クレジット:
UTokyo Online Education 学術俯瞰講義 2018　美馬秀樹

東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

人工知能と自然言語処理を利用した人文知の構造化

東京大学 大学総合教育研究センター
産業技術総合研究所人工知能研究センター
美馬 秀樹

講義内容

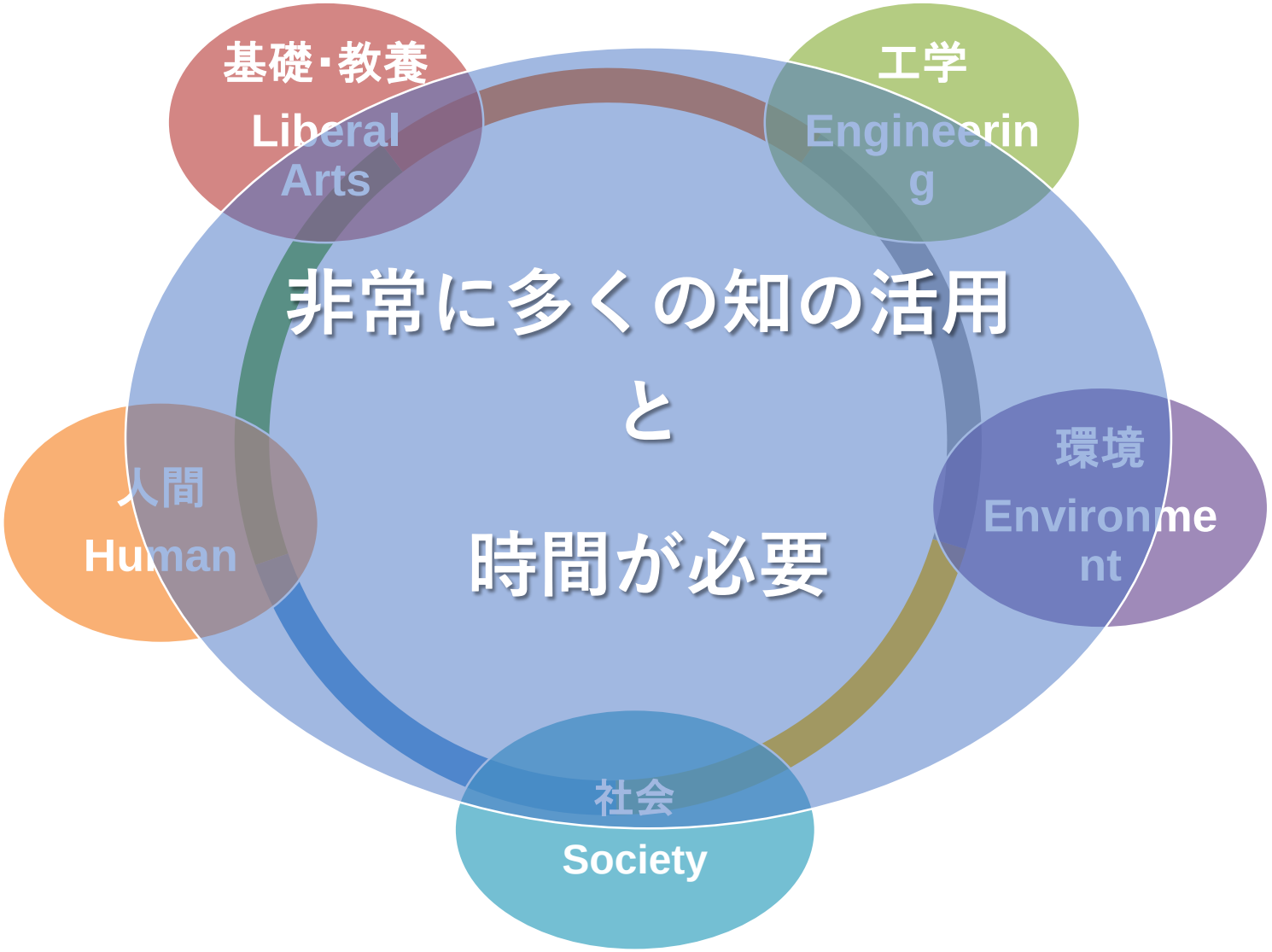
I：学際的教育、研究とデジタルヒューマニティーズ

II：人工知能、自然言語処理、デジタルアーカイブ化

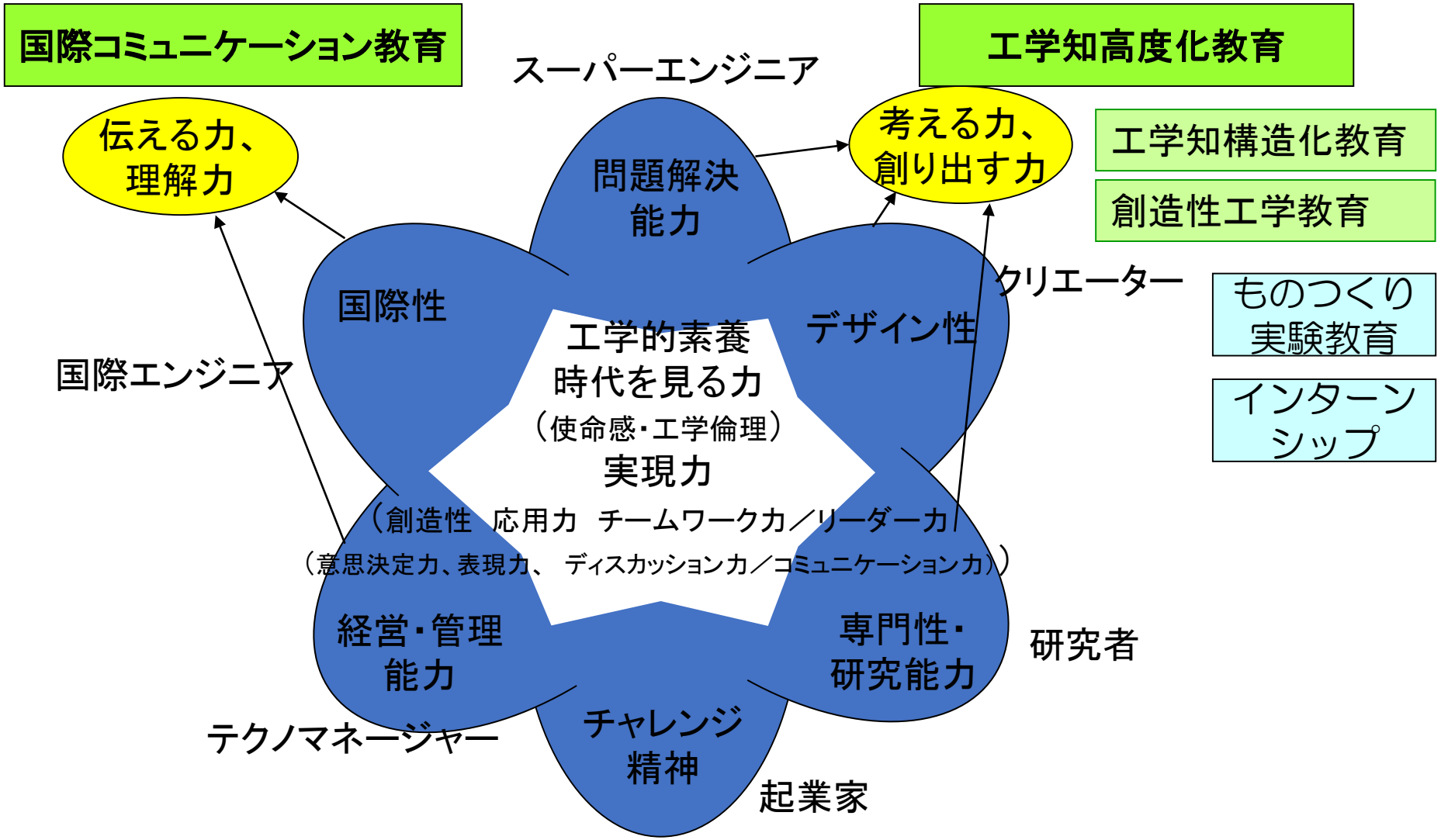
III：知の構造化の背景と岩波書店「思想」のデジタル化

IV：現状と今後

ものづくりと学際性
Creative Engineering

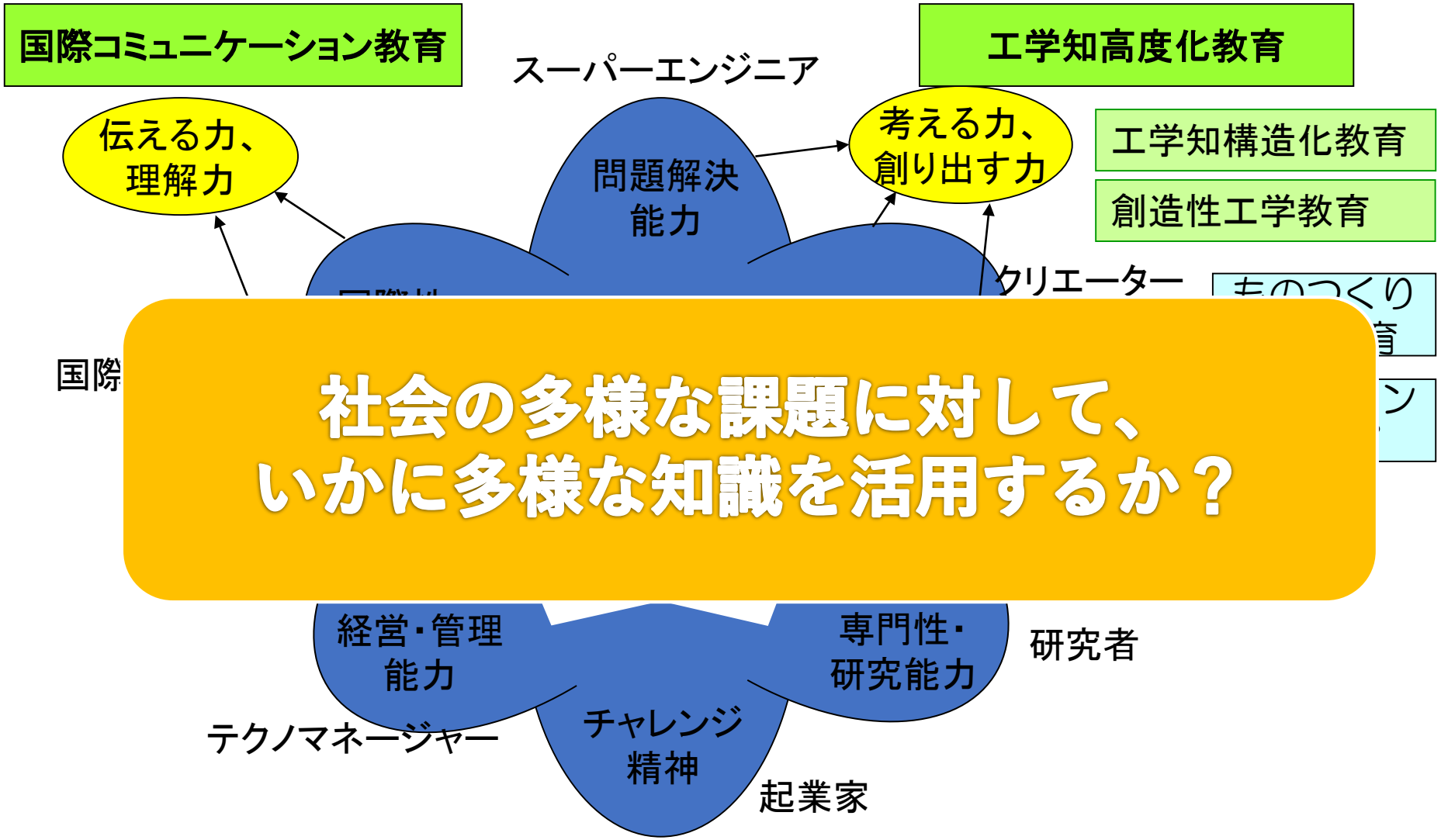


多様でグローバルな人材のモデル



中島尚正編：「工学は何を目指すのか　東大工学部は考える」、東京大学出版会、2000　のP．261に、加筆

多様でグローバルな人材のモデル



中島尚正編：「工学は何を目指すのか　東大工学部は考える」、東京大学出版会、2000　のP. 261に、加筆

学際的教育、研究の必要性

「分野、組織、時勢を越えて、知を活用できるようにする」

文理融合、医工連携

FinTech、EdTech

etc.

背景

- 約１９００万 ／ 毎月6万
  - 医学分野文献データベースに登録された文献数（MEDLINE）／ 毎月の増加数

学問における知識・情報の
幾何級数的な増大
ー量の問題ー

- 約９００／約１２０００
  - 東京大学 工学部の授業数／全学の授業数

学問の専門領域の細分化、
従来の学問体系では対応できない
複雑な問題の登場
ー質の問題ー

CVDの論文の数

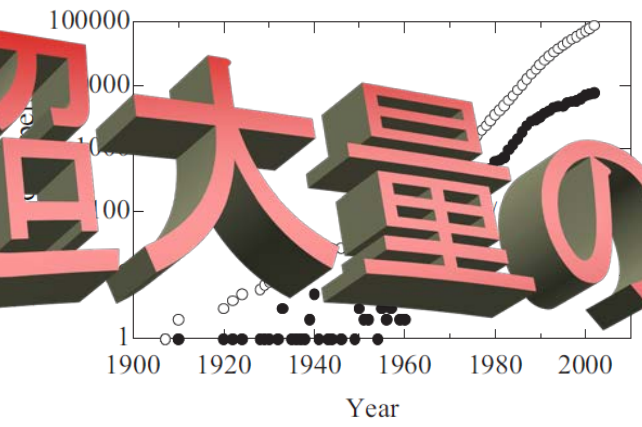
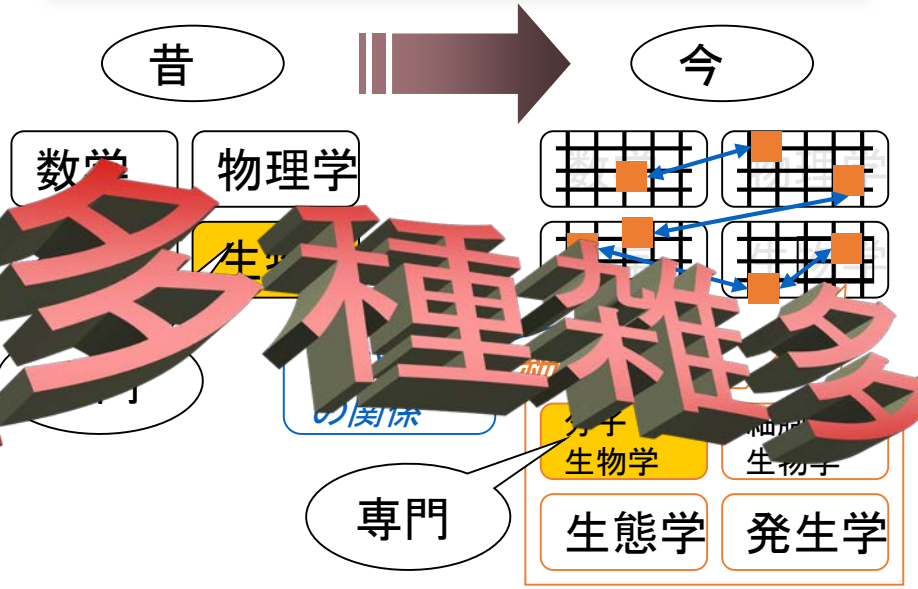


Fig. 1. Semi-log plot of the number of papers related to CVD. The closed circles are the number of papers published in that year. The open circles are the cumulative number of papers published up to and including that year.



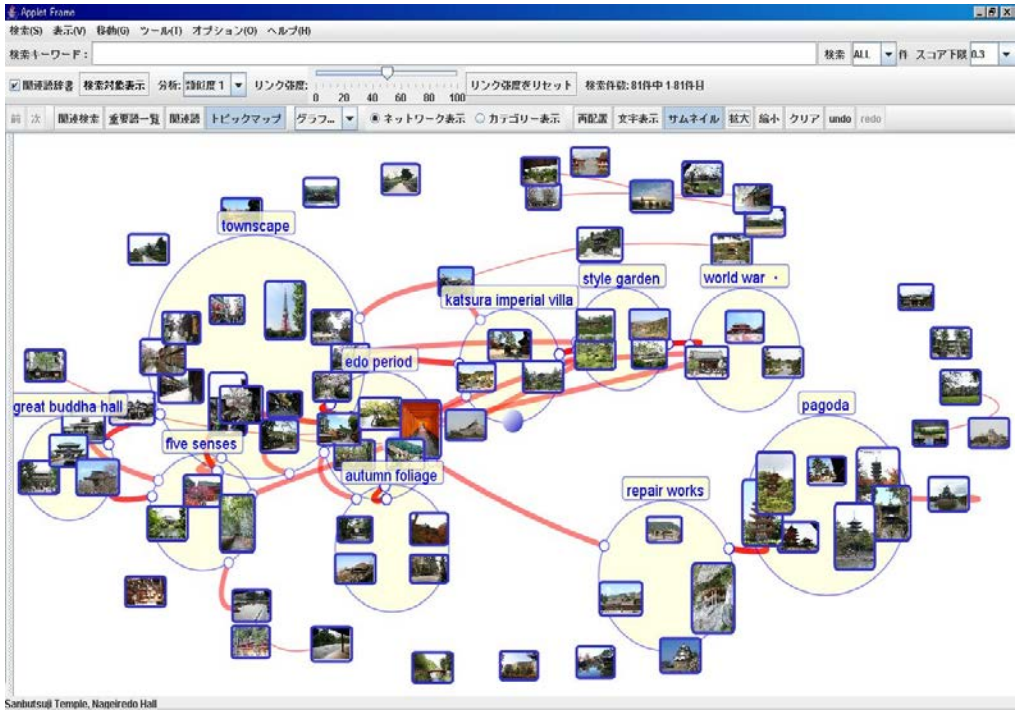
超大量の知 多種雑多

Y. Takeda, Y. Kajikawa, K. Matsushima
Citation Network of CVD Research: Research Topics and Journals
Chemical Vapor Deposition（2007）

知の活用の技術

東京大学　美馬秀樹

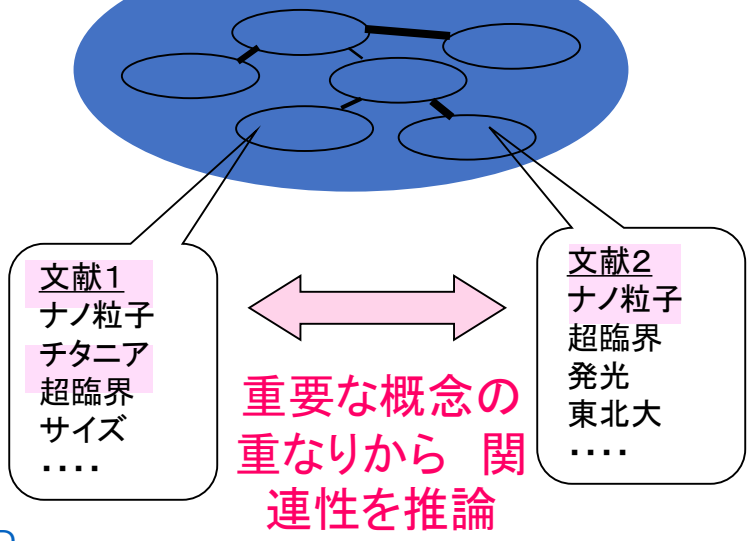
1. 知識を収集・蓄積し、
2. 知識を分析・分類し、
3. 知識を可視化する。



テキストを解析し、

重要な概念を抽出し

関連性の認識を行う



知の活用の技術

東京大学　美馬秀樹

1. 知識を収集・蓄積し、
2. 知識を分析・分類し、
3. 知識を可視化する。

情報技術の活用

テキストを解析

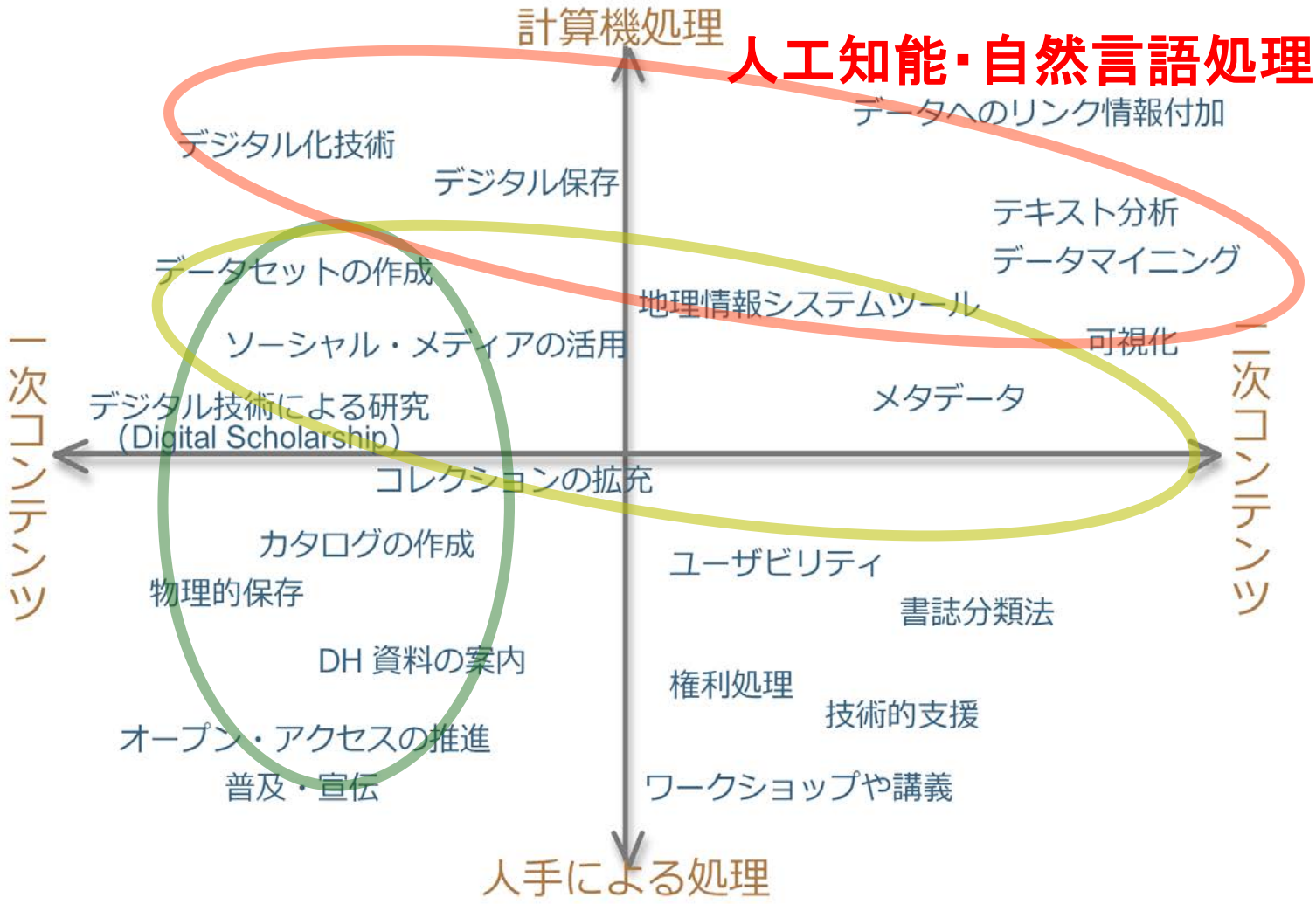
の認識を行う

文献1
ナノ粒子
チタニア
超臨界
サイズ
・・・・

重要な概念の
重なりから　関
連性を推論

文献2
ナノ粒子
超臨界
発光
東北大
・・・・

DHの構造



Sula, Chris Alen, 2013, "Digital Humanities and Libraries: A Conceptual Model," *Journal of Library Administration*, 53(1) より

人工知能でできること

- 探索と最適化
  - 路線検索
  - 将棋、チェス
  - 巡回セールスマン問題
  - 最適配置
- 認識
  - 五感の認識＋α
    - 音声認識
    - 文字認識、顔認識、物体認識
    - 味覚、におい、触覚
  - 認知
    - 言葉の理解
    - 感情の理解
- 分類
  - 行動モデリングと予測
  - リコメンデーション

類推 →
似た情報を同様に処理

計算機の発展 →
大量の情報を高速に処理

最適化の例(本郷三丁目～駒場東大前)

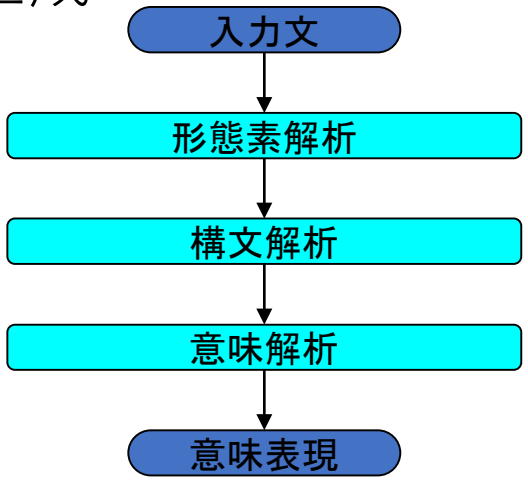
- 経路探索
  可能なルートから最適なものを選ぶ
  - 評価軸
    - 運賃
    - 時間
    - 乗換回数
    - CO2量　etc.

NAVITIME
https://www.navitime.co.jp/

(c)OpenStreetMap contributors

自然言語処理（NLP）

- **計算機を用いて言語の理解を行う**
  - 形態素解析　－単語（形態素）に区切る
  - 構文解析　－語構成、文の構成（主語、述語等）
  - 意味解析　－意味表現の生成



- **アプリケーション**
  - 仮名漢字変換
  - 機械翻訳
  - 用語（概念）抽出
  - 全文検索システム

計算機の発展 → 大量のテキストを高速に処理

自然言語処理と人工知能

- 「本郷三丁目から駒場東大前までの電車」

現在地	行き先	手段
本郷三丁目	東大駒場前	電車

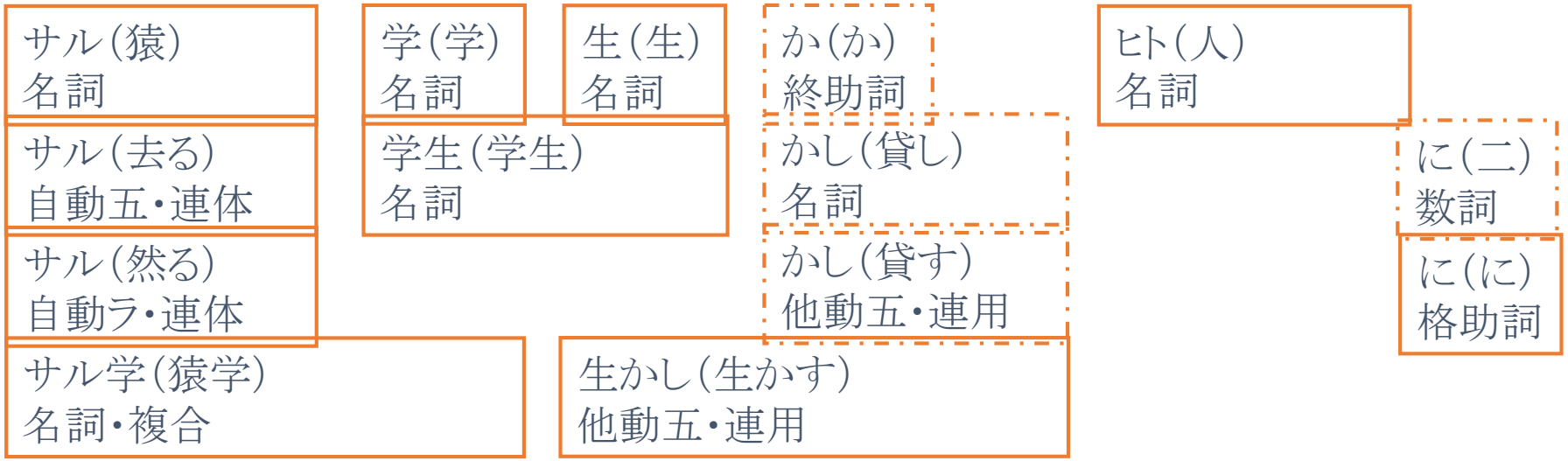
形態素解析

- 文を形態素（単語）に分け、品詞等の属性情報を同定する
  例：「本郷三丁目から駒場東大前までの電車」

形態素の表示				
表記	**よみ**	**品詞**	**基本形**	**全情報**
本郷三丁目	ほんごうさんちょうめ	名詞	本郷三丁目	名詞,地名,*,本郷三丁目,ほんごうさんちょうめ,本郷三丁目
から	から	助詞	から	助詞,格助詞,*,から,から,から
駒場東大前	こまばとうだいまえ	名詞	駒場東大前	名詞,地名,*,駒場東大前,こまばとうだいまえ,駒場東大前
まで	まで	助詞	まで	助詞,副助詞,*,まで,まで,まで
の	の	助詞	の	助詞,助詞連体化,*,の,の,の
電車	でんしゃ	名詞	電車	名詞,名詞,*,電車,でんしゃ,電車

形態素解析の例

- 例文(新聞の見出しから):
  サル学生かしヒトに平和を



構文解析の例

日本語パーザ オンラインデモ on Enju 2.4

東京大学 知の構造化センター美馬研究室で研究を進めている日本語CCGパーザのオンラインデモサイトです。

文を入力するとXML形式で解析結果を見ることができます。

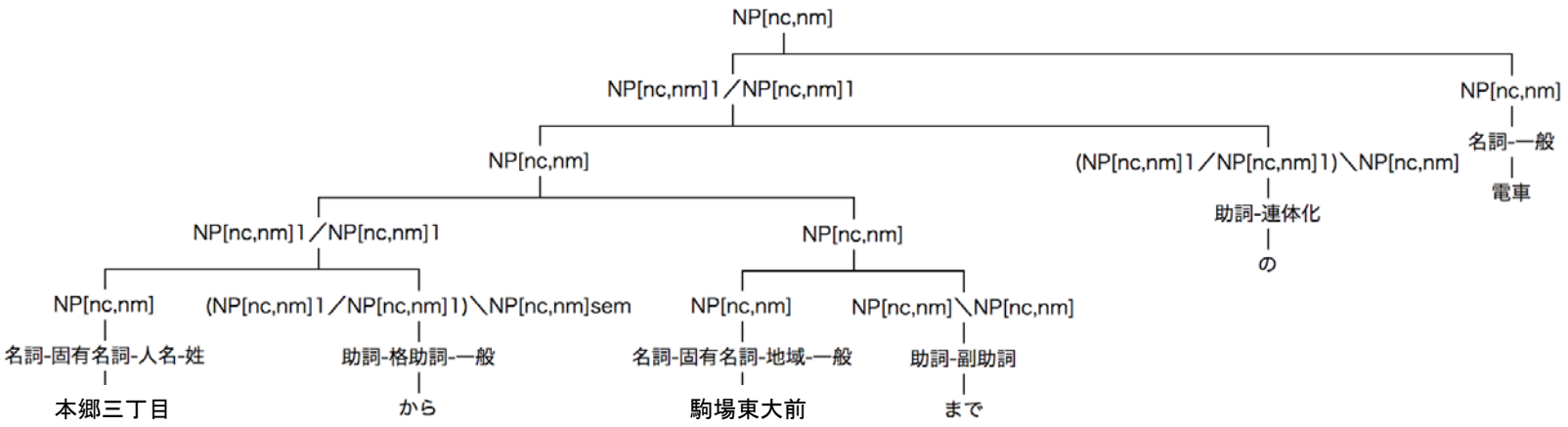
本郷三丁目から駒場東大前までの電車　解析　クリア

例文

- 気象庁は強風や高波に注意するよう呼びかけています。
- 前半のショートプログラムで首位に立った浅田選手は、２０日に行われたフリーで出場した１０人の最後に滑りました。

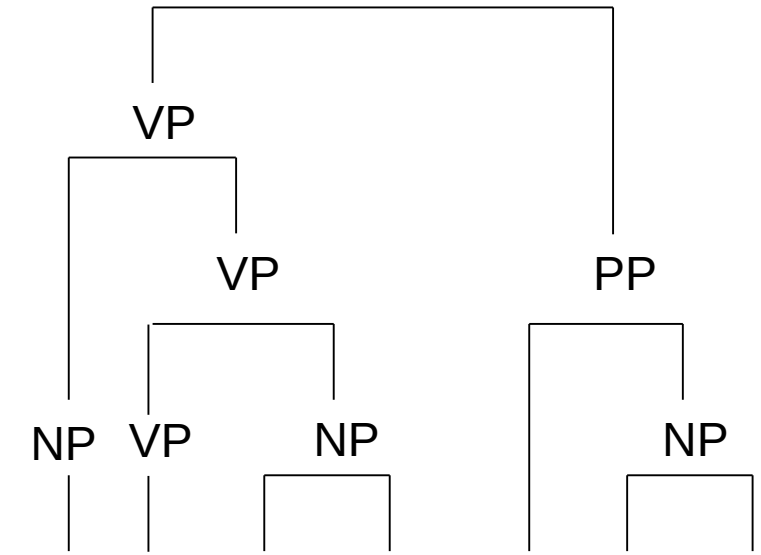
Parse status: success

Click to change the view.



構文解析とあいまい性

- 文の統語的構造を同定する
  - 係り受け関係
- 例：「I saw a man in the park」

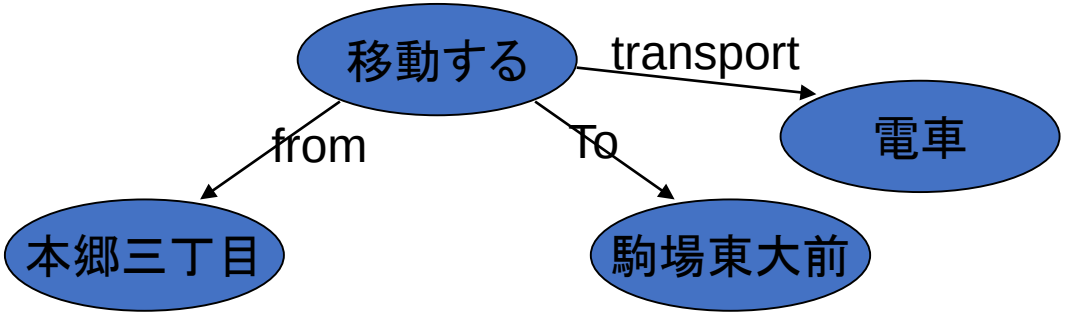


文法ルール

S ← VP PP
VP ← VP NP
VP ← V
NP ← det N
NP ← N
PP ← P NP

意味解析

- 文の構造から意味表現を同定する（構造化）
- 文の意味とは
  - 意味ネットワーク



  - フレーム
    - (移動する
      (from X?)
      (to Y?)
      (transport Z?)
)

実体化 →

(移動する-001
(from 本郷三丁目-001)
(to 駒場東大前-001)
(transport 電車-002)
)

構造化

現在地	行き先	交通手段
本郷三丁目-001	東大駒場前-001	電車-002

実検索

最適化の例(本郷三丁目～駒場東大前)

NAVITIME
https://www.navitime.co.jp/

- 経路探索
  可能なルートから最適なものを選ぶ
  - 評価軸
    - 運賃
    - 時間
    - 乗換回数
    - CO2量　etc.

(c)OpenStreetMap contributors

「電車で駒場東大前まで本郷三丁目から行きたい」→ 同じ結果

意味のあいまい性

- 「apple」
  - 果物？　コンピュータ会社？
- 「Time flies like an arrow」
  - time – 時間、計る
  - fly – 飛ぶ、ハエ
  - like – 好き、～のように
- 「ぼくはキツネ」
  - 状況により様々な意味

文脈（コンテクスト）を理解すること

- 「僕はキツネ！」

イラスト：いらすとや

岩波「思想」の構造化プロジェクト

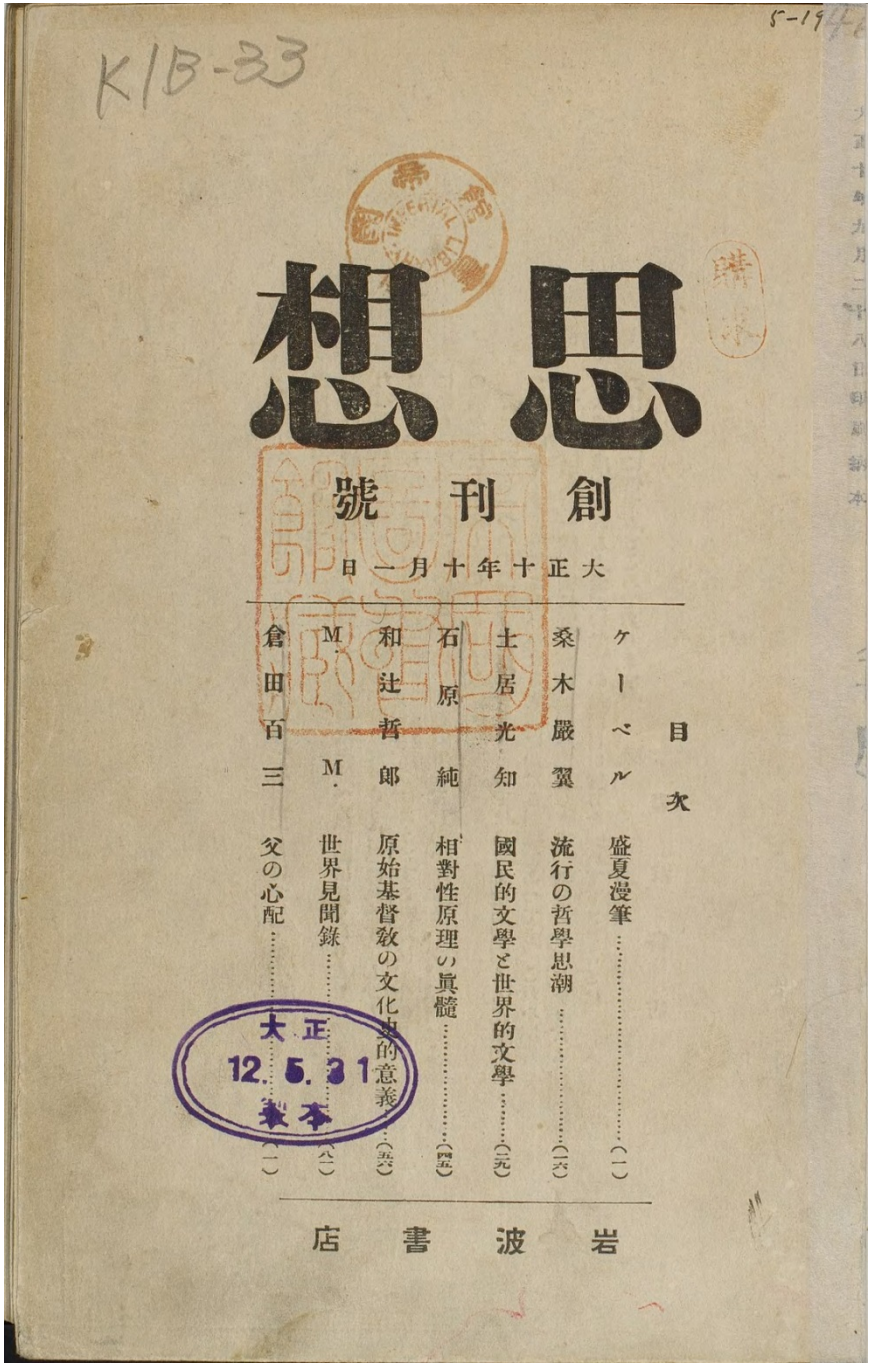
- 構想
  - 日本を代表する思想・哲学ジャーナル『思想』（1921年創刊）の80年分（約900号、約9,000論文、 約16万ページ、かつ1986年までは活版印刷）
  - **デジタル化と「知の構造化」**を行う

情報技術を活用して分析をしたい

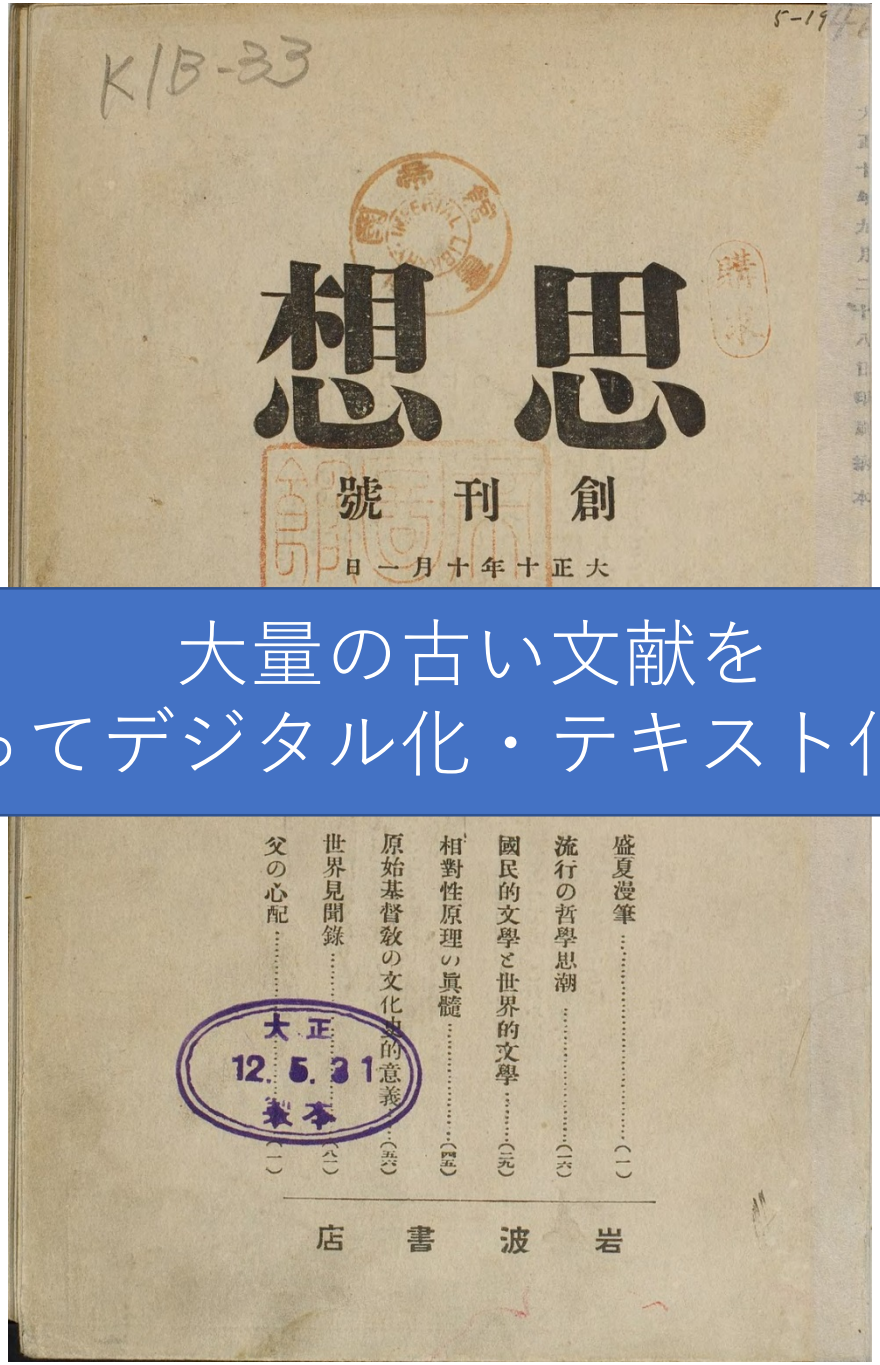
デジタルのテキストデータがない

そもそもデジタルのデータがない

- 目的
  - 『思想』という知の集積と20世紀日本の哲学・思想史を明らかにする
  - 文献のデジタル化に関する方法論を確立する



『思想』創刊号、岩波書店、1921年



大量の古い文献を
どうやってデジタル化・テキスト化するか？

『思想』創刊号、岩波書店、1921年

テキストのデジタルアーカイブ

- コンピュータで処理できるテキストを構築

→検索や自然言語処理による解析が可能

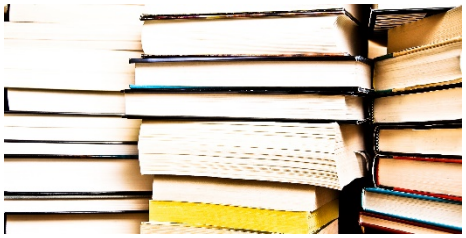
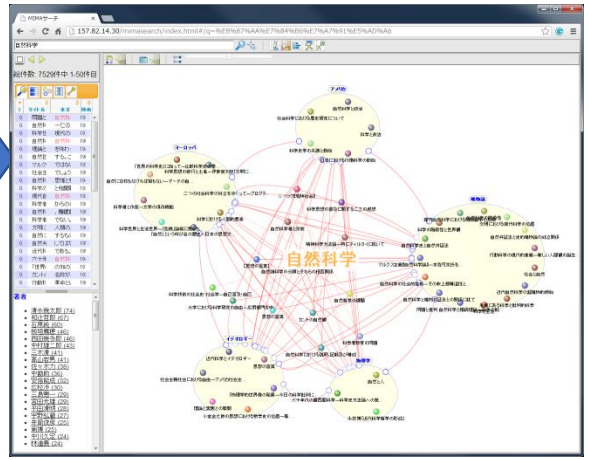


Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

分析システム



大規模なデジタルテキスト化

- **大量**の文献を**高精度**にテキスト化したい

大量の書籍・出版物

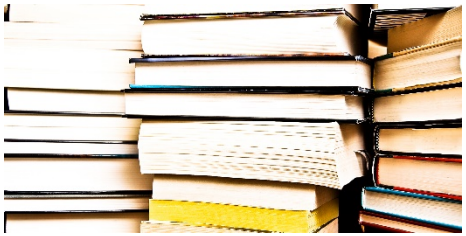


Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0

?

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

大規模なデジタルテキスト化

大量の書籍・出版物

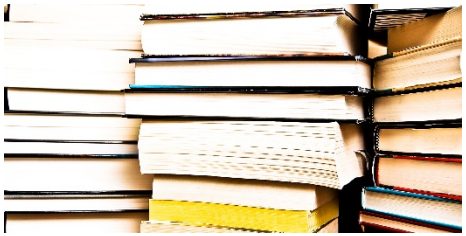


Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0

©いらすとや

人手入力
利点: ほぼ100%の精度
欠点: コストが高い

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

大規模なデジタルテキスト化

大量の書籍・出版物

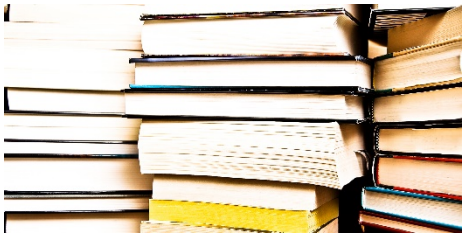


Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0

©いらすとや

音声認識
利点: 手入力よりも低コスト
欠点: 精度が低い

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

大規模なデジタルテキスト化

大量の書籍・出版物

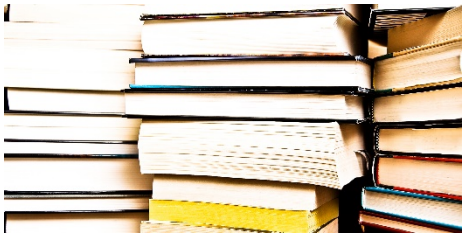
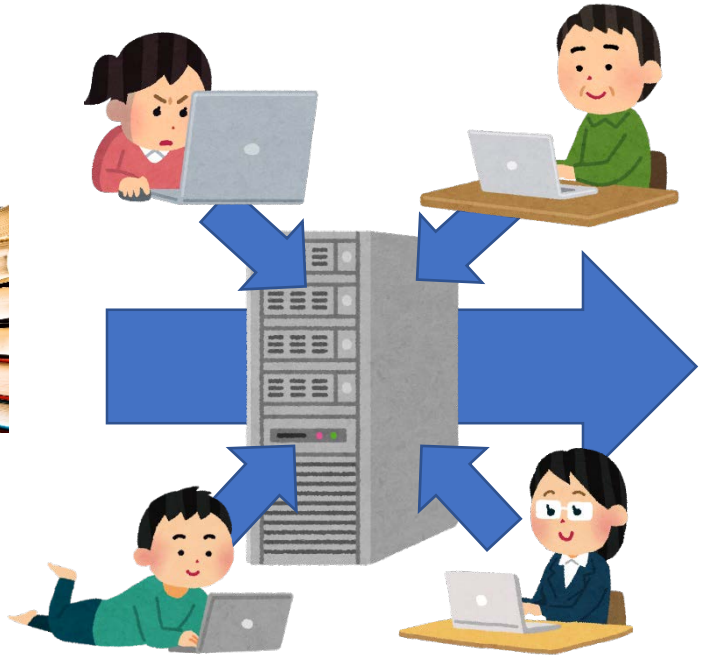


Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0



©いらすとや

手入力(クラウドソーシング)
利点: 高精度・低コスト
欠点: 質の統一が困難
データの管理

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

クラウドソーシング例
NDLラボ 翻デジ

翻デジ2014
http://lab.ndl.go.jp/dhii/omk2/

検索 +

アイテム | コレクションをブラウジング | Scripto | 翻刻テクスト

はじめに

2014年3月24日
一般財団法人人文情報学研究所主席研究員・東京大学大学院情報学環特任准教授・国立国会図書館研究員 永崎研宣

以下の説明の簡単な要約:
現状では、近代デジタルライブラリー(以下、近デジ)をテキストデータ化して色々便利に使えるようにするためのプラットフォームとなっております。関係者それぞれに多忙ななかでの片手間の運営となっておりますのでおおらかな気持ちで接していただけますと幸いです。徐々に色々と改良・発展していく予定です。それから、Mediawiki APIを使える方には少し面白いかもしれません。使い方はこのページの下の方にあります。

- このシステムは日本デジタル・ヒューマニティーズ学会(JADH)の分科会であるSIG-Transcribe JPが提供しています。
- このシステムは国立国会図書館次世代システム開発研究室が運営するサーバ上で動作しています。
- このシステムはジョージ・メイソン大学で開発されたOmeka(Omekaの説明の日本語訳)というGLAM向けコンテンツマネジメントシステム(CMS)を核として構築されています。データの蓄積はMediawikiに行なわれるようになっています。
- このシステムでは、各資料のページ毎にデジタル翻刻を行っていただきますと、**デジタル翻刻テクストの各頁から近デジの当該頁に自動的にリンクが張られる**ようになっており、翻刻テクストの頁から容易に実際の本文を画像で確認できるようになっています。したがって、翻刻テクストとしての正確性がそれほど高くなくとも検索用途等としてある程度利用可能なものを作成することができるということになっております。
- このシステムは以下のことを**目標としています**。
  - **近代デジタルライブラリー等の画像でしか提供されていない日本語**

最近追加されたアイテム

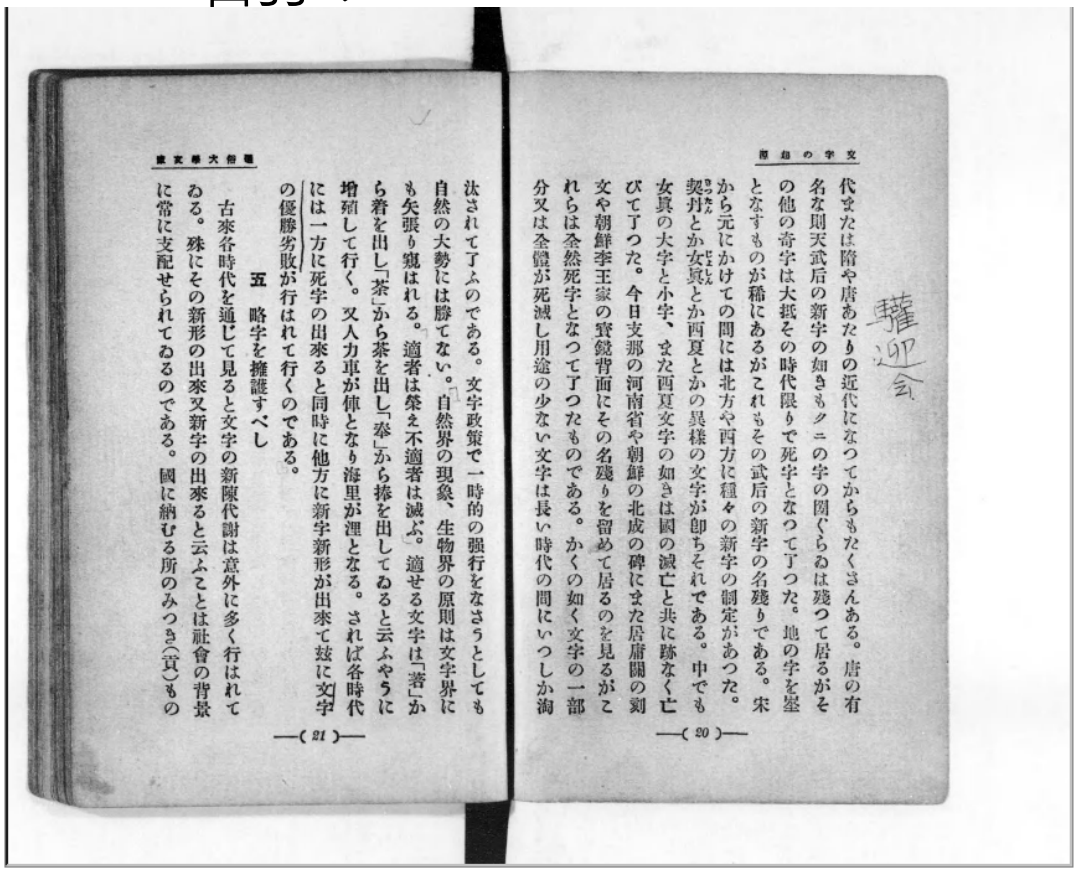
道後温泉誌

天竺徳兵衛

あさどり

すべてのアイテムを見る

クラウドソーシング例
NDLラボ 翻デジ



国立国会図書館デジタルコレクションより引用

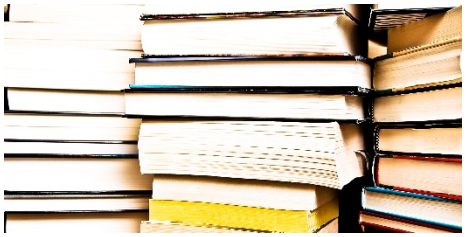
この頁を電子翻刻 [履歴]

新字旧字混在 ▾　仮名遣い混在 ▾　Wikiのみ ▾

代または隋や唐あたりの近代になつてからもたくさんある。唐の有名な則天武后の新字の如きもクニの字の圀ぐらゐは殘つて居るがその他の奇字は大抵その時代限りで死字となつて了つた。地の字を[＃「山＋水＋土」、20-3]となすものが稀にあるがこれもその武后の新字の名殘りである。宋から元にかけての間には北方や西方に種々の新字の制定があつた。契丹とか

大規模なデジタルテキスト化

大量の書籍・出版物



スキャン+ 文字認識
利点: コストが低い
短時間で処理可能
欠点: 精度が低い
→技術発展により向上が望める

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

デジタルテキスト化

- 印刷された文字をコンピュータで処理可能なデジタルテキストに変換
  - 画像化: スキャン
  - テキスト化: OCR(Optical Character Recognition)

書籍・出版物

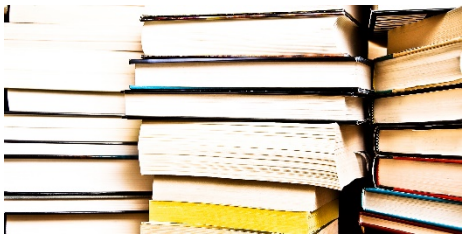
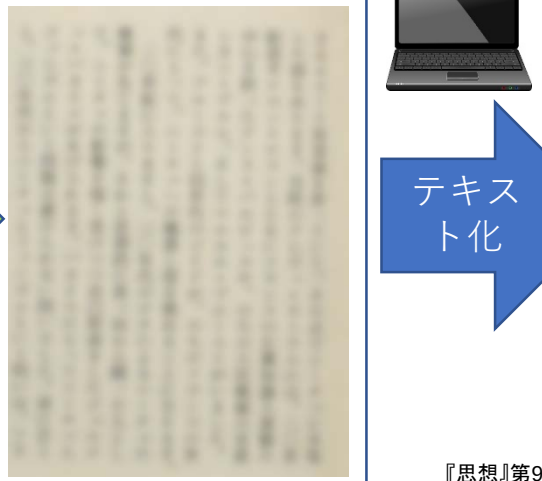


Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0

画像化

画像データ



テキスト化

テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

『思想』第919号(2000年12月)
鵜飼哲・長原豊・三島憲一「座談会　ニーチェ　その魅力と魔力」p13より引用

デジタル画像化

- 対象物のスキャンによりデジタルデータ化
  - スキャナの種類(スキャン形態による分類)
    - フラットベッドスキャナ
      - 本をガラスの台に固定し、
        下から光を当てて読み取る
    - シートフィードスキャナ
      - 自動原稿送り装置で
        原稿を送り、読み取る
    - カメラスキャナ(スタンドスキャナ)
      - 原稿を専用の台に置き、カメラで撮影
      - 本の非破壊スキャンが可能

著作権等の都合により
ここに挿入されていた画像を
削除しました

フラットベッドスキャナの画像

http://cweb.canon.jp/imageformula/lineup/p215ii/index.html
Canon DR－p215

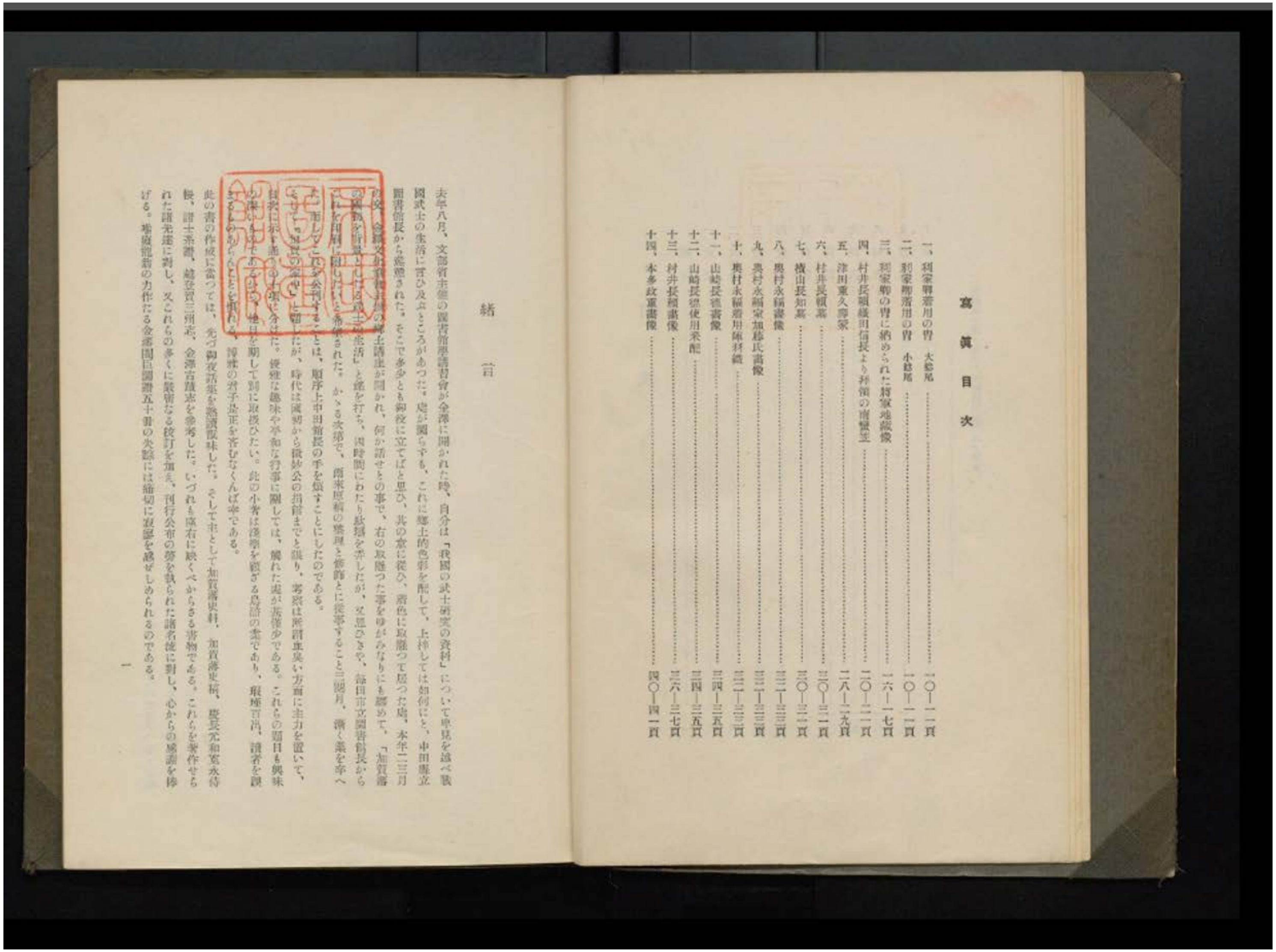
カメラスキャナ

キルタス社製BookScanner（APT BookScan 2400RA）
出典：株式会社ブライトホープhttp://www.bright-hope.co.jp/cnt_02.html

自動的に本をめくり
上にある2台のカメラで撮影

本を痛めることなく
自動でスキャンが可能

カメラスキャナによるスキャン例



平らな台に置いて見開きを撮影

国立国会図書館デジタルコレクション
岡本勇著『加賀の家中』(昭和10年)

カメラスキャナ？

- Scansnap SV600
  - 実はカメラではなくラインセンサーでスキャン
  - ラインセンサーと
    ライトが首を振る
  - 機能としてはカメラ
    カメラスキャナと同様

FUJITSU
ScansnapSV600
充実したソフトウェア
http://scansnap.fujitsu.com/jp/product/sv600/function.html

Book Turner

CASIO
https://web.casio.jp/bookturner/

テキストデータの取得

- 印刷された文字をコンピュータで処理可能なデジタルテキストに変換
  - 画像化: スキャン
  - テキスト化: OCR(Optical Character Recognition)

『思想』第919号(2000年12月)
鵜飼哲・長原豊・三島憲一「座談会　ニーチェ　その魅力と魔力」p13より引用

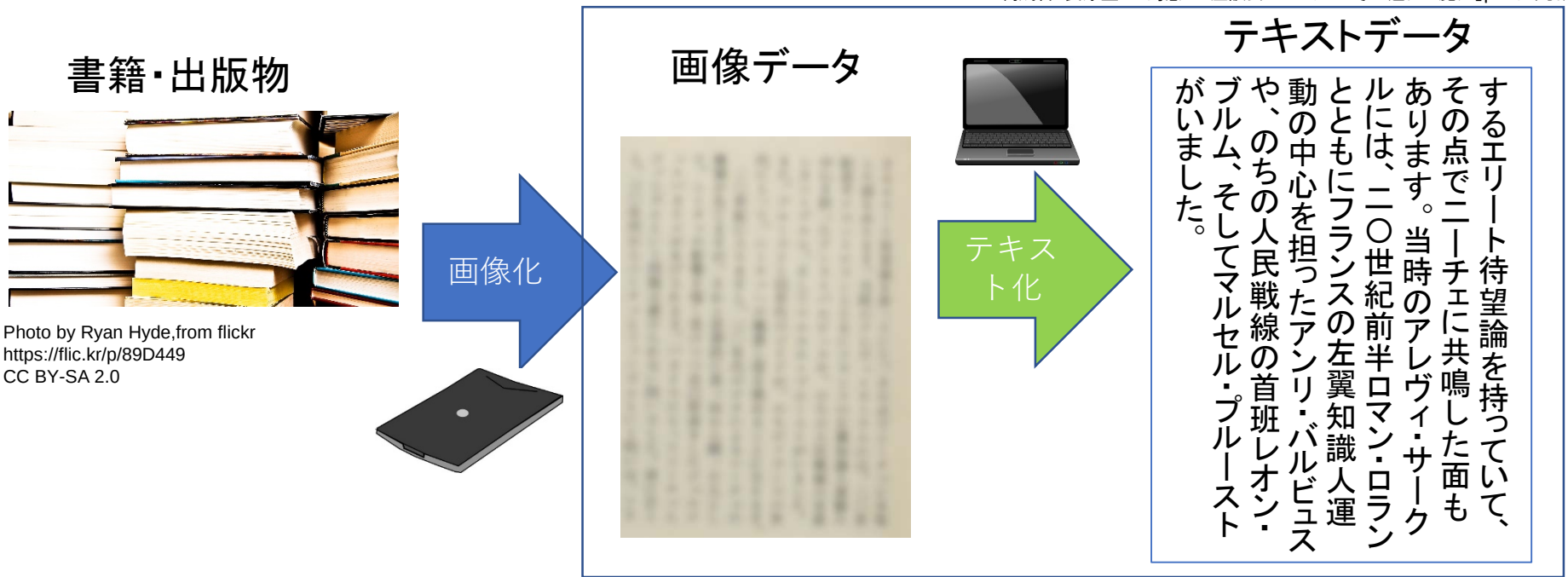
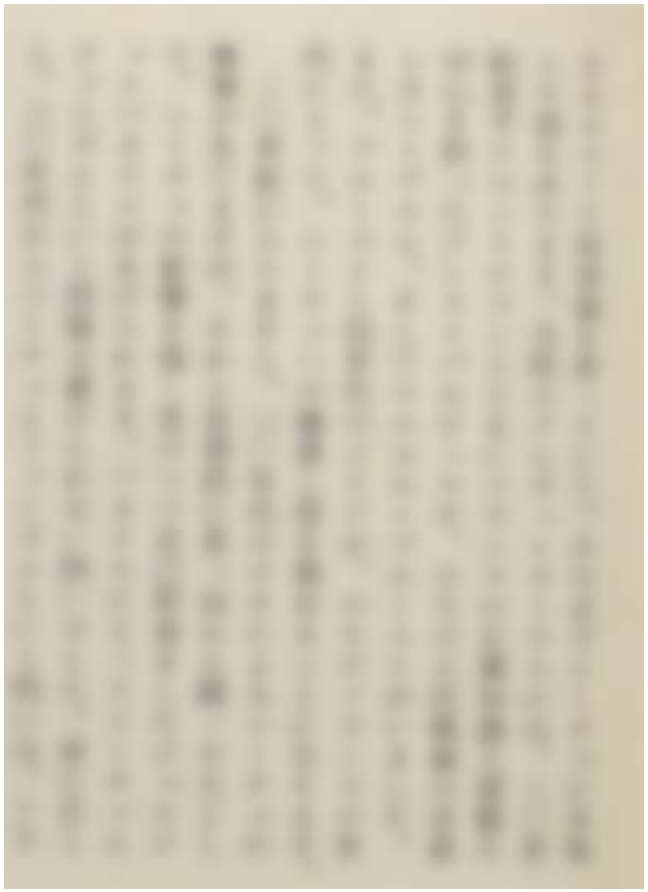


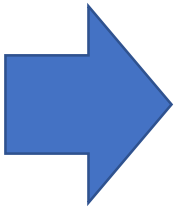
Photo by Ryan Hyde,from flickr
https://flic.kr/p/89D449
CC BY-SA 2.0

デジタルテキスト化: OCR(Optical Character Recognition)

『思想』第919号(2000年12月)
鵜飼哲・長原豊・三島憲一「座談会　ニーチェ　その魅力と魔力」p13より引用

- デジタル画像から文字を認識する





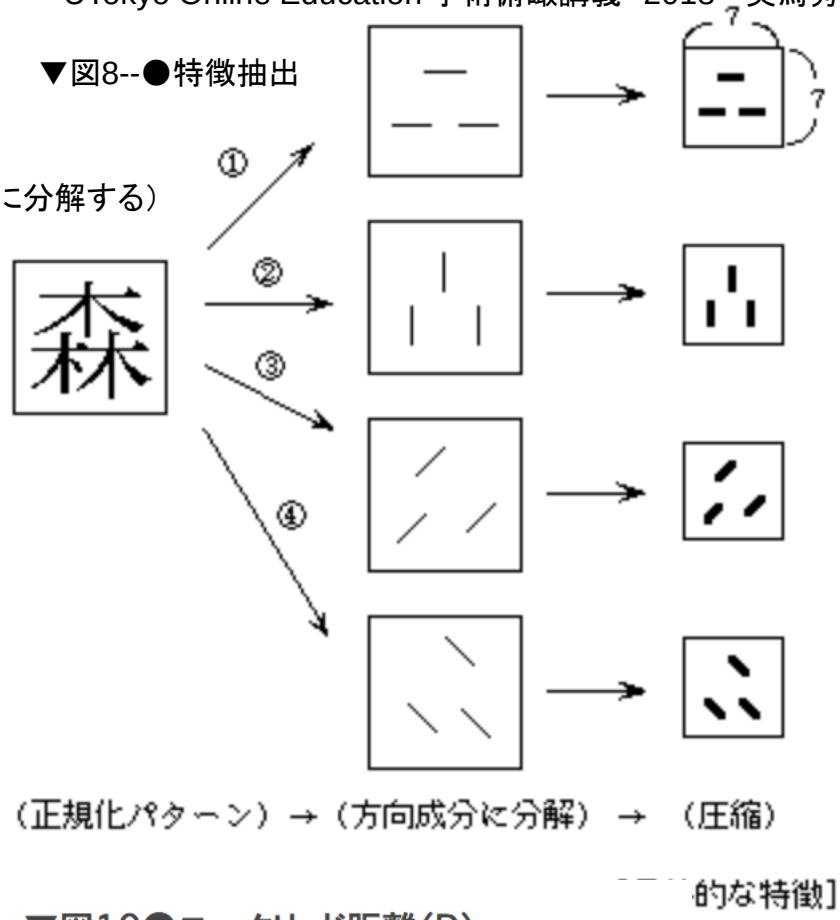
するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。また、ブルーストと同世代のジイドが、のちのフランスの世代にとって、ニーチェへの橋渡し役を務めることになります。二〇世紀に入りますと、三〇年代のナチによるニーチェの横領が生じますが、それと思想的に真っ向から闘った人として、ニーチェの影響を強く受けつつ自己形成をしたジョルジュ・バタイユがあげられます。バタイユにとってニーチェとファシズムという問題は避けられない問いでした。逆に言うと、三〇年代からニーチェとファシズムという問いは、フラ

OCRシステム

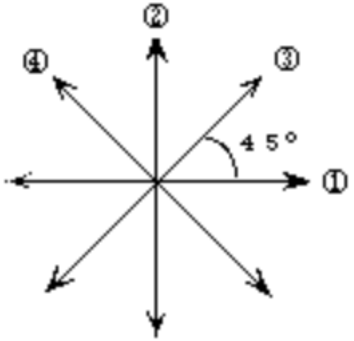
- 市販ソフト
  - メディアドライブ WinReaderPRO、 e.Typist
  - ABBYY FineReader
  - パナソニック 読取革命
    ・・・
- サービス
  - Google Cloud Vision API
  - Evernote API
    ・・・

OCRの仕組み

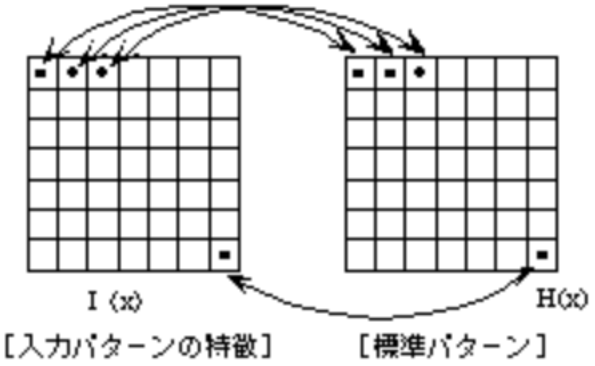
▼図8--●特徴抽出



▼図9●4つの方向（1～4の方向に近い線分に分解する）



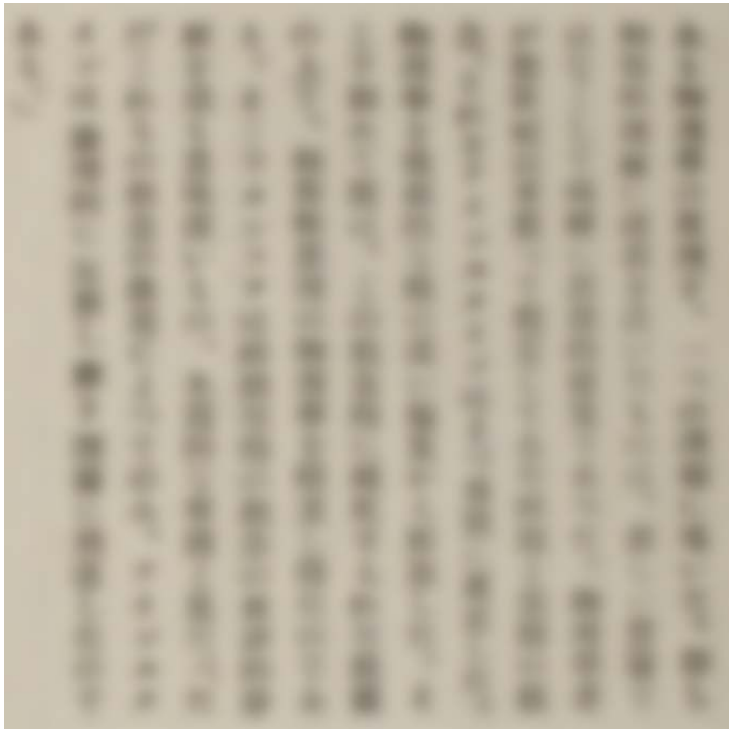
▼図10●ユークリッド距離（D）



出典:
https://mediadrive.jp/technology/techocr05.html
図8-10

古い文献のデジタル化

- デジタルアーカイブの対象は古い文献が多い



ゐる物理學の原理を、一つの調和に導いだ。即ち相對性理論に礎石を置いたものは、新しい實驗ではなぐして純粹し思想的探究であつた。物理堅者が數世紀以來默つて暇定してゐた時間と筌間の概念、それをアインスタイソはまつ勇敢に更正した。物理學を傳統的な根の深い偏見から解放した。そこで初めて彼は、この概念的に純化せられた根據の上で、相對牲原理の物理學を樹立し得たのである。そこでカシラアは時間塞間の概念の批評的分析を最も意康深いもの、本質的な契機と見た。「ただこれらの概念の改造によつてのみ、アイソスタイソは論理的に反對し難き理論に到達したのである。」

古い文献では精度が低下

『思想』創刊号、1921年
MM「世界見聞録」p82より引用

OCR精度の向上による言語処理結果の向上

- OCR精度

  95%　から　99.85%へ

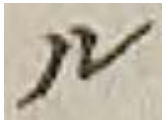
OCR精度95%での用語抽出結果

	1930年	1931年	1932年	1933年	1934年	1935年	1936年	1937年	1938年	1939年
1	自然 科学	日本	日本	**道 穂**	日本	自己 自身	西田 哲学	**道 穂**	アメリカ	自己 同一
2	資本 主義	カント	ギルヘルム	フランス	宗 敦	自己 限定	日本	自己 自身	英国	日本
3	日本	アメリカ	**ブルジ**	自然 科学	フランス	**道 穂**	自己 自身	日本	日本	**道 穂**
4	フランス	自己 自身	カント	カント	資本 主義	日本	自己 限定	アントニ	自己 同一	プラトン
5	**道 穂**	**ブルジ**	**宗 敦**	日本	**ブルジ**	**根 附**	寺田 先生	**圭 体**	フランス	**宗 歓**
6	ブルジ	自己 限定	自然 科学	**ブルジ**	昭和 八	自然 科学	フランス	職業 生活	イギリス	ピュタゴラス
7	**回顧**	**宗 敦**	ファウスト	アジア	**道 穂**	フランス	自己 矛盾	フランス	ヨーロッパ	自己 自身
8	エンゲルス	フランス	資本 主義	スペンサー	昭和 九	独逸	ノエシス	表現 活動	自己 自身	支那
9	唯物 弁証法	エンゲルス	自己 自身	ドイツ	ドイツ	**宗 敦**	**道 穂**	ベルグソン	**自 由圭 義**	絶対 矛盾
10	**宗 敦**	メーリング	絶対 精神	ゲシヒテ	自然 科学	**口 子**	カント	自己 否定	自由 主義	吉利

道徳、宗教、ブルジョア等の認識間違いが多数みられる

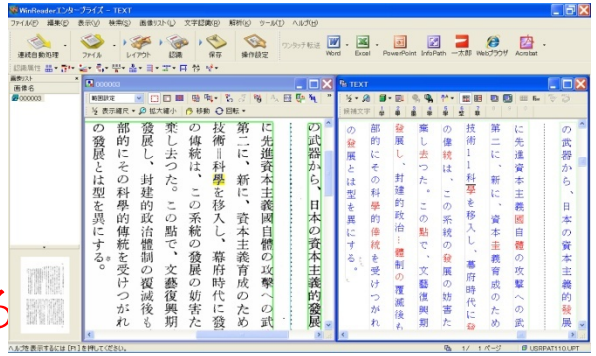
精度低下の原因

- 異体字・旧字体
  例: 教⇔教　学⇔學
- フォントの違い



「と」:一画目が斜め、二画目の入りが大きい
→「ご」や「ざ」と誤る
「感」:「心」の部分が小さい
→「戚」や「咸」と誤る
「ル」:右側が上がっている
→「〆」と誤る

OCRによる電子化と課題



- 文字認識精度が低い
  - 旧字体の認識精度の問題
  - 強調部分、ルビ、外来語部分の認識
  - レイアウト解析に失敗し文字化けする

  →近代文語論文の特徴抽出を行い精度を上げる

- 人手による作業コストの増大
  - レイアウト解析が不十分
    - タイトル、著者、脚注部分等のメタ情報を識別しなければならない
    - 論文の区切りを識別（ページの途中で別論文が始まる等）
  - OCRシステム開発会社と協同でOCRの高精度化
    - 誤認識の人手による訂正と学習
    - 異体字の対応、ルールによるレイアウト解析

  →自動化・システム化を行い、作業コストを軽減する

精度99.85%での用語抽出結果

	1930年	1931年	1932年	1933年	1934年	1935年	1936年	1937年	1938年	1939年
1	自然 科学	ヘーゲル	ヘーゲル	自然 科学	ヘーゲル	自己 自身	ヒューマニズム	アントニ	アメリカ	自己 同一
2	マルクス主義	マルクス主義	ゲーテ	唯物 論	マルクス	自己 限定	西田 哲学	ヘーゲル	ヘーゲル	ピュタゴラス
3	唯物 論	弁証 法	マルクス主義	マルクス	フランス	ヘーゲル	ヘーゲル	自己 自身	自己 同一	プラトン
4	ヘーゲル	カント	ギルヘルム	カント	キリスト	アリストテレス	自己 自身	職業 生活	自由 主義	自己 自身
5	エンゲルス	自己 自身	形而 上学	スペンサー	アリストテレス	自然 科学	自己 限定	表現 活動	イギリス	ディオニューソス
6	プロレタリアート	プロレタリアート	弁証 法	存在 論	日本 精神	ロゴス	弁証 法	ベルグソン	自己 自身	キリスト
7	資本 主義	ジャーナリズム	エンゲルス	アリストテレス	自然 科学	プラトン	ノエシス	自己 否定	フランス	モラル
8	トーキー	形而 上学	ファウスト	イデオロギー	エンゲルス	エネルギー	プラトン	プラトン	ヨーロッパ	ロゴス
9	コント	マルクス	ディルタイ	ディルタイ	自己 自身	ノエシス	アリストテレス	弁証 法	ロゴス	エートス
10	観念 論	イデオロギー	カント	形而 上学	カント	ベルグソン	ソクラテス	デカルト	アントニ	ソクラテス

上位には間違いがほとんど見られない

「思想」の構造化プロジェクト

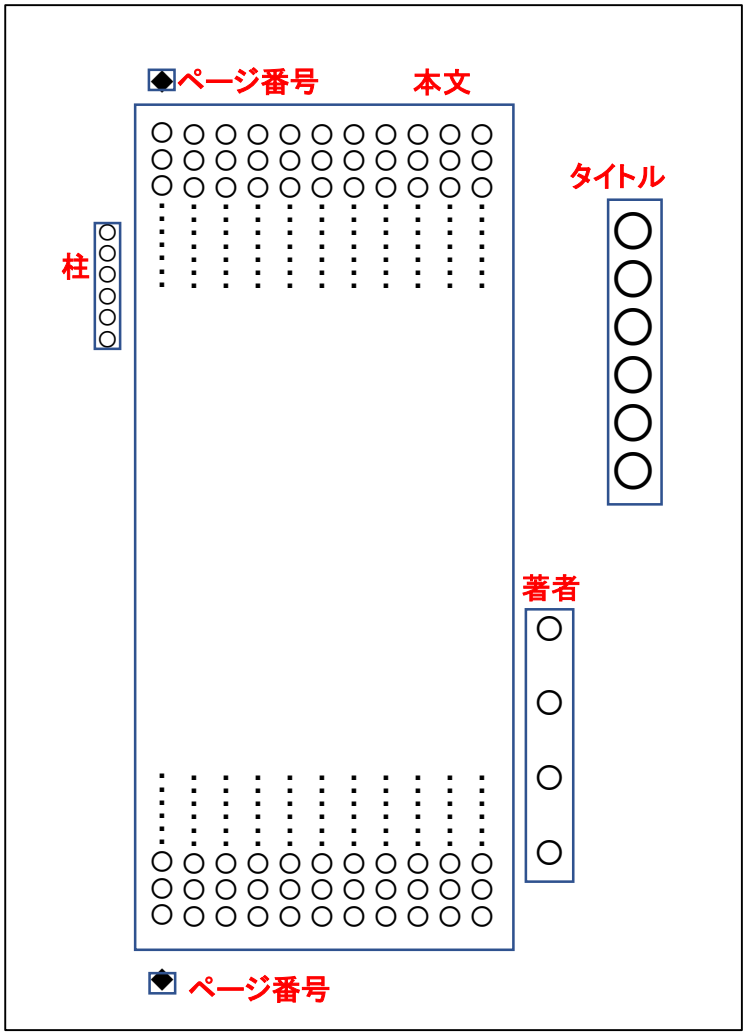
- テストとして20年分(1930-49)をテキスト化
  - OCRシステム開発会社と協同でOCRの高精度化
    - 異体字の対応、ルールによるレイアウト解析

    → 99%程度の精度でテキスト化

    タイトル・著者等も自動的に抽出

→80年分に拡大

レイアウト構造の解析



- 各ブロックに論理ラベルを自動的に付与
- ルールベースで付与
  - 文字サイズ
  - 位置
  - ブロックの順番

レイアウト解析ソフトウェアの評価

- 3号分（250ページ、1631ブロック中）における解析間違い

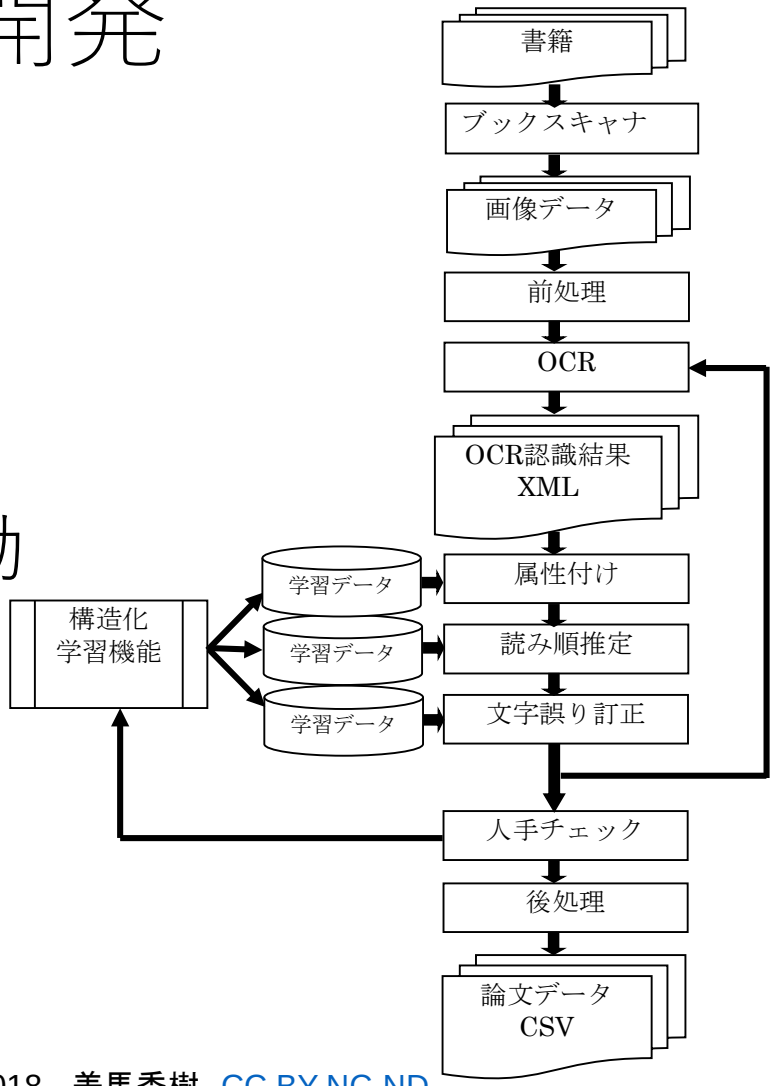
①ページ番号や欄外の見出しを本文の一部と誤認識する	51件
②属性付けの間違い（タイトル、著者）	2件
③レイアウト解析の間違い	4件
④ルビや句読点、汚れの存在によるレイアウト解析の間違い	5件
⑤広告や編集後記の存在による区切りの誤認識	5件
⑥特殊なレイアウトの存在による区切りの誤認識	3件

合計　70件

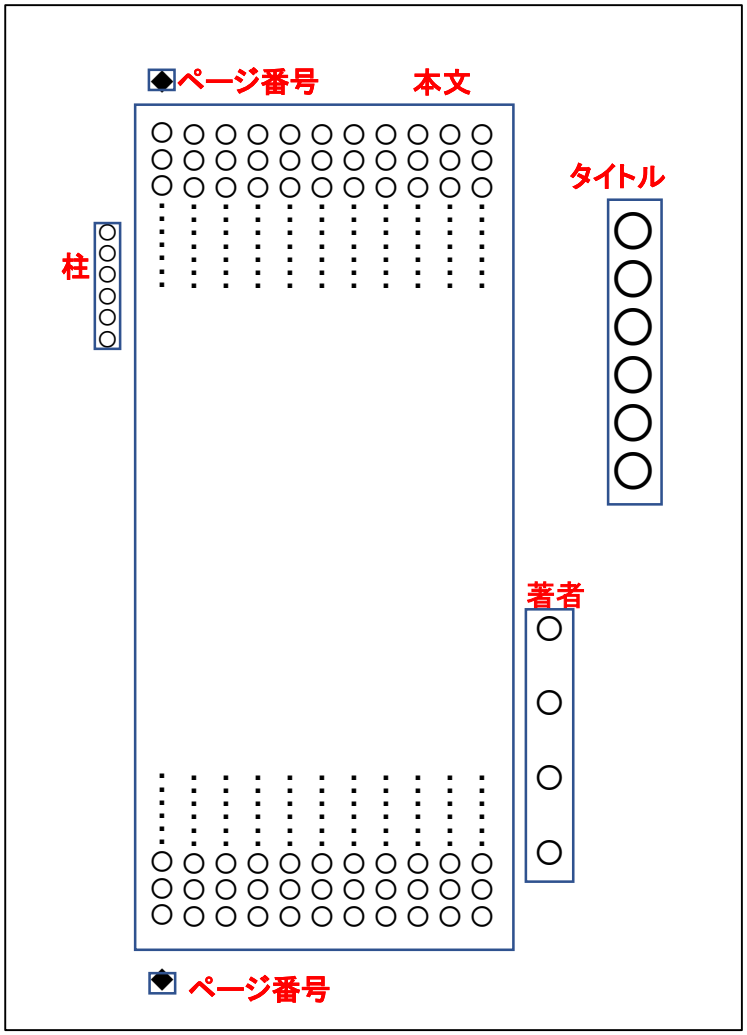
全1631ブロック中、 1561ブロックを正しく認識　(認識率95.7%)

現在の取り組み
人工知能（機械学習）を用いた自動化システムの開発

- ルールベースの限界
  - 『思想』に特有のルールを追加すると…
    - ◎『思想』の認識精度は向上
    - ×他の文献には適用できない
- 機械学習を用いた新たな自動化システムの開発
  - 属性付け
  - 読み順推定
  - 統計的文字誤り訂正



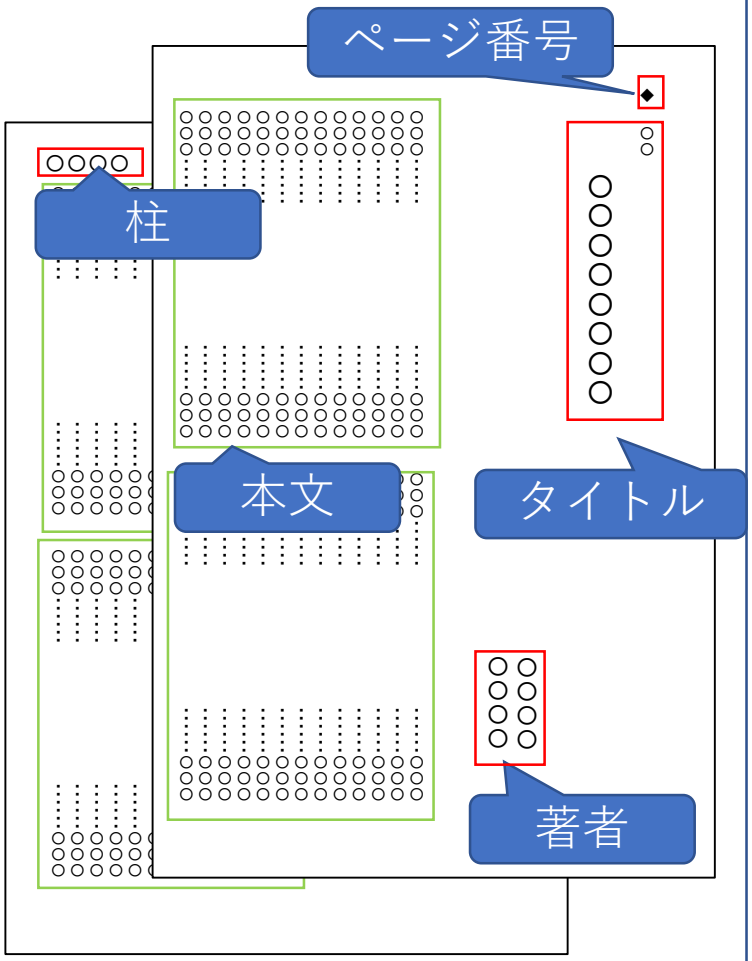
論理レイアウト構造の認識



- 各ブロックに論理ラベルを自動的に割り当てる
- 機械学習手法を利用
  - 様々なレイアウトに対応
  - 割り当てに使用する特徴
    - 文字サイズ
    - 位置
    - 周辺の余白
    - ・・・

機械学習を用いたレイアウト属性付け

レイアウト属性の正解データ



機械学習

特徴量による分類器

特徴量

視覚的特徴量
- 文字サイズ
- ブロックサイズ
- ブロック位置
- 余白長さ

言語的特徴
- 「名詞」割合
- 「人名」割合

精度(F値)	ルール	機械学習
タイトル	0.913	0.960
著者	0.966	0.985
柱	0.974	0.979
ページ番号	0.994	0.995
本文	0.985	0.992

人手で作成したルールに比べ高精度

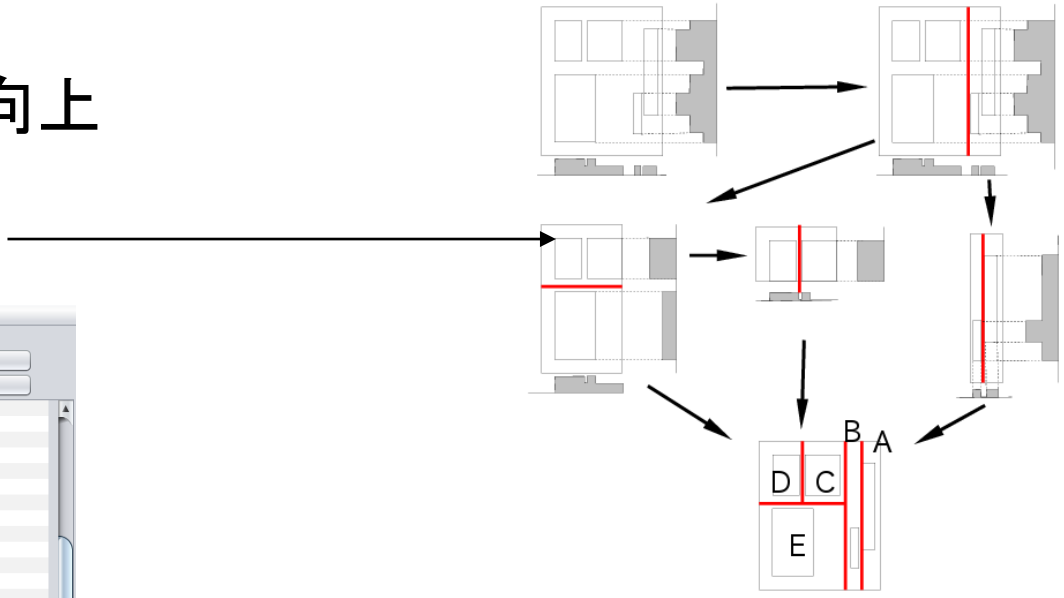
ブロック読み順の推定

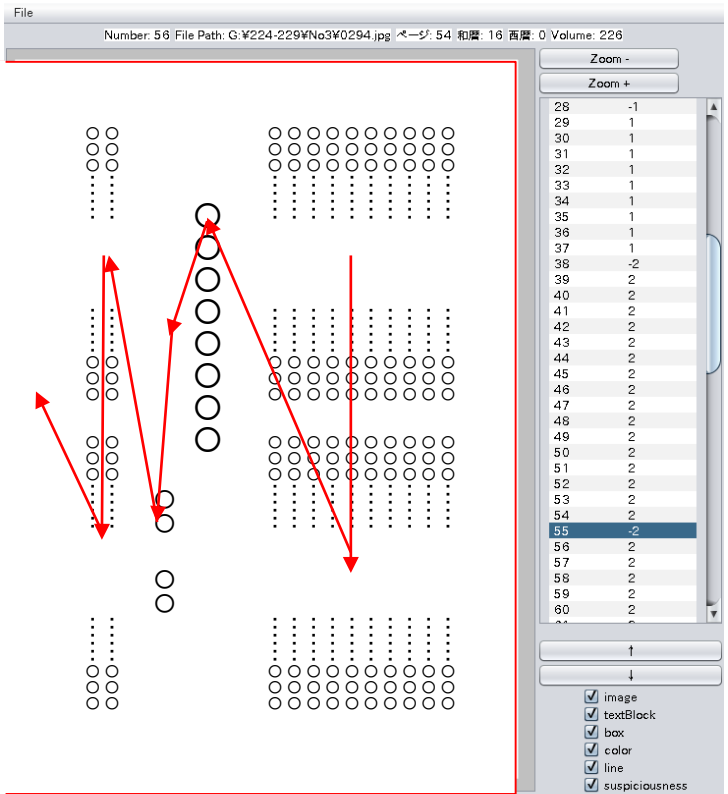
OCR結果のブロック間の順番を自動で推定

　自動読み上げ

　テキスト検索の精度向上

機械学習による推定

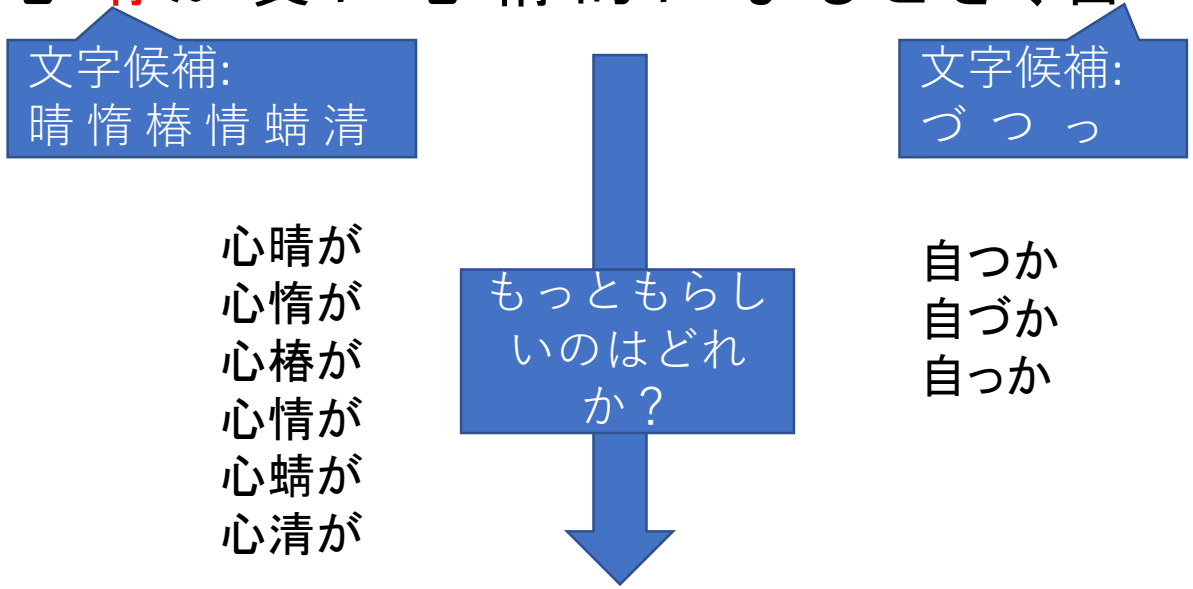




人手による修正インタフェース

言語情報を用いたOCR誤りの訂正

OCR結果 心 晴 が 眞 に 心 情 的 に な る と き 、自 つ か ら ･･･



訂正結果 心 情 が 眞 に 心 情 的 に な る と き 、自 づ か ら ･･･

年代別文字認識精度

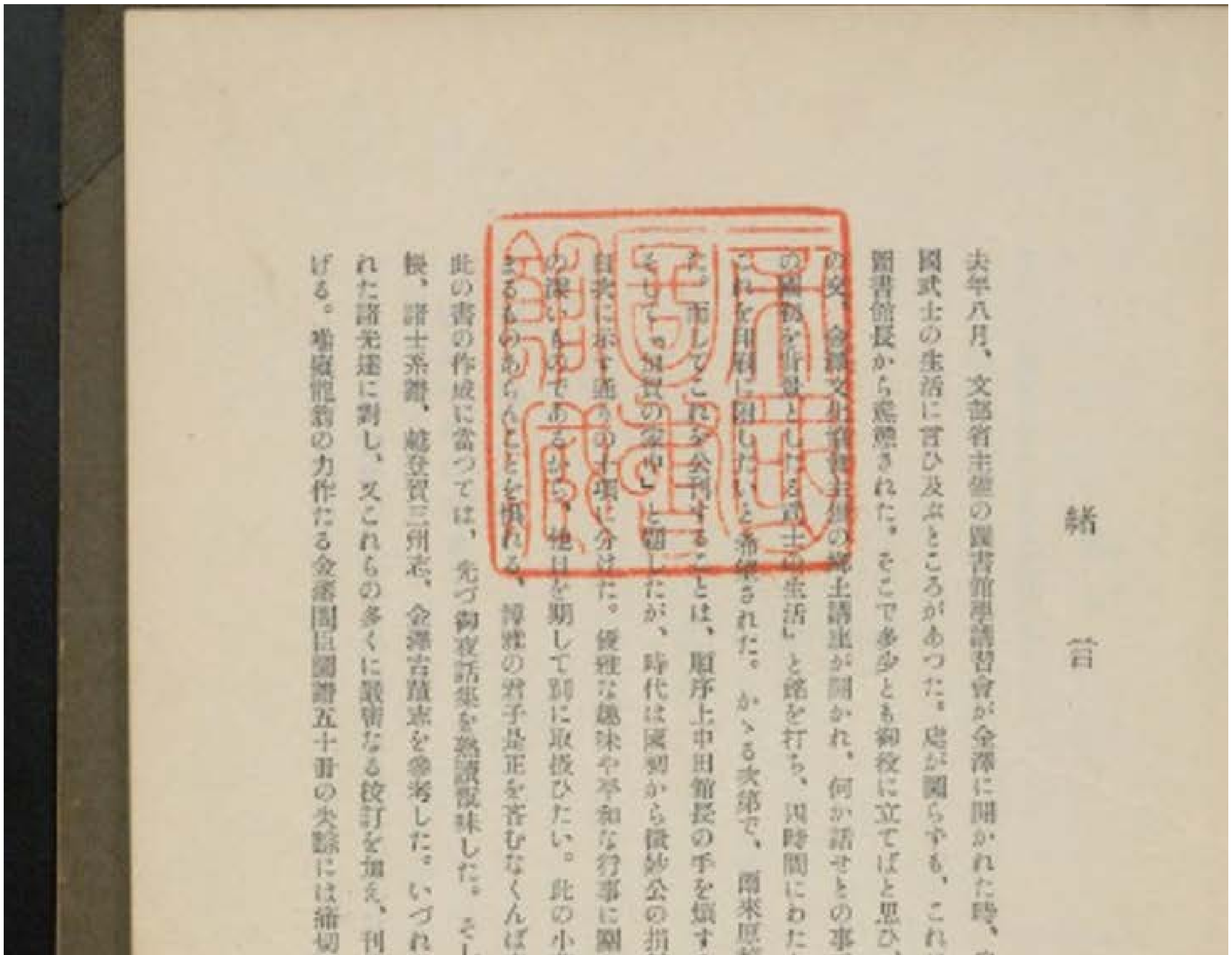
各年代1論文を人手でチェック

年	データ番号	文字数	誤認識数	認識精度
2000	0010-1	15998	31	99.81%
1990	0010-1	41638	215	99.48%
1980	0010-1	32229	83	99.74%
1970	0010-1	29476	846	97.13%
1960	0010-1	3090	61	98.02%
1950	0010-1	6066	170	97.19%
1940	0010-1	8316	52	99.37%
1930	0010-1	4294	348	91.90%
1921	0010-1	9218	1078	88.31%

古い年代で大きく精度が低下

実際に画像を見てみると・・・

国立国会図書館デジタルコレクション
岡本勇著『加賀の家中』（昭和10年）



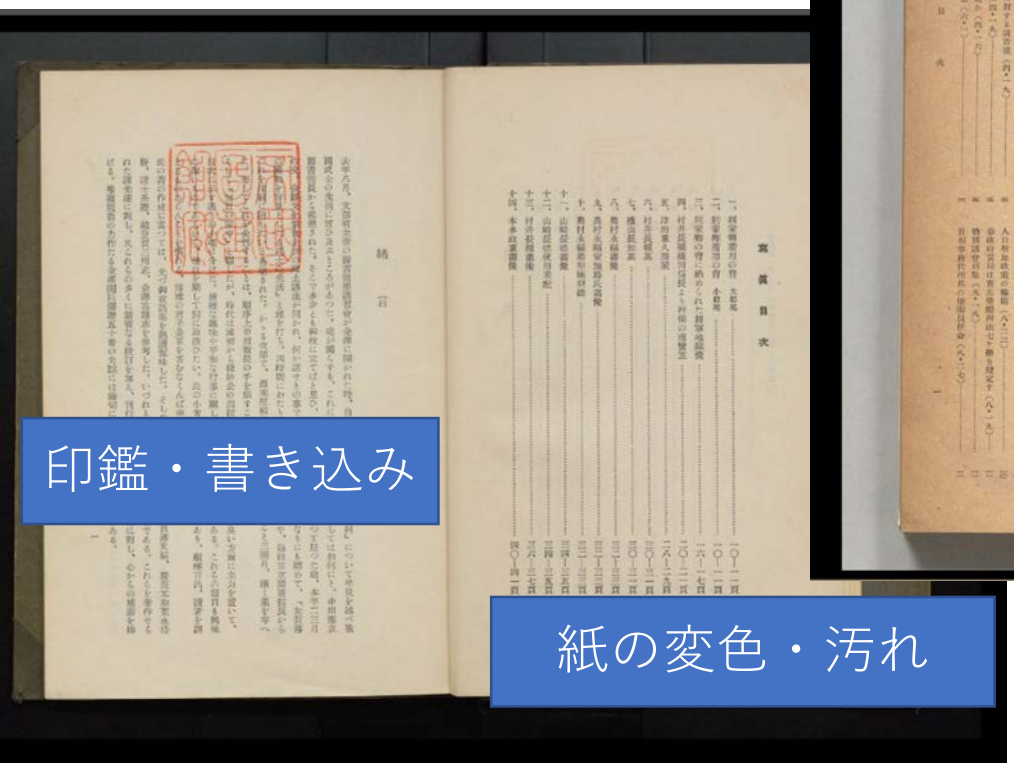
緒　言

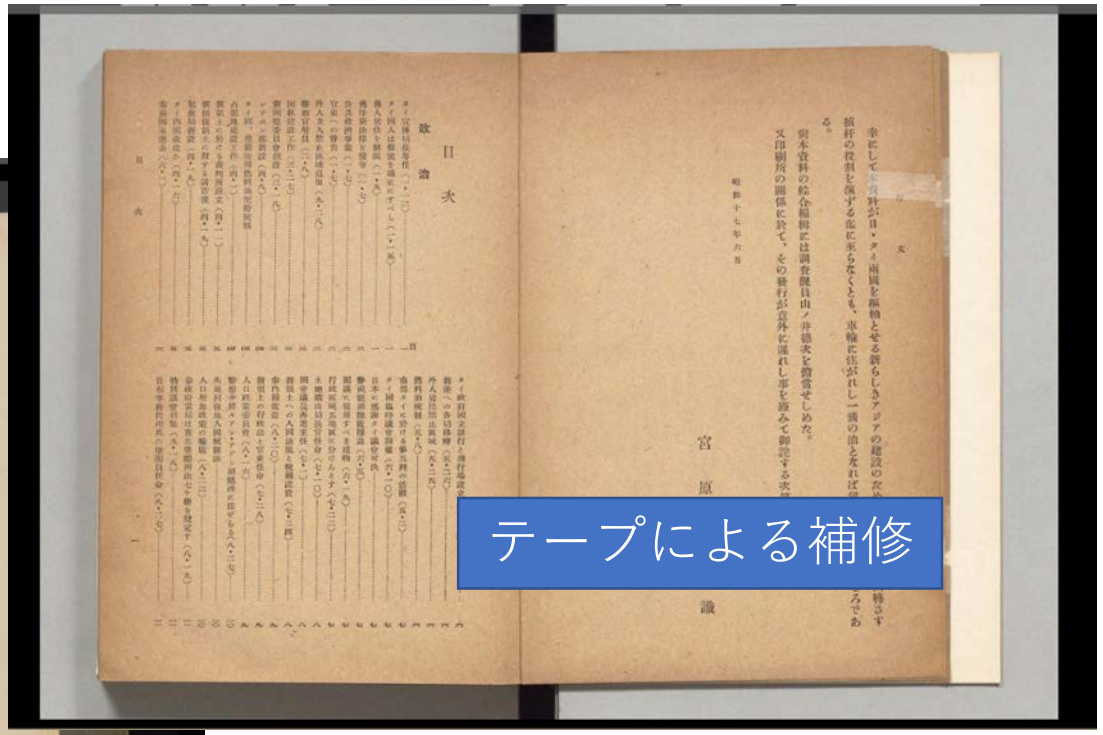
去年八月、文部省主催の圖書館學講習會が金澤に開かれた時、
國武士の生活に言ひ及ぶところがあつた。處が圖らずも、これ
圖書館長から慫慂された。そこで多少とも御役に立てばと思ひ
の友、金澤文化[illegible]の鄕土講座が開かれ、何か話せとの事
の[illegible]武士の生活」と銘を打ち、四時間にわた
これを印刷に附したいと希望された。かゝる次第で、爾來原
た[illegible]之れを公刊することは、順序上中田館長の手を煩す
そもそも[illegible]家中」と題したが、時代は寛[illegible]から微妙公の指
自然に[illegible]項に分けた。優雅な趣味や平和な行事に關
の深い[illegible]である[illegible]他日を期して別に取扱ひたい。此の小
[illegible]を懼れる。博雅の君子是正を吝むなくんば
此の書の作成に當つては、先づ御夜話集を熟讀玩味した。そ
検、諸士系譜、越登賀三州志、金澤古蹟志を參考した。いづれ
れた諸先進に對し、又これらの多くに周密なる校訂を加へ、刊
げる。[illegible]の力作たる金澤[illegible]圖譜五十冊の大部には縮切

精度低下の原因

- 原本状態の悪さ

国立国会図書館デジタルコレクション
岡本勇著『加賀の家中』(昭和10年)





国立国会図書館デジタルコレクション
『一九四一年タイ国政治経済情勢』(1942年)
http://dl.ndl.go.jp/info:ndljp/pid/1879315

現状では対応が不可能

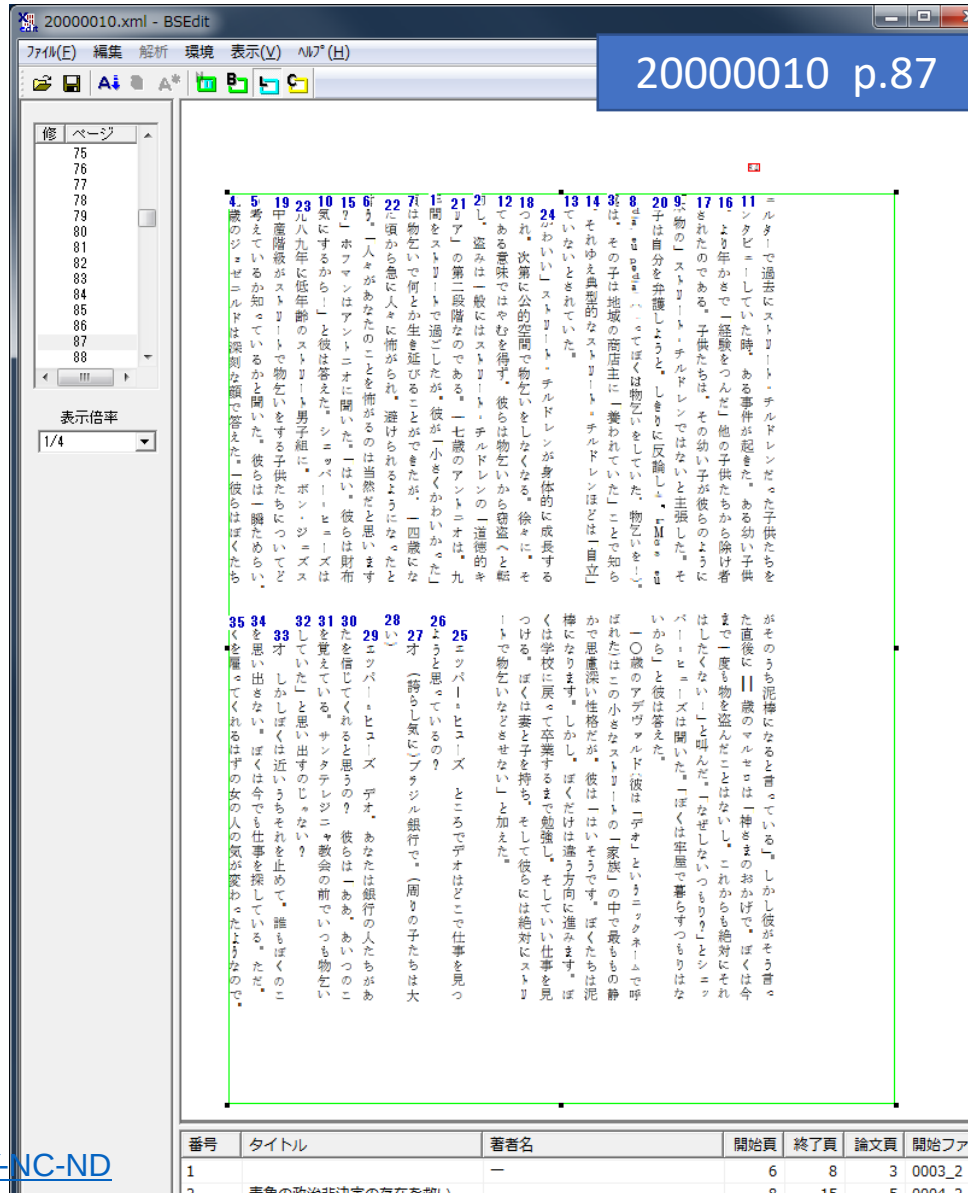
その他、OCRの課題

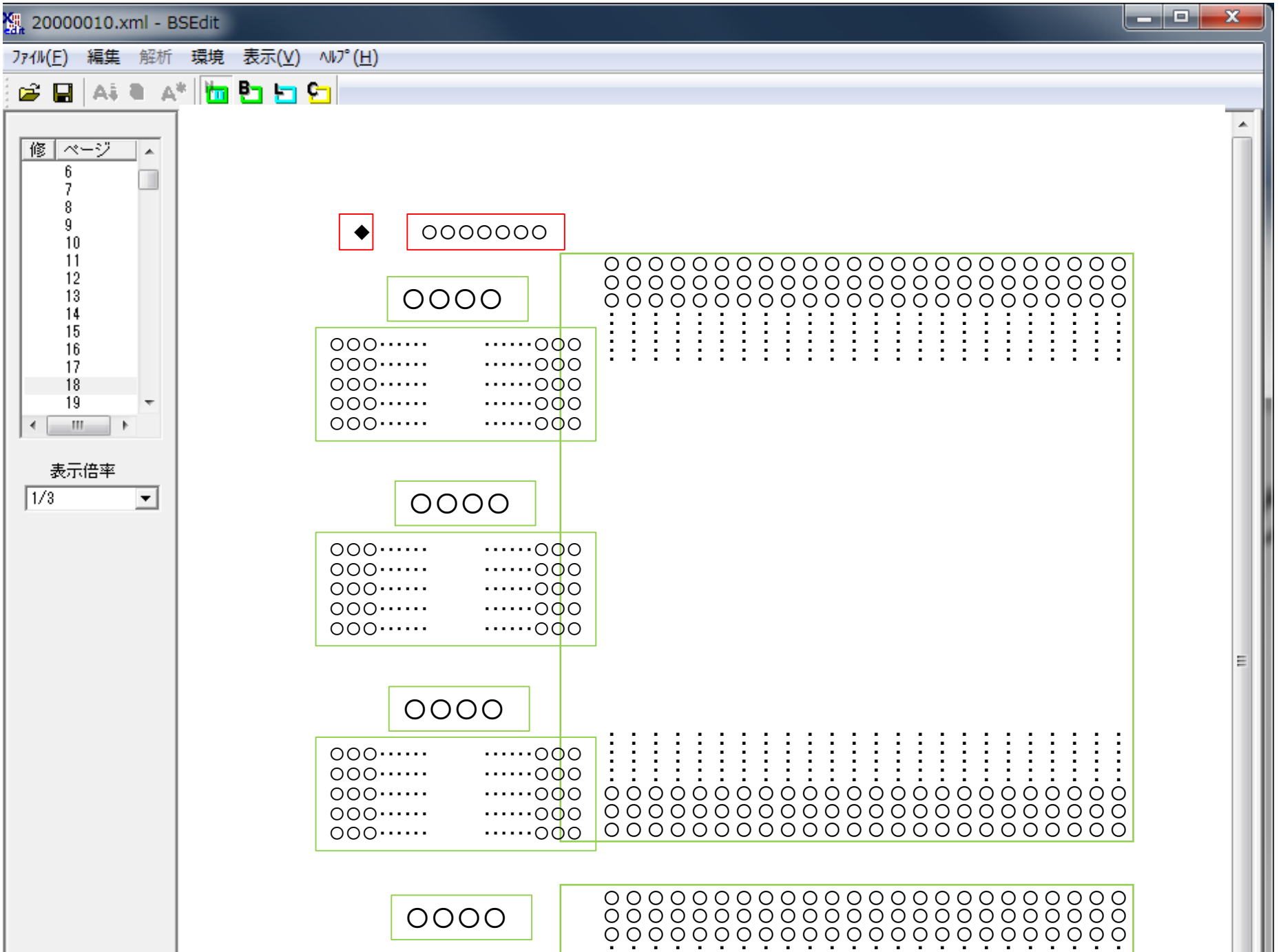
OCR処理:ページ端・背景の誤認識

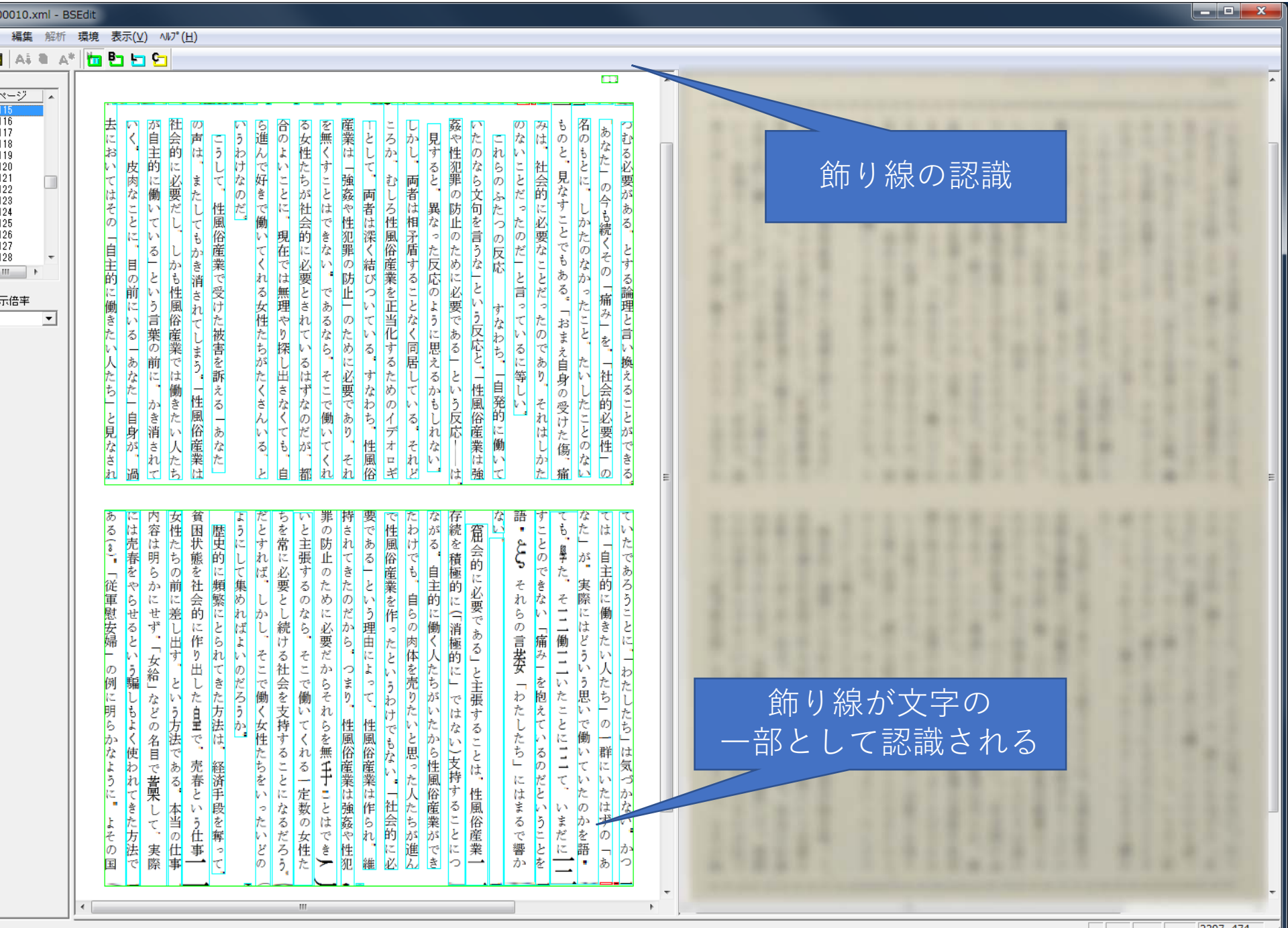
- ページの端や背景を文字と誤認識する
- 誤認識が他の　　ブロックに影響する
  - ブロック境界誤り
  - 行順番誤り

SSEdit でのブロック再構成時に発生

→前処理でブロックを削除




20000010.xml - BSEdit
ファイル(F)
編集
解析
環境
表示(V)
ヘルプ(H)
修
ページ
6
7
8
9
10
11
12
13
14
15
16
17
18
19
表示倍率
1/3
○○○○○○○
○○○○
○○○……
……○○○
○○○○
○○○○
○○○○


飾り線の認識
飾り線が文字の
一部として認識される

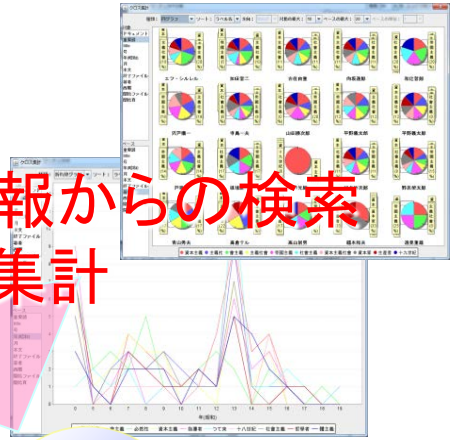
MIMAサーチ「思想」構造化デモ

「思想」の構造化のサイクル

データベース

検索

膨大な情報からの検索
と特徴の集計



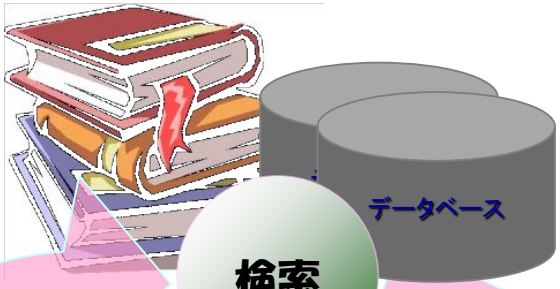
集計
分析

知識の可視化

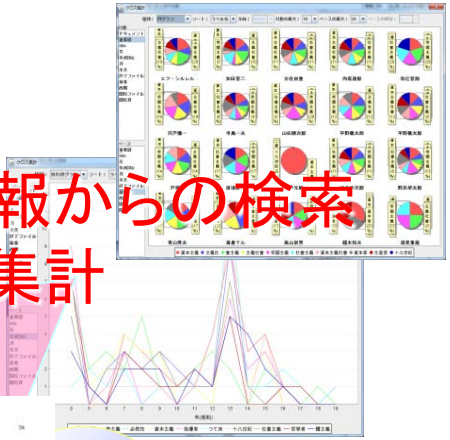
可視化

共有化
共創

「思想」の構造化のサイクル

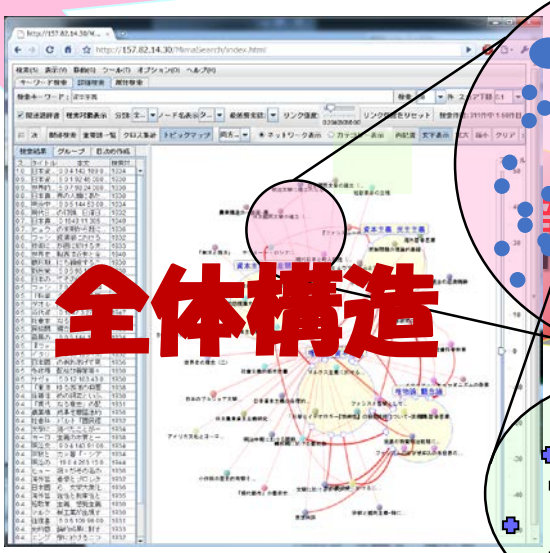


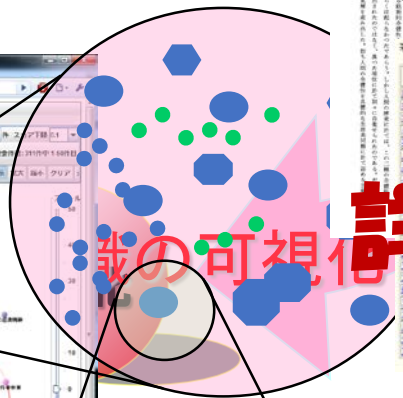
膨大な情報からの検索と特徴の集計

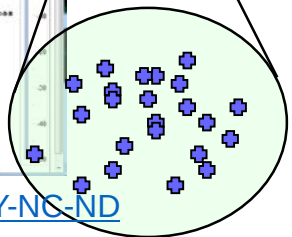


共有化
共創

集計
分析



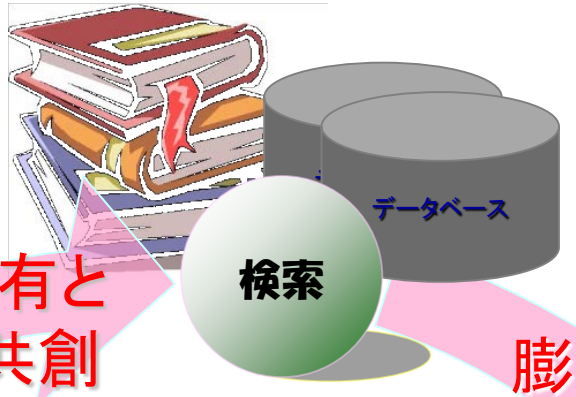




全体構造

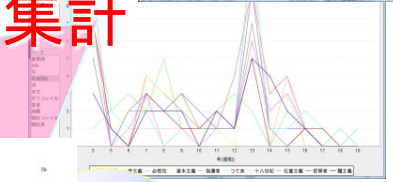
詳細構造

「思想」の構造化のサイクル



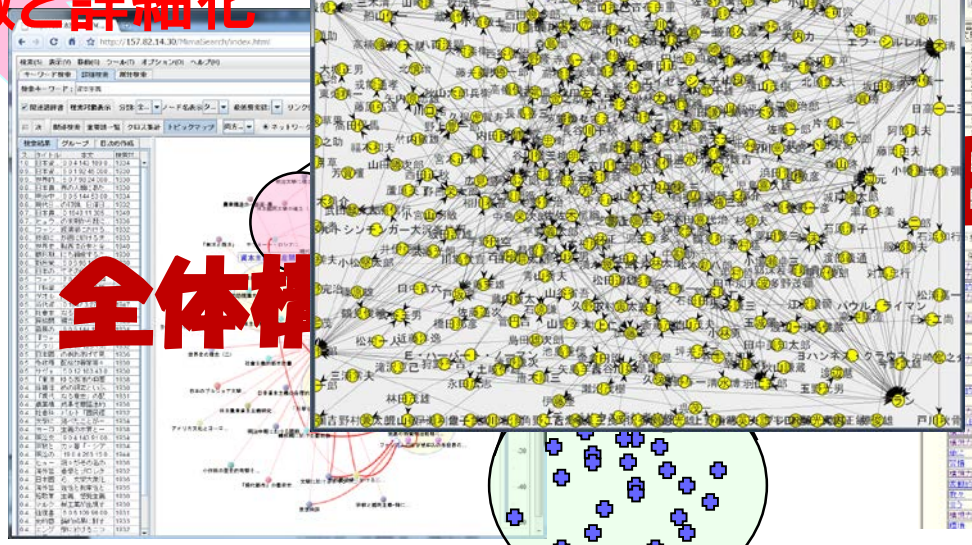
知識の共有と
再利用、共創

膨大な情報からの検索
と特徴の集計



共有化
共創

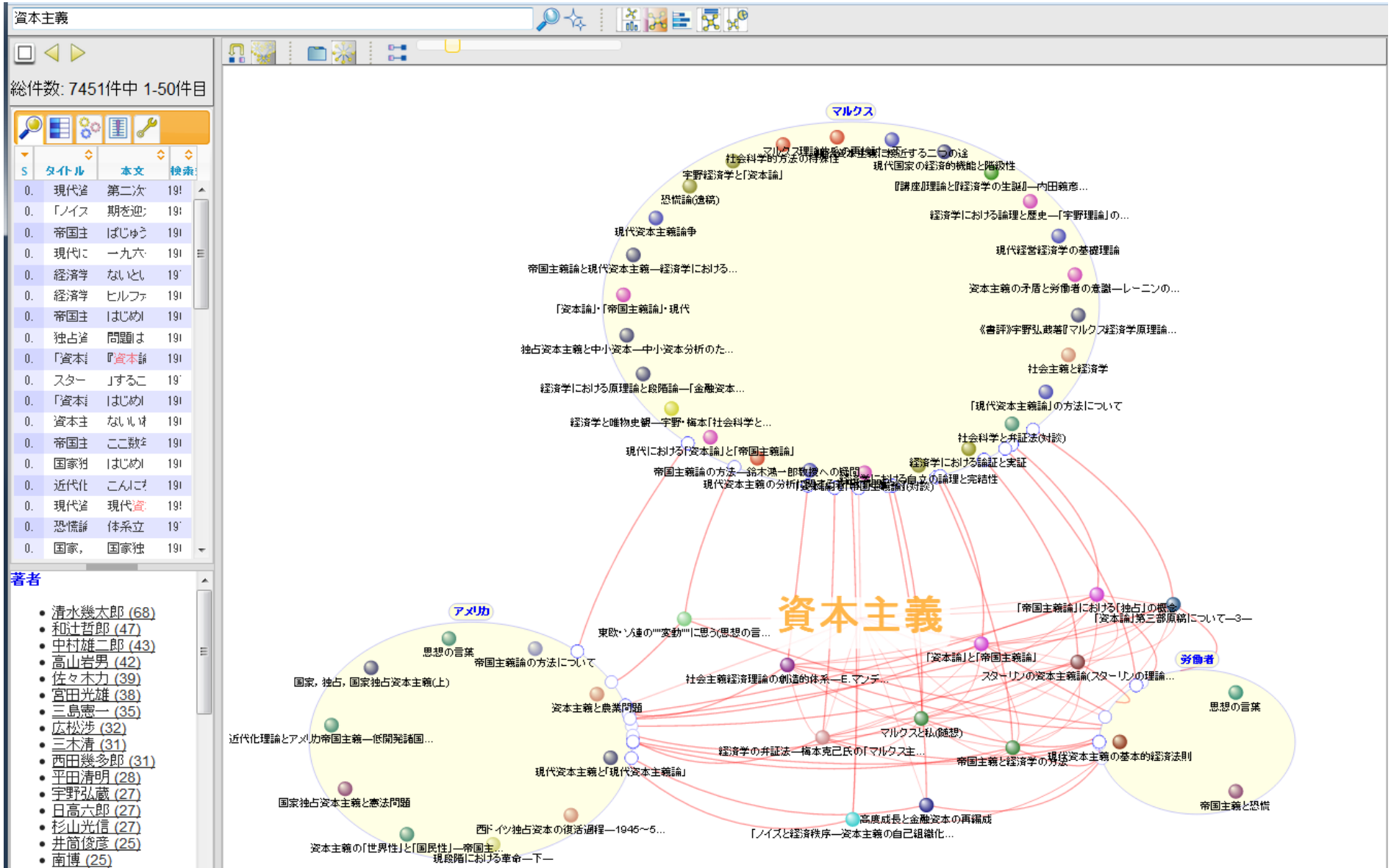
様々な視点からの知識
の俯瞰と詳細化



「思想」構造化システム

- MIMAサーチ
- 係り受け関係検索
- 言及関係検索

MIMAサーチ



MIMAサーチ 詳細画面

Title 如何にして科学の発展は可能であるか

92号(1930年1月) p. 25

Author 三木清

1/20

科学の発展に関して普通に行はれてゐる見解は次の如くである、科学は「進歩」(Fortschritt)の過程にある。しかもそれは不断に、連続的に進歩してゐる。そしてこの点に於て科学は他の種々なる文化形態に対して特殊な位置を占めてゐる。例へば芸術については、或る時代の芸術作品が初めて新しい技術的手段または遠近法の法則を刑ゐるに到つたからといつて、そのためにそれが、それ以前の、かかる手段及び法則の知識を全く欠いてゐた芸術作品よりも、純粋に芸術的に一層高く立つてゐるとは云はれない。各の芸術作品はそれぞれの「完成」(邑審巨頻)である、従つてそれは決して打ち克たれるものでなく、また決して旧びるものでもない

もっと読む

係り受け関係検索

著者:三木清 (「思想」著者リスト) キーワード: 開始年: 終了年: 検索

◉係り受け単位 ○キーワード単位

係り受け単位での検索結果
(→キーワード単位で 著者: "三木清" で検索)

著者	係り元	係り先	頻度	スコア	用例
三木清	構想力	論理	118	171.84	用例
	言う	こと	708	143.33	用例
	こと	出来る	630	139.72	用例
	此の	もの	137	132.25	用例
	もの	もの	332	124.81	用例
	意味	於く	185	121.60	用例
	もの	為る	476	120.62	用例
	こと	因る	281	116.69	用例
	主観的	もの	69	111.39	用例
	意味	有する	71	103.86	用例
	考える	こと	108	100.02	用例
	我々	出来る	70	99.41	用例
	此の	場合	100	98.56	用例
	我々	取る	56	98.20	用例
	存在	仕方	92	96.87	用例
	ところ	もの	100	96.59	用例
	斯かる	もの	64	96.24	用例
	これ	反する	67	95.29	用例
	構想力	於く	41	94.95	用例
	於く	言う	120	94.51	用例
	客観的	もの	66	93.05	用例
	こと	無い	191	91.79	用例
	もの	考える	139	88.93	用例
	もの	有る	237	88.59	用例
	於く	為る	133	87.52	用例

三木清 著作リスト

- ボルツアノーの「命題自体」(1923/12 26号)
- 消息一通(1924/3 29号)
- パスカルと生の存在論的解釈(1925/5 43号)
- 愛の情念に関する説―パスカル覚書(1925/8 46号)
- パスカルの方法(1925/11 49号)
- パスカルの「賭」(1925/12 50号)
- 解釈学的現象学の基礎概念(1927/1 63号)
- 人間学のマルクス的形態(1927/6 68号)
- マルクス主義と唯物論(1927/8 70号)
- プラグマチズムとマルキシズムの哲学(1927/12 74号)
- ヘーゲルの歴史哲学(1929/4 83号)
- 社会と自然(1929/8 87号)
- 紹介と検討 現象学叙説―山内氏の著作の読後に(1929/9 88号)
- 弁証法に於ける自由と必然(1929/10 89号)
- 如何にして科学の発展は可能であるか(1930/1 92号)
- 存在の歴史性(1931/12 115号)
- 拙著批評に答ふ(1932/9 124号)
- 倫理と人間(1933/6 133号)
- ヘルデル大辞典について(1934/9 148号)
- 表現に於ける真理(1935/9 160号)
- 西田哲学の性格について(1936/1 164号)
- ヒューマニズムの哲学的基礎(1936/10 173号)
- ヒューマニズムの哲学的基礎(完)(1936/11 174号)
- 神話―構想力の論理に就いて(其一)(1937/5 180号)
- 神話(中)構想力の論理に就いて其二(1937/6 181号)
- 制度(一)―構想力の論理に就いて(其四)(1937/8 183号)
- 制度(二)―構想力の論理に就いて(其五)(1937/9 184号)
- 制度(三)―構想力の論理に就いて(其六)(1937/10 185号)
- 技術―構想力の論理に就いて其七(一)(1938/2 189号)
- 技術―構想力の論理に就いて其八(二)(1938/3 190号)
- 技術―構想力の論理に就いて其九(三)(1938/5 192号)
- 経験(一)構想力の論理に就いて 続一(1939/9 208号)
- 経験(二)―構想力の論理に就いて 続二(1940/8 219号)
- 経験(三)―構想力の論理に就いて(続三)(1940/11 222号)
- 経験(四)―構想力の論理に就いて(続四)(1940/12 223号)
- 経験(五)―構想力の論理に就いて(続五)(1941/1 224号)
- 経験―構想力の論理に就いて続六(六)(1941/2 225号)
- 充足的経験論(1941/4 227号)
- 経験(七)―構想力の論理に就いて(続七)(1942/4 239号)
- 経験―構想力の論理に就いて続八(八)(1943/3 250号)
- 経験―構想力の論理に就いて続九(九)(1943/4 251号)

言及関係検索

著者: 三木清 開始年: 終了年: 検索 (「思想」著者リスト)

三木清 著作リスト(著作数: 44)

01:ボルツアノーの「命題自体」(1923/12 26号)
02:消息一通(1924/3 29号)
03:パスカルと生の存在論的解釈(1925/5 43号)
04:愛の情念に関する説―パスカル覚書(1925/8 46号)
05:パスカルの方法(1925/11 49号)
06:パスカルの「賭」(1925/12 50号)
07:解釈学的現象学の基礎概念(1927/1 63号)
08:人間学のマルクス的形態(1927/6 68号)
09:マルクス主義と唯物論(1927/8 70号)
10:プラグマチズムとマルキシズムの哲学(1927/12 74号)
11:ヘーゲルの歴史哲学(1929/4 83号)
12:社会と自然(1929/8 87号)
13:紹介と検討 現象学叙説―山内氏の著作の読後に(1929/9 88号)
14:弁証法に於ける自由と必然(1929/10 89号)
15:如何にして科学の発展は可能であるか(1930/1 92号)
16:存在の歴史性(1931/12 115号)
17:拙著批評に答ふ(1932/9 124号)
18:倫理と人間(1933/6 133号)
19:ヘルデル大辞典について(1934/9 148号)
20:表現に於ける真理(1935/9 160号)
21:西田哲学の性格について(1936/1 164号)
22:ヒューマニズムの哲学的基礎(1936/10 173号)
23:ヒューマニズムの哲学的基礎(完)(1936/11 174号)
24:神話―構想力の論理に就いて(其一)(1937/5 180号)
25:神話(中)構想力の論理に就いて其二(1937/6 181号)
26:制度(一)―構想力の論理に就いて(其四)(1937/8 183号)
27:制度(二)―構想力の論理に就いて(其五)(1937/9 184号)
28:制度(三)―構想力の論理に就いて(其六)(1937/10 185号)
29:技術―構想力の論理に就いて其七(一)(1938/2 189号)
30:技術―構想力の論理に就いて其八(二)(1938/3 190号)
31:技術―構想力の論理に就いて其九(三)(1938/5 192号)
32:経験(一)構想力の論理に就いて 続一(1939/9 208号)
33:経験(二)―構想力の論理に就いて 続二(1940/8 219号)
34:経験(三)―構想力の論理に就いて(続三)(1940/11 222号)
35:経験(四)―構想力の論理に就いて(続四)(1940/12 223号)
36:経験(五)―構想力の論理に就いて(続五)(1941/1 224号)
37:経験―構想力の論理に就いて続六(六)(1941/2 225号)
38:充足的経験論(1941/4 227号)
39:経験(七)―構想力の論理に就いて(続七)(1942/4 239号)
40:経験―構想力の論理に就いて続八(八)(1943/3 250号)
41:経験―構想力の論理に就いて続九(九)(1943/4 251号)
42:経験―構想力の論理に就いて続一〇(一〇)(1943/5 252号)
43:経験―構想力の論理に就いて続一一(一一)(1943/6 253号)
44:経験―構想力の論理に就いて続一二(一二)(1943/7 254号)

著者間の言及関係:
本文から人物名の
言及を抽出

三木清についての『思想』での言及	言及文献数	言及頻度
中村雄二郎(総文献数:43)	9	47
上村忠男(総文献数:14)	6	31

中村雄二郎
制度論的視角と日本型思想(1962/457 7号)
「思想」の思想史―1―(1966/504 6号)
「思想」の思想史―2―(1966/505 7号)
「思想」の思想史(6・完)―日本における「思想」のあり方をめぐって(1966/510 12号)
【思想の言葉】(1977/640 10号)
現代思想と西田幾多郎(1985/738 12号)
西田幾多郎の宗教論と歴史論―深層の知・制度・弁証法(1987/756 6号)
〈場所の論理〉の彼方へ―述語的世界と制度(1994/835 1号)
西田哲学と日本の社会科学(1995/857 11号)

上村忠男
ヨーロッパ的諸学の危機におけるヴィーコ―或る出会いのなかから(1987/752 2号)

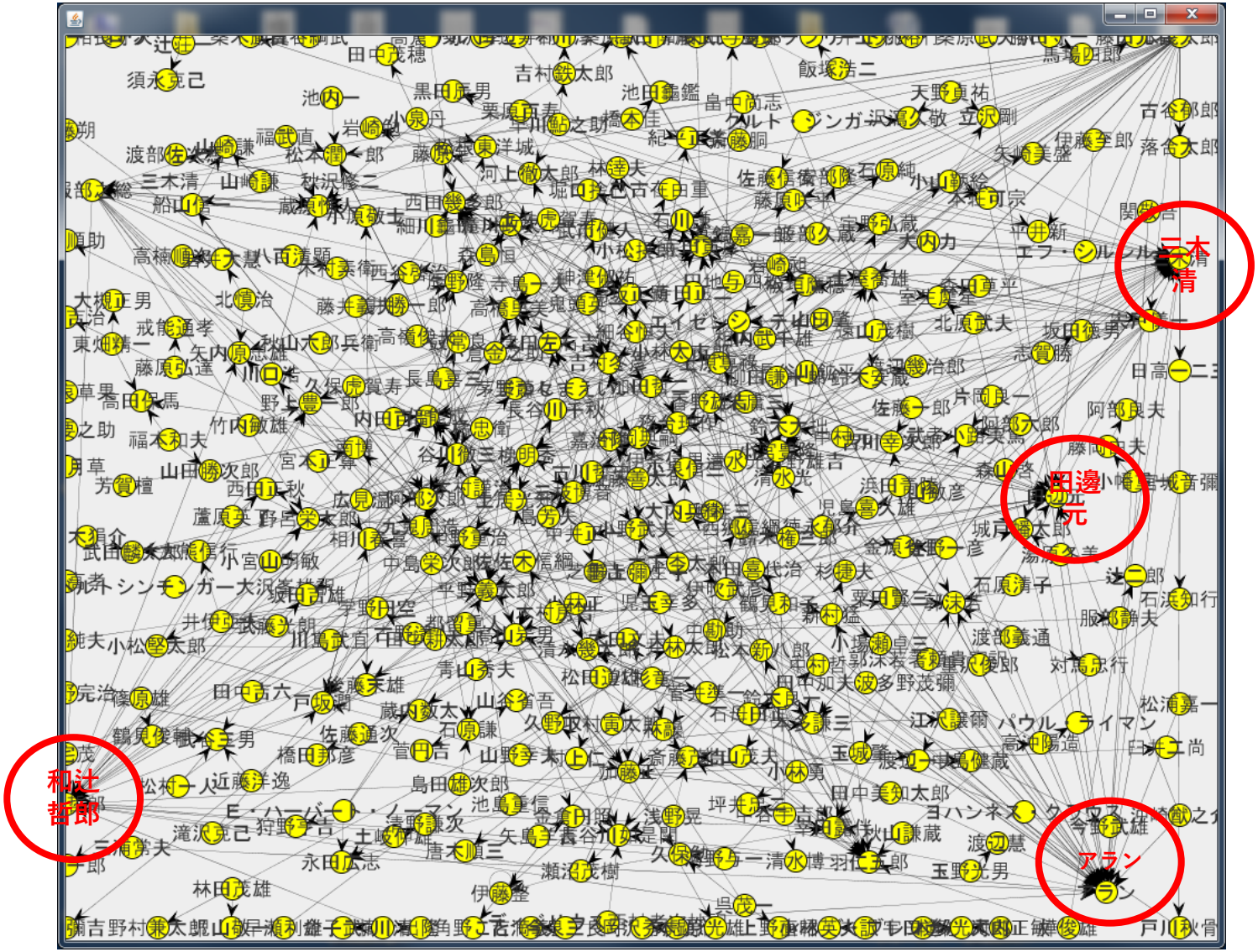
三木清の『思想』著者についての言及	言及文献数	言及頻度
ベルグソン	14	120

ベルグソン
人間学のマルクス的形態(1927/68 6号)
プラグマチズムとマルキシズムの哲学(1927/74 12号)
倫理と人間(1933/133 6号)
表現に於ける真理(1935/160 9号)
西田哲学の性格について(1936/164 1号)
ヒューマニズムの哲学的基礎(完)(1936/174 11号)
神話(中)構想力の論理に就いて其二(1937/181 6号)
制度(一)―構想力の論理に就いて(其四)(1937/183 8号)
制度(二)―構想力の論理に就いて(其五)(1937/184 9号)
制度(三)―構想力の論理に就いて(其六)(1937/185 10号)
技術―構想力の論理に就いて其九(三)(1938/192 5号)
経験(一)構想力の論理に就いて 続一(1939/208 9号)
充足的経験論(1941/227 4号)
経験―構想力の論理に就いて続九(九)(1943/251 4号)

三木清の他の人物についての言及	言及文献数	言及頻度
カント	30	347

カント
ボルツアノーの「命題自体」(1923/26 12号)
消息一通(1924/29 3号)
愛の情念に関する説―パスカル覚書(1925/46 8号)
パスカルの方法(1925/49 11号)
人間学のマルクス的形態(1927/68 6号)
マルクス主義と唯物論(1927/70 8号)
ヘーゲルの歴史哲学(1929/83 4号)
社会と自然(1929/87 8号)
紹介と検討 現象学叙説―山内氏の著作の読後に(1929/88 9号)
弁証法に於ける自由と必然(1929/89 10号)
存在の歴史性(1931/115 12号)
拙著批評に答ふ(1932/124 9号)
倫理と人間(1933/133 6号)
表現に於ける真理(1935/160 9号)

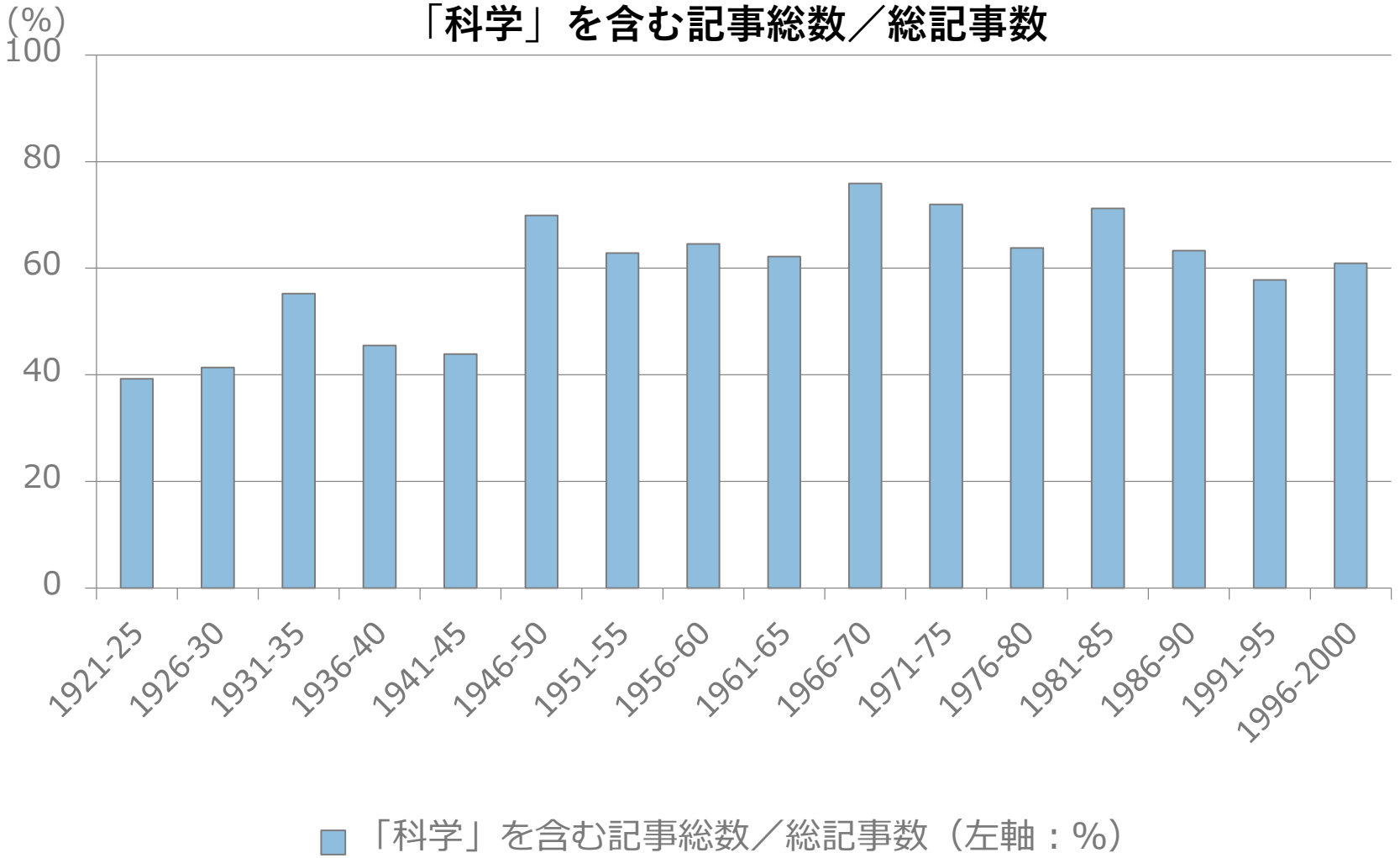
著者間の言及関係の可視化



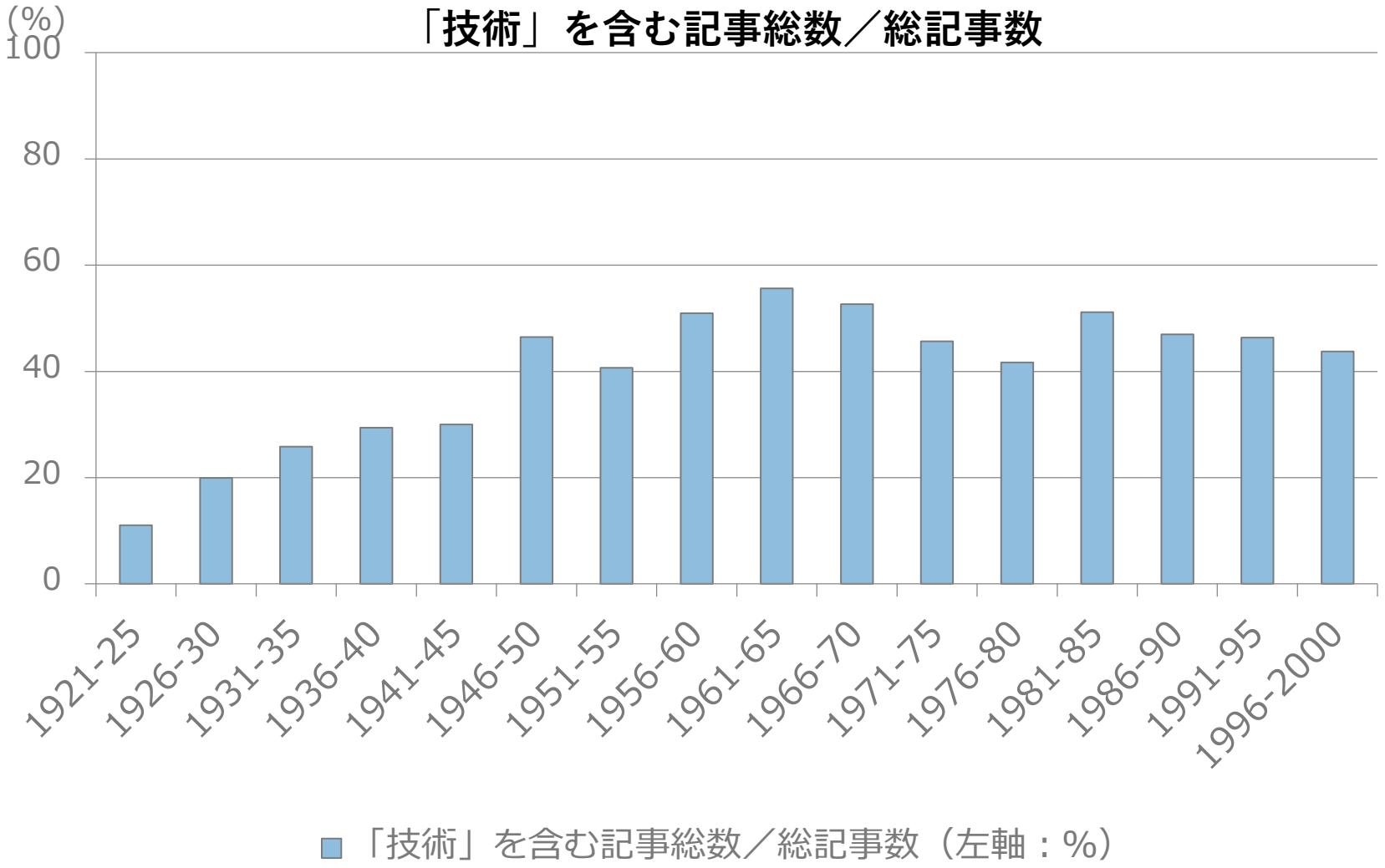
『思想』の分析例

- 使用される言葉・概念の年代による変遷
  - 対象1: 「科学」「技術」「科学技術」
    - 似た概念の言葉の使われ方の違い
  - 対象2: 原子力の戦争利用と平和利用
    - 戦争利用：「核兵器」「原子爆弾」「核実験」等
    - 平和利用：「原子力発電」「原発」等

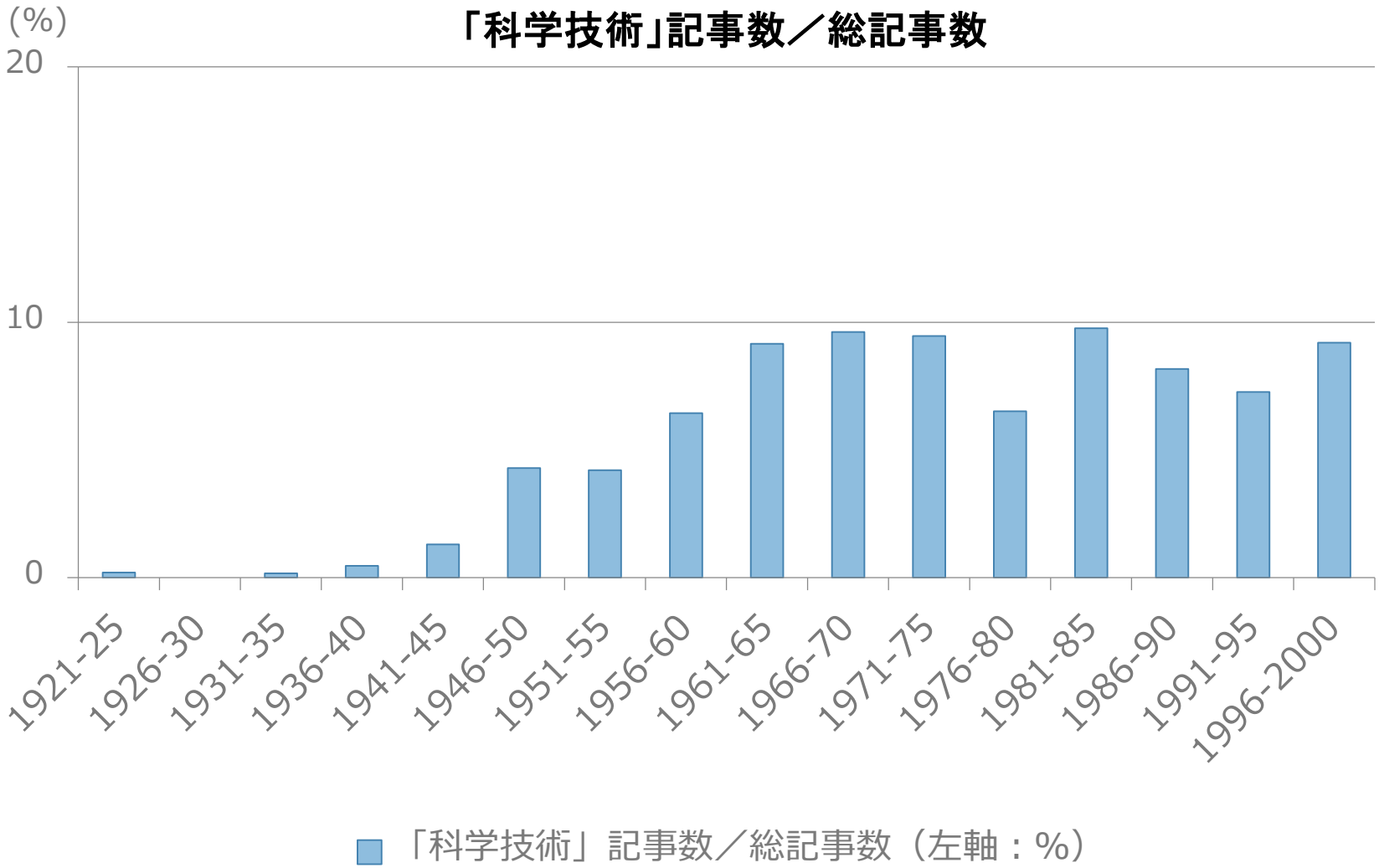
科学の登場頻度の推移



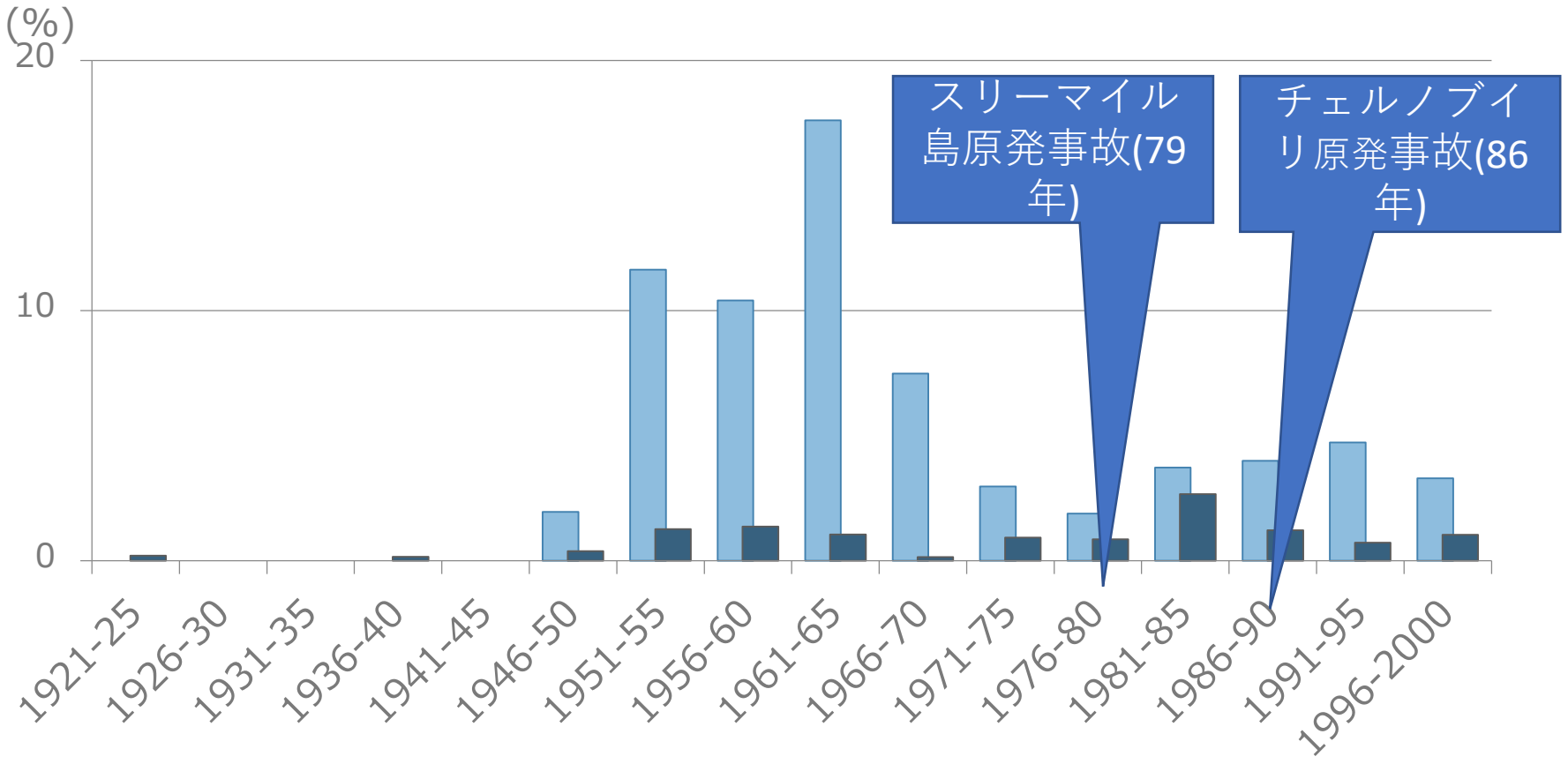
技術の登場頻度の推移



“科学技術”の登場頻度の推移



核兵器・原子力



講義／ワークショップの様子

世論調査　文学・思想・雑誌　政治思想史　読書・出版文化　論壇・メディア

出版流通

ナショナリズム

社会意識

技術開発チームとの対話

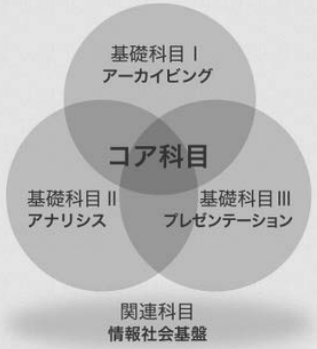
http://dh.iii.u-tokyo.ac.jp/program.html

大学院教育への展開

- デジタルヒューマニティーズ
  知の構造化論

プログラム構成

本プログラムは、〈コア科目〉、〈基礎科目〉、〈関連科目〉によって構成されます。〈コア科目〉はデジタル・ヒューマニティーズの中核をなすもので、2つの「必修科目」が含まれます。〈基礎科目〉は、I)アーカイビング、II)アナリシス、III)プレゼンテーションという3つの要素からなり、領域を横断して理論と方法を学ぶことができます。〈関連科目〉では、これらの科目に関係する情報社会基盤の知識を得ることができます。



コア科目とは

デジタル・ヒューマニティーズの中核をなす科目です

知の構造化論

講義題目	開講研究科	科目番号	担当者	単位	時間割	教室割
知識と情報処理、 人文学とコンピュータ、 臨床医療とコンピュータ	学際情報学府	4990110	美馬 秀樹 荒牧 英治	2.0	夏・水・4限 (14:50〜16:20)	福武ホールB2階・ 福武ラーニングスタジオ1

授業の目標・概要(クリックで詳細を表示)

本講義では、大量に蓄積された知識に対し、様々な構造化技術を用いて価値の創出を行う「知の構造化」を実践することを目的とする。知の構造化では、単なる検索にとどまらず、コンピュータを使って知識の抽出と分析を行い、分野や時勢を越えて各要素の間の関係を明らかにすることでその活用を促す。そのために、自然言語処理、人工知能、Web工学等の最先端技術を駆使する手法を学び、実社会で生じる様々な課題に対する分野を超えた知の適用の議論を進める。これにより、構造の抽出と、構造から機能を導出するプロセスを体現する。

人文学オープンデータ共同利用センター

https://codh.rois.ac.jp/

人文学オープンデータ共同利用センター（Center for Open Data in the Humanities / CODH）は、情報学・統計学の最新技術を用いて人文学データへのアクセスを改善する研究開発を進めるとともに、オープンサイエンスの考え方に基づき多くの人々が参加できるデータプラットフォームを構築することで、データ駆動型の人文学研究や超学際的な人文学研究など、情報技術を用いた新しい人文学の方法論を開拓します。[もっと詳しく..]

お知らせ

まとめ

- 文理融合による教育／研究の可能性
  - 可視化、対話による共創、知見の拡大、共有

- 資料体へのアクセス・公開
  - 教育支援、教材づくり
  - 海外の研究者の支援
    - 近代の言語資源の価値化

