平成23年度学術ポータル担当者研修　2011年8月24日

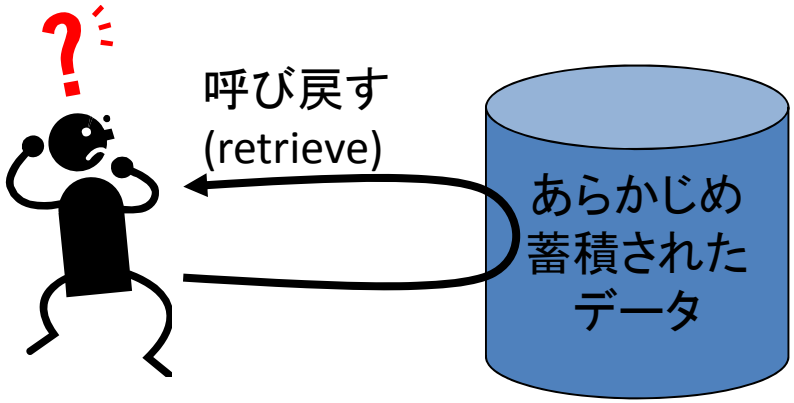
学術情報流通を実現する技術（1）
--要素技術(検索、DB等の基盤技術に特化した話を中心に)--

物質・材料研究機構 科学情報室　高久雅生
Code4Lib JAPAN コアメンバー

Twitter: @tmasao

情報検索とは（由来）

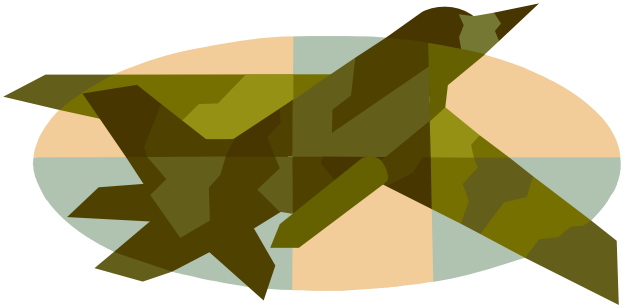
- **情報検索**
  - **IR: i**nformation (storage and) **r**etrieval
    - 情報(information) を呼び戻すこと(retrieval)
    - 元は**i**nformation storage and **r**etrieval　情報の蓄積と検索
  - 1950年にムアーズ(Calvin N. Mooers)が初めて定義
  - 1960年代に広く使われるようになる
  - (search: これも「検索」と訳すが。。。)



retriever（レトリーバー）:
獲物をくわえて戻って
くるように訓練された猟犬

データベースの起源

- 1950年代
- 米国国防総省が戦力に関する情報を保管、集中管理するためコンピュータを使ったライブラリーを開発
- データの基地(data base)から由来



データベースの定義(1)

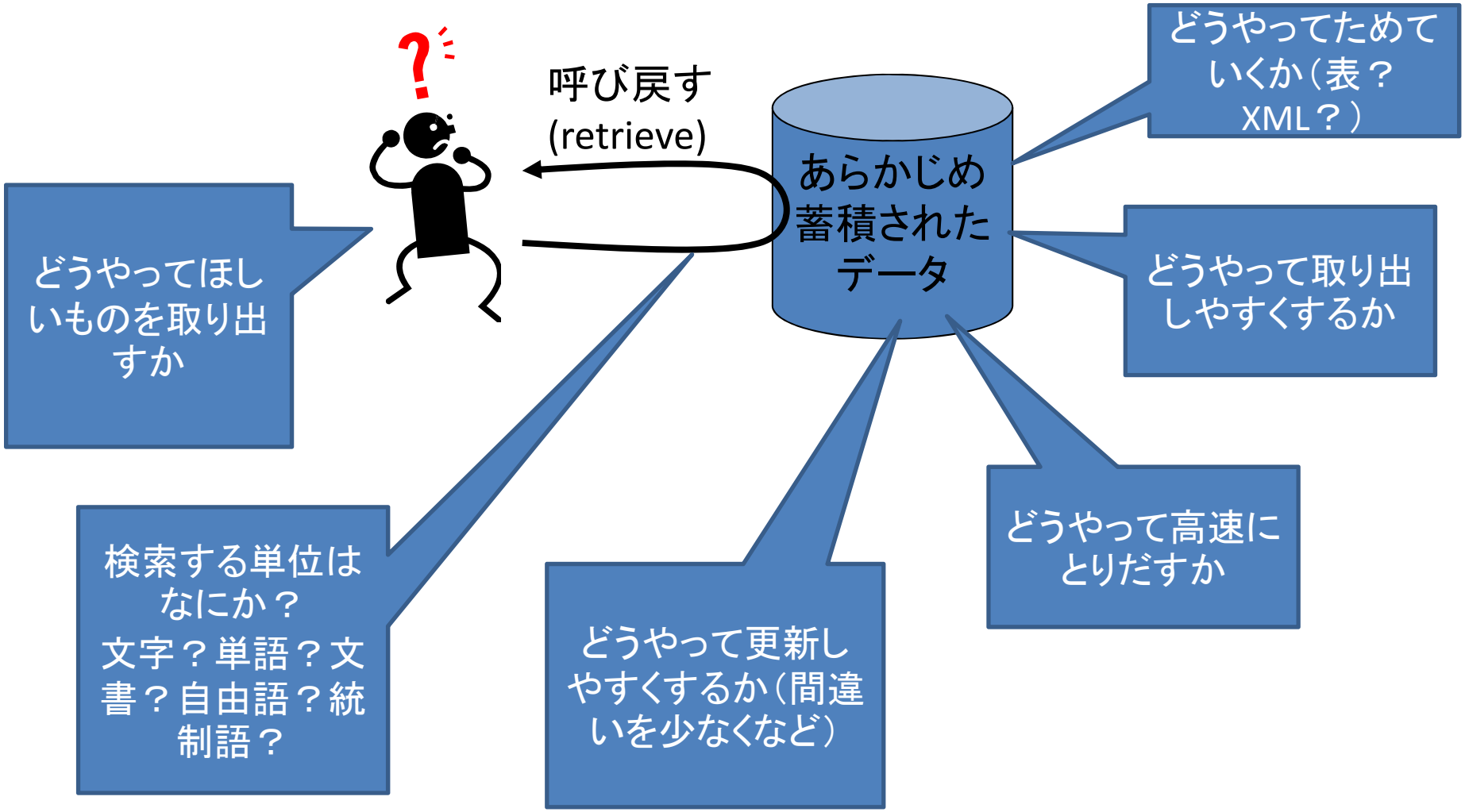
- 著作権法二条十の三
  - 論文、数値、図形その他の情報の集合物であって、それらの情報を電子計算機を用いて検索することができるように体系的に構成したもの
- 日本工業規格(JIS)
  - 適用業務分野で使用するデータの集まりであって、データの特性とそれに対応する実態の間の関係とを記述した概念的な構造によって編成されたもの(X0017)
  - 特定の規則に従って電子的な形式で、一か所に蓄積されたデータの集合であって、コンピュータでアクセス可能なもの(X0807)

データベースの定義(2)
--データベースの特徴--

- コンピュータを用いて検索できることが重要。情報が電子メディアに蓄積され、コンピュータを使用して検索できる状態になっている。
- データや情報がコンピュータ処理できるように体系的に整理され、統合化・構造化されて蓄積・保存されており、必要な情報だけを部分的に取り出せる。
- 蓄積情報の検索や更新が容易に行えるよう、効率化を図ったものである。

※ ちなみに、ヨーロッパにおけるデータベースの定義では、
コンピュータを使用するかしないか、電子的であるかどうかについては特に限定していない

検索にまつわる様々な観点（一部）

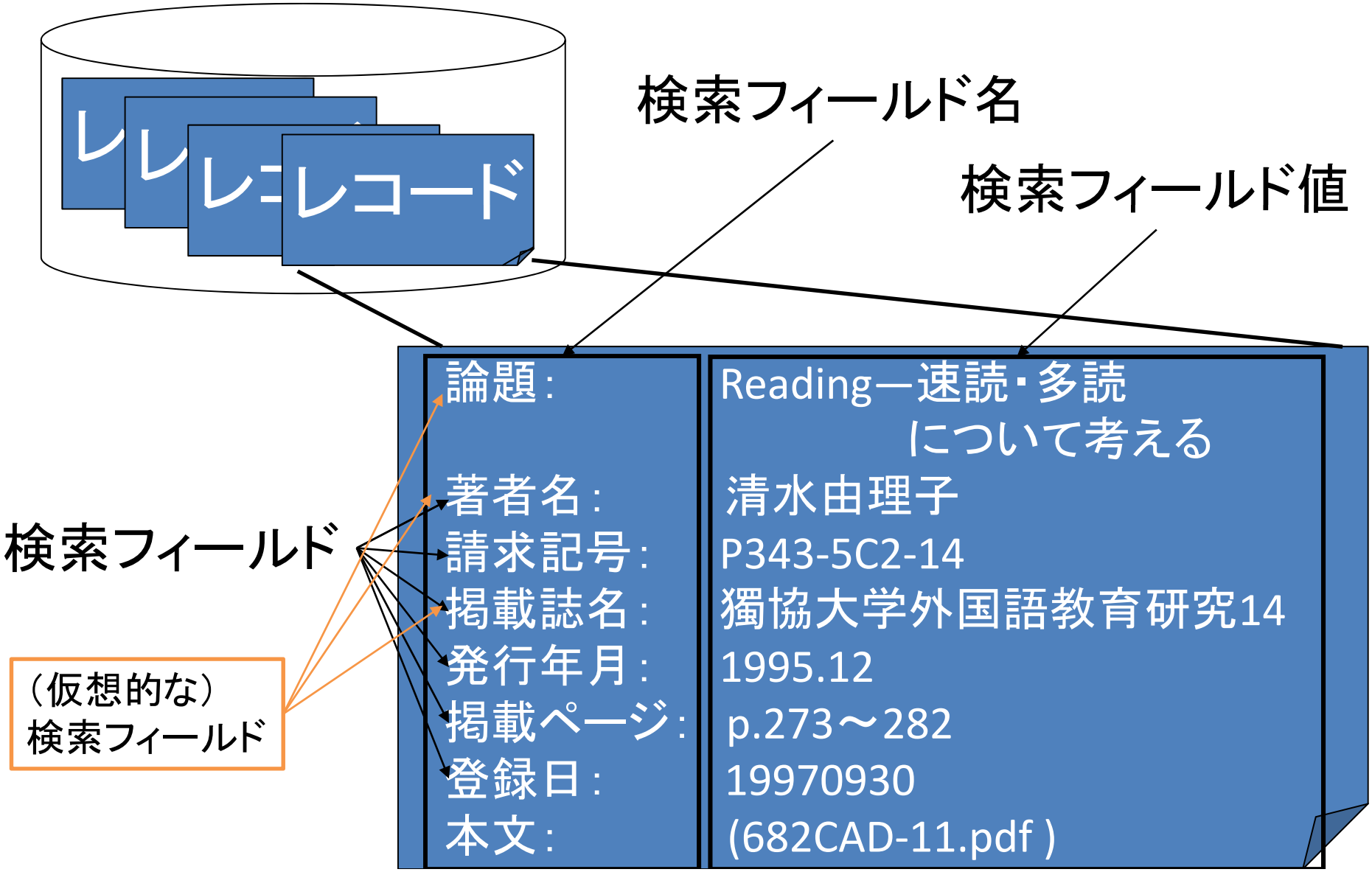


どうやってためていくか？表？XML?
-- データベースの種類 --

- リレーショナルデータベース（RDB）
  - 表としてデータを扱う
- オブジェクトデータベース
  - オブジェクトとしてデータを扱う
- XMLデータベース
  - XMLとしてデータを扱う
- 全文データベース

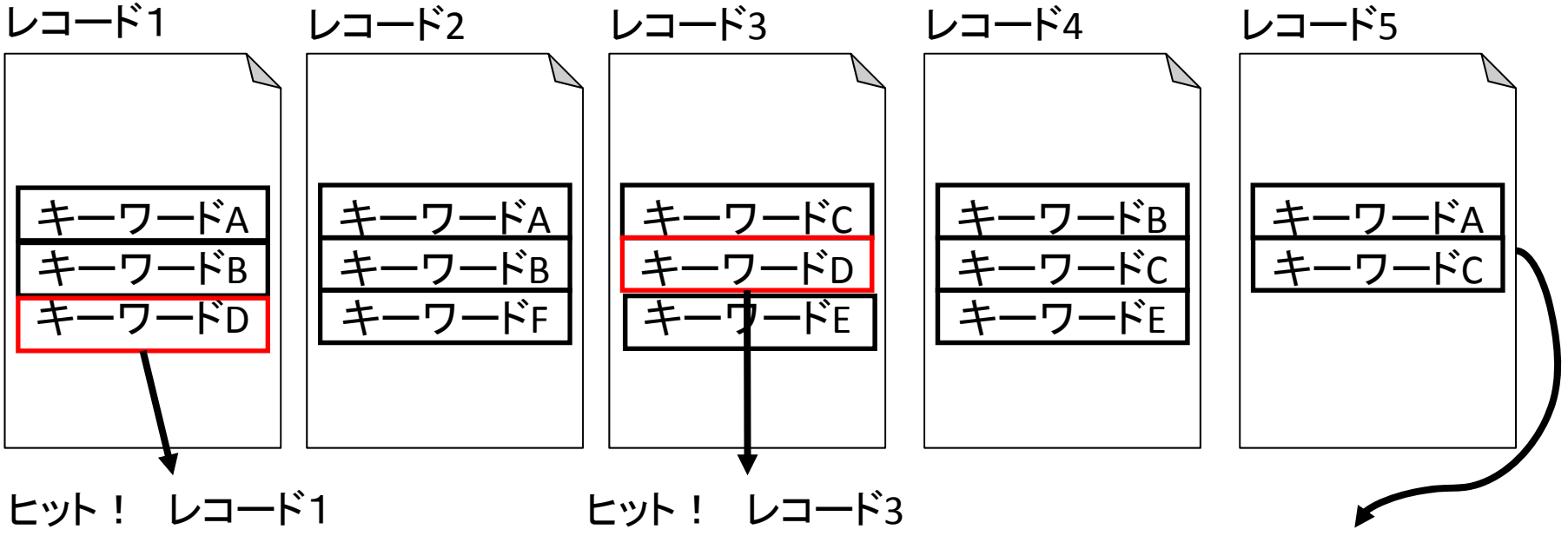
  - いろいろある、なんでもRDBにすればよいというものではない
  - 書誌データ、全文データはRDBには向いていない

どうやって取り出しやすくするか？
-- レコードと検索フィールド--



どうやって高速にとりだすか(1)
--なにも仕掛けがないと。。。--

- レコードを最初から最後まで順番に検索
- レコードが多くなると時間がかかってしまう



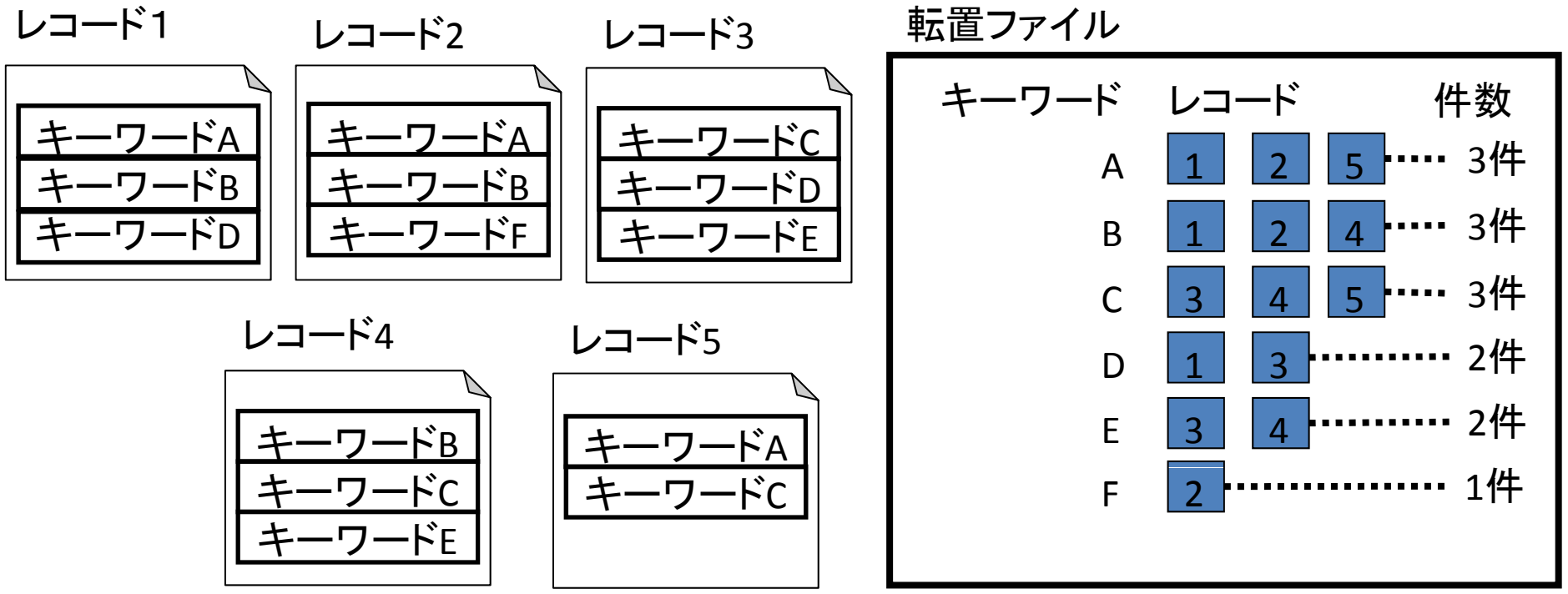
例:「キーワードD」を検索する

どうやって高速にとりだすか(2)
-- インデックス を用意する--

- 高速にデータにアクセスするために必要
- インデックス方式
  - 転置ファイル（Inverted file; インバーテッドファイル）
    - もっともよく使われる方式
  - TRI木
  - Suffix Array
  - ・・・いろいろある

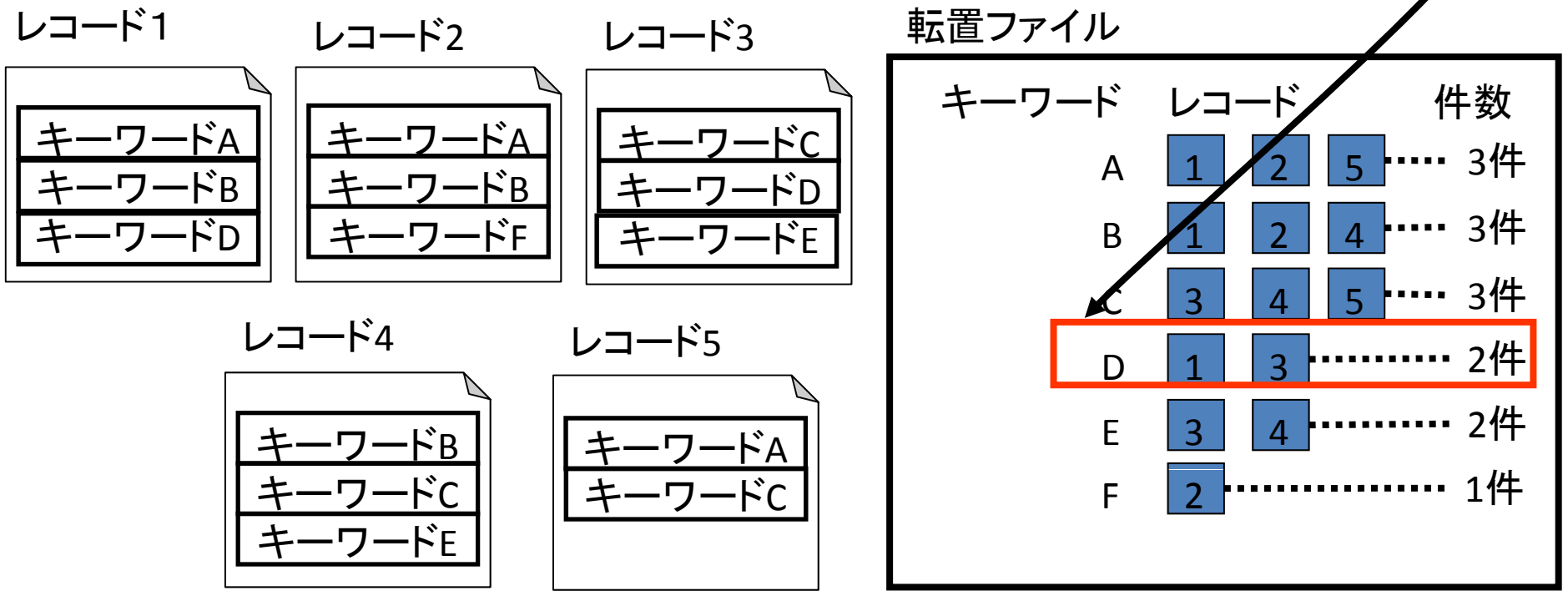
転置ファイル(1)

- 検索キーごとにレコードを集計したファイルを作成する方式
- Inverted file（転置ファイル、倒置ファイル）ともいう



転置ファイル(2)

- 例:「キーワードD」を検索する
- 欠点:あらかじめ転置ファイルを作らなければならない、ファイルの容量増



転置ファイル(3)
--どうやってキーワードを取り出す？--

レコード例(1):「http://nyti.ms/po6J1Z うーむ、Aaron Swartzが昨年9月のJSTORからの大量ダウンロード容疑で逮捕・起訴とは、絶句。#librahack」

http://twitter.com/tmasao/status/9360716986396416

- http://nyti.ms/po6J1Z
- うーむ
- Aaron
- Swa
- が
- 昨年
- 9
- 月
- の
- JSTOR
- から
- の
- 大量
- 逮捕
- ・
- 起訴
- とは
- 絶句
- #librahack

意味のある語単位で切る＝形態素解析
（日本語の文章の意味（かかり受けや品詞など）を解析して意味のある語で区切るようにする）

転置ファイル(4)
--どうやってキーワードを取り出す？--

レコード例(1):「http://nyti.ms/po6J1Z うーむ、Aaron Swartzが昨年9月のJSTORからの大量ダウンロード容疑で逮捕・起訴とは、絶句。#librahack」

http://twitter.com/tmasao/status/9360716986396416

- http://nyti.ms/po6J1Z
- うー
- ーむ
- Aaron
- Swartz
- が昨
- 年9
- 9月
- 月の
- JSTOR
- から
- らの
- の大
- ウン
- ンロ
- ロー
- ード
- ド容
- 容疑
- 捕・
- ・起
- 起訴
- 訴と
- とは
- は絶
- 絶句

N-gram: 強引に、2文字づつなどで強制的に切る方法

転置ファイル(5)

レコード例(2):「要はライブラリアンに必要なのは、国内文学読んだり、月9のドラマを見たりするんじゃなくて、存在をかけて勉強することじゃないかなあ。本音を言えば、悪いが、みんな遊びすぎ。専門家を名乗るなら、もっと勉強したほうがいい。」

http://twitter.com/arg/status/20007995359

形態素解析

- 要は
- ライブラリアン
- に
- 必要
- な
- の
- は
- 、
- 国内
- 文学
- 読ん
- だり
- 、
- 月
- 9
- の
- ドラマ
- を
- 見
- たり
- する
- ん
- じゃ
- なく
- て
- 、
- 存在
- を
- かけ
- て
- 勉強
- する
- こと
- じゃ
- ない
- か
- なあ
- 。
- 本音
- を
- 言え
- ば
- 、
- 悪い
- が
- 、
- みんな
- 遊び
- すぎ
- 。
- 専門
- 家
- を
- 名乗る
- なら
- 、
- もっと
- 勉強
- し
- た
- ほう
- が
- いい
- 。

転置ファイル (6)

レコード例(1)

http://nyti.ms/po6J1Z
うーむ
Aaron
Swartz
が
昨年
9
月
の
JSTOR
から
の
大量
ダウンロード
容疑
で
逮捕
・
起訴
とは
絶句
#librahack

レコード例(2)

要は
ライブラリアン
に
必要
な
のは
、
国内
文学
読ん
だり
、
月
9
の
ドラマ
を
見
たり

ひとまとめにして、
辞書順に並べ替え

#librahack	1	とは	1	昨年	1
9	1,2	な	2	絶句	1
Aaron	1	なあ	2	専門	2
JSTOR	1	ない	2	存在	2
Swartz	1	なく	2	逮捕	1
http://nyti.ms/po6J1Z	1	なら	2	大量	1
、	1,2	に	2	読ん	2
。	2	の	1,2	必要	2
・	1	は	2	文学	2
いい	2	ば	2	勉強	2
	1	ほう	2	本音	2
	2	みんな	2	名乗る	2
	2	もっと	2	遊び	2
から	1	を	2	容疑	1
が	1,2	ん	2	要は	2
こと	2	ダウンロード	1		
し	2	ドラマ	2		
じゃ	2	ライブラリアン	2		
すぎ	2	悪い	2		
する	2	家	2		
た	2	起訴	1		
たり	2	月	1,2		
だり	2	見	2		
て	2	言え	2		
で	1	国内	2		

転置ファイル (7)

#librahack	1
9	1,2
Aaron	1
JSTOR	1
Swartz	1
http://nyti.ms/po6J1Z	1
、	1,2
。	2
・	1
いい	2
うーむ	1
か	2
かけ	2
から	1
が	1,2
こと	2
し	2
じゃ	2
すぎ	2
する	2
た	2
たり	2
だり	2
て	2
で	1

とは	1
な	2
なあ	2
ない	2
なく	2
なら	2
に	2
の	1,2
は	2
ば	2
ほう	2
みんな	2
もっと	2
を	2
ん	2
ダウンロード	1
ドラマ	2
ライブラリアン	2
悪い	2
家	2
起訴	1
月	1,2
見	
言え	2
国内	2

昨年	1
絶句	1
専門	2
存在	2
逮捕	1
大量	1
読ん	2
必要	2
文学	2
勉強	2
本音	2
名乗る	2
遊び	2
容疑	1
要は	2

「ダウンロード」で検索する場合

- 検索する際には、ひとまとめにした単語リストに対して照合する。
- 単語リストの長さをTとした場合、平均で $Log_2(T)$ 回の照合で検索できる（＝2分探索）

⇒ 対数をとるので、飛躍的な速度が実現できる。

⇒ 100語で 平均3.3回

⇒ 100万語で平均20回

転置ファイル(8)
--どうやって検索する？--

- レコード例:
  - (1)「http://nyti.ms/po6J1Z うーむ、Aaron Swartzが昨年9月のJSTORからの大量ダウンロード容疑で逮捕・起訴とは、絶句。#librahack」
  - (2)「要はライブラリアンに必要なのは、国内文学読んだり、月9のドラマを見たりするんじゃなくて、存在をかけて勉強することじゃないかなあ。本音を言えば、悪いが、みんな遊びすぎ。専門家を名乗るなら、もっと勉強したほうがいい。」
- Q: [Swartz] -> [Swartz] -> Hit! (1)
- Q: [専門家] -> [専門][家] -> Hit! (2)
- Q: [9月] -> [9][月] -> Hit? (1) (2)
- Q: [#librahack] -> Hit???
- Q: [アーロン・シュワルツ] -> No Hit?

インデックスを使った
検索(仕組み/人)
のところで、
それぞれ工夫必要

まとめ

- 情報検索～データベースとは？
- 検索にまつわる観点
- レコードの蓄積からインデックスの構築、検索まで
- 転置ファイルを例に

- 参考文献:
  - 情報アクセスの新たな展開ー情報検索・利用の最新動向. 日本図書館情報学会編. 勉誠出版, 2009, 216p. ISBN: 978-4-585-00278-9
  - Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition). Ricardo Baeza-Yates, Berthier Ribeiro-Neto. ACM Press Books, 2011, 944p. ISBN: 978-0321416919