人文学における オープンデータの活用

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
永崎研宣

「オープンデータ」

- 「*自由に使えて再利用もでき、かつ誰でも再配布できるようなデータのことだ。従うべき決まりは、せいぜい「作者のクレジットを残す」あるいは「同じ条件で配布する」程度である。*」
  - http://opendatahandbook.org/ja/what-is-open-data/index.html
  - CC BY あるいは CC BY SA ?

人文学におけるデータの活用

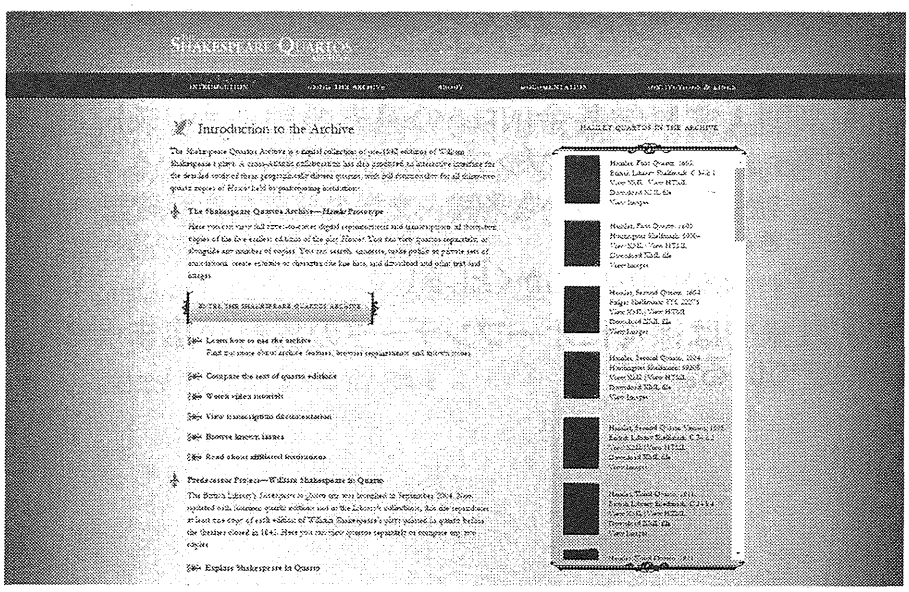
- 多くは非オープンデータ。
  - 著作権保護されている資料のため個人で入力or OCR or 何らかの方法で入手した(主にテクスト)データ
    - データ自体の配布・再利用不能
    - ここから出てきた成果の検証も不可能
  - 一定の条件下で利用を許可されたデータ
    - 学術利用のみ
    - 再配布不可
  - 一定の条件下で再配布も許可されたデータ
    - 学術利用のみ
    - 商用利用不可

人文学向けのオープンデータ？

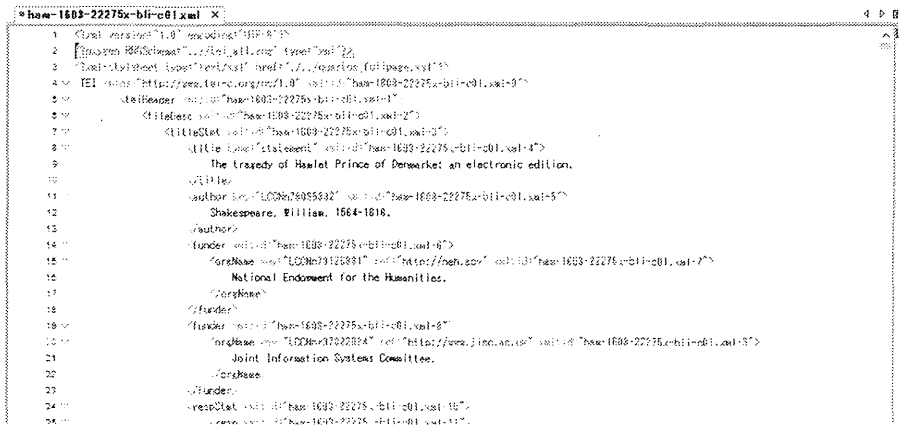
- 米国人文科学基金(NEH)の過去の助成金リスト
  - Data.govからオープンデータとして公開されている
  - 助成金のデータから米国人文学の研究動向が確認できる
  - ・・・というのはともかくとしまして。
- 欧米では徐々にオープンデータの資料が広まってきている。

欧米の人文学向けオープンデータ

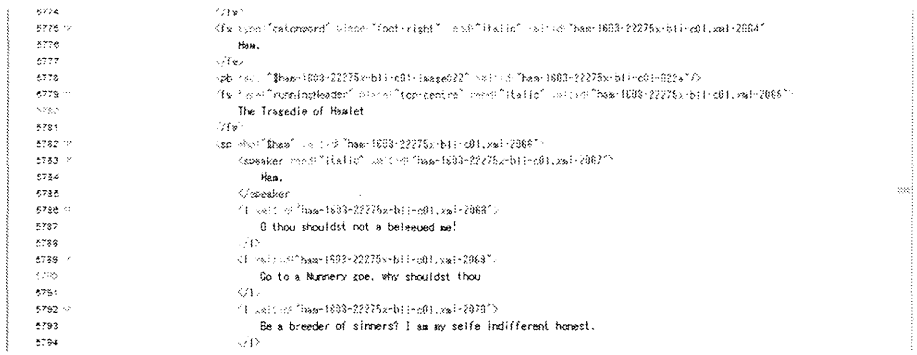
- プラットフォームレベル
  - Europeana
  - Hathitrust / Google Books / Internet Archive
  - 統合検索エンジンとしてのDPLA
  - Gallica @BnF
- 個別の資料レベル
  - シェイクスピア・アーカイブ
    - 各地の図書館の連携により30以上の版本の構造化テクスト作成＋画像公開
    - ⇒専門家による深い作り込み/ CC BY-NC
  - 各種クラウドソーシング翻刻プロジェクト



TEI/XMLによる詳細なマークアップ



内容に基づく様々な構造化



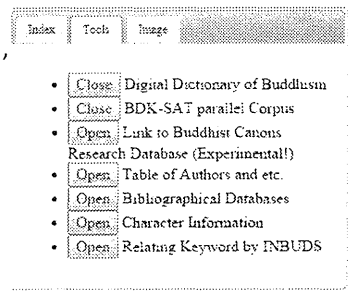
- 「ある登場人物の発言すべて」「Aに続くBの発言のみをすべて」などの様々な情報をTEI/XMLの構造を利用して簡単に抽出可能。

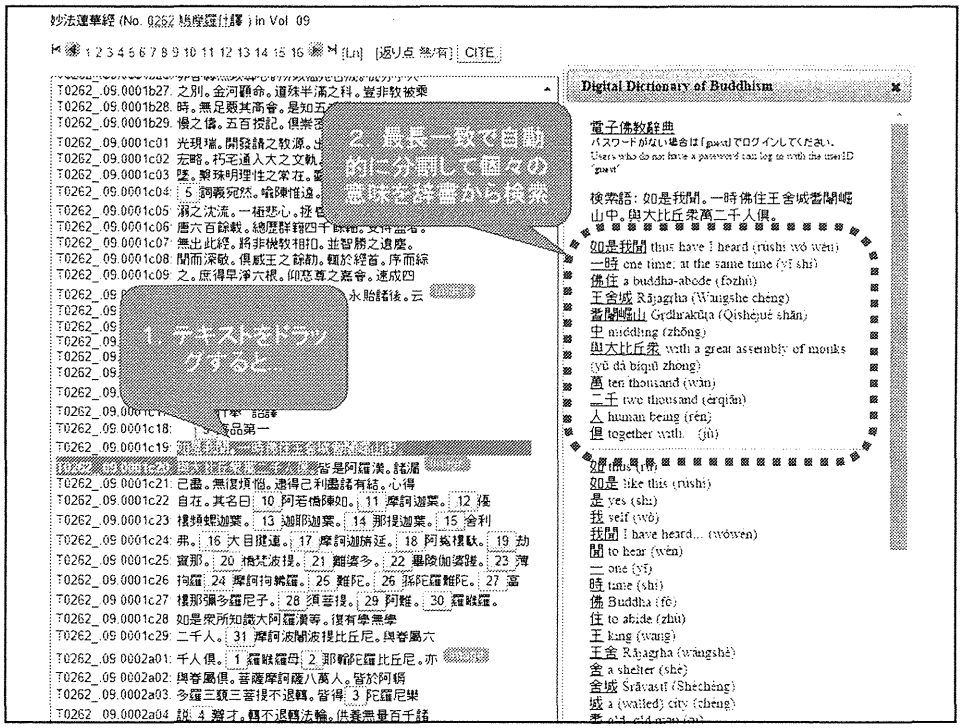
一方、我が国では...

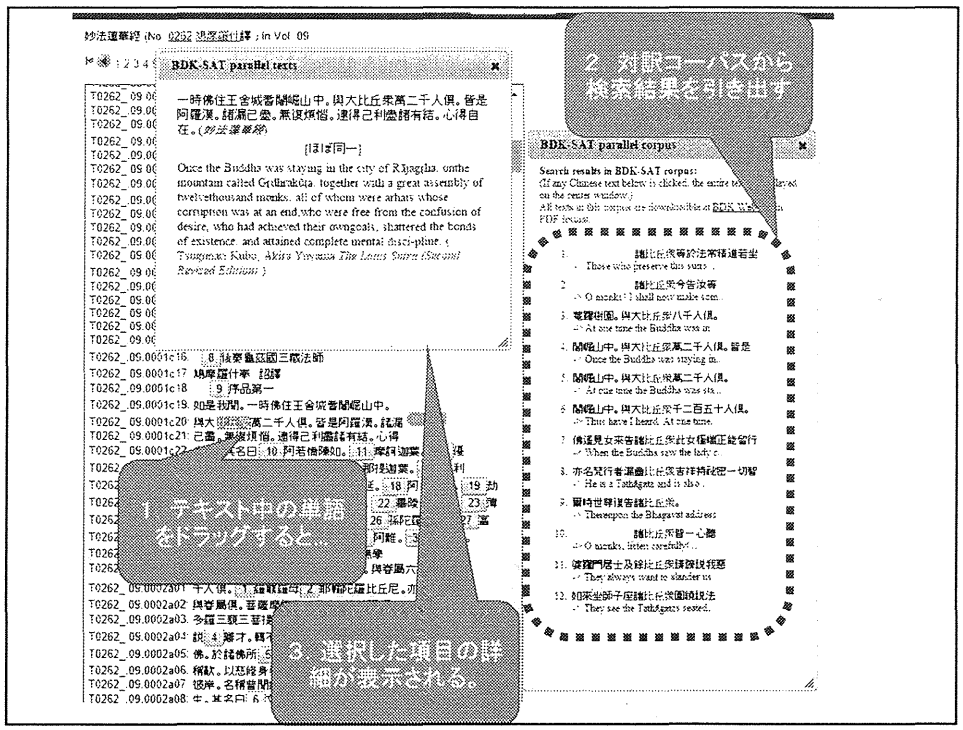
- オープンデータと呼べるものはきわめて少ない
  - 「東寺百合文書」は今のところ例外的
    - CC BYで公開
  - 青空文庫は素晴らしいオープンデータだが研究者向けとまでは言えない。
  - 日本語版Wikipedia は成果の発信に有効
- それでも可能性は様々に垣間見える
  - 「翻デジ」の可能性
  - 仏典のデータベースを参考に。
    - 一部、オープンデータを活用している。

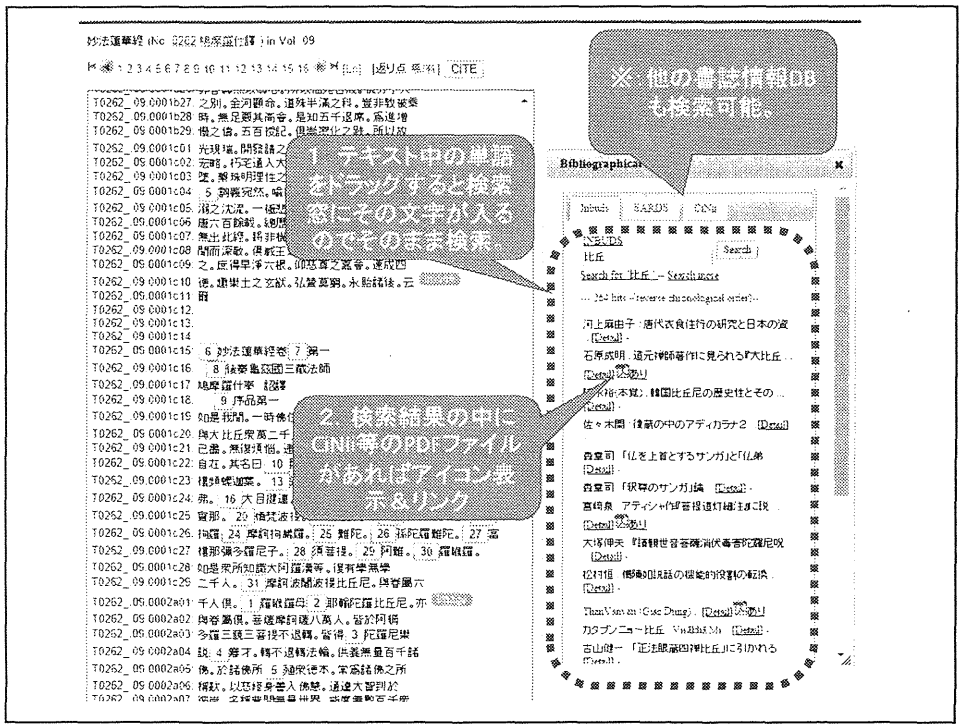
例：大蔵経テキストデータベース

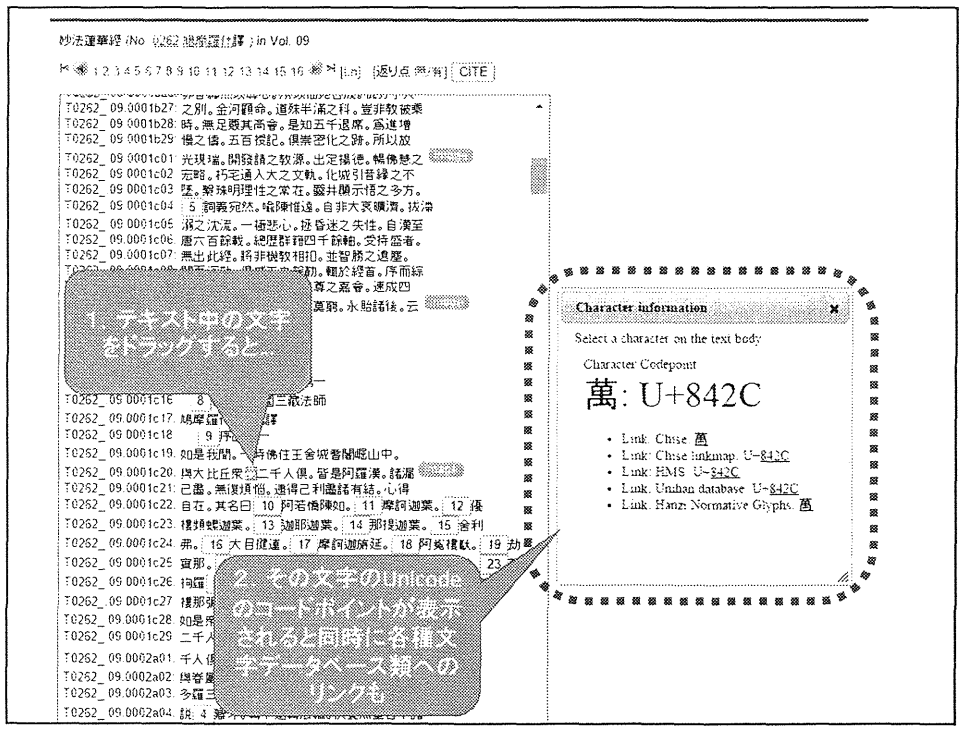
- 様々な情報資源と連携することで利便性を提供している。
  - 電子仏教辞典
  - 漢文・英訳対照コーパス（仏教伝道協会）
  - 漢字データベース関連
    - CHISE（京都大学）, HNG（北海道大学等）, Unihan, HMS...
  - 書誌データベース
    - INBUDS（日本印度学仏教学会）, SARDS3（ドイツ・ハレ大学）, CiNii（NII）
  - 目録データベース
    - BCRD (コロンビア大学)
  - 文書画像データベース
    - 近デジ、国デコ（NDL）、Gallica（フランス）、BL（イギリス）、国文学研究資料館DB、その他各大学画像DB

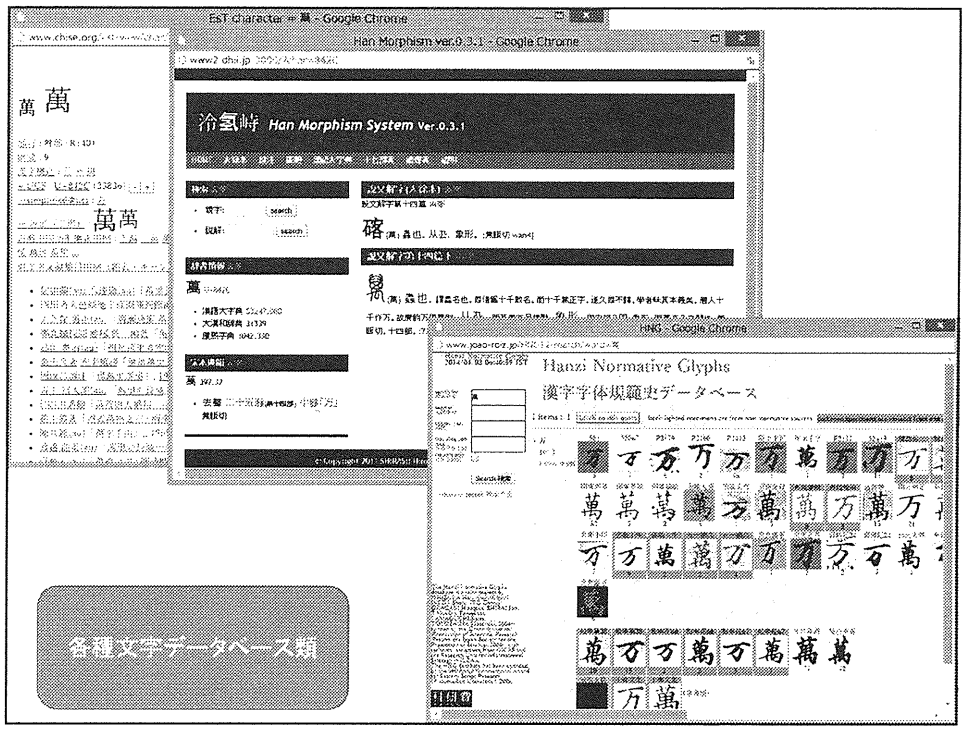



1. テキストをドラッグすると…
2. 最長一致で自動的に分割して個々の意味を辞書から検索


1. テキスト中の単語をドラッグすると…
2. 対訳コーパスから検索結果を引き出す
3. 選択した項目の詳細が表示される。

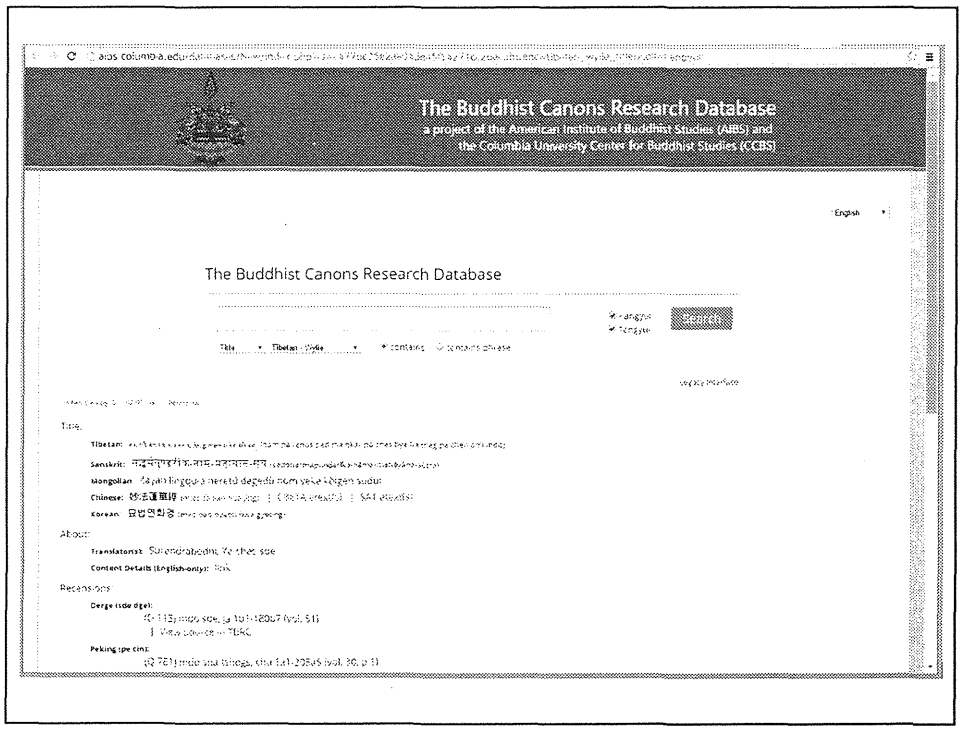

※他の書誌情報DBも検索可能。
1. テキスト中の単語をドラッグすると検索窓にその文字が入るのでそのまま検索。
2. 検索結果の中にCiNii等のPDFファイルがあればアイコン表示&リンク


1. テキスト中の文字をドラッグすると、
2. その文字のUnicodeのコードポイントが表示されると同時に各種文字データベース類へのリンクも
Character information
Select a character on the text body
Character Codepoint
萬: U+842C
Link: Chise 萬
Link: Chise linkmap. U+842C
Link: HMS U+842C
Link: Unihan database U+842C
Link: Hanzi Normative Glyphs. 萬

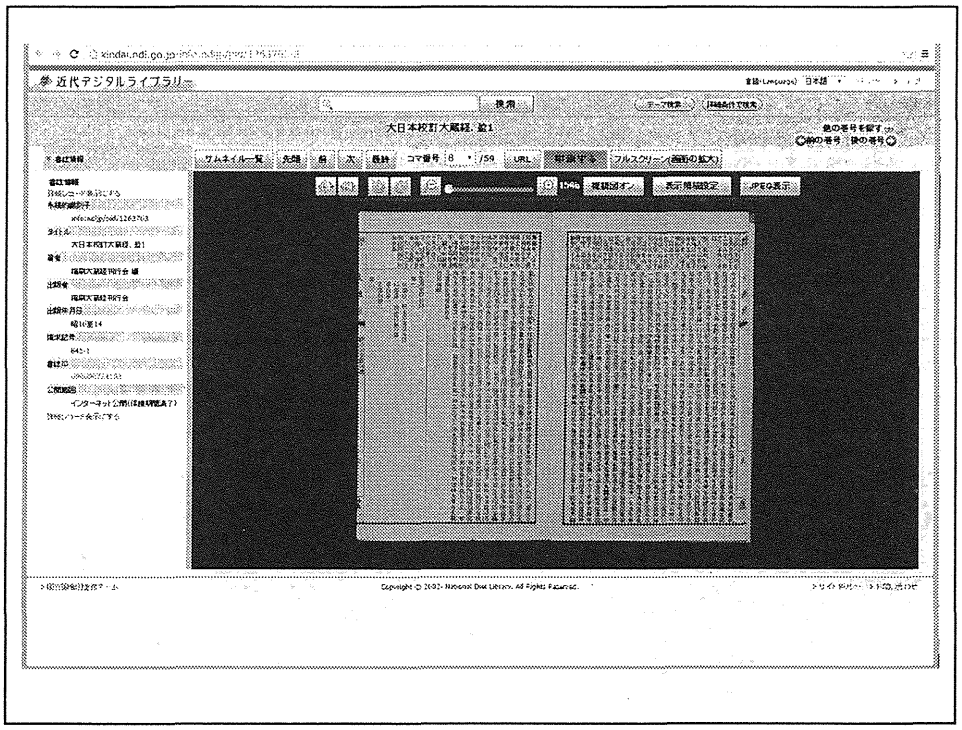

各種文字データベース類
Han Morphism System Ver.0.3.1
Hanzi Normative Glyphs
漢字字体規範史データベース

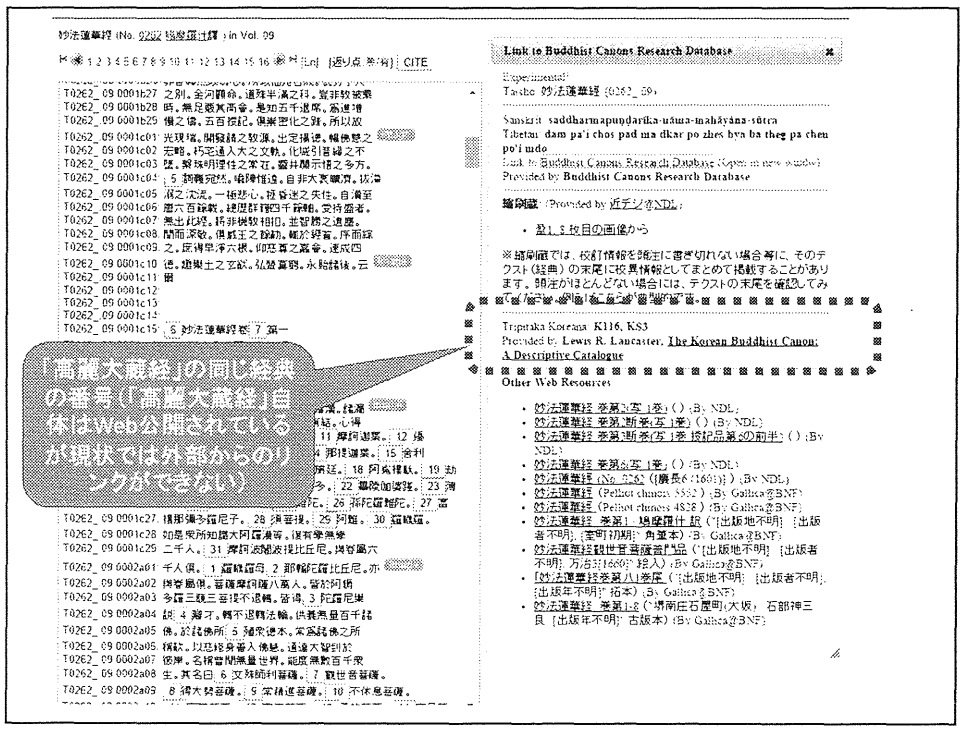

テキストを選択した時
点で一次資料やテ
ペット大蔵経DBへの
リンクがリストされる
Link to Buddhist Canons Research Database
Other Web Resources

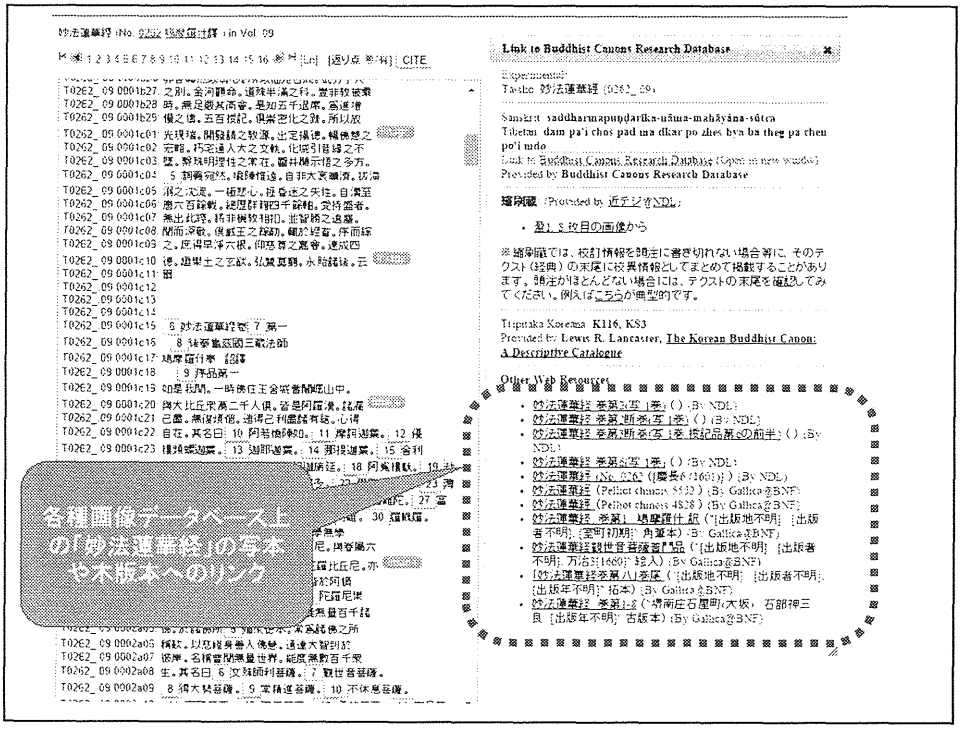

サンスクリット語、チベット語訳での『妙法蓮華経』のタイトルとチベット大蔵経DBへのリンク
Link to Buddhist Canons Research Database
Taisho 妙法蓮華経 (No. 0262, 09)
Sanskrit: saddharmapuṇḍarīka-nāma-mahāyāna-sūtra
Tibetan: dam pa'i chos pad ma dkar po zhes bya ba theg pa chen po'i mdo
Link to Buddhist Canons Research Database (Open in new window)
Provided by: Buddhist Canons Research Database
Triptaka Koreana: K116, KS3
Provided by: Lewis R. Lancaster, The Korean Buddhist Canon: A Descriptive Catalogue
Other Web Resources
妙法蓮華経 卷第六(写 1巻) ( ) (By NDL)
妙法蓮華経 卷第1-断卷(写 1巻) ( ) (By NDL)
妙法蓮華経 卷第1-断卷(写 1巻 授記品第六の前半) ( ) (By NDL)
妙法蓮華経 卷第六(写 1巻) ( ) (By NDL)
妙法蓮華経 (No. 0262) ([鎌倉]6 (1601)) ( ) (By NDL)
妙法蓮華経 (Pelliot chinois 5532) (By Gallica@BNF)
妙法蓮華経 (Pelliot chinois 4838) (By Gallica@BNF)
妙法蓮華経 卷第1, 18摩訶什 訳 (「出版地不明」 [出版者不明]; [室町初期]" 角筆本) (By Gallica@BNF)
妙法蓮華経観世音菩薩普門品 (「出版地不明」 [出版者不明]; 万治3[1660]" 経入) (By Gallica@BNF)
[妙法蓮華経卷第八]卷尾 (「出版地不明」 [出版者不明]; [出版年不明]" 祐本) (By Gallica@BNF)
妙法蓮華経 卷第1-8 (「堺南庄石屋町(大坂) 石部神三良 [出版年不明]" 古版本) (By Gallica@BNF)

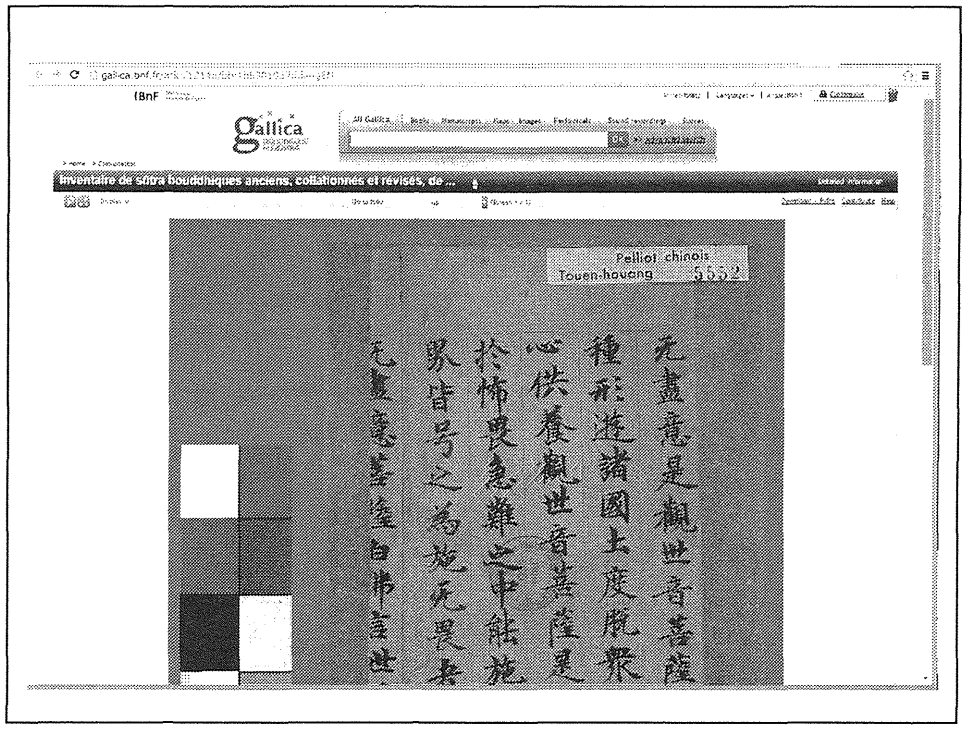

The Buddhist Canons Research Database
a project of the American Institute of Buddhist Studies (AIBS) and the Columbia University Center for Buddhist Studies (CCBS)
The Buddhist Canons Research Database

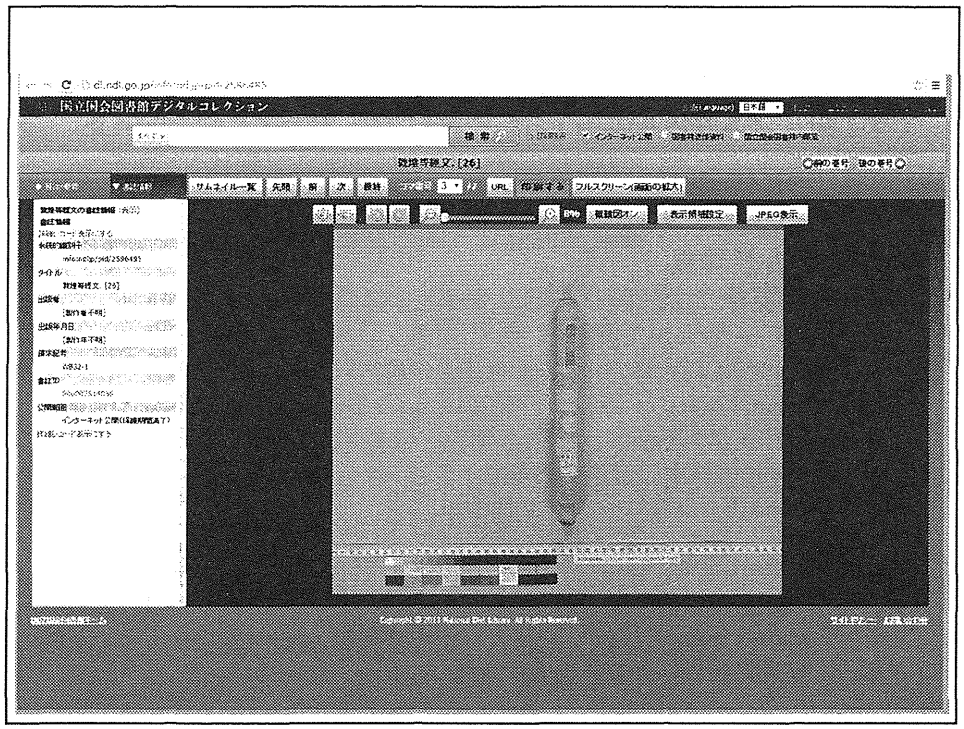

NDL近代デジタルライブラリー上の「縮刷大蔵経」の同じ経典へのリンク
Link to Buddhist Canons Research Database
Taisho 妙法蓮華経 (0262_09)
Sanskrit saddharmapuṇḍarika-nāma-mahāyāna-sūtra
Tibetan dam pa'i chos pad ma dkar po zhes bya ba theg pa chen po'i mdo
Link to Buddhist Canons Research Database (Open in new window)
Provided by Buddhist Canons Research Database
縮刷蔵
Provided by Lewis R. Lancaster, The Korean Buddhist Canon: A Descriptive Catalogue
Other Web Resources

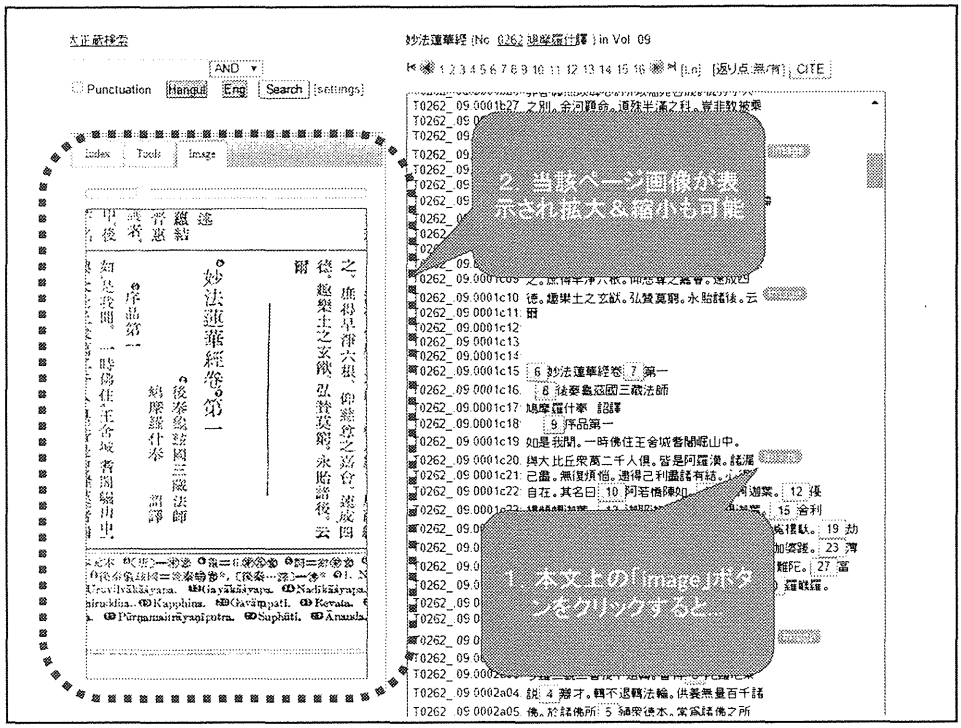

近代デジタルライブラリー
大日本校訂大蔵経. 第1


「高麗大蔵経」の同じ経典の番号（「高麗大蔵経」自体はWeb公開されているが現状では外部からのリンクができない）
Link to Buddhist Canons Research Database
Taishō 妙法蓮華経 (0262_09)
Sanskrit saddharmapuṇḍarīka-nāma-mahāyāna-sūtra
Tibetan dam pa'i chos pad ma dkar po zhes bya ba theg pa chen po'i mdo
Link to Buddhist Canons Research Database (Open in new window)
Provided by Buddhist Canons Research Database
Triptaka Koreana K116, KS3
Provided by: Lewis R. Lancaster, The Korean Buddhist Canon: A Descriptive Catalogue
Other Web Resources
妙法蓮華経 巻第六(写 1巻) ( ) (By: NDL)
妙法蓮華経 巻第六断巻(写 1巻) ( ) (By NDL)
妙法蓮華経 巻第六断巻(写 1巻 授記品第六の前半) ( ) (By NDL)
妙法蓮華経 巻第六(写 1巻) ( ) (By NDL)
妙法蓮華経 (No. 0262) ([慶長6 (1601)]) ( ) (By NDL)
妙法蓮華経 (Pelliot chinois 5532) (By Gallica@BNF)
妙法蓮華経 (Pelliot chinois 4828) (By Gallica@BNF)
妙法蓮華経 巻第1 鳩摩羅什 訳 (写[出版地不明]: [出版者不明]: [室町初期]: 角筆本) (By: Gallica@BNF)
妙法蓮華経観世音菩薩普門品 (写[出版地不明]: [出版者不明]: 万治3[1660]: 捻入) (By Gallica@BNF)
[妙法蓮華経巻第八]巻尾 (写[出版地不明]: [出版者不明]: [出版年不明]: 拓本) (By Gallica@BNF)
妙法蓮華経 巻第1-8 (写増南庄右屋町(大坂), 石部神三良 [出版年不明]: 古版本) (By Gallica@BNF)


各種画像データベース上の「妙法蓮華経」の写本や木版本へのリンク
Link to Buddhist Canons Research Database
Taishō 妙法蓮華経 (0262_09)
Sanskrit saddharmapuṇḍarīka-nāma-mahāyāna-sūtra
Tibetan dam pa'i chos pad ma dkar po zhes bya ba theg pa chen po'i mdo
Link to Buddhist Canons Research Database (Open in new window)
Provided by Buddhist Canons Research Database
Triptaka Koreana K116, KS3
Provided by: Lewis R. Lancaster, The Korean Buddhist Canon: A Descriptive Catalogue
Other Web Resources
妙法蓮華経 巻第六(写 1巻) ( ) (By: NDL)
妙法蓮華経 巻第六断巻(写 1巻) ( ) (By NDL)
妙法蓮華経 巻第六断巻(写 1巻 授記品第六の前半) ( ) (By NDL)
妙法蓮華経 巻第六(写 1巻) ( ) (By NDL)
妙法蓮華経 (No. 0262) ([慶長6 (1601)]) ( ) (By NDL)
妙法蓮華経 (Pelliot chinois 5532) (By Gallica@BNF)
妙法蓮華経 (Pelliot chinois 4828) (By Gallica@BNF)
妙法蓮華経 巻第1 鳩摩羅什 訳 (写[出版地不明]: [出版者不明]: [室町初期]: 角筆本) (By: Gallica@BNF)
妙法蓮華経観世音菩薩普門品 (写[出版地不明]: [出版者不明]: 万治3[1660]: 捻入) (By Gallica@BNF)
[妙法蓮華経巻第八]巻尾 (写[出版地不明]: [出版者不明]: [出版年不明]: 拓本) (By Gallica@BNF)
妙法蓮華経 巻第1-8 (写増南庄右屋町(大坂), 石部神三良 [出版年不明]: 古版本) (By Gallica@BNF)


BnF
gallica
Inventaire de sûtra bouddhiques anciens, collationnés et revisés, de ...
Pelliot chinois
Touen-houang
9502


国立国会図書館デジタルコレクション

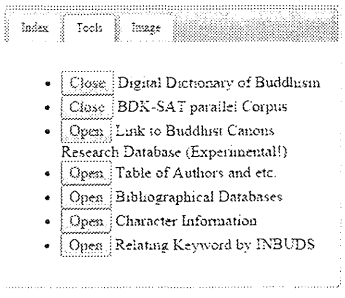


再掲：大蔵経テキストデータベース

- 様々な情報資源がある。

「オープンデータ」と呼べそうなものは…？

  - 電子仏教辞典
  - 漢文・英訳対照コーパス（仏教伝道協会）
  - 漢字データベース関連
    - CHISE（京都大学），HNG（北海道大学等），Unihan, HMS…
  - 書誌データベース
    - INBUDS（日本印度学仏教学会），SARDS3（ドイツ・ハレ大学），CiNii（NII）
  - 目録データベース
    - BCRD (コロンビア大学)
  - 文書画像データベース
    - 近デジ、国デコ（NDL）、Gallica（フランス）、BL（イギリス）、国文学研究資料館DB、その他各大学画像DB



「オープンアクセス」では？

- PDFによる論文公開が分野としては徐々に広がってきている
  - 仏教学では:
    - 日本印度学仏教学会
      - Journal@rticle + NII-ELS ⇒ JSTAGE or JSTAGE Lite
    - それ以外にも:
      - 主にNII-ELS経由
  - 日本語学会:
    - CiNiiにてOA
  - ⇒ NII-ELSの多大な貢献

  - Open Library of Humanitiesの取り組み
  - OpenEdition.orgの取り組み

データ作成者への評価

- 「校訂テクスト」「目録」「事典」等とのアナロジーはどうか
  - 同分野の研究者コミュニティからの評価は、冬の時代を迎えつつある現在、どの程度有効か。
  - 他の研究分野からも評価される枠組みが必要
  - 人件費/謝金はどのようにカバーし得るか
  - プロの編集は必要か・必要ならどうするか
- データジャーナル的なものが評価指標たり得るか
  - データの質についてはどう評価するか
  - プログラムの動作確認等と異なり、人文学向けデータの内容の確認の自動化はかなり困難。

終わりに

- 人文学向け資料では「オープンデータ」はまだまだ少ない。
  - 特に日本では弱いような印象があるが引き続き調査したい。
    - エコシステムとして機能するに至っていない？
  - オープンデータでなくとも研究用途ではそれなりに有益。
    - しかし、商用利用を禁ずることで広がりを阻害している面もあるかもしれない。
- 我が国での人文学向け資料における「オープンデータ」を活用したソリューションの必要性