



Rescuing “lost” data

Using Text Mining to apply
thesaurus-based Indexing to
digitized Print Material

What did we lose?

- We didn't really!
- What we did have was 49 bound volumes of *Biological Abstracts*® for 1926 to 1968
- We digitised the content using offshore keying
- But we needed to make the data searchable with the BIOSIS controlled vocabulary

12866-12875 PISCES, ETC. [Biol. Ab. 9(6)] 1426

(Collembola), occurring in European caves, with key to 22 spp.; *O. armatus* ab. *multituberculata* (p.132), Germany; *O. rubricus* f. *dentifera* (p.144), Hungary; *O. strasseri** (p.147) and *O. varicostriolatus** (p.161), Italy; *O. ambulans retospinatus* (*O. r. Stach*) (p.183); *O. denisi** (p.191), Poland; *O. paucituberculatus** and *O. granulatus**, descr.; *O. camzianus** (p.200), Italy; *O. cavernicolus** (p.212), Austria.

CYCLOSTOMATA, ELASMOBRANCHII, AND PISCES
H. WALTON CLARK, Associate Editor
(See also in this issue Entries 10843, 10847, 10848, 10855, 10866, 10871, 10880, 10947, 11335, 11397, 11464, 12397, 12399, 12403, 12407, 12408, 12414, 12415, 12417, 12426, 12457, 12490, 12492, 12505, 12512)

12866. BARTON, E. A. Chalk streams and water-meadows. xiii+128p. 12 pl. John Murray: London, 1932. Pr. 7s.6d.—This is an angler's account of the natural history of trout and chub in the chalk streams of southeastern England, with comments and observations on the night sight and senses in freshwater fishes.—*C. A. Kofoid.*

12867. BORODIN, N. A. A new Australian fish, *Copeia* 1933 (3): 141-142, 1933.—*CONGROGADOIDES* (p.241), resembling *Congrogadus*, erected for *C. splinter* (p.193), W. Australia.—*F. H.*

12868. BRYANT, WILLIAM J. New fishes from the Triassic of Pennsylvania. *Proc. Amer. Phil. Soc.* 73 (5): 319-326, 8 pl. 1934.—A new locality (North Wales, Pennsylvania) for fossil fishes is reported. In addition to the fishes descr. below, dinosaur tracks and plant remains were also found. **CARINACANTHUS** (p.320) (Elasmobranchii, Cestraciontidae), type *C. jepseni** (p.320). The remains of Triassic sharks, other than detached teeth and spines, are extremely rare, and this new elasmobranch has the additional interest of being the first such discovery in the Triassic rocks of eastern N. Amer., and seems to indicate either marine or estuarine conditions, a theory of deposition now abandoned, at least for the Connecticut Valley Trias. *Coelacanthus newarkii** (p.323) (Coelacanthini). Both new spp. are from the Lockatong Formation of the Newark group.—*H. S. Lull.*

12869. CRECCHIA-RISPOLI, C. Di un nuovo genere di "Pristidae" del Cretaceo Superiore della Tripolitania. [A new genus of Pristidae from the Upper Cretaceous of Tripolitania.] *Mem. R. Accad. Italia Cl. Sci. Fis., Mat., Nat. Estratto* 4(1): 1-6, 1 pl. 1933.—**DALFIATZA** (p.1) erected for *D. strumeni** (p.1), Mesostichian, tooth.

Erin, *Isle of Man* 45, p.138-149. In: *Proc. and Trans. Liverpool Biol. Soc.* 46, 1931/1932.—1513 fish were examined. Tables show variability in the number of vertebrae, ranges of body-lengths, ages, and phases of sexual maturity.—*F. LaMonte.*

12873. FRASER-BRUNNER, A. A new species of eel of the genus *Ophichthus* Aul. *Ann. and Mag. Nat. Hist.* 13(76): 465-468, 1 fig. 1934.—*O. serripolus* Richardson, emended descr.; *O. cyclorhinus** (p.486), Queensland.

12874. GÄNDOLFI-ROSTFOLD, A. Les Ophichthes de 8 Anguilles du Caumasse (Grison). *Rev. Suisse Zool.* 40(2): 273-279, 1 pl. 1933.—Eels occurring in Lake Cauma, Switzerland, are thought to have been planted not later than 1887. Since eels do not breed in fresh water it is very likely that these specimens are a part of those introduced in 1887. The ophichthes are more opaque than those found in eels living under normal conditions but are not otherwise different. There is considerable variation in form and size of the ophichthes of the same eel. In the right and left molths of the same eel.—*C. L. Turner.*

12875. GINSBURG, ISAAC. A revision of the genus *Garmannia*. *Bull. Bingham Oceanogr. Coll.* 4(5): 1-3, 3 fig. 1933.—A review of the literature on the genus *Garmannia* shows that the common species on the E. coast of the U. S. have been confused generally and were largely separated hitherto by geographical lines. The intraspecific ranges of variation in the structure of the 3 common spp. of the W. Atlantic are studied in detail largely based on material from the E. coast of the U. S. The species may be distinguished definitely by morphological characters aided to some extent by color differences. The geographic ranges of the separate spp. are overlapping over wide extents of territory; *G. robustum**

The approach

- Use entity extraction to obtain candidate terms from the titles and abstracts
- Map the extracted entities to the BIOSIS vocabulary
- Output the resulting indexing as XML for loading to the Content Management System
- After an RFP process we selected TEMIS & MONDECA as the base technology
- Project kick-off 23 January 2006

Step One – Entity Extraction

- Candidate Term Extraction
 - Off-the-shelf Skill Cartridges™
 - Biological Entity Relationships
 - Medical Entity Relationships
 - Text Mining 360° (Locations)
 - IUPAC Recognizer
 - Custom Skill Cartridges™
 - Enzymes (48,723 terms)
 - Diseases (42,978 terms)
 - Geopolitical (3,066 terms)
 - Anatomical (6,784 terms)
 - Organisms (2,602,844 terms)
 - Geological time (826 terms)

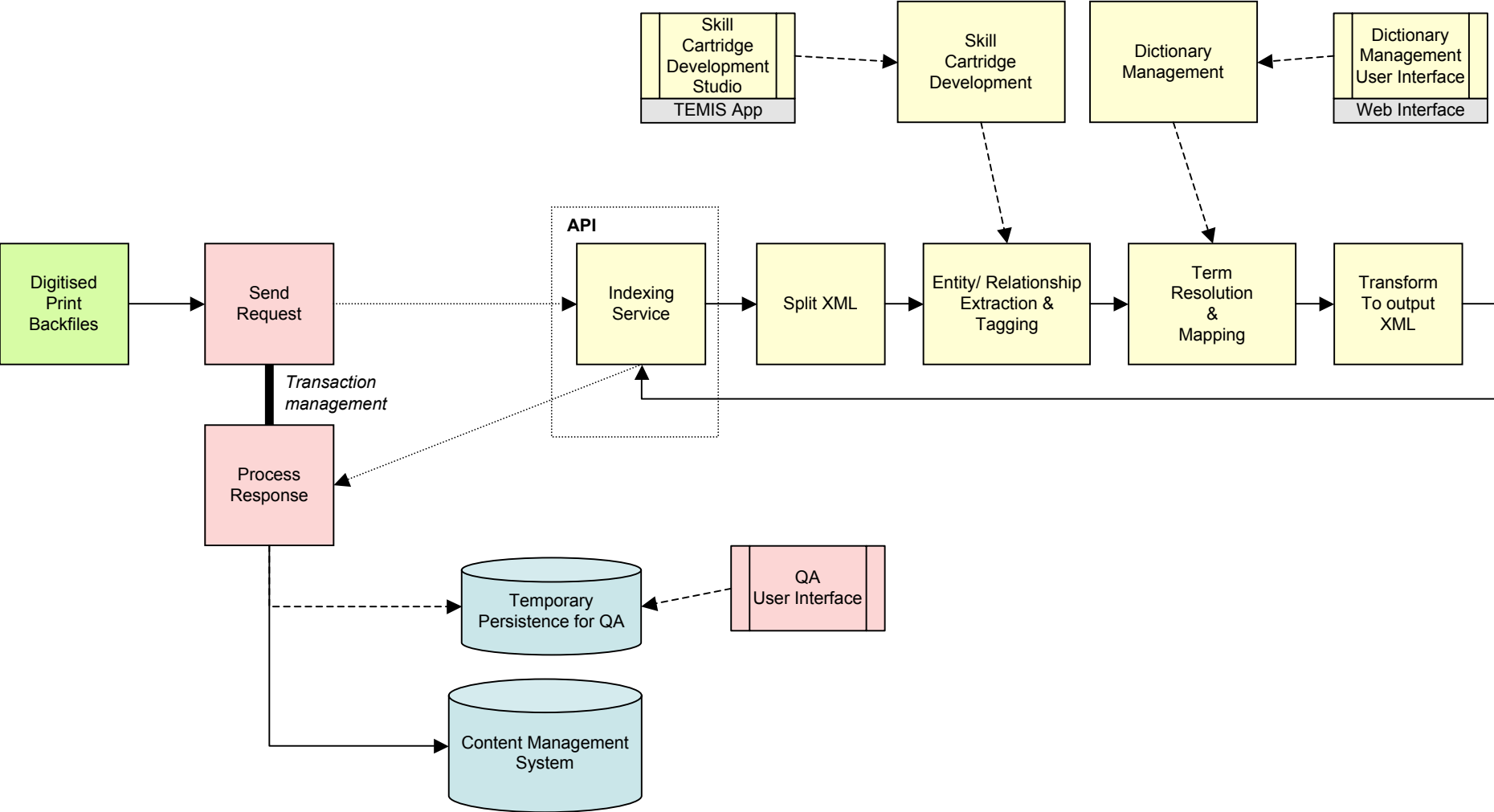
Step two - Refining

- Candidate terms mapped to 6 Term Types
 - (Bio)Chemicals
 - Diseases
 - Geological Time
 - Geopolitical Location
 - Organisms
 - Parts and Structures
- Disambiguation
 - Voting algorithm where term appears in > 1 Term Type
 - Casing
 - Dictionary customisation
 - Filtering

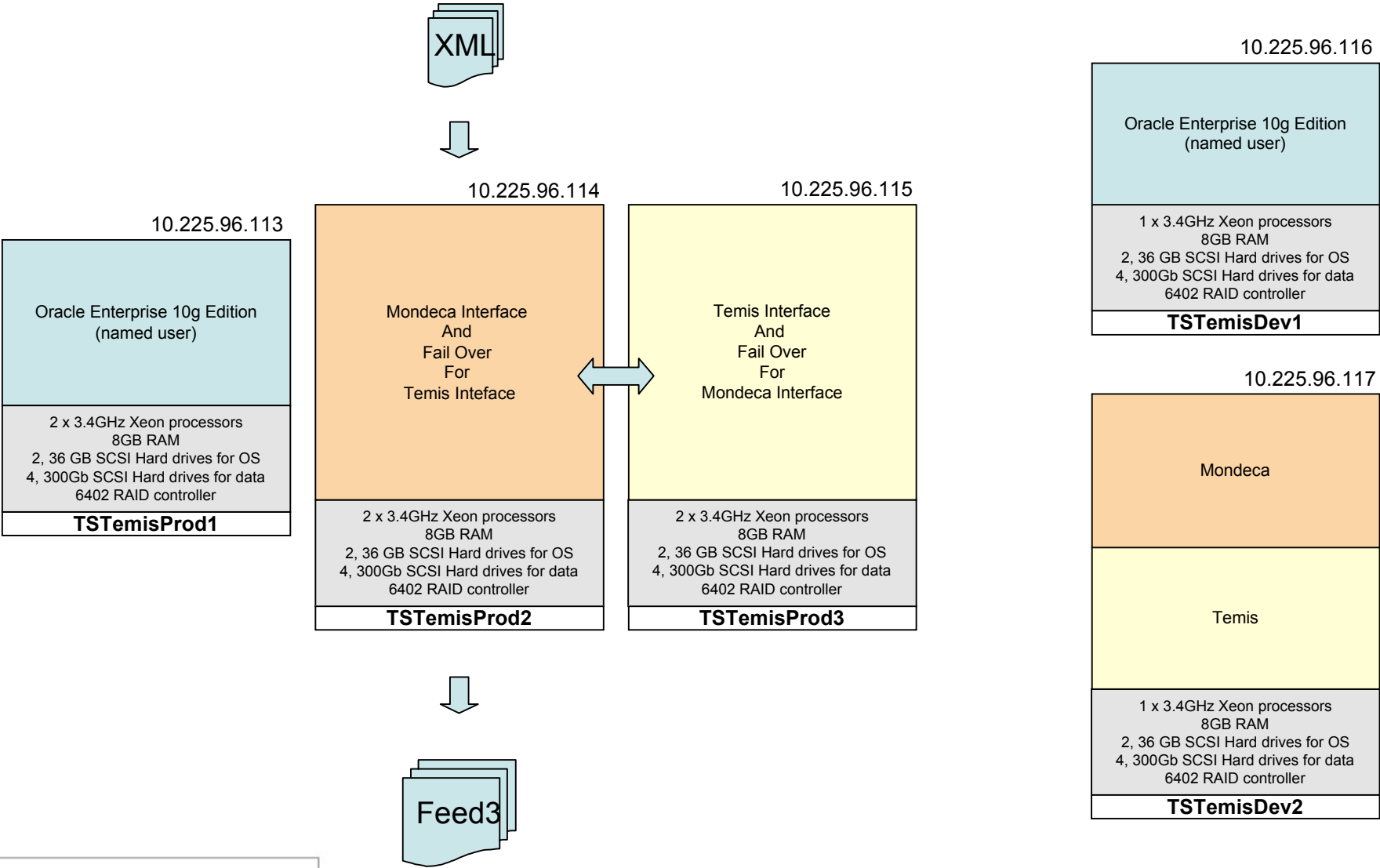
Step three – Term Mapping

- Using Intelligent Topic Manager from MONDECA
 - Mapped print volume headings to BIOSIS Major Concepts & Concept Codes
 - Mapped candidate terms to BIOSIS vocabulary
- Using existing logic in current process
 - Added metadata e.g. Taxa Notes, Super Taxa and Chemical Registry Numbers

Logical Architecture



Physical Architecture



Top Line Results

- Project completed to schedule
 - 26 June (5 months)
- 1.9M indexed records
- All records received at least required minimal indexing
- Throughput: 500ms per item

Sample Output

BIOSIS Previews®

WELCOME HELP GENERAL SEARCH SEARCH HISTORY ADVANCED SEARCH

Full Record

Record 1 of 1 (Set #4)

Accession Number: PREV19350900012868

Document Type: Article

Title: New fishes from the **Triassic** of Pennsylvania

Author(s): [BRYANT, WILLIAM L.](#)

Source: PROC AMER PHIL SOC 73 ((5)) : 319-326 1934

Abstract: A new locality (North Wales, Pennsylvania) for fossil fishes is reported. In addition to the fishes descr. below, dinosaur tracks and plant remains were also found. CARINACANTHUS (p.320) (Elasmobranchii, Cestraciontidae), type *C. jepseri** (p. 320). The remains of Triassic sharks, other than detached teeth and spines, are extremely rare, and this new elasmobranch has the additional interest of being the first such discovery in the Triassic rocks of eastern N. Amer., and seems to indicate either marine or estuarine conditions, a theory of deposition now abandoned, at least for the Connecticut Valley Trias. *Coelacanthus newarki** (p.323) (Coelacanthini). Both new spp. are from the Lockatong Formation of the Newark group. || ABSTRACT AUTHORS: R. S. Lull

MAJOR CONCEPTS: [Systematics and Taxonomy](#)

CONCEPT CODE: [62510_Chordata: general](#)

Taxonomic Data:

SUPER TAXA	TAXA NOTES	Organism Classifier	Organism Name
Pisces, Vertebrata, Chordata, Animalia	Animals, Chordates, Fish, Nonhuman Vertebrates, Vertebrates	Chondrichthyes [85202]	elasmobranch sharks
Monocotyledones, Angiospermae, Spermatophyta, Plantae	Angiosperms, Monocots, Plants, Spermatophytes, Vascular Plants	Orchidaceae [25375]	Trias
Pisces, Vertebrata, Chordata, Animalia	Animals, Chordates, Fish, Nonhuman Vertebrates, Vertebrates	Osteichthyes [85206]	Coelacanthus
Vertebrata, Chordata, Animalia	Animals, Chordates, Fish, Nonhuman Vertebrates, Vertebrates	Pisces [85200]	fish
Plantae	Plants	Plantae [11000]	plant
Vertebrata, Chordata, Animalia	Animals, Chordates, Nonhuman Vertebrates, Reptiles, Vertebrates	Reptilia [85400]	dinosaur

Geographic Data:

Term	GEOPOLITICAL TERMS	ZOOGEOGRAPHICAL REGION
Pennsylvania	USA ; North America	Nearctic region
Wales	British Isles ; UK ; Europe	Palearctic region
Connecticut	USA ; North America	Nearctic region

Output This Record

Bibliographic Fields

PRINT E-MAIL SAVE

EXPORT TO REFERENCE SOFTWARE

SAVE TO MY EndNote Web

[Sign in to access EndNote Web]

Or add it to the Marked List for later output and more options.

ADD TO MARKED LIST

[0 records marked]

Create Citation Alert

CREATE CITATION ALERT

Receive e-mail alerts on future citations to this record. (Requires registration.)

LINKS

View in Web of Science

[Citing Articles](#)

Full Record

Record 6 of 15 [SUMMARY](#)

Accession Number: PREV19381200016071

Document Type: Article

Title: The induction of brooding behavior in the jewel fish

Author(s): [NOBLE, C. K.](#); [KUMPF, K. F.](#); [BILLINGS, V. N.](#)

Source: ENDOCRINOLOGY 23 ((3)) : 353-359 1938

Abstract: Brooding behavior in the cichlid, *Hemichromis bimaculatus*, with spawning experience can be produced experimentally by treatment with a no. of substances. Most effective, in order, when a large series of fish is tested are corpus luteum, proluton and prolactin. Ant. pituitary extract, fresh fish pituitary, thyroxin, desiccated thyroid and phenol in 0.1 or 0.5% soln. are effective to a much less extent. Sexually mature fish without spawning experience cannot be induced to brood with any of the above substances. Previous brooding experience is not essential but greatly increases the response following treatment. At the beginning of their breeding cycle fish require less prolactin than those about to spawn; castrates require less than normals; [male][male] have a lower threshold to prolactin than do [female][female]. | | ABSTRACT AUTHORS: D. Permar

MAJOR CONCEPTS: [Physiology](#)

CONCEPT CODE: [12002, Physiology - General](#)

Taxonomic Data:

SUPER TAXA	TAXA NOTES	Organism Classifier	Organism Name
Pisces, Vertebrata, Chordata, Animalia	Animals, Chordates, Fish, Nonhuman Vertebrates, Vertebrates	Osteichthyes [85206]	<i>Hemichromis bimaculatus</i>
Vertebrata, Chordata, Animalia	Animals, Chordates, Fish, Nonhuman Vertebrates, Vertebrates	Pisces [85200]	fish

Chemical Data:

Chemical Name	CAS Registry No.
phenol	108-95-2
thyroxin	51-48-9
prolactin	9002-62-4

Parts and Structures Data:

Term	ORGAN SYSTEMS
corpus luteum	endocrine system ; reproductive system
pituitary	endocrine system
thyroid	endocrine system

Output This Record

Bibliographic Fields ▼

 ?
[\[Sign in to access EndNote Web\]](#)
 Or add it to the Marked List for later output and more options.
 ?
 [0 records marked]

Create Citation Alert

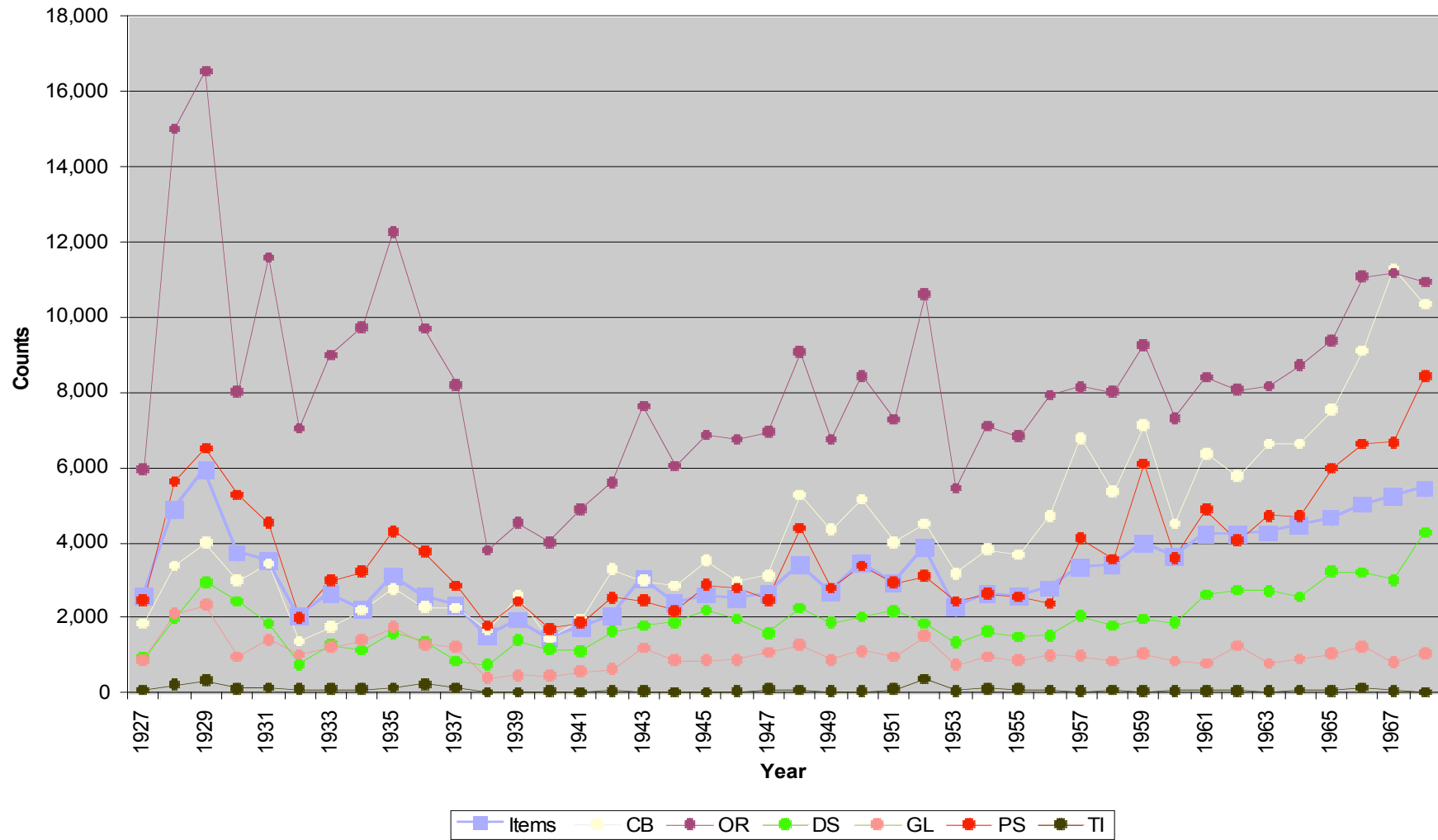
Receive e-mail alerts on future citations to this record. (Requires registration.)

View in Web of Science

[Citing Articles](#)

Frequency distribution of extracted terms (1 issue per volume)

Extracted TT



CB-Chemicals/Biochemicals OR-Organisms DS-Diseases GL-Geopolitical Location PS-Parts & Structures TI-Geological Time

Quality Metrics

Term Type	Rating
(Bio)Chemicals	Poor (+)
Diseases	Very Good
Geopolitical Location	Average

Next Steps

- Entity Extraction
 - Improve chemical name extraction
 - Implement new term recognition
 - Develop additional Skill Cartridges
- Refining
 - Apply contextual data for disambiguation
 - Apply categorization to disambiguation
- Term Mapping
 - Refine vocabulary management
 - Bulk uploading
 - Reporting
 - Multiple vocabularies
 - Improve management of very large vocabularies

Credits

- Stefan Geißler, TEMIS
- Thomas Francart, MONDECA
- Joel Hammond, TS
- Bruce Kiesel, TS
- Dennis Chaney, TS
- Dom Greco, TS