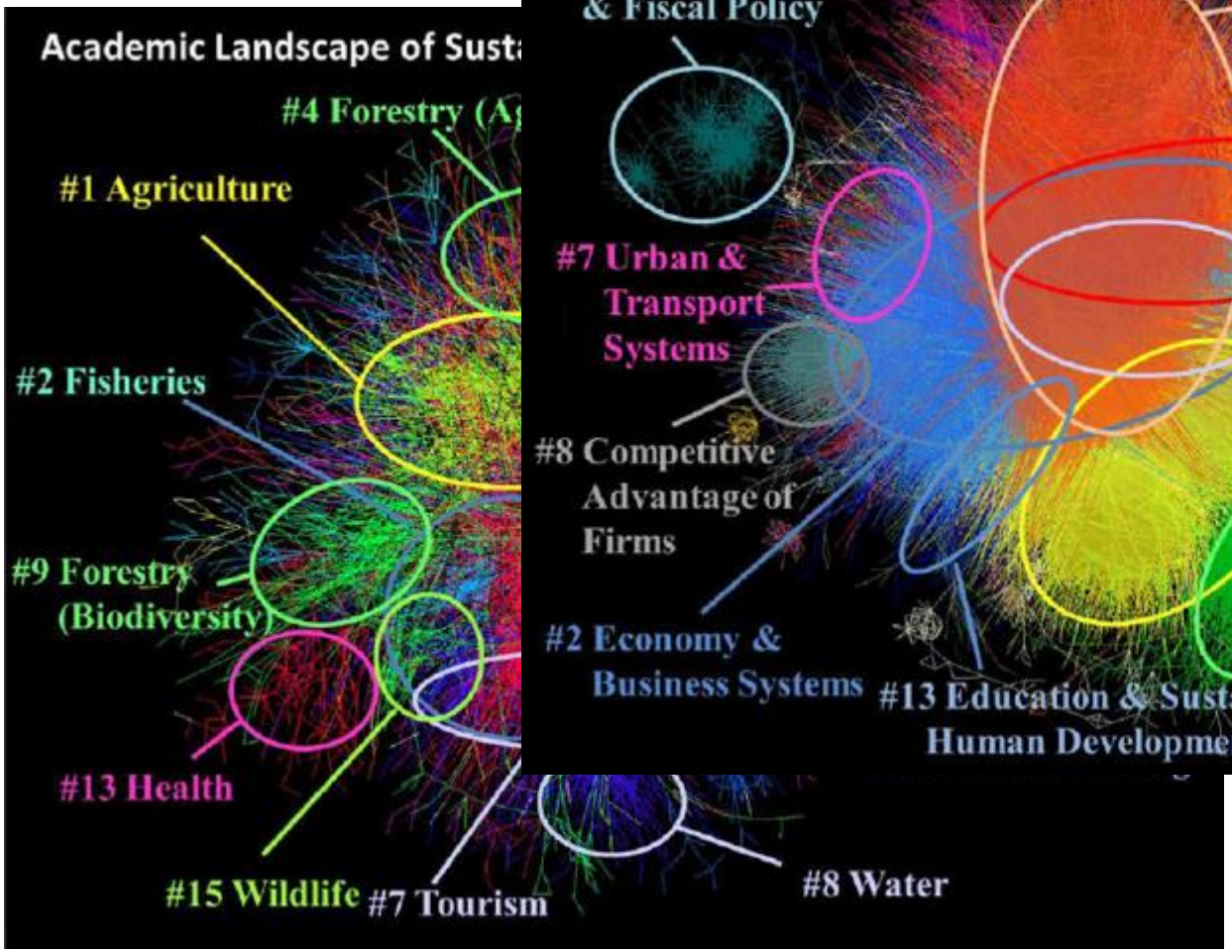
学術俯瞰システムの使い方
How to Use Academic Landscape System

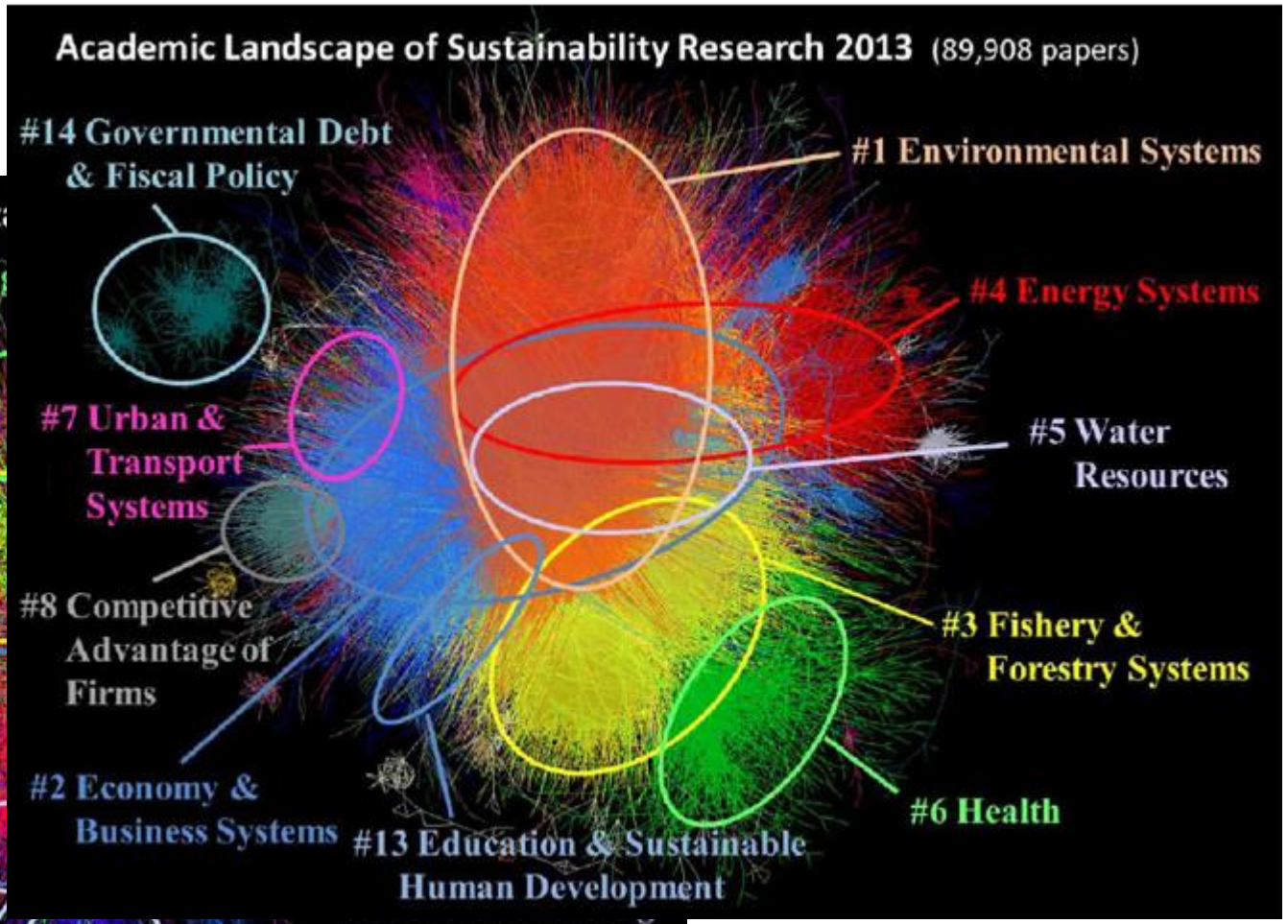
東京大学
イノベーション政策研究センター
IPRC, the University of Tokyo

2015-01-21

俯瞰図

Sustainability 2007 and 2013





Kajikawa et al., 2014. Sustainability science: the changing landscape of sustainability research.

Agenda

- 学術俯瞰システムの概要

  Overview of Academic Landscape System
- 論文のデータの入手

  How to Get your Dataset of Papers
- 学術俯瞰システムの操作方法

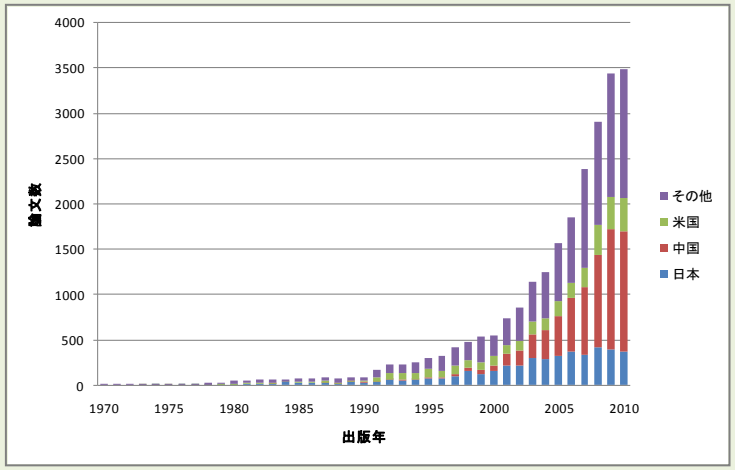
  How to Operate the System

「学術俯瞰システム」とは

学術俯瞰システムは、大規模な書誌情報をテキストマイニングとネットワーク分析により自動分析し、**学術分野の俯瞰**を可能にするシステムです。

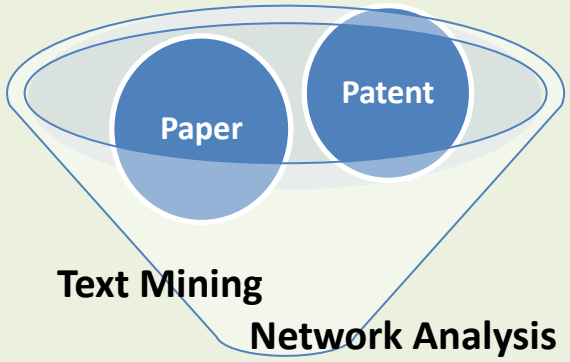
The Academic Landscape System enables us to **overview science and technology** with text-mining and network analysis of huge bibliographic information.

Background：知識の爆発の現状（光触媒の例）



80年代は日本が3-4割。現在は日本が1割、中国が4割。

専門家も全てを把握しきれないほど、
情報が増えている。



Overviewing Academic Fields

On the System

- Overview Map
- Graph Tool
- Heat Map

Out of the System

You can download results and add your original analysis.

学術俯瞰システムのしくみ

The Concept of Citation Network

Paper A

PHYSICAL REVIEW E 69, 066133 (2004)

Fast algorithm for detecting community structure in networks

M. E. J. Newman

Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109-1120, USA

(Received 22 September 2003; revised manuscript received 22 March 2004; published 18 June 2004)

Many networks display community structure—groups of vertices within which connections are dense but between which they are sparser—and sensitive computer algorithms have in recent years been developed for detecting this structure. These algorithms, however, are computationally demanding, which limits their application to small networks. Here we describe an algorithm which gives excellent results when tested on computer-generated and real-world networks and is much faster, typically thousands of times faster, than previous algorithms. We give several example applications, including one to a network of collaborations between more than 50 000 physicists.

DOI: 10.1103/PhysRevE.69.066133 PACS number(s): 89.75.Hc,

I. INTRODUCTION

There has in recent years been a surge of interest within the physics community in the properties of networks of many kinds, including the Internet, the world wide web, citation networks, transportation networks, software call graphs, email networks, food webs, and social and biochemical networks [1–4]. One property that has attracted particular attention is that of "community structure": the vertices in networks are often found to cluster into tightly knit groups with a high density of within-group edges and a lower density of between-group edges. Girvan and Newman [5,6] proposed a computer algorithm based on the iterative removal of edges with high "betweenness" scores that appears to identify such structure with some sensitivity, and this algorithm has been employed by a number of authors in the study of such diverse systems as networks of email messages, social networks of animals, collaborations of jazz musicians, metabolic networks, and gene networks [5–11]. As pointed out by Newman and Girvan [6], the principal disadvantage of their algorithm is the high computational demands it makes. In its simplest and fastest form, it runs in worst-case time $O(m^2n)$ on a network with n edges and n vertices, or $O(n^3)$ on a sparse network (one for which n scales with n in the limit of large n, which covers essentially all networks of current scientific interest, with the possible exception of food webs). With typical computer resources available at the time of writing, this limits the algorithm's use to networks of a few thousand vertices at most, and substantially less than this for interactive applications. Increasingly, however, there is interest in the study of much larger networks; citation and collaboration networks can contain millions of vertices [12,13], for example, while the world wide web numbers in the billions [14].

In this paper, therefore, we propose another algorithm for detecting community structure. The algorithm operates on different principles from that of Girvan and Newman (GN), but, as we will show, gives qualitatively similar results. The worst-case running time of the algorithm is $O[(m+n)n]$, or $O(n^2)$ on a sparse graph. In practice, it runs to completion on current computers in reasonable times for networks of up to a million or so vertices, bringing within reach the study of

communities in ma
been considered intr

Our algorithm is
any network, the Gi
produces *some* divisi
gardless of whether
sion. To test whethe
define a quality fun
Let e_{ij} be one-hal
that connect vertice
the total fraction of
will be the diagona
fraction of edges tha
half). Then $\Sigma_i e_{ij}$ is
groups. All other ed
values of this sum i
communities is good
order 1. On its own,
of community struct
if we put all vertice
trivial and not partic
ture.

A more useful m
culate the sum $\Sigma_i e_{ij}$
would take if edges
gives a score of ze
single community,
ings.

Let a_i be the fract
to vertices in group
by noting that $a_i 2
together at random,
connect vertices wi
larity to be

If a particular divi
edges than would be

1539-3755/2004/69(6)/066133(5)/$22.50 69 066133-1

FAST ALGORITHM FOR DETECTING COMMUNITY... PHYSICAL REVIEW E 69, 066133 (2004)

FIG. 5. Cumulative distribution function of the sizes of communities found in one of the subnetworks of the physics collaboration graph, as described in the text. The dotted line represents the slope the plot would have if the distribution followed a power law with exponent −1.6.

−1.6, although this conclusion should be treated with caution as there is significant deviation from a perfect power law [20].

Narrowing our focus still further to the particular one of these communities that contains the present author, we find the structure shown in the right panel of Fig. 4. Feeding this one last time through the algorithm, it breaks apart into communities that correspond closely to individual institutional research groups, the author's group appearing in the corner of the figure, highlighted by the dashed box. One could pursue this line of analysis further, identifying individual groups, iteratively breaking them down, and looking, for example, at the patterns of collaboration between them, but we leave this for later studies.

IV. CONCLUSIONS

In this paper we have described an algorithm for extracting community structure from networks, which has a considerable speed advantage over previous algorithms, running to completion in a time that scales as the square of the network size. This allows us to study much larger systems than has previously been possible. Among other examples, we have applied the algorithm to a network of collaborations between more than 50 000 physicists, and found that the resulting community structure corresponds closely to the traditional divisions between specialties and research groups in the field.

We believe that our method will not only allow for the extension of community structure analysis to some of the very large networks that are now being studied for the first time, but will also provide a useful tool for visualizing and understanding the structure of these networks, whose daunting size has hitherto made many of their structural properties obscure.

ACKNOWLEDGMENTS

The author thanks Leon Danon, Pablo Gleiser, David Lusseau, and Douglas White for providing network data used in the examples. This work was supported in part by the National Science Foundation under Grant No. DMS-0234188.

[1] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[4] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[5] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[6] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[7] D. Wilkinson and B. A. Huberman, Proc. Natl. Acad. Sci. U.S.A. **101**, 5241 (2004).
[8] P. Holme, M. Huss, and H. Jeong, Bioinformatics **19**, 532 (2003).
[9] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giral, and A. Arenas, Phys. Rev. E **68**, 065103 (2003).
[10] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, in *Proceedings of the First International Conference on Communities and Technologies*, edited by M. Huysman, E. Wenger, and V. Wulf (Kluwer, Dordrecht, 2003).
[11] P. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).
[12] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[13] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[14] J. Kleinberg and S. Lawrence, Science **294**, 1849 (2001).
[15] B. Everitt, *Cluster Analysis* (John Wiley, New York, 1974).
[16] J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. (Sage, London, 2000).
[17] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[18] D. Lusseau, Proc. R. Soc. London, Ser. B **270**, S186 (2003).
[19] The criterion for deciding correct classification is as follows. We find the largest set of vertices that are grouped together by the algorithm in each of the four known communities. If the algorithm puts two or more of these sets in the same group, then all vertices in those sets are considered incorrectly classified. Otherwise, they are considered correctly classified. All other vertices not in the largest sets are considered incorrectly classified. This criterion is quite harsh—there are cases in which one might consider some of the vertices to have been identified correctly, where this method would not. Even with this harsh definition, however, our algorithm performs well, and a laxer definition would only make its performance more impressive.
[20] This power law is different from the one observed in an email network by Guimerà *et al.* [9]. They studied the histogram of community sizes over all levels of the dendrogram; we are looking only at the single level corresponding to the maximum value of Q.

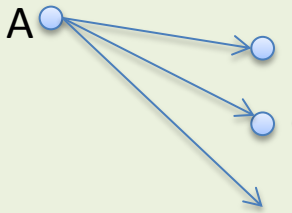
066133-5

Paper B

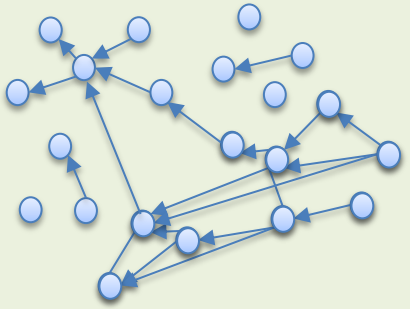
Paper C

…

Namely, relations between papers are shown as a networks.

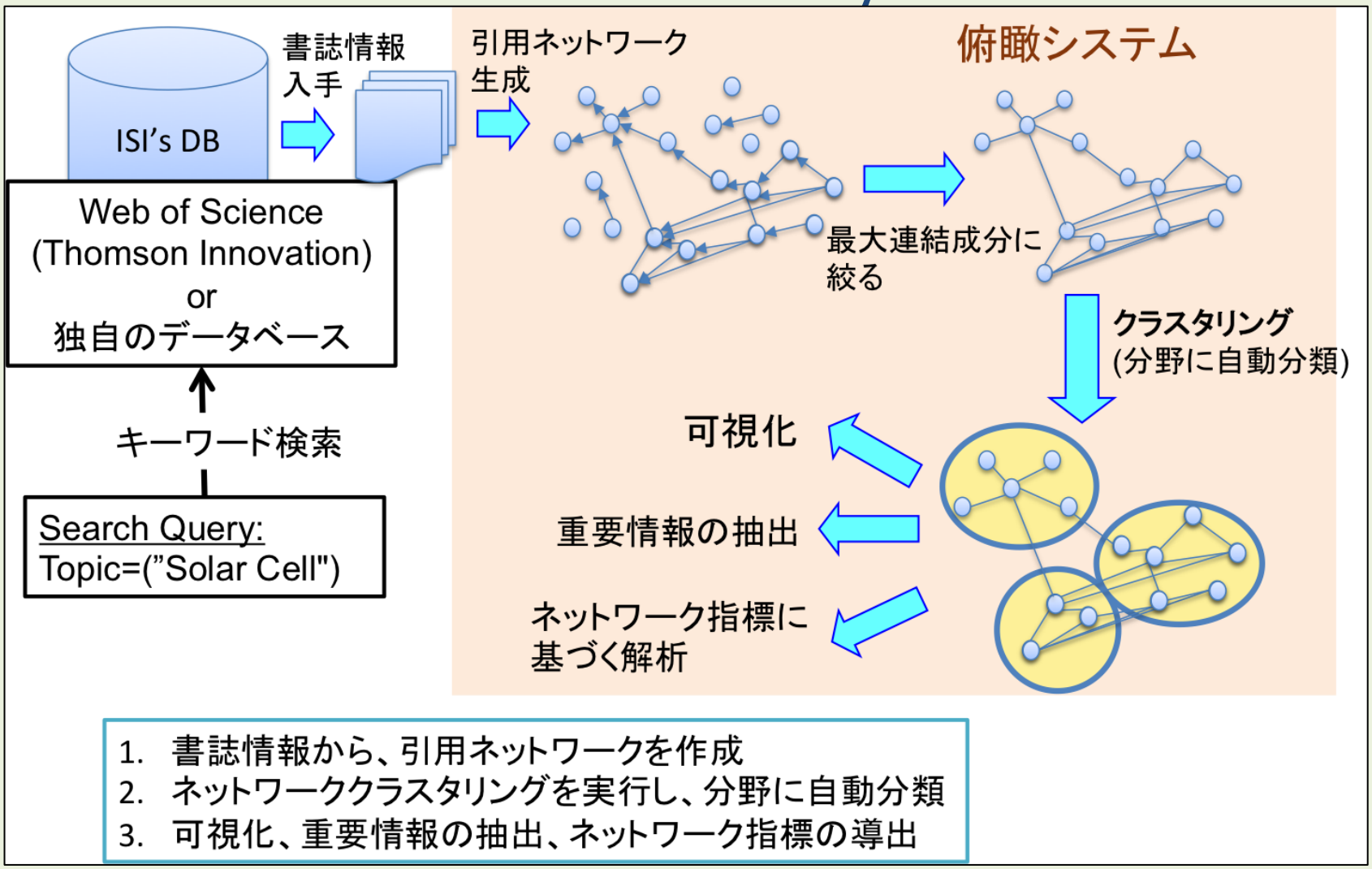


↓More papers…



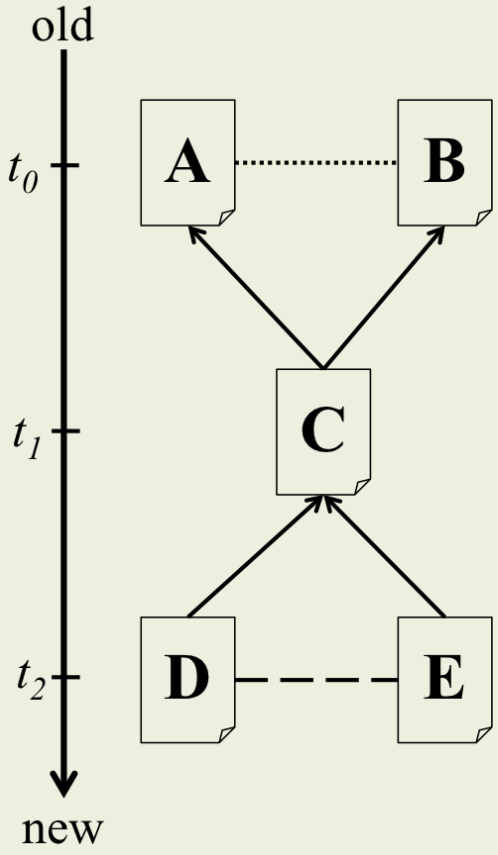
学術俯瞰システムのしくみ

The Flow of Analysis



1. 書誌情報から、引用ネットワークを作成
2. ネットワーククラスタリングを実行し、分野に自動分類
3. 可視化、重要情報の抽出、ネットワーク指標の導出

Type of Citations



Direct citation

- Edge between two documents when *paper C* cites *paper A* directly
- An earlier paper is cited by a new paper

Co-citation

- Edge between two documents cited by the same paper(s)
- Considering the number of citations

Bibliographic coupling

- Edge between two documents citing the same paper(s)
- Considering the number of citations

direct-citation
co-citation
bibliographic coupling

	Direct vs Co-citation vs Bibliographic
Nodes	Direct > Bibliographic > Co-citation
Edges	Bibliographic > Co-citation > Direct
Year	Direct = Bibliographic > Co-citation

Agenda

- 学術俯瞰システムの概要

  Overview of Academic Landscape System
- 論文のデータの入手

  How to Get your Dataset of Papers
- 学術俯瞰システムの操作方法

  How to Operate the System

最新データのアップロード

Update and Upload the Latest Data

- 論文の場合（学内LAN）

Web of Scienceから最新データを入手する
- 東京大学　GACoS, or
- “Web of Science” で検索

- 特許の場合
  - Thomson Innovationのデータに対応しています
  - 入手困難な場合は、イノベーション政策研究センターにご相談ください

最新データのアップロード

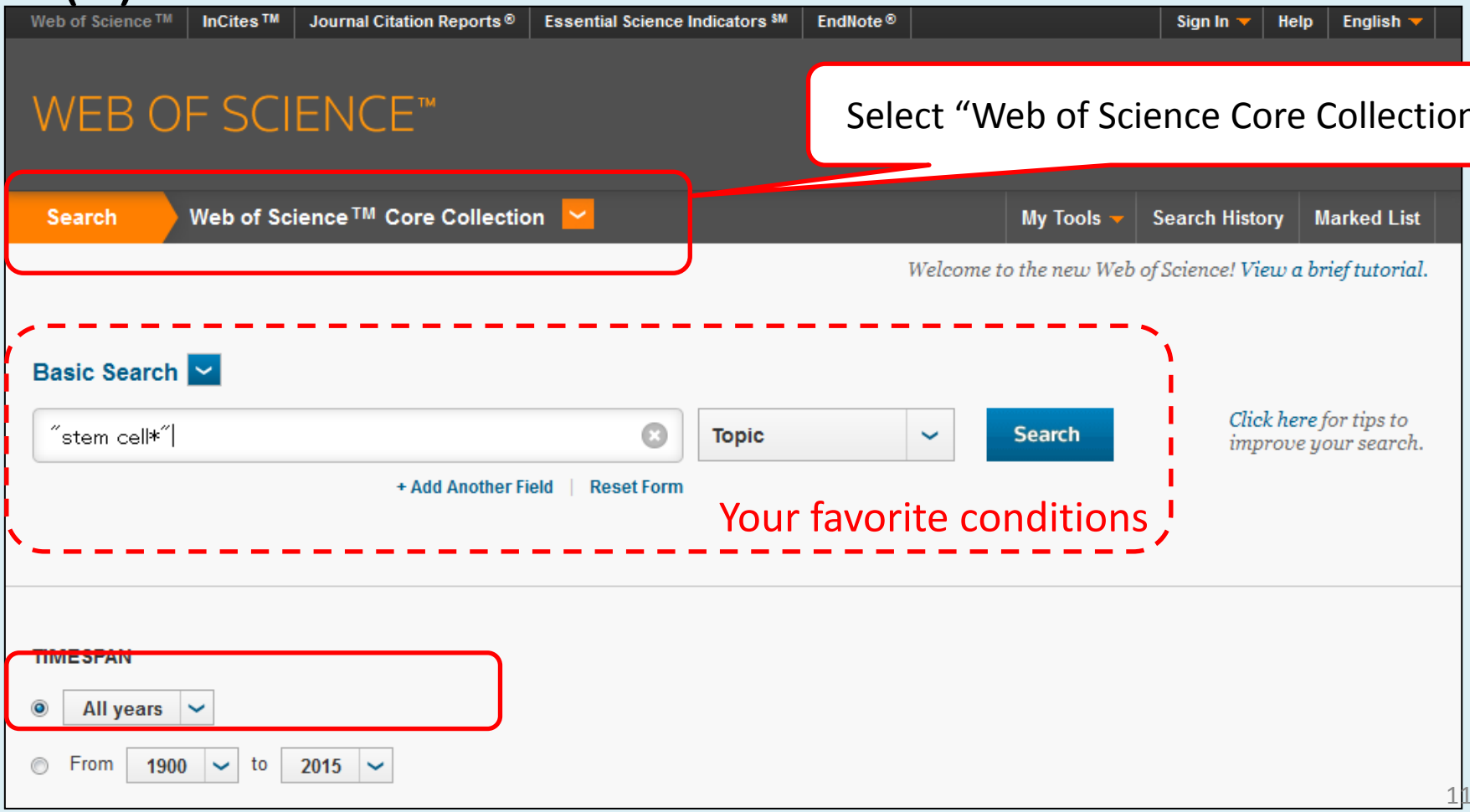
Update and Upload the Latest Data

(1) Search “Web of Science” within UT LAN, and

最新データのアップロード

Update and Upload the Latest Data

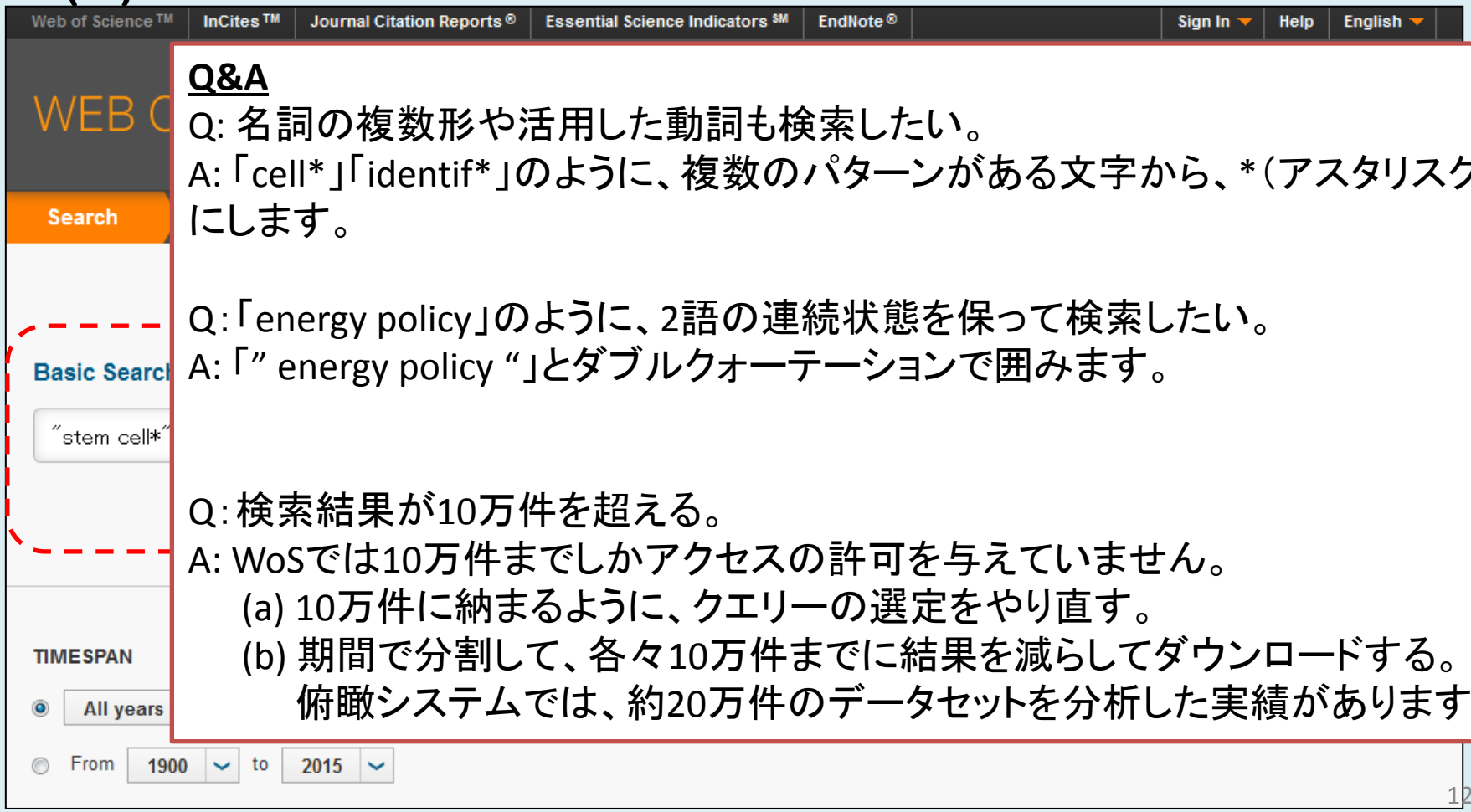
(2) Search on Web of Science



最新データのアップロード

Update and Upload the Latest Data

(2) Search on Web of Science



Q&A

Q: 名詞の複数形や活用した動詞も検索したい。
A:「cell*」「identif*」のように、複数のパターンがある文字から、*（アスタリスク）にします。

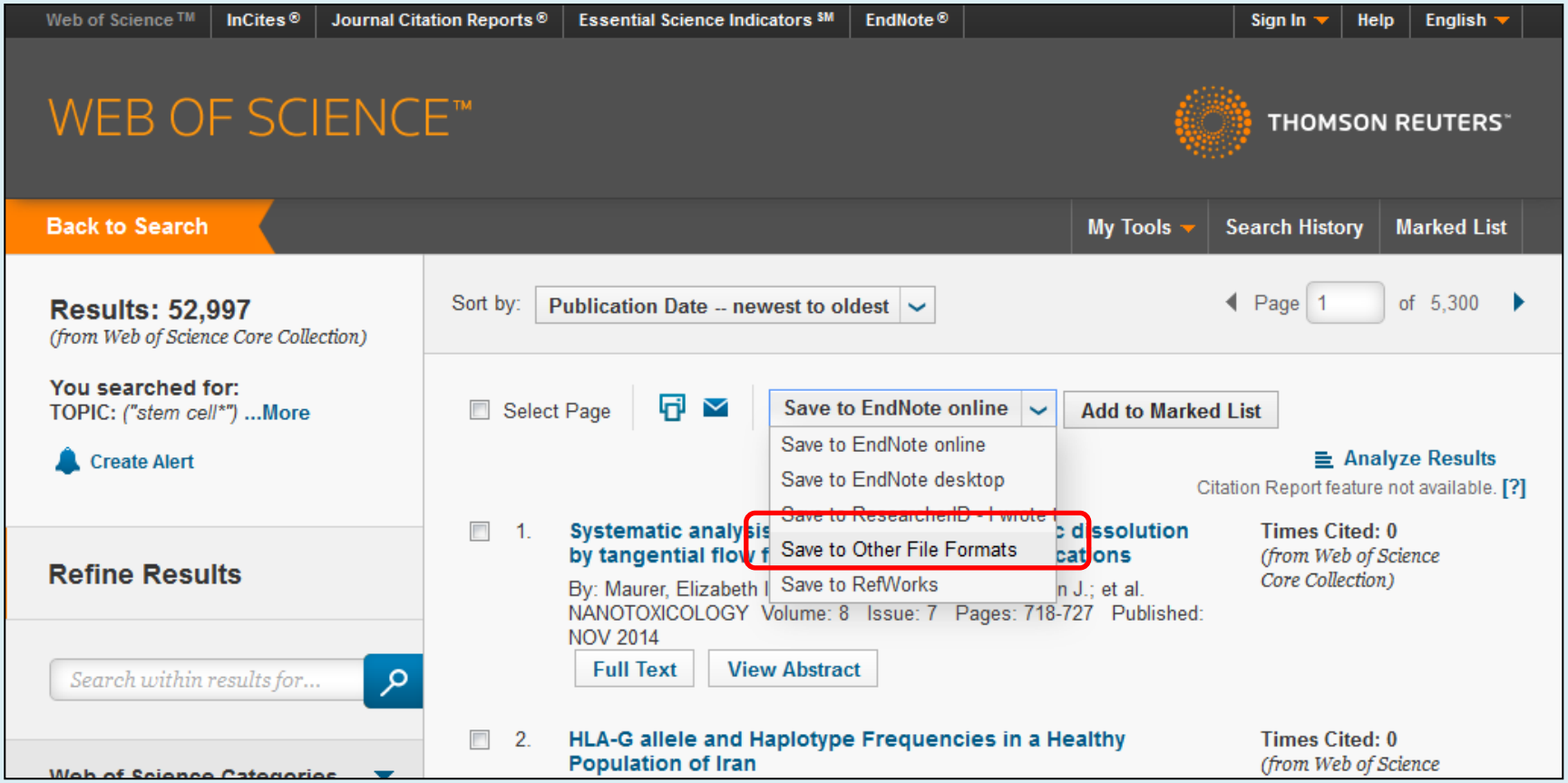
Q:「energy policy」のように、2語の連続状態を保って検索したい。
A:「” energy policy “」とダブルクォーテーションで囲みます。

Q:検索結果が10万件を超える。
A: WoSでは10万件までしかアクセスの許可を与えていません。
(a) 10万件に納まるように、クエリーの選定をやり直す。
(b) 期間で分割して、各々10万件までに結果を減らしてダウンロードする。
俯瞰システムでは、約20万件のデータセットを分析した実績があります。

最新データのアップロード

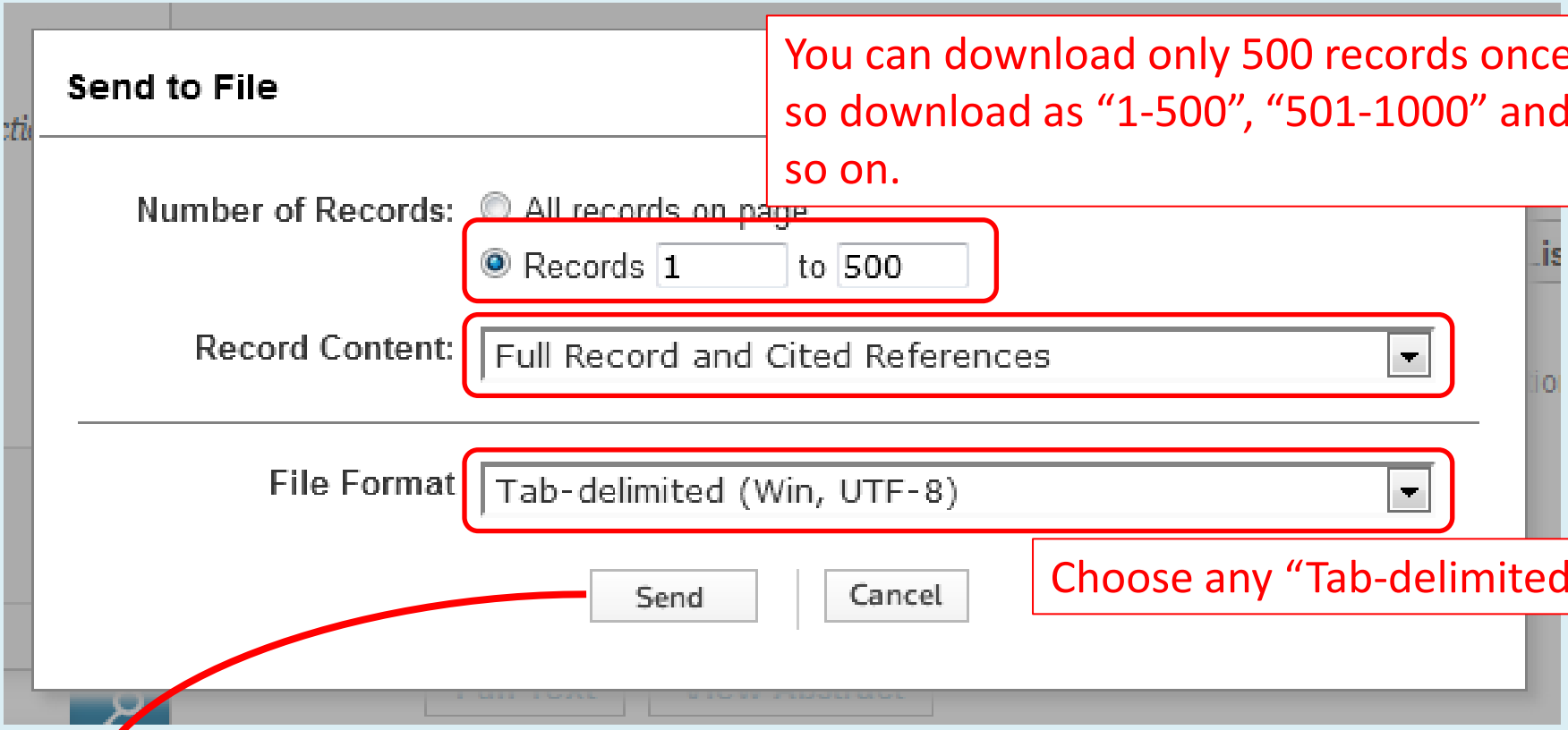
Update and Upload the Latest Data

(3) Select “Save to Other File Formats”



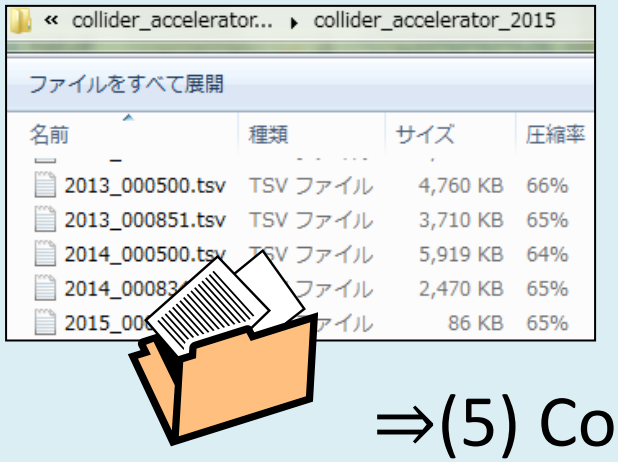
最新データのアップロード

Update and Upload the Latest Data



(4) Make a directory on your PC, and
put all downloaded files into the directory.
File’s extension: txt or tsv

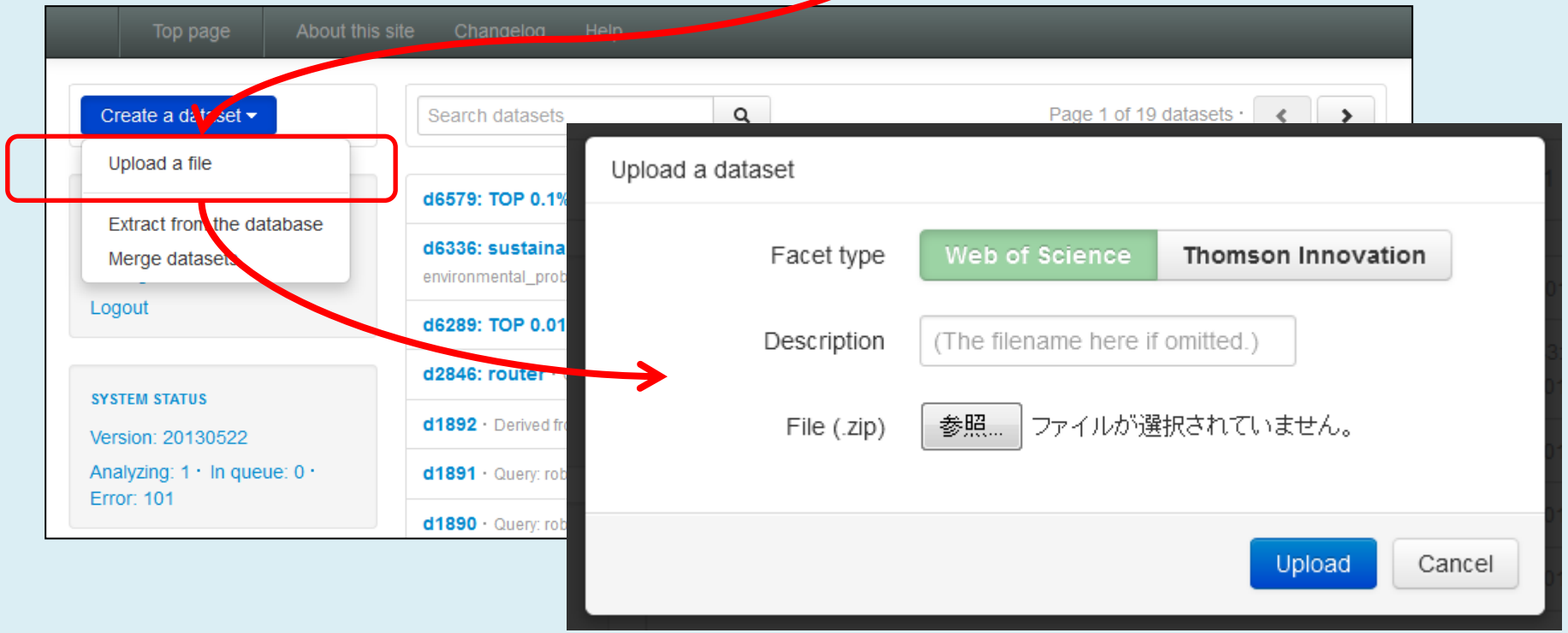
Examples of file names: 000500.tsv, 001000.tsv, …



最新データのアップロード

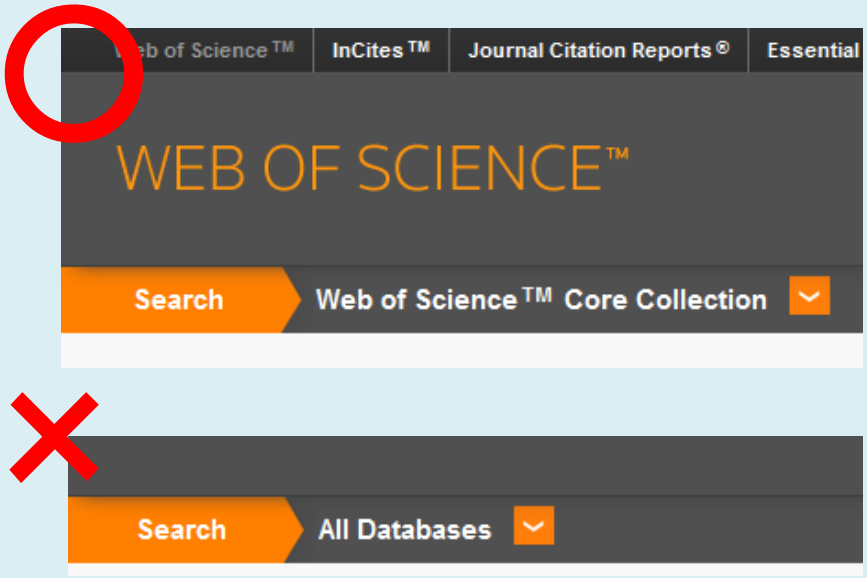
Update and Upload the Latest Data

⇒(5) Compress into a zip file, and upload it.



俯瞰システムでの分析に失敗するとき（データセットに起因するFAQ）

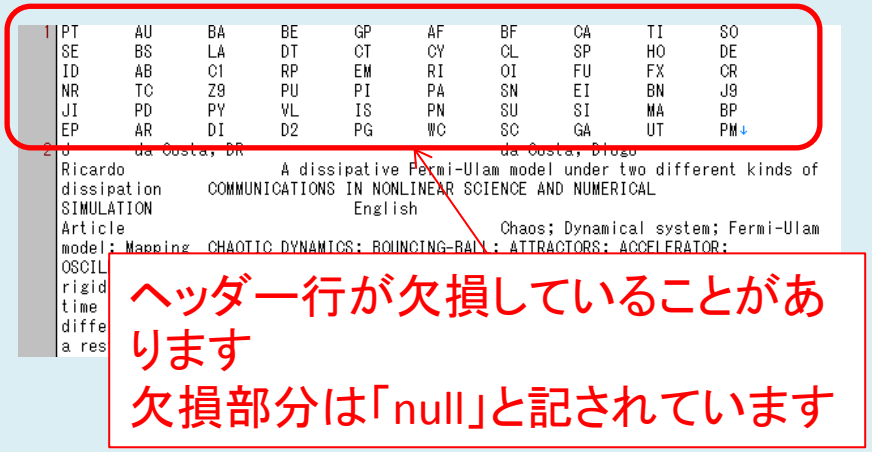
Q.論文数は大量にあるのに、引用がないと指摘される



A.“Web of Science Core Collection” を選んでいますか？

- 「横断検索」（All Databases）は、引用関係データが含まれません

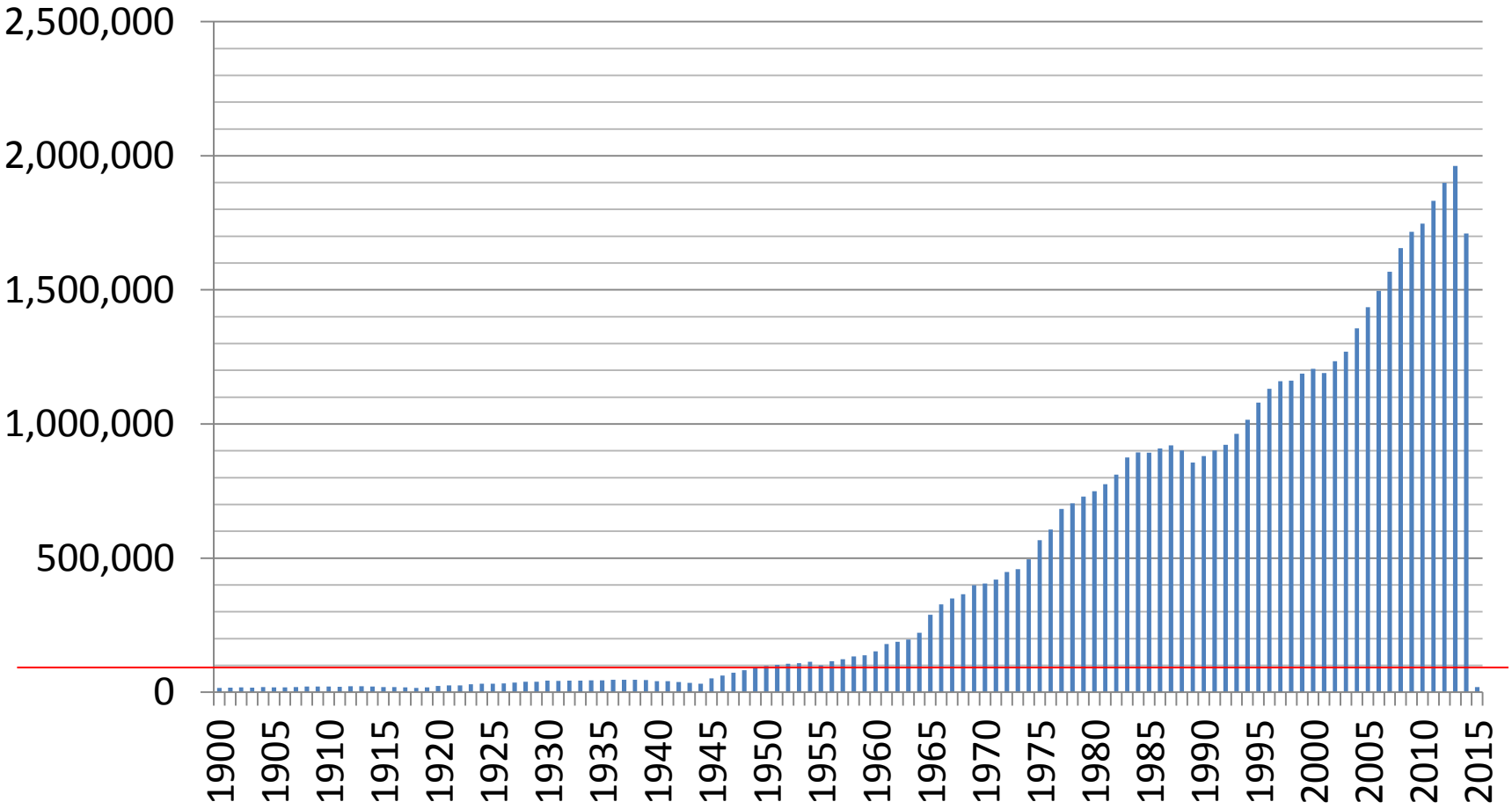
Q.「データセットに記法がおかしいファイルが含まれています」とエラーになる



A.壊れていないファイルからコピー＆ペーストして修復してください

- WoS内部で既に壊れていますので、再ダウンロードに効果はありません

Web of Scienceの論文数（2015-01-02）



注意事項
Precautions

- アクセス権限　Access Authority
  - 論文本文は権限のない人に渡さないようにしてください。
    Don’t give papers to unauthorized people.
  - アクセス権限のないデータを利用する場合 if you are not authorized:
    - アクセス権限のある方と共同研究を行ってください。
      your research must be a collaborative research with people who is authorized.

	坂田・森研所属者 IPRC所属者	東京大学所属者	その他
データベースから抽出	○	×	×
Web of Scienceから取得してアップロード	○	○	×
Thomson Innovation （特許データベース）	○	×	×

* 論文等で公表される場合、または、企業で使用される場合は、
権限者（坂田先生・森先生）に共同研究を相談してください。
If you publish your paper using the system, your paper must be a collaborative research .
授業や研究の文献調査など、権限内の非公開での利用は問合せ不要です。
You have already been permitted in private use.

アカウントの作成
How to make your account

- アカウントを作成します。Staffs make your account.
  - 必要な情報 Please e-mail your data to us.
    - アカウント名 account
      - 半角アルファベット、半角数字、@/./+/-/_ で30文字以下。
        You can use 30 or less English characters, which are alphabets, number, @/./+/-/_.
      - アカウント名に関しては、他の利用者から見える場合があります。
        Other users can see your account. Don’t use your secret data.
    - パスワード Password
    - 連絡先メールアドレス E-mail address

- URL of the System
  - http://academic-landscape.com/