored in reference management software (EndNote, Reference Manager, etc).

Traditional Duplicate Detection

- Source data is assumed to be incomplete, inconsistently fielded, and have artifacts (noise) in fields
- Basic techniques look at first author, title, year
- Naïve techniques find 50-60% of matches at best
- Weighted qualification by page, volume, or issue can improve to >80% duplicate detection

Rathbone et al, *Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module*. Syst Rev. 2015; 4(1): 6. DOI: 10.1186/2046-4053-4-6

The Problems With AU/TI Matching

- Redundant Papers (publication of variations of a paper in multiple journals, presentation of a paper at multiple conferences)
- Orthographic standards for names, titles
(John Willmore, J Willmore, Willmore-J-A)
- Common names (Smith; Yang; Li; Singh)
- Heavy publication activity can cause ‘collisions’
COVID-19 project at TechMiningForGlobalGood.org
61,278 publications in PubMed as of Oct 7, 2020

How the Pros Do It

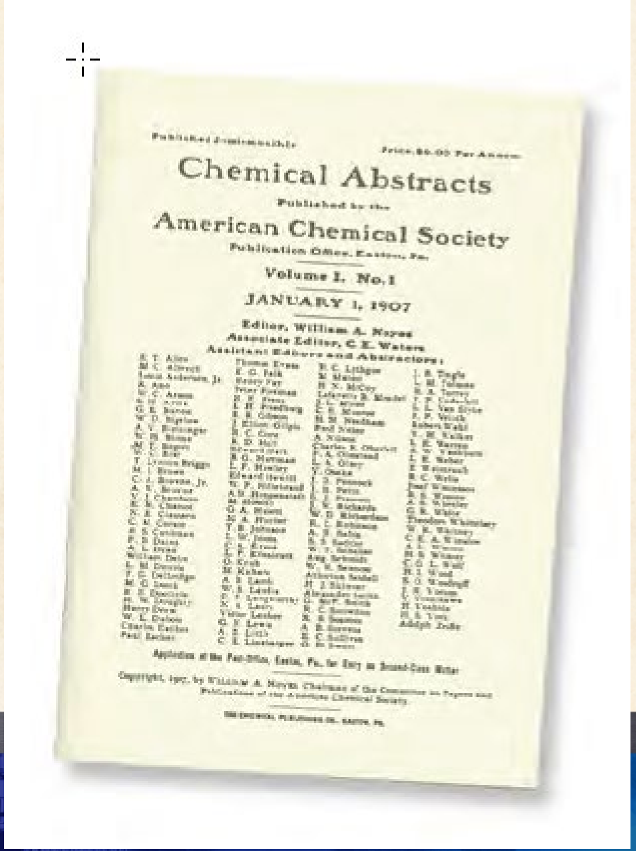
- Search aggregators often do much better in duplicate detection [Citation Needed]
- There is not a significant difference in algorithms
- Key differences are based on the data available
  - Knowing where the data comes from
  - Having a complete feed of content

When You Know The Source

- Knowing the source (e.g. MEDLINE, EMBASE) lets you de-duplicate based on AN
- Or extract content based on knowledge of the structure and tendencies of the publisher
- And use additional identifiers, not often included in typical display formats (e.g. PMID, PMCID, PII)
- *If the data tells you the answer, listen!*

DOI – The Great Hope

- “In our field, we have 100% coverage with DOI – de-duplication is not a problem” – private discussion



Published Semi-monthly [illegible]

Chemical Abstracts

Published by the

American Chemical Society

Publication Office, Easton, Pa.

Volume I. No. 1

JANUARY 1, 1907

Editor, William A. Noyes

Associate Editor, C. E. Waters

Assistant Editors and Abstractors:

[illegible]

Application at the Post-Office, Easton, Pa., for Entry as Second-Class Matter

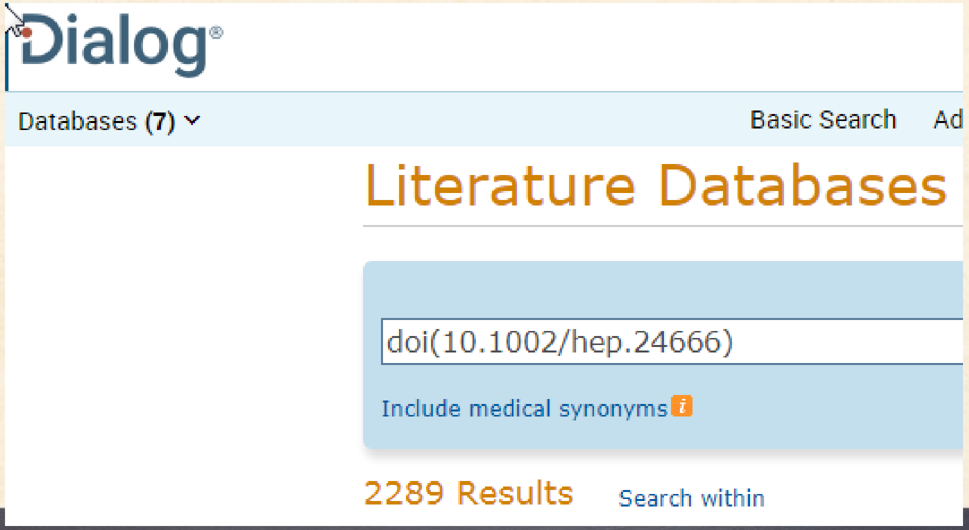
Copyright, 1907, by William A. Noyes, Chairman of the Committee on Papers and Publications of the American Chemical Society

[illegible]

Source: Introduction to CAS

DOI - missed it by that much

- The DOI system would in theory allow for clear duplicate detection
- Use is not universal across journals
- Use is not consistent – one DOI for an entire conference proceeding



doi(10.1002/hep.24666)

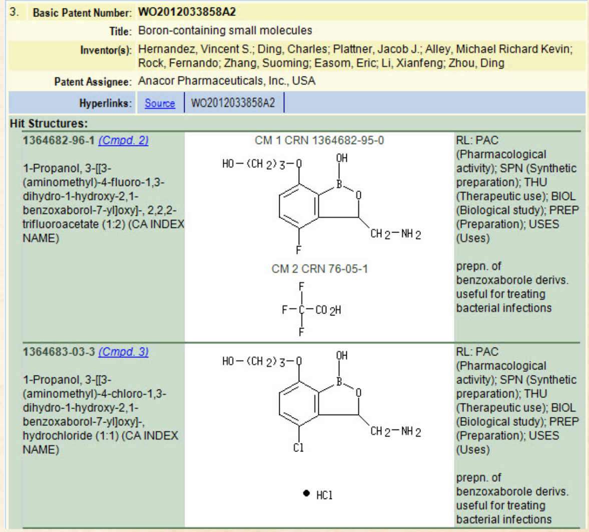
2289 Results

Identifiers Which are Close

- DOI + first page number
- ISSN + Year + first page + first author
- Electronic-ISSNs, ISBNs treated similarly

Content Differences Among Duplicates

- Dependent on the retrieval path (platform, format) as well as the publisher / database
- Indexing is often valuable primarily for search but can be the most important aspect of a result
- Availability of dates varies greatly (e.g. first available date)



3. Basic Patent Number:	**WO2012033858A2**
Title:	Boron-containing small molecules
Inventor(s):	Hernandez, Vincent S.; Ding, Charles; Plattner, Jacob J.; Alley, Michael Richard Kevin; Rock, Fernando; Zhang, Suoming; Easom, Eric; Li, Xianfeng; Zhou, Ding
Patent Assignee:	Anacor Pharmaceuticals, Inc., USA
Hyperlinks:	Source | WO2012033858A2

Hit Structures:

1364682-96-1 *(Cmpd. 2)* 1-Propanol, 3-[[3-(aminomethyl)-4-fluoro-1,3-dihydro-1-hydroxy-2,1-benzoxaborol-7-yl]oxy]-, 2,2,2-trifluoroacetate (1:2) (CA INDEX NAME)	CM 1 CRN 1364682-95-0 CM 2 CRN 76-05-1	RL: PAC (Pharmacological activity); SPN (Synthetic preparation); THU (Therapeutic use); BIOL (Biological study); PREP (Preparation); USES (Uses) prepn. of benzoxaborole derivs. useful for treating bacterial infections
1364683-03-3 *(Cmpd. 3)* 1-Propanol, 3-[[3-(aminomethyl)-4-chloro-1,3-dihydro-1-hydroxy-2,1-benzoxaborol-7-yl]oxy]-, hydrochloride (1:1) (CA INDEX NAME)		RL: PAC (Pharmacological activity); SPN (Synthetic preparation); THU (Therapeutic use); BIOL (Biological study); PREP (Preparation); USES (Uses) prepn. of benzoxaborole derivs. useful for treating bacterial infections

Conclusion

- Duplicates are a feature, not a bug
- Different techniques can handle duplicates in different contexts
- Give thought to the selection of databases and display formats

Software for
Business Intelligence

BizInt Smart Charts

bizint.com/PIUG