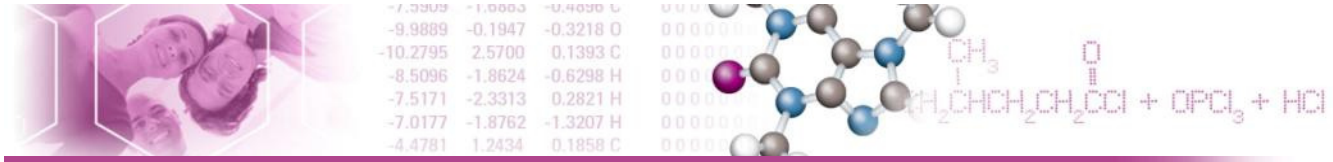


Progresses in Automated Chemical Structure Recognition in Text and Images

V. Eigner-Pitto and H. Saller.
InfoChem GmbH, Landsberger Strasse 408, Munich, 81241, Germany

Historical Importance of Chemisches Zentralblatt


Pharmaceutisches
Central Blatt.
14. Januar 1830.
1.
Ankündigung.

1830 Chemisches Zentralblatt 1969

First and oldest abstracts journal in chemistry

Covers chemical literature from 1830 to 1969

Describes the „birth“ of chemistry as science (vs. alchemy)

1840 1907 Chemical Abstracts …

Biggest and single abstracts source in chemistry

Currently >31 million papers and patents

Content 1840-1906 added retrospectively

1772 1817 Gmelin Handbook …

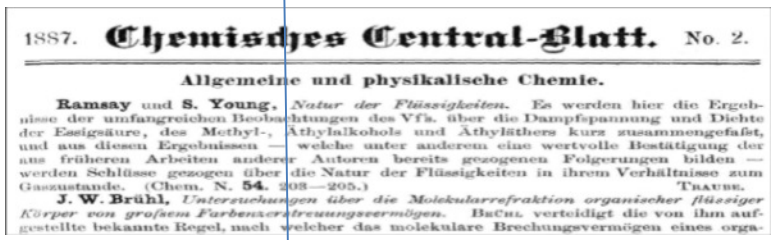
1771 1881 Beilstein Handbook …

Chemisches Zentralblatt: History (1)

1st year: 544 pages/400 abstracts After 6 months weekly


Pharmaceutisches
Central Blatt.
14. Januar 1830.

„Chemisches Zentralblatt“


1887. Chemisches Central-Blatt. No. 2.

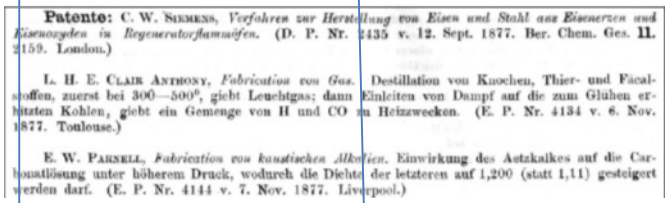
Allgemeine und physikalische Chemie.

Ramsay und **S. Young**, *Natur der Flüssigkeiten.* Es werden hier die Ergebnisse der umfangreichen Beobachtungen des Vfs. über die Dampfspannung und Dichte der Essigsäure, des Methyl-, Äthylalkohols und Äthyläthers kurz zusammengefaßt, und aus diesen Ergebnissen — welche unter anderem eine wertvolle Bestätigung der aus früheren Arbeiten anderer Autoren bereits gezogenen Folgerungen bilden — werden Schlüsse gezogen über die Natur der Flüssigkeiten in ihrem Verhältnisse zum Gaszustande. (Chem. N. **54**. 203—205.) Traube.

J. W. Brühl, *Untersuchungen über die Molekularrefraktion organischer flüssiger Körper von großem Farbenzerstreuungsvermögen.* Brühl verteidigt die von ihm aufgestellte bekannte Regel, nach welcher das molekulare Brechungsvermögen eines orga-

1st edition of the General Index

The editorial office moves from Leipzig to Berlin

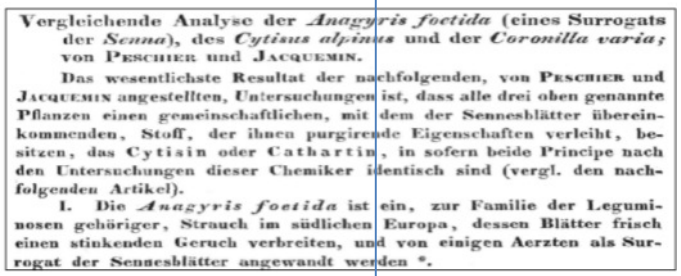

Patente: C. W. Siemens, *Verfahren zur Herstellung von Eisen und Stahl aus Eisenerzen und Eisenoxyden in Regeneratorflammöfen.* (D. P. Nr. 2435 v. 12. Sept. 1877. Ber. Chem. Ges. **11**. 2159. London.)

L. H. E. Clair Anthony, *Fabrication von Gas.* Destillation von Knochen, Thier- und Fäcalstoffen, zuerst bei 300—500°, giebt Leuchtgas; dann Einleiten von Dampf auf die zum Glühen erhitzten Kohlen, giebt ein Gemenge von H und CO zu Heizzwecken. (E. P. Nr. 4134 v. 6. Nov. 1877. Toulouse.)

E. W. Parnell, *Fabrication von kaustischen Alkalien.* Einwirkung des Aetzkalkes auf die Carbonatlösung unter höherem Druck, wodurch die Dichte der letzteren auf 1,200 (statt 1,11) gesteigert werden darf. (E. P. Nr. 4144 v. 7. Nov. 1877. Liverpool.)

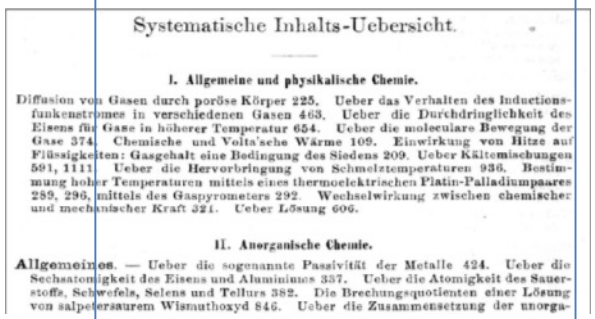
Introduction of Patent abstracts

1830 1850 1856 1862 1879 1883 1897 1902 1919 1924 …


Vergleichende Analyse der *Anagyris foetida* (eines Surrogats der *Senna*), des *Cytisus alpinus* und der *Coronilla varia*; von Peschier und Jacquemin.

Das wesentlichste Resultat der nachfolgenden, von Peschier und Jacquemin angestellten, Untersuchungen ist, dass alle drei oben genannte Pflanzen einen gemeinschaftlichen, mit dem der Sennesblätter übereinkommenden, Stoff, der ihnen purgirende Eigenschaften verleiht, besitzen, das Cytisin oder Cathartin, in sofern beide Principe nach den Untersuchungen dieser Chemiker identisch sind (vergl. den nachfolgenden Artikel).

I. Die *Anagyris foetida* ist ein, zur Familie der Leguminosen gehöriger, Strauch im südlichen Europa, dessen Blätter frisch einen stinkenden Geruch verbreiten, und von einigen Aerzten als Surrogat der Sennesblätter angewandt werden *.


Systematische Inhalts-Uebersicht.

I. Allgemeine und physikalische Chemie.

Diffusion von Gasen durch poröse Körper 225. Ueber das Verhalten des Inductionsfunkenstromes in verschiedenen Gasen 463. Ueber die Durchdringlichkeit des Eisens für Gase in höherer Temperatur 654. Ueber die moleculare Bewegung der Gase 374. Chemische und Volta'sche Wärme 109. Einwirkung von Hitze auf Flüssigkeiten: Gasgehalt eine Bedingung des Siedens 209. Ueber Kältemischungen 591, 1111. Ueber die Hervorbringung von Schmelztemperaturen 936. Bestimmung hoher Temperaturen mittels eines thermoelektrischen Platin-Palladiumpaares 289, 296, mittels des Gaspyrometers 292. Wechselwirkung zwischen chemischer und mechanischer Kraft 321. Ueber Lösung 606.

II. Anorganische Chemie.

Allgemeines. — Ueber die sogenannte Passivität der Metalle 424. Ueber die Sechsatomigkeit des Eisens und Aluminiums 337. Ueber die Atomigkeit des Sauerstoffs, Schwefels, Selens und Tellurs 382. Die Brechungsquotienten einer Lösung von salpetersaurem Wismuthoxyd 846. Ueber die Zusammensetzung der unorga-

CZ splitts in two parts: „technical“ and „scientific“

CZ is re-unified in one journal of abstracts

Title changes in „Chemisch-Pharmaceutisch Central-Blatt“

Introduction of a systematic table of contents

Holding passes to the *Deutschen Chemischen Gesellschaft* Introduction of Patent Register

Chemisches Zentralblatt: History (2)

Introduction of Formula Index

Formelregister

der organischen Verbindungen,

geordnet nach M. M. Richters Formelsystem.

Diejenigen Verbindungen, bei denen nicht mit Kursivschrift auf den Registrierort im Sachregister hingewiesen ist, finden sich lediglich im Formelregister. Vergl. auch Vorwort für das Sach- und Formelregister (C. 1925. II. 2581).

C_1-Gruppe.

— 1 1 —

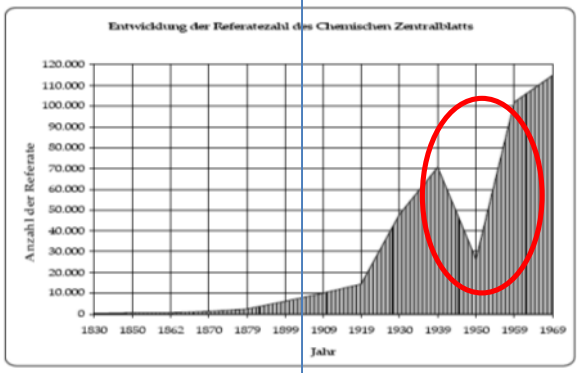
CH_4 s. Methyl.
CH_4 s. Methan.
CO s. Kohlenoxyd.
CO_2 s. Kohlensäure (Kohlendioxyd).
CCl_4 s. Kohlenstofftetrachlorid (Tetrachlorkohlenstoff).
CBr_4 s. Kohlenstofftetrabromid (Tetrabromkohlenstoff).
CJ_4 s. Kohlenstofftetrajodid.
CF_4 s. Kohlenstofftetrafluorid.
CS_2 s. Schwefelkohlenstoff.
$[COCl]_2$ Verb. $[COCl]_2$ aus C_2Cl_6 II 1121.
$COCl_2$ s. Phosgen.
$COBr_2$ s. Bromphosgen.
COS s. Kohlenoxysulfid.
CO_4Cl_2 Trichlormethylperchlorat I 1600.
CO_2N_2 s. Methan, dinitro.
CNCl s. Chlorcyan.
CNBr s. Bromcyan.
CNJ s. Jodcyan.
CNF s. Fluorcyan.
CCl_3F s. Methan, chlortrifluor.
CCl_2F_2 s. Methan, dichlordifluor.
CCl_2S_2 s. Thiophosgen.
CCl_3F s. Methan, fluortrichlor.
OCl_2S Thiocarbonylperchlorid I 2994.

Double production of CZ in East and West Germany

Berlin wall does not disturb production of CZ

Introduction of SRD „Schnell Referat Dienst“ („quick abstract service)

… 1925 | 1940-45 | 1947-49 | 1950 | 1961 | 1967 | 1969



Difficulties in collecting information

Reunification of the journal under East and West German organizations

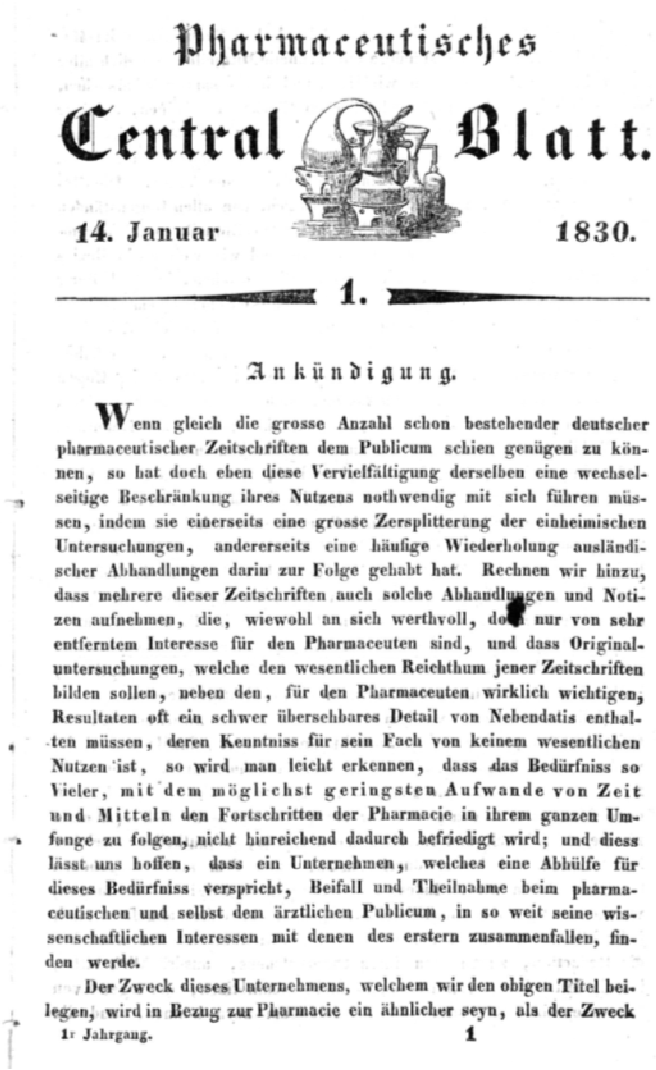
CHEMISCHES ZENTRALBLATT

Nr. 1 · 8. Januar 1964
Referat 0001—3013
Seite 1—304

Geschichte, Unterricht

GDR office declares unable to afford production of SRD and of journal Chemisches Zentralblatt ceases publication

Chemisches Zentralblatt: Content



Pharmaceutisches

Central Blatt.

14. Januar 1830.

1.

Ankündigung.

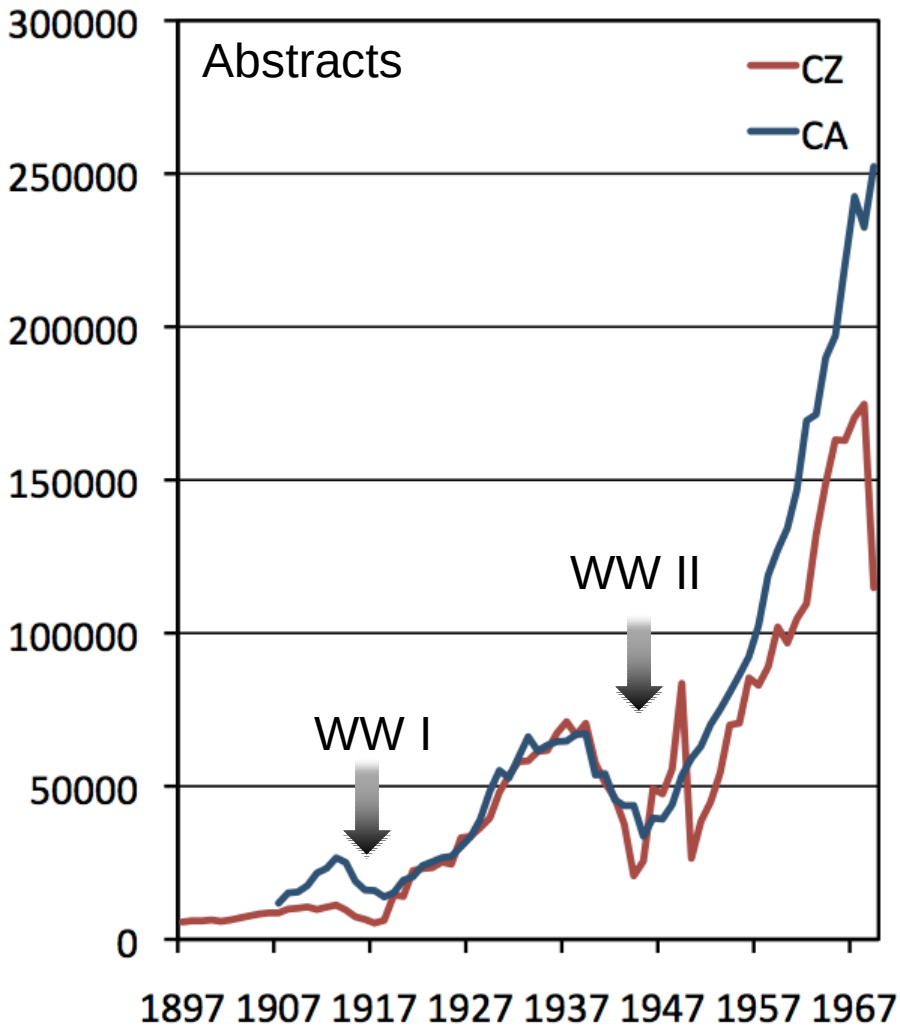
Wenn gleich die grosse Anzahl schon bestehender deutscher pharmaceutischer Zeitschriften dem Publicum schien genügen zu können, so hat doch eben diese Vervielfältigung derselben eine wechselseitige Beschränkung ihres Nutzens nothwendig mit sich führen müssen, indem sie einerseits eine grosse Zersplitterung der einheimischen Untersuchungen, andererseits eine häufige Wiederholung ausländischer Abhandlungen darin zur Folge gehabt hat. Rechnen wir hinzu, dass mehrere dieser Zeitschriften auch solche Abhandlungen und Notizen aufnehmen, die, wiewohl an sich werthvoll, doch nur von sehr entferntem Interesse für den Pharmaceuten sind, und dass Originaluntersuchungen, welche den wesentlichen Reichthum jener Zeitschriften bilden sollen, neben den, für den Pharmaceuten wirklich wichtigen, Resultaten oft ein schwer übersehbares Detail von Nebendatis enthalten müssen, deren Kenntniss für sein Fach von keinem wesentlichen Nutzen ist, so wird man leicht erkennen, dass das Bedürfniss so Vieler, mit dem möglichst geringsten Aufwande von Zeit und Mitteln den Fortschritten der Pharmacie in ihrem ganzen Umfange zu folgen, nicht hinreichend dadurch befriedigt wird; und diess lässt uns hoffen, dass ein Unternehmen, welches eine Abhülfe für dieses Bedürfniss verspricht, Beifall und Theilnahme beim pharmaceutischen und selbst dem ärztlichen Publicum, in so weit seine wissenschaftlichen Interessen mit denen des erstern zusammenfallen, finden werde.

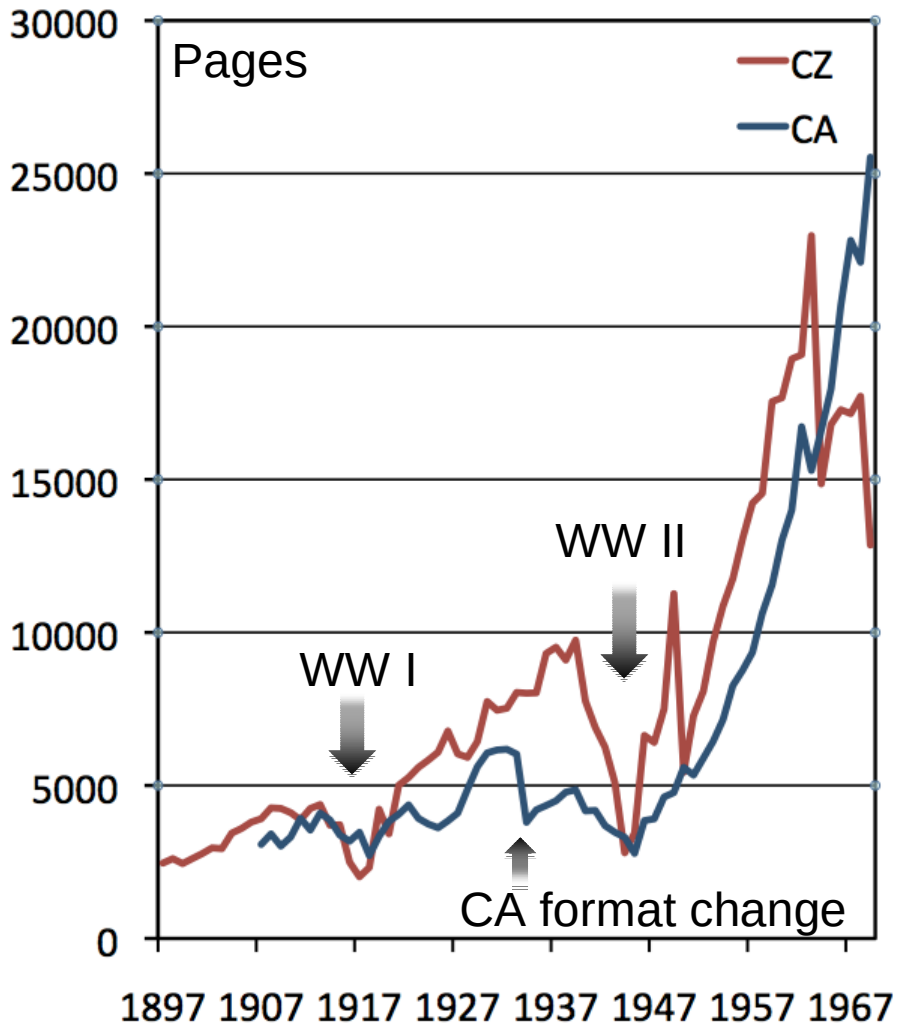
Der Zweck dieses Unternehmens, welchem wir den obigen Titel beilegen, wird in Bezug zur Pharmacie ein ähnlicher seyn, als der Zweck

1r Jahrgang. 1

- 140 years publication
- 900,000 pages
  - 700,000 contain ca. 2 million abstracts
  - 200,000 are indexes
  - 120,000 cover time period 1830-1907

Chemisches Zentralblatt vs. CA: Quantity



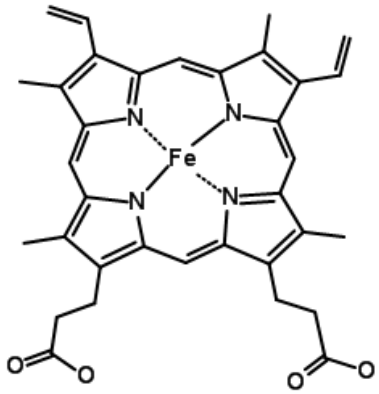


Chemisches Zentralblatt vs. CA: Quality

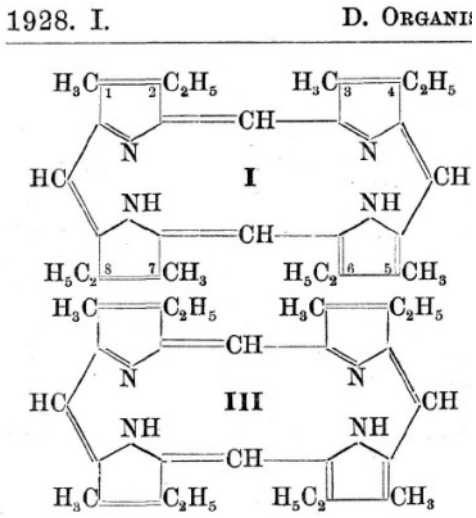
- Many textbooks on chemical literature claim better quality of Chemisches Zentralblatt than CA for pre-WW II
  - H. Skolnik, *The literature matrix of chemistry*, 1982: „outstanding A/I service“
  - R.E. Maizell, *How to find chemical information*, 3rd ed. 1998, citing E.J. Crane, „[..] has value because of [..] good abstracts“
  - M. Mücke, *Die chemische Literatur*, 1982, „Zwar war CA zahlenmässig [..] dem Chemischen Zentralblatt überlegen, doch war dies gerade umgekehrt, was die Qualität der Referate betraf.“
  - R.T. Bottle, J.F. Rowland, *Information Sources in Chemistry*, 4th ed. 1993, „Before WW II, many chemists regarded CZ as superior in coverage to CA; its abstracts were longer and more informative [...]“

Chemisches Zentralblatt vs. CA: Quality

Example: Hans Fischer, Georg Stangler, Synthese des Mesoporphyrings, Mesohämins und über die Konstitution des Hämins, Justus Liebigs Ann. Chem. 459(1927), 53-98.



	CZ I(1928), 528	CA 22:11339 (1928), 1363
Length (pages)	7.5	1
Length (words)	3,882	690
Length (chars)	24,308	4,695
Compounds	~ 120	~ 70
Structure formulas	Yes	No

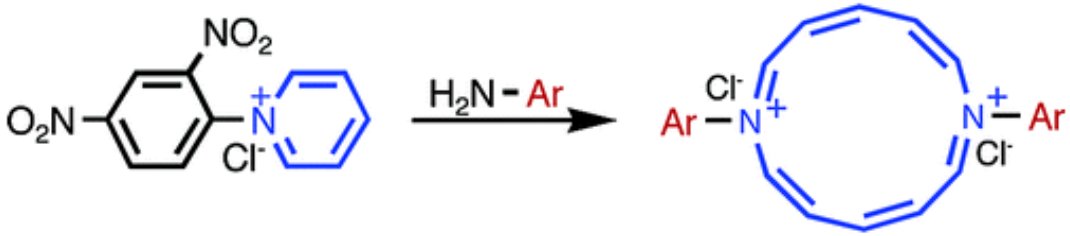


Chemisches Zentralblatt: Conclusion

- Covers chemical literature from 1830 to 1969
- Very good abstract quality
  - Better quality (length, details) than CA for pre-WW II period 1907-1940
- Contains also important patent information
- Invaluable information in indexes (e.g. synonyms of ancient chemical names)
- Only comprehensive abstract journal on the market up to 1907
  - More comprehensive than CA for 19th century literature
  - Complements Beilstein and Gmelin handbooks for 19th century literature

Importance of Chemisches Zentralblatt: Example

Org. Lett., 2006, 8 (19), pp 4279–4281



N-Substituted diaza[12]annulenes are obtained by one -pot reaction of *N*-(2,4-dinitrophenyl)pyridinium chloride with amines in moderate to high yields. The ^{1}H NMR spectrum reveals that diamagnetic ring current is generated in the diaza[12]annulene ring. The *N*-substituted diaza[12]annulenes are electrochemically active in solution.

The authors have retracted this paper on November 15, 2007 (Org. Lett. 2007, 24, 5139)

Chemisches Zentralblatt., 1904, 2, 1145

Th. Zincke, *Über Dinitrophenylpyridiniumchlorid und dessen Umwandlungsprodukte.* Das Dinitrochlorbenzol-(2,4,1) verbindet sich mit Pyridin direkt zu einem gut charakterisierten quaternären Ammoniumsalz, dem *Dinitrophenylpyridiniumchlorid*, $C_5H_5N^v(Cl)C_6H_3(NO_2)_2$. Diese Pyridinverb. reagiert in sehr eigenartiger Weise mit Alkali und mit verschiedenen primären und sekundären Aminen; durch Alkali wird sie in einen schön rot gefärbten Körper umgewandelt, der die Zus. der freien Ammoniumbase besitzt, während die Amine mit grofser Leichtigkeit eine Spaltung unter B. von Dinitranilin herbeiführen.

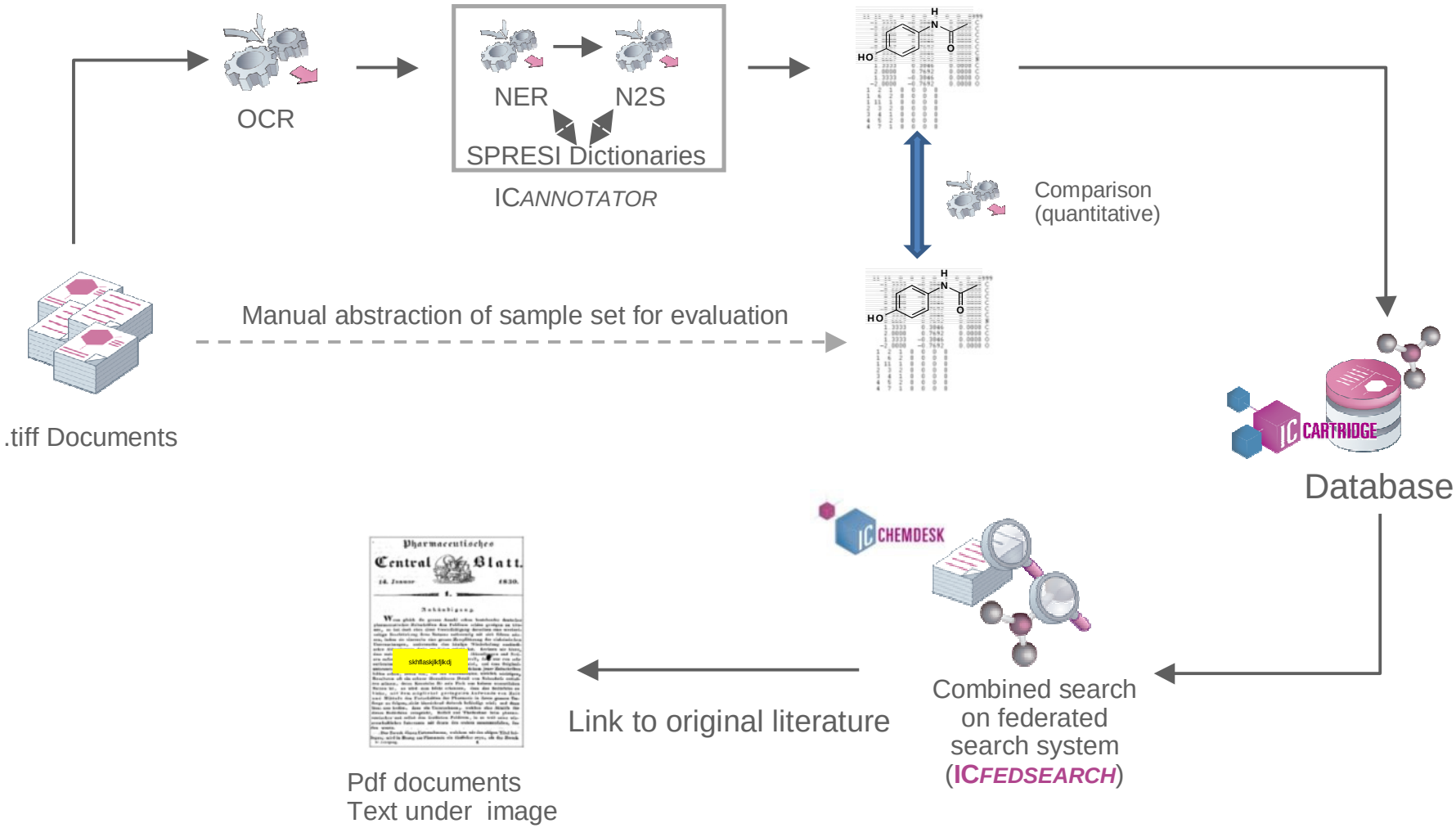
Chemisches Zentralblatt: Digitalization

- Relevant for documentation of prior art
- Continuous and growing demand of the information
- FIZ Chemie Berlin has scanned the whole work and offers a full text searchable database for the web and the dataset for integration in Intranets
- ETH Zurich has bought the digitalized raw material (pdfs with OCRed text in the background) from FIZ and is creating a database offering full text search
- CAS has performed automatic translation (German → English) of the 1897-1907 volumes and included them in CAplus

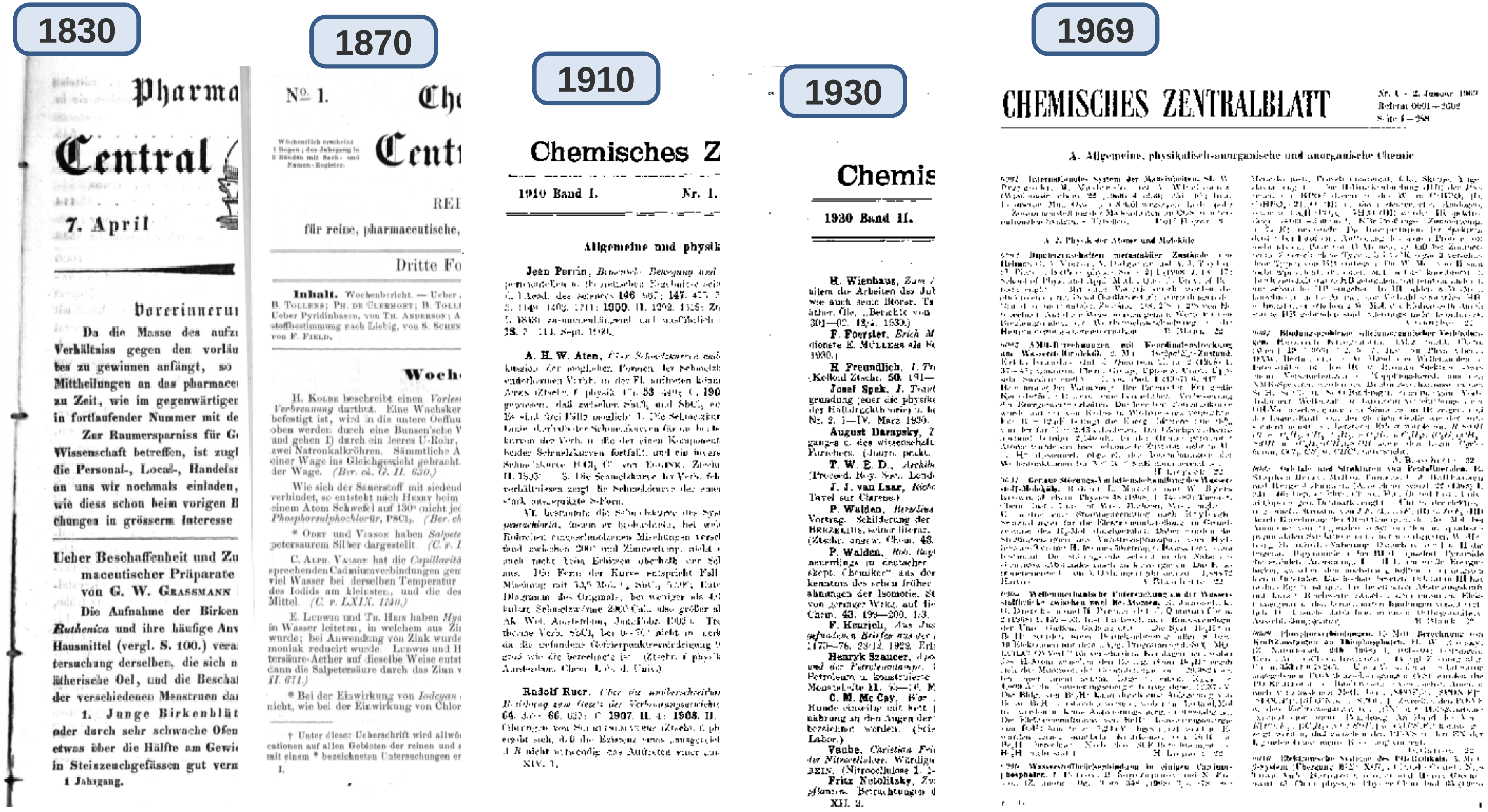
InfoChem Motivation

- Text search in Chemisches Zentralblatt:
  - Abstracts in German language
  - High number of old German chemical names
- Chemists think in structures!!!
- Language independent structure search would help ALL scientists to access this historical source and to use the relevant information of this art
- Required technology for structure search projects
- Optimized German-English dictionaries
- 30 million SPRESI names

Overview of Approach and Applied Technology

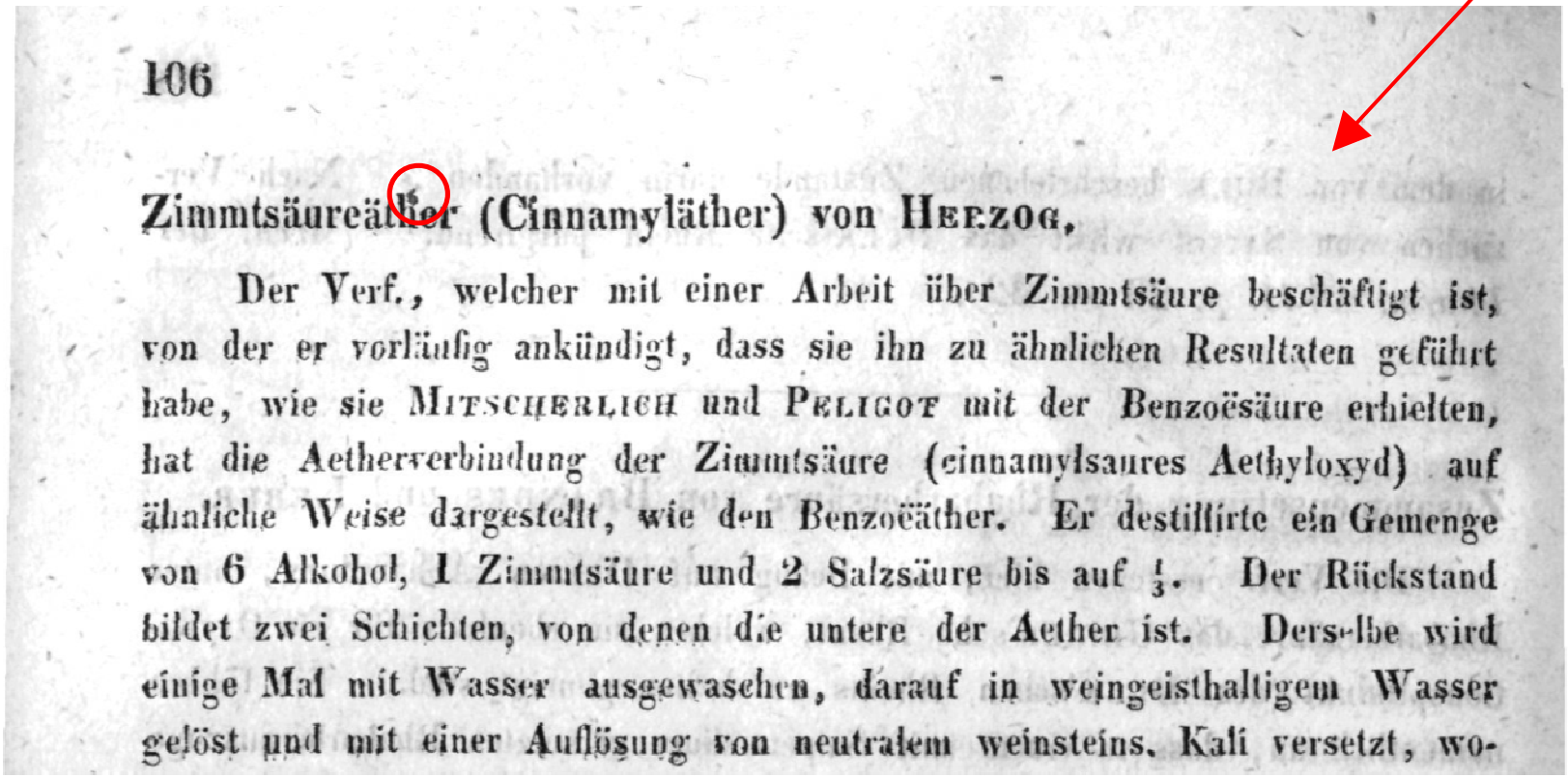


Challenges OCR (1)



Challenges OCR (2)

- Bad quality of original source: dirty (blotted, stained) pages
  print from back page



106

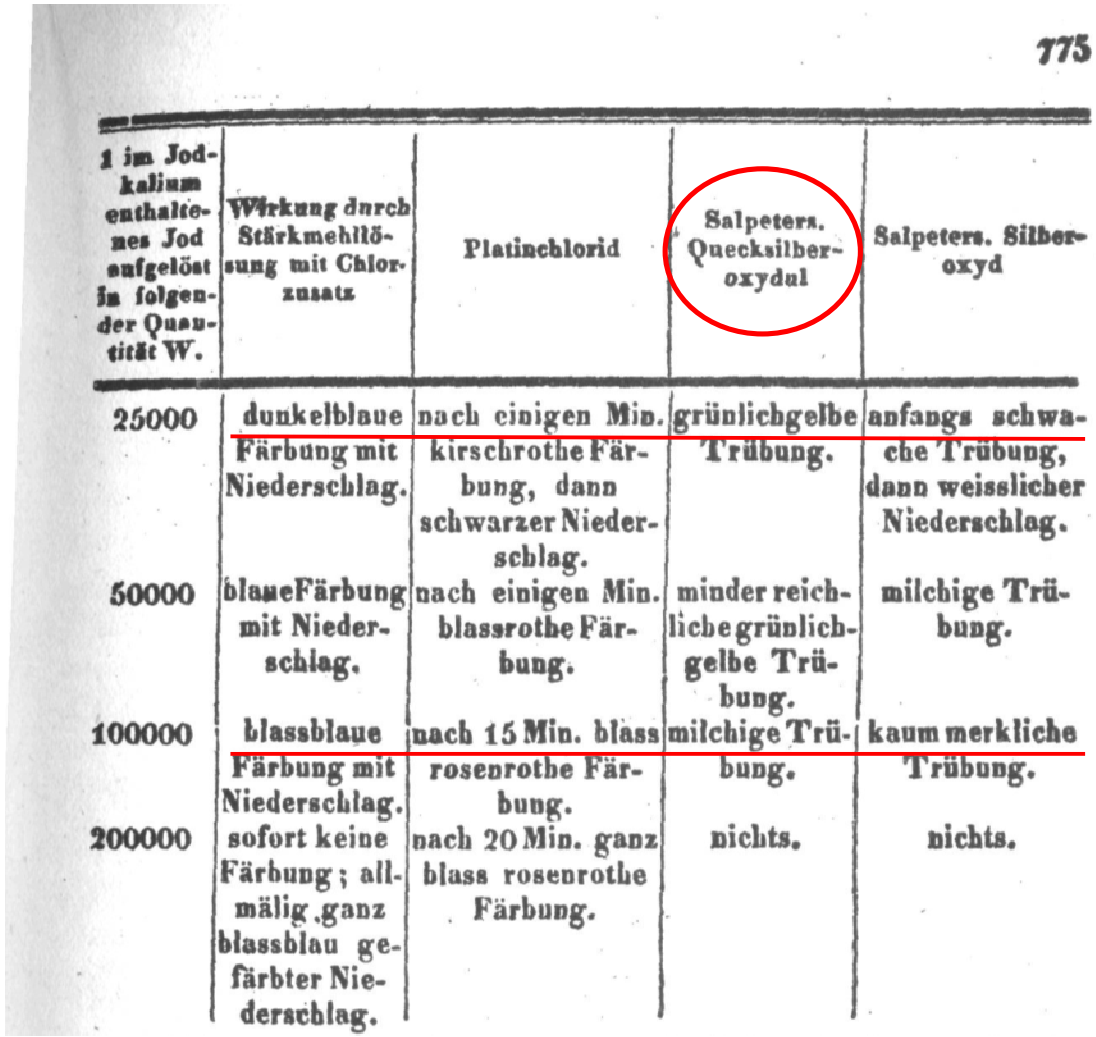
Zimmtsäureäther (Cinnamyläther) von Herzog.

Der Verf., welcher mit einer Arbeit über Zimmtsäure beschäftigt ist, von der er vorläufig ankündigt, dass sie ihn zu ähnlichen Resultaten geführt habe, wie sie Mitscherlich und Peligot mit der Benzoësäure erhielten, hat die Aetherverbindung der Zimmtsäure (cinnamylsaures Aethyloxyd) auf ähnliche Weise dargestellt, wie den Benzoëäther. Er destillirte ein Gemenge von 6 Alkohol, 1 Zimmtsäure und 2 Salzsäure bis auf $\frac{1}{3}$. Der Rückstand bildet zwei Schichten, von denen die untere der Aether ist. Derselbe wird einige Mal mit Wasser ausgewaschen, darauf in weingeisthaltigem Wasser gelöst und mit einer Auflösung von neutralem weinsteins. Kali versetzt, wo-

Challenges OCR (3)

- Tables: extremely small fonts,

  not recognizable
  begin / end of columns

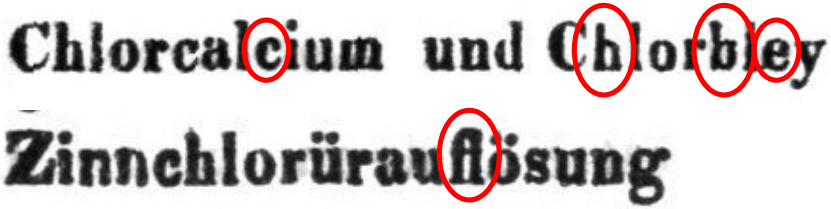


775

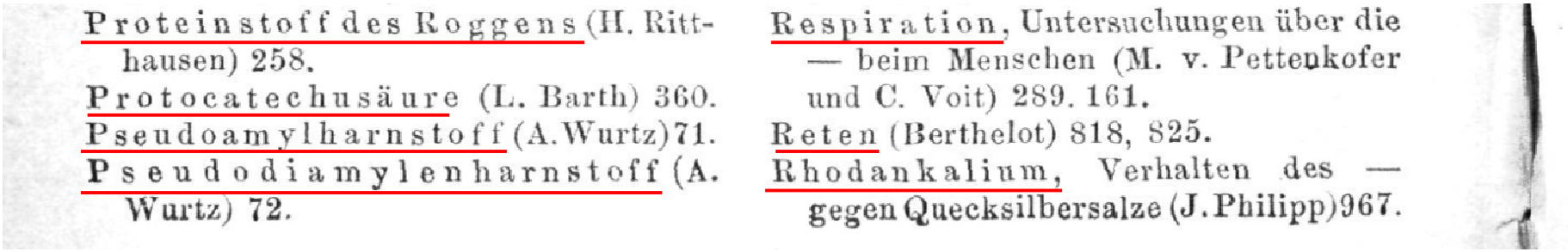
1 im Jodkalium enthaltenes Jod aufgelöst in folgender Quantität W.	Wirkung durch Stärkmehllösung mit Chlorzusatz	Platinchlorid	Salpeters. Quecksilberoxydul	Salpeters. Silberoxyd
25000	dunkelblaue Färbung mit Niederschlag.	nach einigen Min. kirschrothe Färbung, dann schwarzer Niederschlag.	grünlichgelbe Trübung.	anfangs schwache Trübung, dann weisslicher Niederschlag.
50000	blaueFärbung mit Niederschlag.	nach einigen Min. blassrothe Färbung.	minder reichliche grünlichgelbe Trübung.	milchige Trübung.
100000	blassblaue Färbung mit Niederschlag.	nach 15 Min. blass rosenrothe Färbung.	milchige Trübung.	kaum merkliche Trübung.
200000	sofort keine Färbung; allmälig ganz blassblau gefärbter Niederschlag.	nach 20 Min. ganz blass rosenrothe Färbung.	nichts.	nichts.

Challenges OCR (4)

- Ambiguous old fonts (h=b; c=e; ligations)



- Spaced text



Specific rules, large German dictionaries and extensive training are applied to correct systematic mistakes of standard OCR process

Main Challenges Annotation

- Obsolete German language
  - Schwefelsaures Natrium, Chlorür, Bromür
- Historical names
  - Pelopeum → Columbium → Niobium
- Different spelling for the same name:
  - Dibrom… ⇔ Bibrom…
  - Ätzkali ⇔ Aetzkali
  - Zimmtsäure ⇔ Zimtsäure

Solutions in Annotation Process

- Correction of German-specific grammar
- Translation in English of not available chemical names
- Research in old sources:
  - Beilstein
  - Brockhaus Encyclopedia
  - German-English dictionaries of chemistry
  - Meyers Encyclopedia
  - Pierer Encyclopedia
  - References to very old books, journals, articles
    - “Naturwissenschaftliche Exzerpte und Notizen Mitte 1877 bis Anfang 1883” by Karl Marx

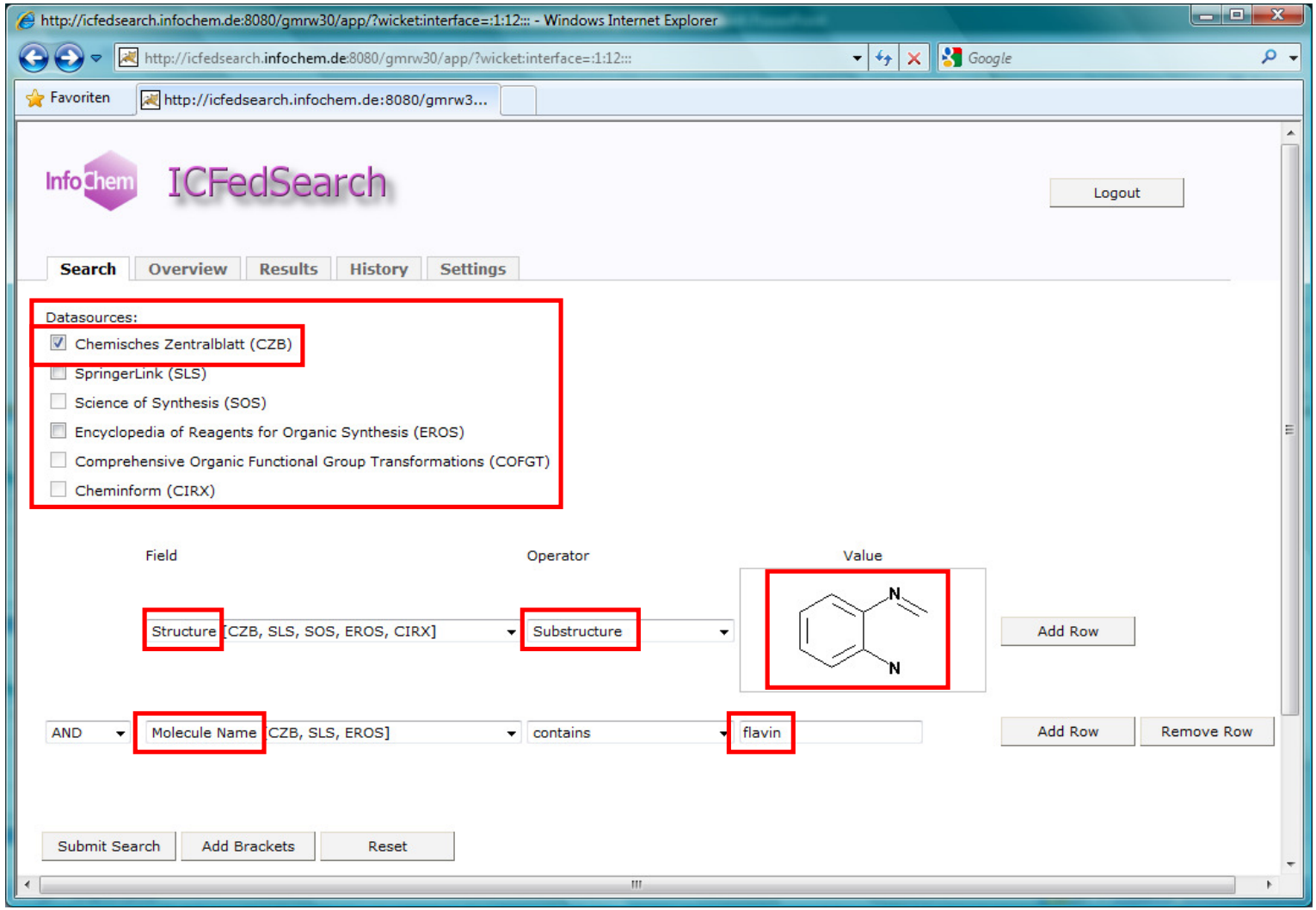
Results Annotation Chemisches Zentralblatt

- 120,000 pages covering time period 1830-1907
- 2.4 million chemical names with associated structure
  - 98,000 unique names
  - 47,000 unique structures

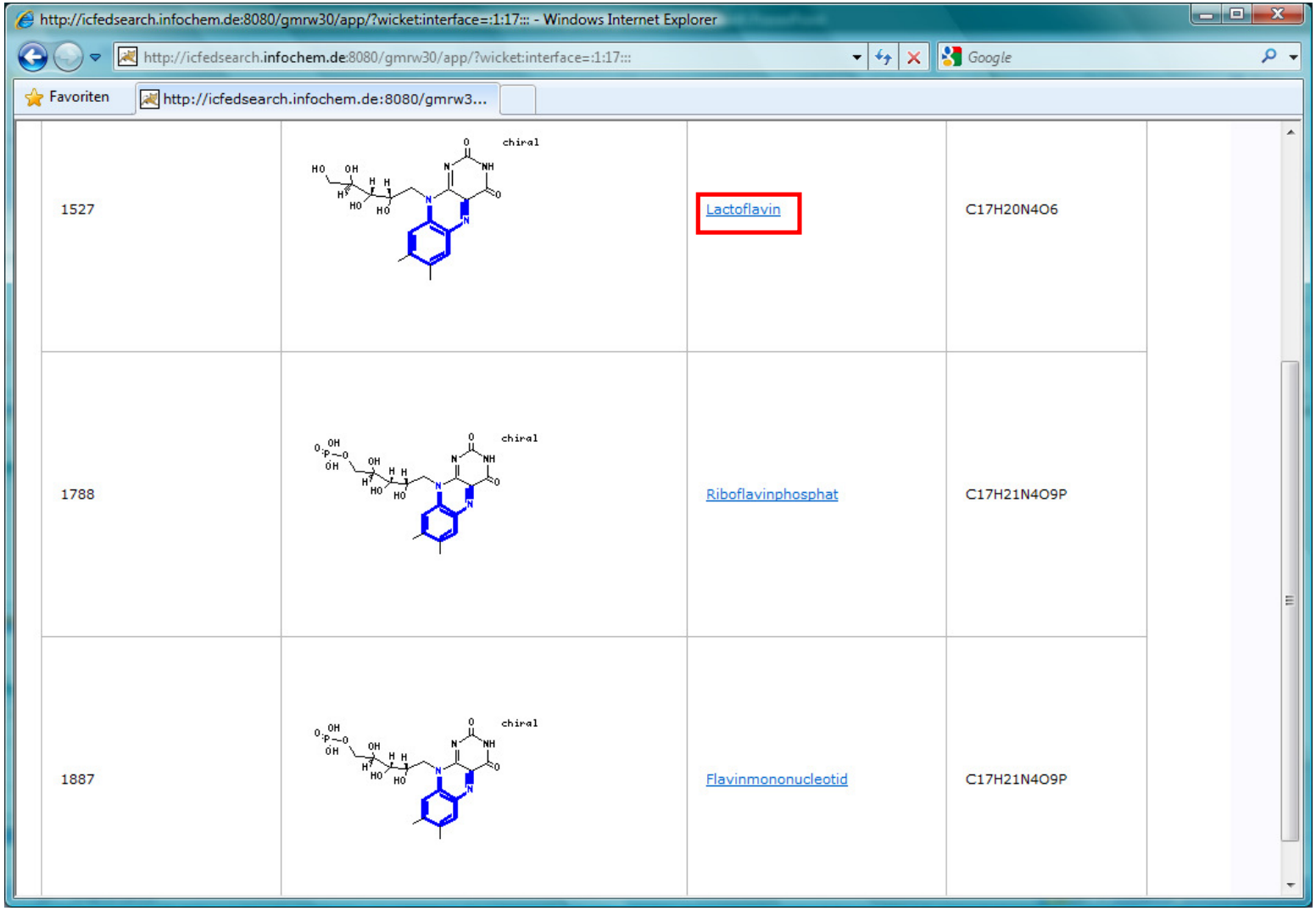
Quantitative comparison with manually abstracted sample set

- Recall 51%
- Precision 87%

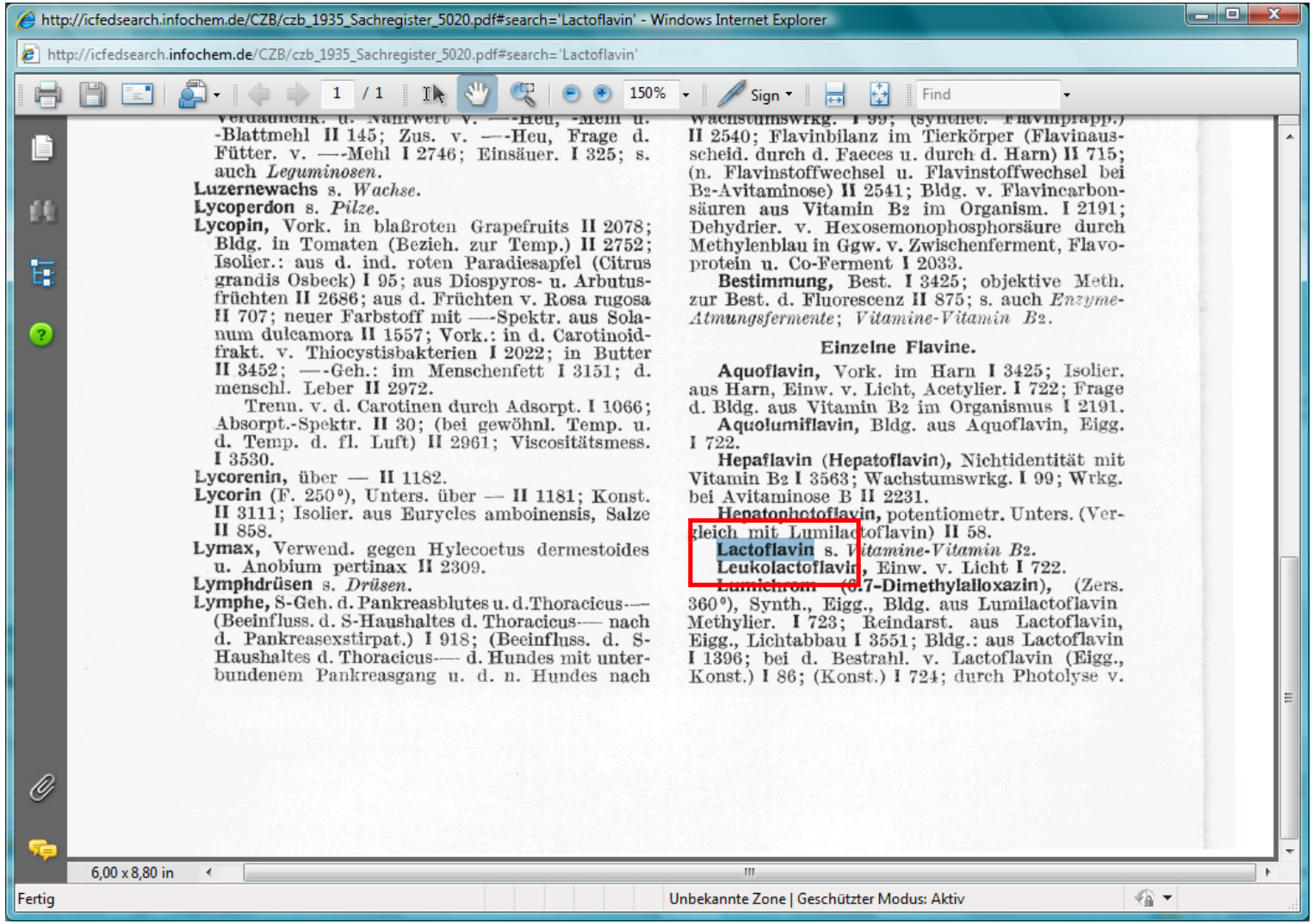
Federated Search Prototype



Federated Search Prototype



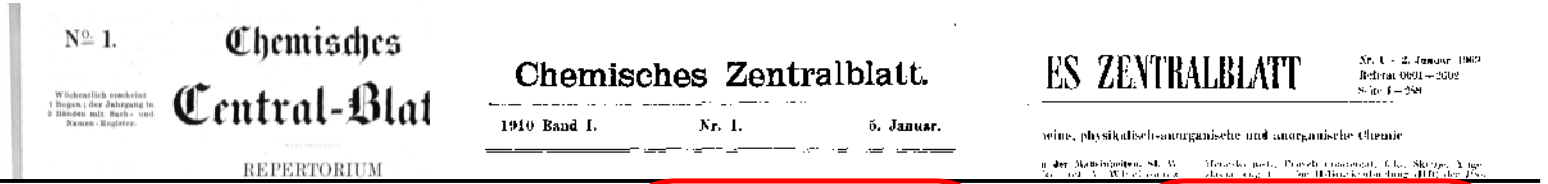
Federated Search Prototype



Summary

- Described history, content and importance nowadays of Chemisches Zentralblatt
- Illustrated how the challenges of OCR and annotation process have been solved
- Time period 1830-1907 contains 98,000 unique names and 47,000 unique structures
- Quantitative comparison proves over 50% recall and nearly 90% precision
- Generated structure searchable Chemisches Zentralblatt database is integrated in IC*FEDSEARCH*

Outlook



Chemisches Zentralblatt:	Phase 1, Q2 2009	Phase 2, Q4 2009
Pages:	120,000	900,000
Time period:	1830-1907	1830-1969
Unique names:	98,000	Ca. 1 million
Unique structures:	47,000	Ca. 500,000
Recall:	50%	?

Acknowledgements

- Prof. Dr. Deplanque, Mr. Heineke and FIZ Chemie Team Berlin
- Ms. Langanke
- Dr. Brändle (ETH Zürich)
- InfoChem Team

Thank you!

InfoChem GmbH: www.infochem.de, www.spresi.com, info@infochem.de

Chemisches Zentralblatt: Structure

Abstracts (Referate)

Short reports of national and international progresses in chemistry
→ **Purpose of CZB**

1870-1886
Wochenberichte

1884
Figures are comprised in the abstracts

1964
Numbering of abstracts

Patents (Patente)

1879
Introduction (end of abstracts part)

1921
Patent Numbers country-specific

1938
Patents from 20 Nations abstracted

1964
Numbering of patent abstracts

Indexes (Register)

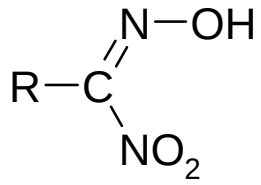
1830	Autorenregister Sachregister
1863	Systematische Register
1897	Patentregister
1924	Patentrückzitateliste
1925	Formelregister
1883	Generalregister

⇒ Aim of all indexes is to facilitate the search of relevant information
⇒They are the first attempt of modern search possibilities

Challenges Annotation (1)

- Names lack position, valence or stoichiometric information
  - Pimarsäure → is it the R or L form?
  - Platinchlorid → in which oxidation state II, III, IV?

- Chemical names that indicate a chemical class or a mixture
  - Nitrolsäure (nitrolic acid) →



NO solution: information for correct structure is not available in the original source