



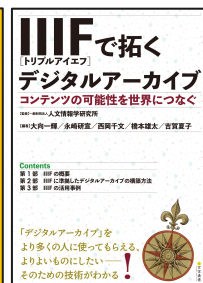
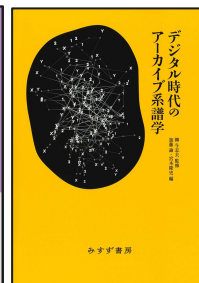
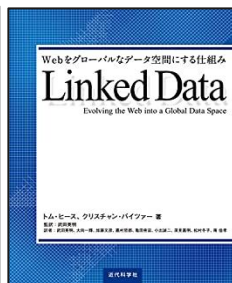
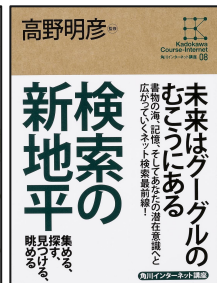
デジタルアーカイブにおける AI活用 現在・課題・展望

大向 一輝

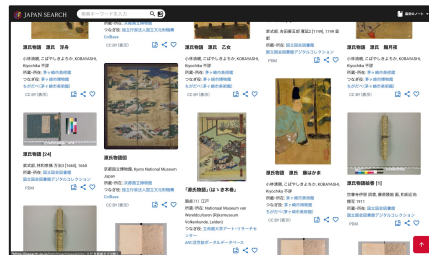
東京大学大学院人文社会系研究科

自己紹介

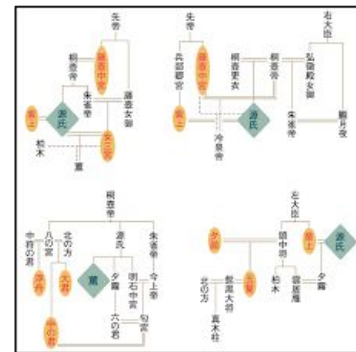
- 1999: 生物の進化を模倣した人工知能技術の研究
- 2002: インターネット上の知識共有・コミュニケーションの研究
- 2009: 学術情報サービスCiNiiの開発責任者(～2019)
- 2019: メディア芸術データベースのプロデューサー(～現在)
- 2019: デジタル人文学の研究教育(～現在)
- 2020: 人工知能学会セマンティックウェブとオントロジー研究会主査(～2022)
- 2020: 人工知能学会編集委員会(～2024)
- 2021: 人工知能学会理事(～2023)
- 2021: デジタルアーカイブ学会理事(～現在)



デジタルアーカイブの構図



いつれの御時にか 女御更衣あまたさぶらひたまひける
なかに いとやむごとなき
際にはあらぬがすぐれて時めきたまふ ありけり...



資料・史料

画像

テキスト

知識

スキャン・撮影

OCR

構造化

メタデータ

画像処理

コンピュータ

テキストマイニング

インターネット

クラウドソーシング

大規模化

共有と協働

透明性

人工知能

AIとは

- 人工知能学会設立趣意書

- 頭脳の働きに代わる機械が欲しいという人類の夢は、大量の数値データに対して複雑な計算を高速に行うという面では、電子計算機により実現された。現在の情報処理技術はこの意味においては、人間の能力をはるかに越えたものといえるが、一方、思考という本質的な面では、全くといっていいほど無力である。人工知能は大量の知識データに対して、高度な推論を的確に行うことを目指したものである。(1990)

- 人工知能 - Wikipedia

- 「『計算』という概念と『コンピュータ』という道具を用いて『知能』を研究する計算機科学の一分野」を指す語。「言語の理解や推論、問題解決などの知的行動を人間に代わってコンピュータに行わせる技術」、または、「計算機による知的な情報処理システムの設計や実現に関する研究分野」ともされる。(参照 2024-03-28)



AIマップ

学習・認識・予測

推論・知識・言語

発見・探索・創造

AIフロンティア

進化・生命・成長

身体・ロボット・運動

人・対話・情動

データサイエンス

数理統計学

脳科学

画像音声
メディア処理

スポーツ脳科学

運動学

身体生理学

身体性

知能ロボット

自動運転

知能メカトロニクス

アフォーダンス

Webインテリジェンス

データ市場

オブジェクト指向

ソフトウェア工学

セマンティック Web

ベイズ推定

エキスパートシステム

知識推論

言語メディア処理

オントロジー

シンボルグラウンディング

機械学習

ニューラルネットワーク

統計的学習

パターン認識

計算論的神経科学・脳科学

コンピュータビジョン

強化学習

音声認識

動画像処理

マルチモーダル処理

ジェスチャー認識

認知ロボティクス

知能ロボティクス

記号創発ロボティクス

サイバーフィジカルシステム

知能ロボット

自動運転

知能メカトロニクス

VR

アフォーダンス

数学基礎論

数理論理学

意味論

ファジィ論理

論理プログラミング

帰納推論

演繹推論

プロダクションシステム

ルールベースシステム

因果推論

仮説推論

知識表現

知識グラフ

知識ベース

知識の利用と共有

AI信頼性

AI公平性

説明可能 AI

Well-being Computing

汎用人工知能

AI倫理

プライバシー

意思決定・合意形成

ヒューマンエージェント
インタラクション

対話処理・対話システム

マルチモーダルインタラクション

知的ユーザインターフェース

WEBインタラクション

クラウドソーシング・ヒューマンコンピューテーション

機械翻訳

ヒューマンコンピュータインタラクション

VR

ヒューマンインタフェース

行動変容

行動経済学

認知神経科学

ヒューマンインタフェース

形式言語理論

数理計画法

制約充足問題・充足可能性問題 (CSP/SAT)

探索

ヒューリスティクス

テキストマイニング

プランニング

バイオインフォマティクス

面層生成

音声生成

対敵的生成
ネットワーク
(GAN)

マテリアルズ・インフォマティクス

人工生命

包摂アーキテクチャ

エージェント

マルチエージェント

認知アーキテクチャ

群知能

ゲーム AI

知識獲得・発見

知識の利用と共有

非タスク指向対話

マルチモーダルインタラクション

オークション

クラウドソーシング・ヒューマンコンピューテーション

VR

ヒューマンインタフェース

行動変容

行動経済学

認知神経科学

グラフ理論

ゲーム理論

情報検索

ビッグデータ解析

データベース

データマイニング

ウェブマイニング

オペレーションズリサーチ

分散協調

数理最適化

シミュレーション

遺伝的アルゴリズム

進化計算

群知能

ゲーム AI

知識共有・管理

知的教育システム (ITS)

会話・談話・意図理解

集合知

情報推薦

情報可視化

ソーシャルメディア

市場デザイン

感性

認知科学

計算社会科学

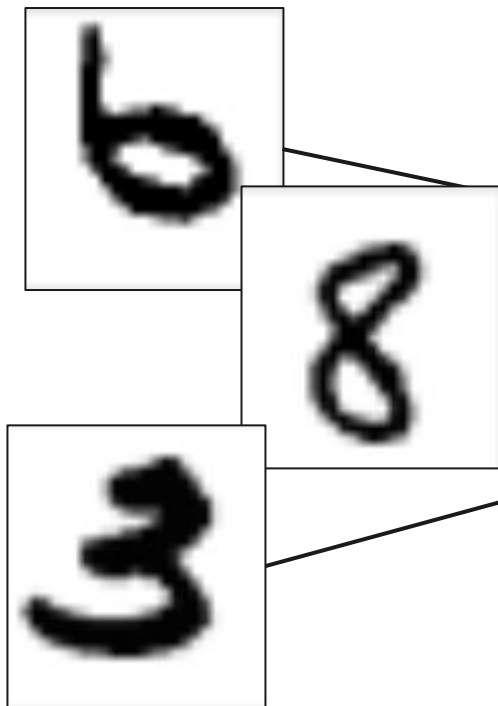
発達心理学

発達科学

生物学

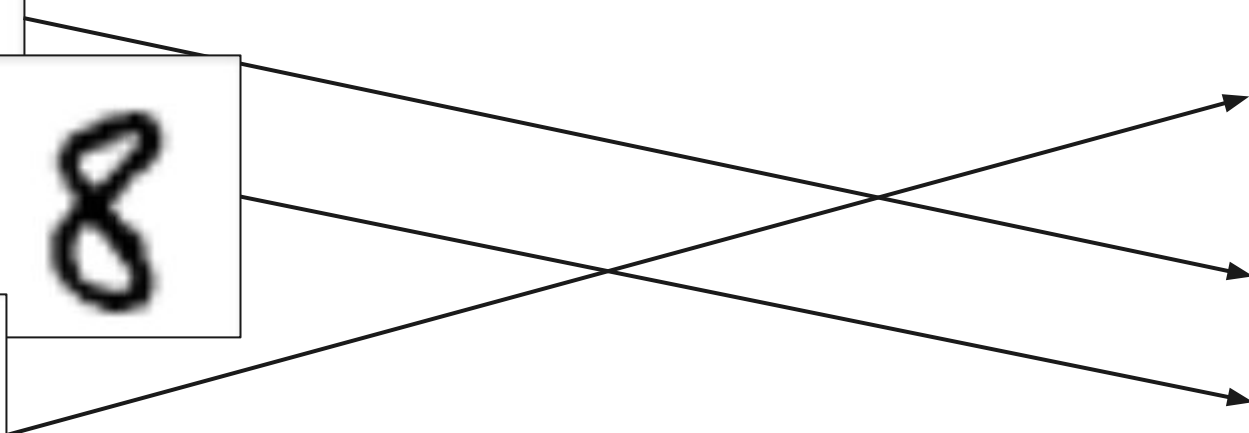
AIによる認識

手書き数字



望ましい認識結果

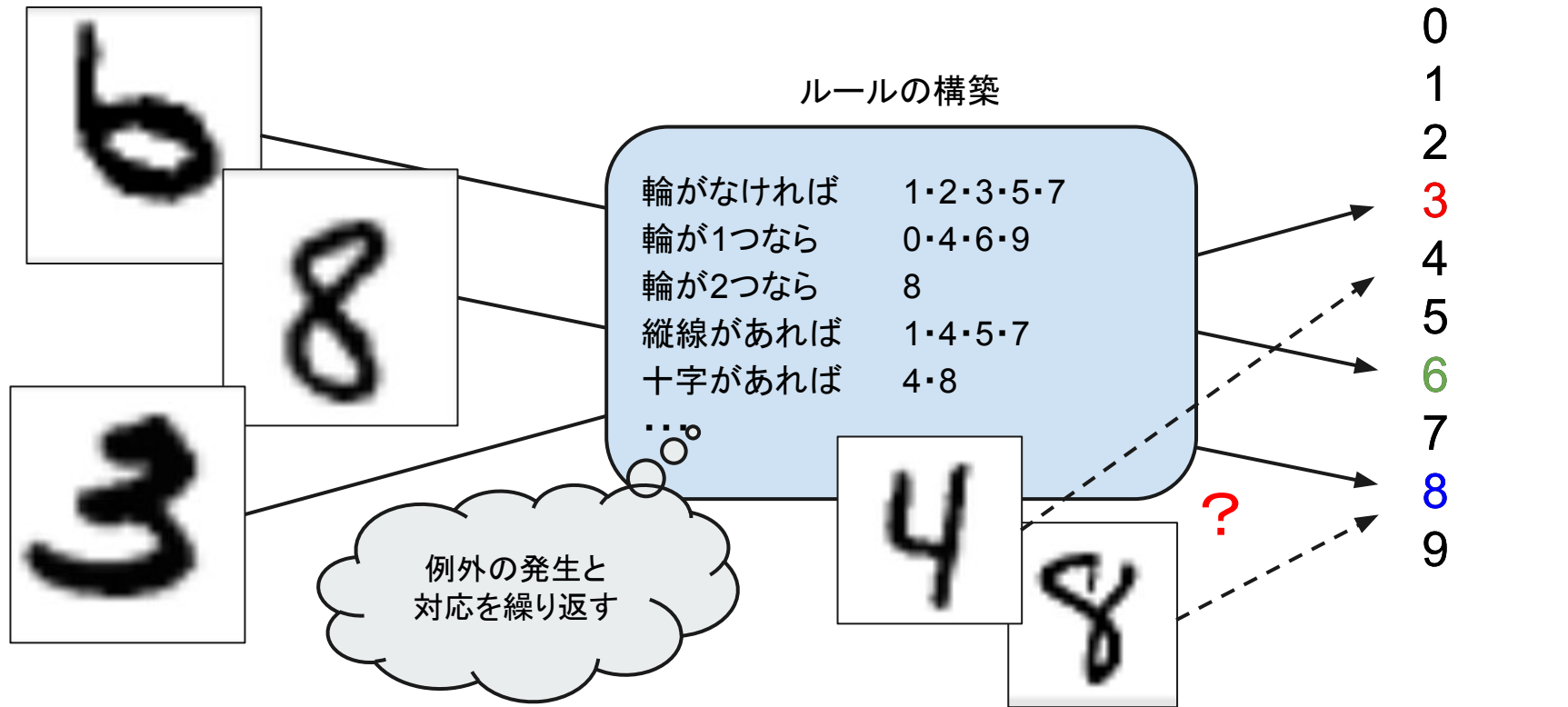
- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9



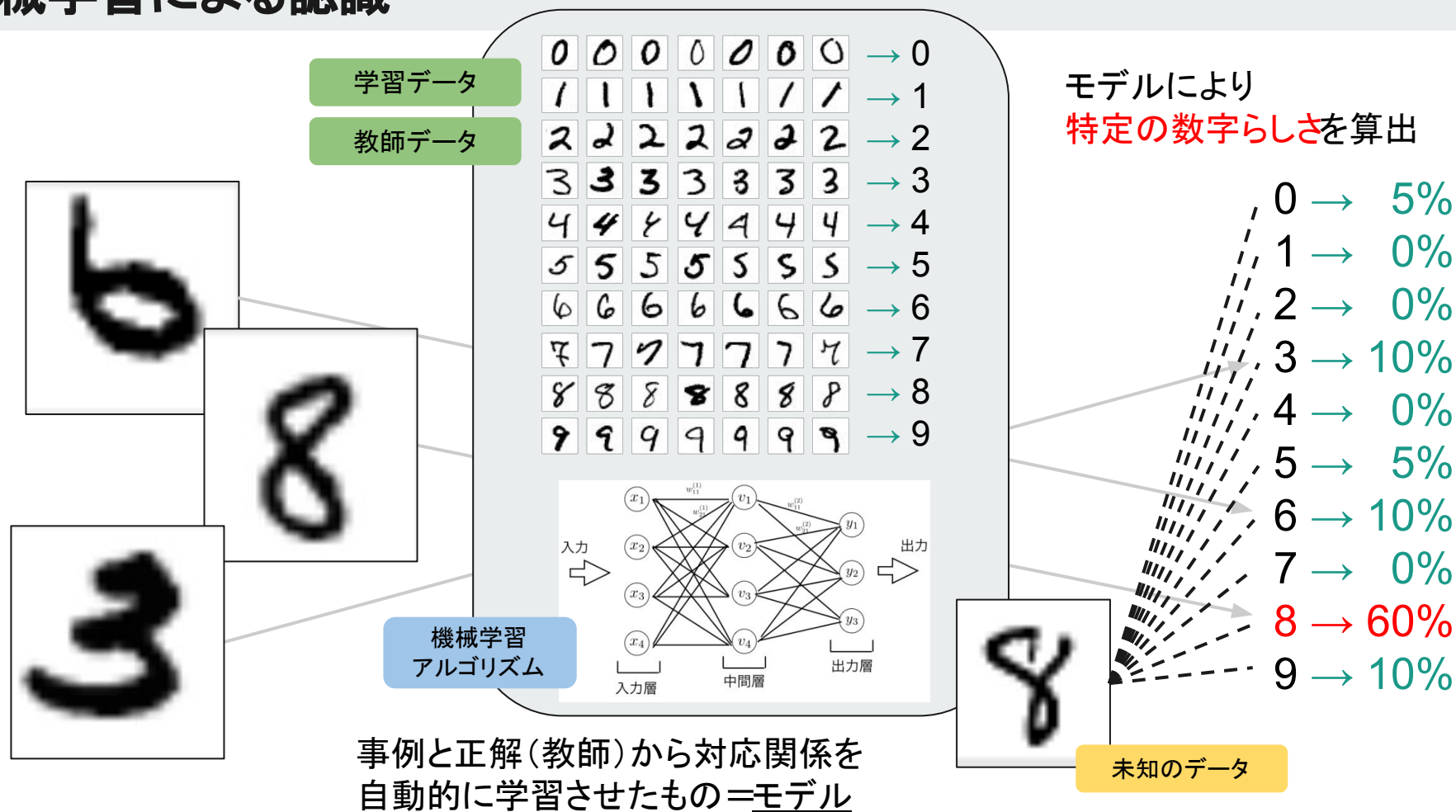
ルールベースによる認識

手書き数字

望ましい認識結果



機械学習による認識



AIにできること

- 経験と知識に基づいて「〇〇らしさ」を把握する技術
 - 経験:すでに存在している大量の情報・データ
 - 知識:論理や手続きによって与えられる、ものごとの関係性

- 例

- 数字らしさ・文字らしさ

- 1・2・3・A・あ・阿・0

- 動物らしさ

- 🐕 · 🐕 · 🐕 · 🐕 · 🐱 · 🐱 · 🐱

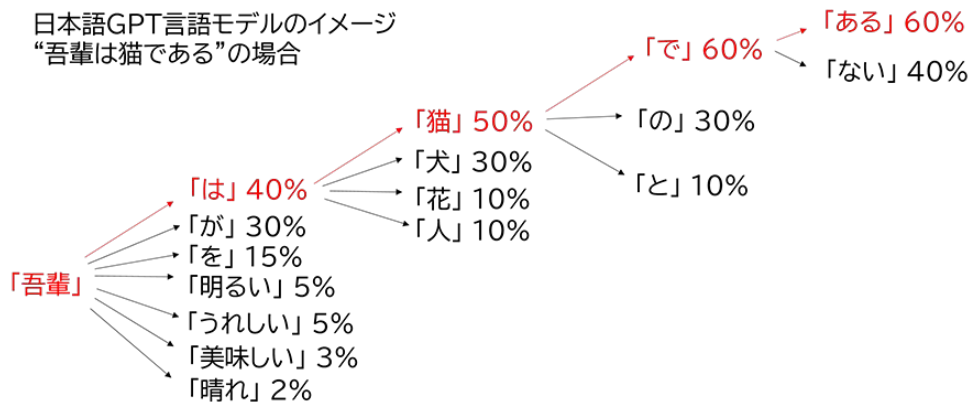
- 言語らしさ

- 意味のある記号のつながり

- 単語

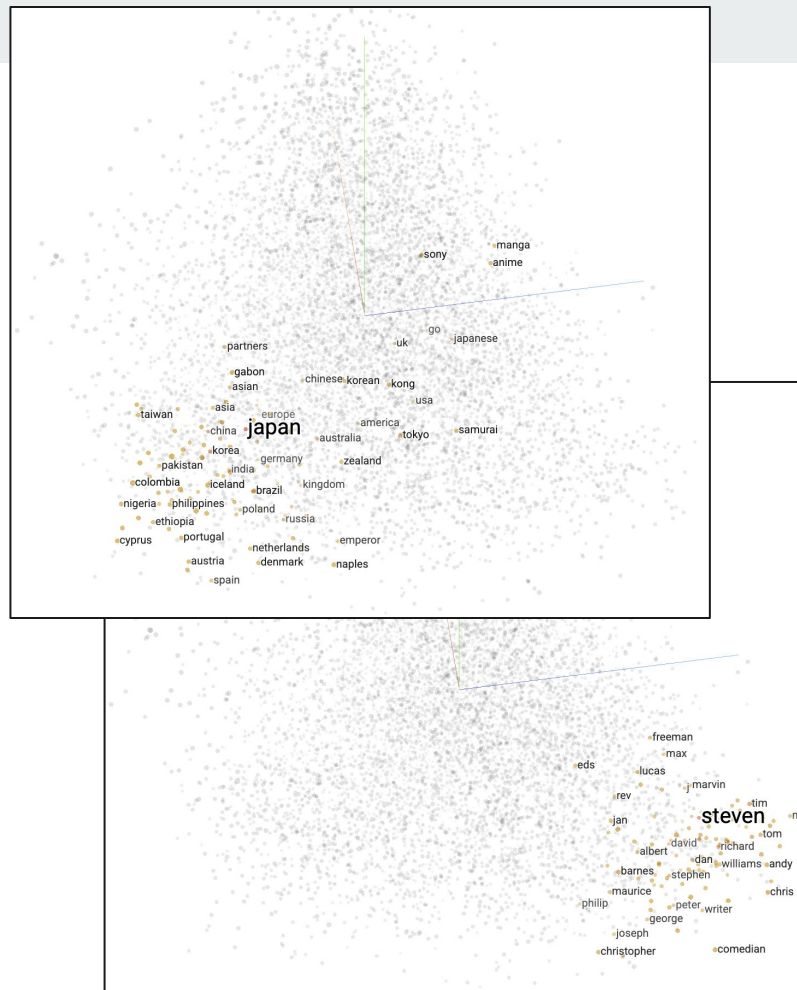
- 流暢な記号のつながり

- 文法・文脈



AIにできること

- 「〇〇らしさ」の把握が可能にすること
 - 認識: 与えられた選択肢から選ぶ
 - 検索: 類似性の高いものを探す
 - 生成: 入力から出力に変換する
 - 要約: 長い文章→短い文章
 - 翻訳: 日本語→英語
 - 対話: 質問→応答
- 数値化(ベクトル化)による汎用性
 - 言語を問わない
 - メディアを問わない
 - テキスト・画像・動画・・・
 - 「距離」を機械的に計算できる
 - 似ているもの・似ていないもの



メタデータの自動入力

🌟 New Record From AI Assistant

Enter the metadata and/or images you want to use to generate the record. The new record draft will be pushed to MDE when ready.

Title*

Enter Text Here

This field is required

Content Note

Enter Text Here

Author

Enter Text Here

ISBN

Enter Text Here

Attachments

Upload up to 4 images of the book (e.g.

The Complete Illustrated Lewis Carroll (99191911200121)

Level [60]: [60] University Level | Record | AI Enriched | 02 | 78 | MARC21 Bib

LDR	00481nam#a2200133##4500
001	99191911200121
005	20241229121337.0
008	241229s1996###x#####000#0#eng#d
020	\$Sa 978-1-95326-897-7
041	\$Sa eng
100 1	\$Sa Carroll, Lewis, \$Sd 1832-1898 \$S0 http://id.loc.gov/authorities/names/n79056546 \$S1 http://viaf.org/loc/viaf/119191911
245 1 3	\$Sa The Complete Illustrated Lewis Carroll
264 1	\$Sa Hertfordshire \$Sb Wordsworth Editions Limited \$Sc 1996
336	\$Sa text \$S2 rdacontent
337	\$Sa unmediated \$S2 rdamedia
338	\$Sa volume \$S2 rdacarrier
520	\$Sa Publisher description: "Lewis Carroll (Charles Lutwidge Dodgson) is famed for his magical story drawings. However, inspired by the insatiable Victorian appetite for party games, tricks and conundrums, he created dozens of other verses, stories, acrostics and puzzles, and the mock heroic nonsense verse 'Jabberwocky'. He was a mathematician, logician, academic author of learned theses, gifted pioneer of portrait photography, and a photographer." \$S7 Generated by AI
500	\$Sa Cover illustration by John Tenniel, 1921.
588	\$Sa Part of the metadata in this record was created with the help of AI Metadata Assistant
650 0	\$Sa Fantasy fiction, English. \$Sv Illustrations.
650 0	\$Sa Children's stories, English. \$Sv Illustrations.
650 0	\$Sa Nonsense literature, English. \$Sv Illustrations.
600 1 0	\$Sa Carroll, Lewis, \$Sd 1832-1898 \$Sv Illustrations.
655 7	\$Sa Illustrated works. \$S2 lcgtf

The Complete Illustrated Lewis Carroll (99191911200121)

Level [60]: [60] University Level | Record | 02 | 48 | MARC21 Bib




LDR	00481nam#a2200133##4500
001	99191911200121
005	20241229121337.0
008	241229s1996###x#####000#0#eng#d
100 1	\$Sa Carroll, Lewis, \$Sd 1832-1898 \$S0 http://id.loc.gov/authorities/names/n79056546 \$S1 http://viaf.org/loc/viaf/119191911
245 1 3	\$Sa The Complete Illustrated Lewis Carroll
264 1	\$Sa Hertfordshire \$Sb Wordsworth Editions Limited \$Sc 1996
336	\$Sa text \$S2 rdacontent
337	\$Sa unmediated \$S2 rdamedia
338	\$Sa volume \$S2 rdacarrier

🌟 Enrich Record With AI Assistant

Enrichment will be done according to the metadata in the record. You may also upload images of the book. In order for the AI assistant to be able to enrich the record, it will be released and locked for editing. The enriched draft will be pushed to MDE when ready.

Title
The Complete Illustrated Lewis Carroll

Attachments
Upload up to 4 images of the book (e.g. back cover, colophon) 📷

	File Name	
1	 Complete illustrated lewis carol copyright page.jpg	...
2	 Complete illustrated lewis carol cover.jpg	...
3	 Complete illustrated lewis carol front cover.jpg	...

Drop a file or [Browse files](#)

Alma AI Metadata Assistant in the Metadata Editor

[https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_\(English\)/Metadata_Management/005Introduction_to_Metadata_Management/The_AI_Metadata_Assistant_in_the_Metadata_Editor](https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/Metadata_Management/005Introduction_to_Metadata_Management/The_AI_Metadata_Assistant_in_the_Metadata_Editor)

類似テキストの検索

- 『古今集』などの主要な歌集は、誰もがその内容を熟知しているものであった。したがって、その中のほんの一部を引用したり、その内容を暗示するような語句を用いることで、その文脈に、つよい意味を与えることができる。(近藤2024)
 - 源氏物語:15366センテンス
 - 古今集: 1063首
- ベクトル化+近傍検索による類似箇所発見
 - 古今集 396
 - 飽かずして別るる涙滝にそふ水まさとやしもは見るらむ
 - 源氏物語 関谷
 - 行くと来とせきとめがたき涙をや絶えぬ清水と人は見るらむ え知りたまはじ
かしと思ふにいとかなし

組み合わせ: 16334058通り

類似画像の検索

- 函館市中央図書館「デジタル資料館」のポスター画像003枚の可視化
 - タイトルの文字列検索(港まつり)
 - 画像自体の類似度
 - タイトルの「意味」の類似度



画像の類似度順

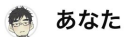


タイトルの類似度順



大規模な資料を対象とした対話

- 東京大学百年史
 - 通史3巻・約3500ページ
 - NDLOCRを用いた全文テキスト化
 - ChatGPTカスタムバージョンの作成
 - 資料に限定した回答をするように指示



あなた

史誌編纂掛が廃止された年と、その理由を教えてください



UTokyo 100 Years Bot

史誌編纂掛は明治二十六年（1893年）三月に廃止されました。廃止の主な理由は、史誌編纂掛の活動に対する政治的な対立と、内閣と神道関係者との間の緊張が高まったことにあります。具体的には、史誌編纂掛のあるメンバーが『古事記』や『日本書紀』の神話を批判したことで神道家からの強い反発を招き、これが対立に発展したこと、また、その研究内容が保守的な立場の人々や内閣からの批判を受けたことが挙げられます。さらに、史誌編纂掛の一部の研究が、大義名分論の立場から史上に名を留める人物を否定的に扱ったことも、保守派からの不満を高める一因となりました。

東京大学百年史 https://www.u-tokyo.ac.jp/adm/history/03_03_j.html

廃止の事情

事績の考証に精力を尽すのが何故いけないのか、編修するにふさわしい人物がいなかったのか、実は修史事業の廃止は、この二点の廃止理由の裏側に潜んでいる。

このころの重野安繹を初め、史誌編纂掛の人々の研究論文として発表されたものなかには、軍記・物語のみに見える事柄・人物を、作者の捏造とする考証が多かった。とくに『大日本史』などで大義名分論の立場から忠臣とされた人物が史上から抹殺されることに対して、保守的な立場の人々は快くこれを迎えてはいなかった。これに乗じて、内閣の内務省局から独立して、旧の如く内閣と対等な立場の神祇官になろうと計る神道関係者は、『古事記』や『日本書紀』の神話を批判する史誌編纂掛の人々の言動に神経を鋭くして、折あるごとに論難を加えていた。とくに、史誌編纂委員のひとり久米邦武の『史学雑誌』の二三二五号（明治二十四年十二月）に連載した「神道は祭天の古俗」を、在野の啓蒙史家田口卯吉が自らの主宰する『史海』八号（明治二十五年一月）にすぐさま転載すると、たちまちこれは史誌編纂掛と神道家の対立に発展していった。

〔三〕宮地正人「久米邦武事件の政治史的考察——天皇制國家の確立と歴史学との関係によせて——」（東京歴史科学研究会編『転換期の歴史学』所収、昭和五十四年）

内務省は、久米の論文が載っている両雑誌を治安妨害の嫌で発行停止とし、文部省も久米を非職とした。久米は辞表を提出して大学を去った。そして、翌二十六年三月、史料編纂掛は廃止に決定し、修史局以来、修史事業とともに生きてきた文科大学講師・史誌編纂委員長の重野も罷免された。もちろん、井上の考え

第三章 帝国大学と教育・学術

1039

大規模な資料を対象とした対話

この画面はGoogle NotebookLMの対話インターフェースのスクリーンショットです。左側の「ソース」パネルには、12個のPDFファイルがリストアップされています。右側の「チャット」パネルには、AIが生成した回答が表示されています。

ソース: これから委員会2012-2024

チャット: 「これからの学術情報システム構築検討委員会」の名称決定過程を説明せよ。

「これからの学術情報システム構築検討委員会」の名称決定過程に関する直接的な記述は、提供された資料には見当たりません。しかし、委員会が設立され、活動を始めた背景や、その目的から名称の意図を推測することは可能です。

- ・**設立の背景:**
 - 資料全体を通じて、**学術情報基盤の将来像**を議論し、その構築に向けた具体的な検討を行う必要性が繰り返し強調されています 1 ...
 - 特に、**電子リソースの増加**、**メタデータ管理の複雑化**、**国際的な連携の必要性**といった、従来の図書館システムでは対応が難しい課題への対応が求められていました 3 ...
 - これらの課題に対処するため、既存の枠組みを超えた**新しい学術情報システムの構築**が不可欠であるという認識が共有されていた 4 ...
- ・**名称の意図:**
 - 「これからの」という言葉は、従来のシステムや考え方にとらわれず、**将来を見据えたシステム構築**を目指すという強い意志を示しています 2 ...
 - 「学術情報システム」という言葉は、**大学図書館における学術情報資源の管理・提供**を包括的に捉え、**目録、ILL、電子リソース管理、デジタルアーカイブ**など、多岐にわたる要素を統合的に扱うことを示唆しています 3 ...
 - 「構築検討委員会」という言葉は、**システムを実際に作るだけでなく、その在り方、必要な機能、運用体制**など、幅広い検討を行う組織であることを明確にしています 1 ...

委員会名称は、これらの要素を踏まえ、**将来の学術情報システムのあるべき姿を議論し、その実現に向けた具体的な検討を行う組織**であることを示すために名付けられたと考えられます。

委員会では、**NACSIS-CAT**のような既存のシステムを単に改善するだけでなく、**新たな情報インフラの構築**を目指しており、この点が名称にも反映されていると解釈できます 5 ...。また、**メタデータの重要性**が強調されている点も、名称を理解する上で重要な要素で

入力を開始します... 40 個のソース

電子ブックの図書館における維持管理の課題を論ぜよ。 NACSIS-CAT データのオープン化に関する課題を分析せよ。 ERDB

NotebookLM は不正確な場合があるので、回答を再確認してください。

デジタルアーカイブと AI

- デジタルアーカイブの構築・活用のさまざまな局面で利用可能
 - データの整形・構造化
 - テキスト化 (OCR・キャプション生成)
 - 固有表現の抽出 (人名・地名・時間情報・概念・・・)
 - 統計・分析
 - 可視化・要約
 - ...
- それぞれのタスクは単一の技術では解決できない
 - 例: 画像→OCR→テキスト→ベクトル化→数値→近傍検索→引き歌
 - タスクを複数の技術モジュールの組み合わせで表現する
 - 要求定義→要件定義→基本設計→詳細設計→開発・実装
 - 情報システムの開発プロセスそのもの
 - ファシリテーターとしてのデジタルアーカイブ専門家の必要性

デジタルアーカイブと AI

- 経験と知識に基づいて「〇〇らしさ」を把握する技術
 - 経験: **すでに存在している**大量の情報・データ
 - 知識: 論理や手続きによって与えられる、ものごとの関係性
- ハルシネーション(口から出任せ)
 - 「〇〇らしさ」には正・誤・真・偽に関するメタ情報は含まれない
 - 自身で解釈可能なものごとをAIに解釈させる
 - 回答の範囲を限定させる
- 存在していないものごとはAIには取り扱えない
 - 本当はあるのに入手できていない?
 - 本当にない?
 - AIの出力から区別することは原理的に不可能
 - デジタルアーカイブの可能性