Legal Implications of Text and Data Mining (TDM)

1982

“Automatically generating logical representations of text passages… by means of an analysis of the coherence structure of the passages.”

Jerry R. Hobbs, Donald E. Walker, and Robert A. Amsler. 1982. Natural language access to structured text. In *Proceedings of the 9th conference on Computational linguistics - Volume 1*(COLING '82), Ján Horecký (Ed.), Vol. 1. Academia Praha, , Czechoslovakia, 127-132. DOI=10.3115/991813.991833 http://dx.doi.org/10.3115/991813.991833

1999

“(semi)automated discovery of trends and patterns across very large datasets”

“Use of large online text collections to discover new facts and trends…”

“(Automating) the tedious parts of the text manipulation process and (integrating) underlying computationally-driven text analysis with human-guided decision making within exploratory data analysis over text”

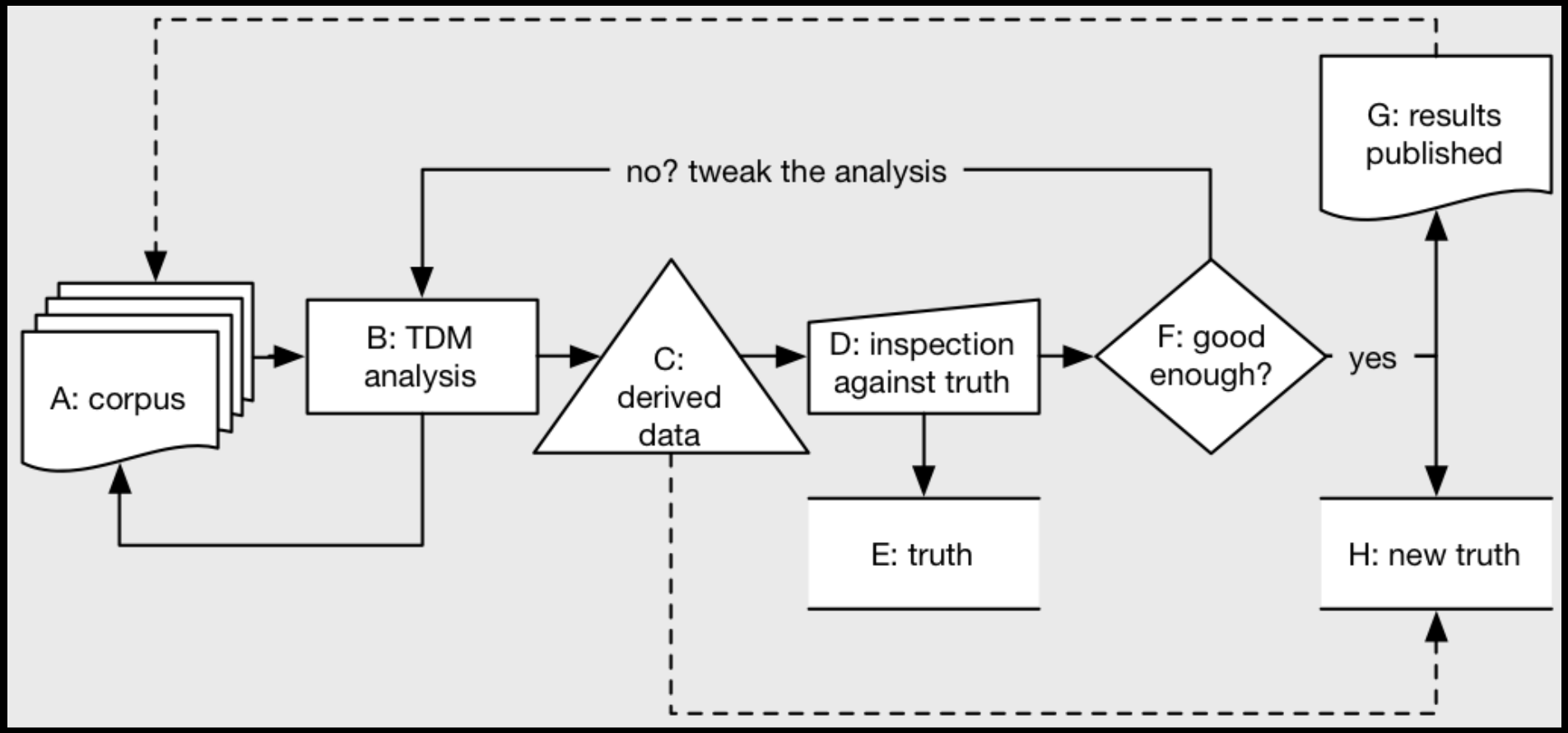
Marti A. Hearst. 1999. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA, 3-10. DOI=10.3115/1034678.1034679 http://dx.doi.org/10.3115/1034678.1034679

2008

“The use of automated methods for exploiting the enormous amount of knowledge available in the biomedical literature.”

Cohen, K. Bretonnel; Hunter, Lawrence (2008). "Getting Started in Text Mining". *PLoS Computational Biology* **4** (1): e20. doi:10.1371/journal.pcbi.0040020. PMC 2217579.PMID 18225946.

TDM Defined



Typical Text and Data Mining (TDM)

1. GeoDeepDive, a system that helps geoscientists discover information and knowledge buried in the text, tables, and figures of geology journal articles

Zhang, C., V. Govindaraju, J. Borchardt, T. Foltz, C. Ré, and S. Peters. 2013. GeoDeepDive: Statistical inference using familiar data-processing languages. SIGMOD '13, New York, New York.

2. Leveraging text mining to improve human curation

Thomas C Wiegers, Allan Peter Davis, K Bretonnel Cohen, Lynette Hirschman and Carolyn J Mattingly. Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). BMC Bioinformatics 2009, 10:326doi:10.1186/1471-2105-10-326

3. Discovering a New Link between Genes and Osteoporosis

Varun K. Gajendran, Jia-Ren Lin, David P. Fyhrie, An application of bioinformatics and text mining to the discovery of novel genes related to bone biology, Bone, Volume 40, Issue 5, May 2007, Pages 1378-1388, ISSN 8756-3282, DOI: 10.1016/j.bone. 2006.12.067. (http://www.sciencedirect. com/science/article/B6T4Y-4MVVSS1 1/2/ df681f901acd33d5f3eceedb36fe441e)

TDM Examples

TDM and Copyright—US

"(TDM is a kind of non-consumptive use) facilitated by new technologies and increasing computer power, that (does) not directly trade on the underlying creative and expressive purpose of the work being used.

Copying may include only the non-copyrightable aspects of the works, such as ideas, facts, or algorithms—in which case fair use need not come into play—or it may entail copying some expressive aspects of the work, but only as a means to a non-consumptive end."

Urban, Jennifer. 2010. Updating Fair Use for Innovators and Creators in the Digital Age

TDM and Copyright—US

Authors Guild, Inc. v. HathiTrust, 902 F. Supp. 2d 445 - Dist. Court, SD New York, 2012;

Judge Baer: “(Defendants’) participation in the (Mass Digitization Project) and the present application of the (HathiTrust Digital Library) are protected under fair use.”

“I cannot imagine a definition of fair use that would not encompass the transformative uses made by Defendants’ MDP and would require that I terminate this invaluable contribution to the progress of science and cultivation of the arts.”

Authors Guild, Inc. et al. v. Google Inc., U.S. District Court, Southern District of New York, No. 05-08136.

Judge Chin: “Google Books provides significant public benefits. It advances the progress of the arts and sciences, while maintaining respectful consideration for the rights of authors and other creative individuals, and without adversely impacting the rights of copyright holders.

Google's actions in providing the libraries with the ability to engage in activities that advance the arts and sciences constitute fair use.”

Not a Lot of Case Law and TDM

TDM and Copyright—US

Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003)

Judge Nelson: “We hold that Arriba’s reproduction of Kelly’s images for use as thumbnails in Arriba’s search engine is fair use under the Copyright Act.”

Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1165 (9th Cir. 2007)

Judge Nelson: We conclude that Google's fair use defense is likely to succeed at trial, and therefore we reverse the district court's determination that Google's thumbnail versions of **Perfect 10's** images likely constituted a direct infringement.

Not a Lot of Case Law and TDM

TDM and Copyright—UK

"Researchers want to use every technological tool available, and they want to develop new ones. However, the law can block valuable new technologies, like text and data mining, simply because those technologies were not imagined when the law was formed. In teaching, the greatly expanded scope of what is possible is often unnecessarily limited by uncertainty about what is legal. Many university academics – along with teachers elsewhere in the education sector – are uncertain what copyright permits for themselves and their students."

Copyright Exceptions for the Digital Age

TDM and Copyright—UK

“29A Copies for text and data analysis for non-commercial research

(1) The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that—

(a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and

(b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).

(2) Where a copy of a work has been made under this section, copyright in the work is infringed if—

(a) the copy is transferred to any other person, except where the transfer is authorised by the copyright owner, or

(b) the copy is used for any purpose other than that mentioned in subsection (1)(a), except where the use is authorised by the copyright owner.

(3) If a copy made under this section is subsequently dealt with—

(a) it is to be treated as an infringing copy for the purposes of that dealing, and

(b) if that dealing infringes copyright, it is to be treated as an infringing copy for all subsequent purposes.

(4) In subsection (3) “dealt with” means sold or let for hire, or offered or exposed for sale or hire.

(5) To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable.”.

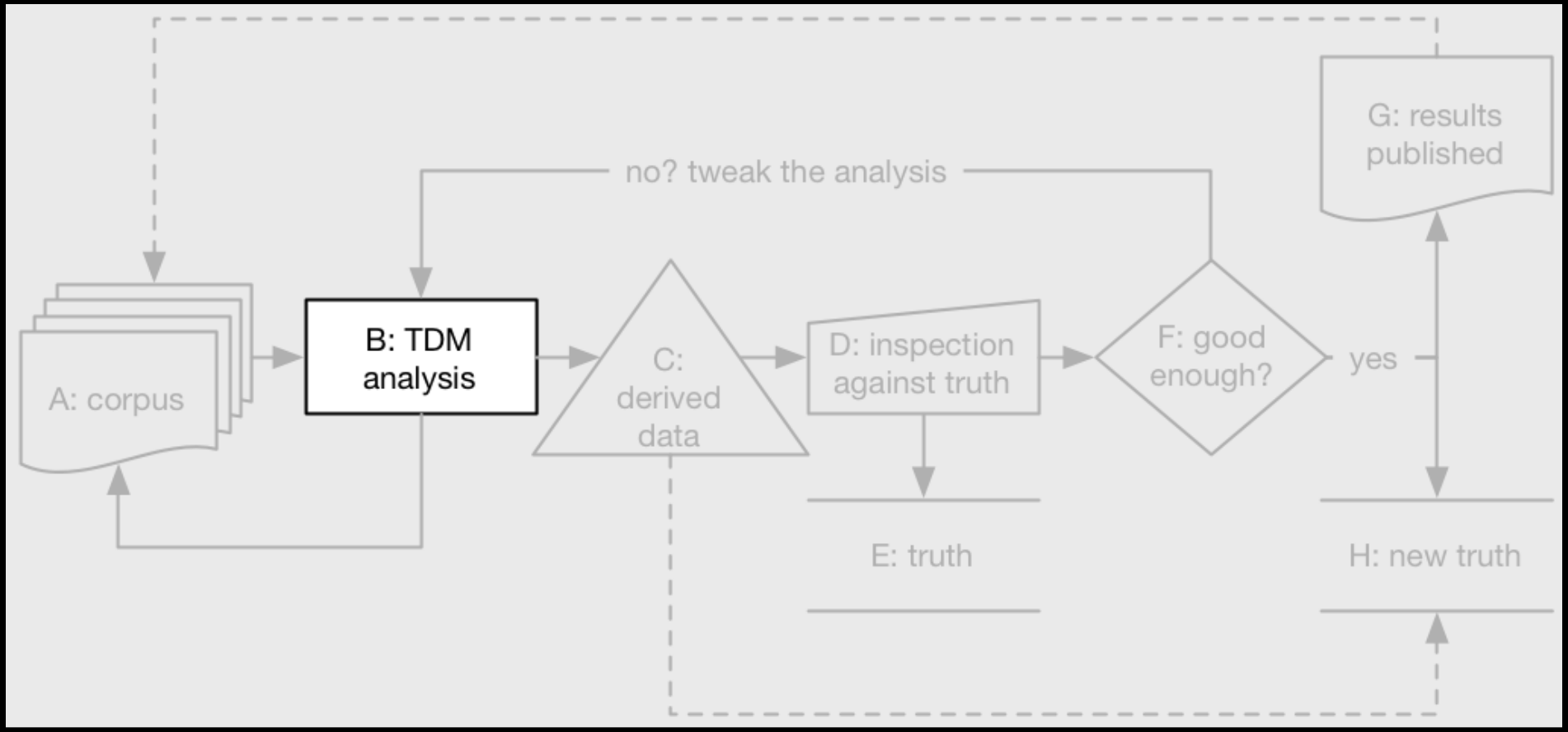
Copyright Exceptions for the Digital Age

TDM and Copyright—Australia

"There is no specific exception in the Copyright Act for text or data mining. Where the text or data mining process involves the copying, digitisation, or reformatting of copyright material without permission, it may give rise to copyright infringement. One issue is whether text mining, if done for the purposes of research or study, would be covered by the fair dealing exceptions. The reach of the fair dealing exceptions may not extend to text mining if the whole dataset needs to be copied and converted into a suitable format. Such copying would be more than a 'reasonable portion' of the work concerned."

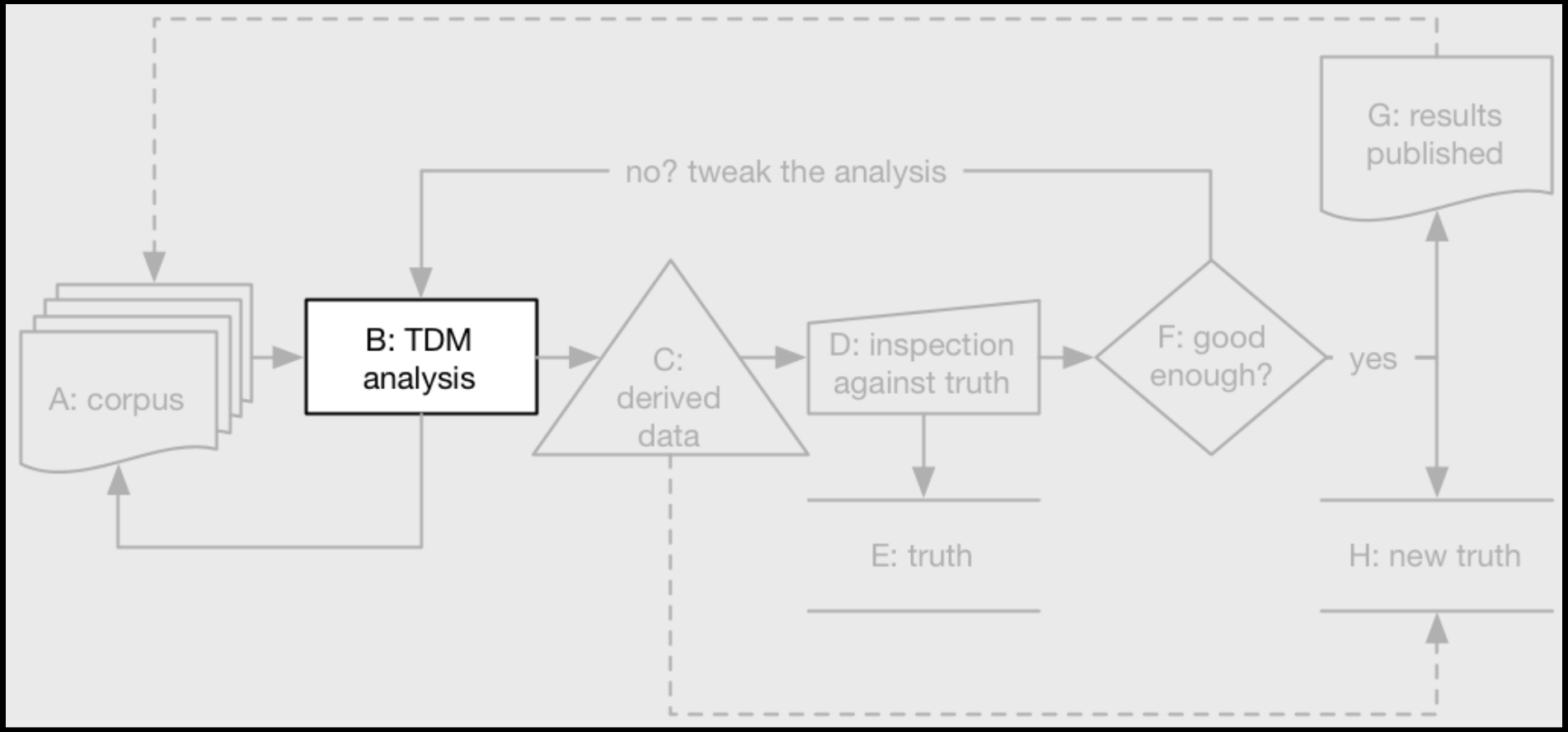
Non Consumptive Use, Australian Law Review Centre

TDM and Copyright—Other Countries

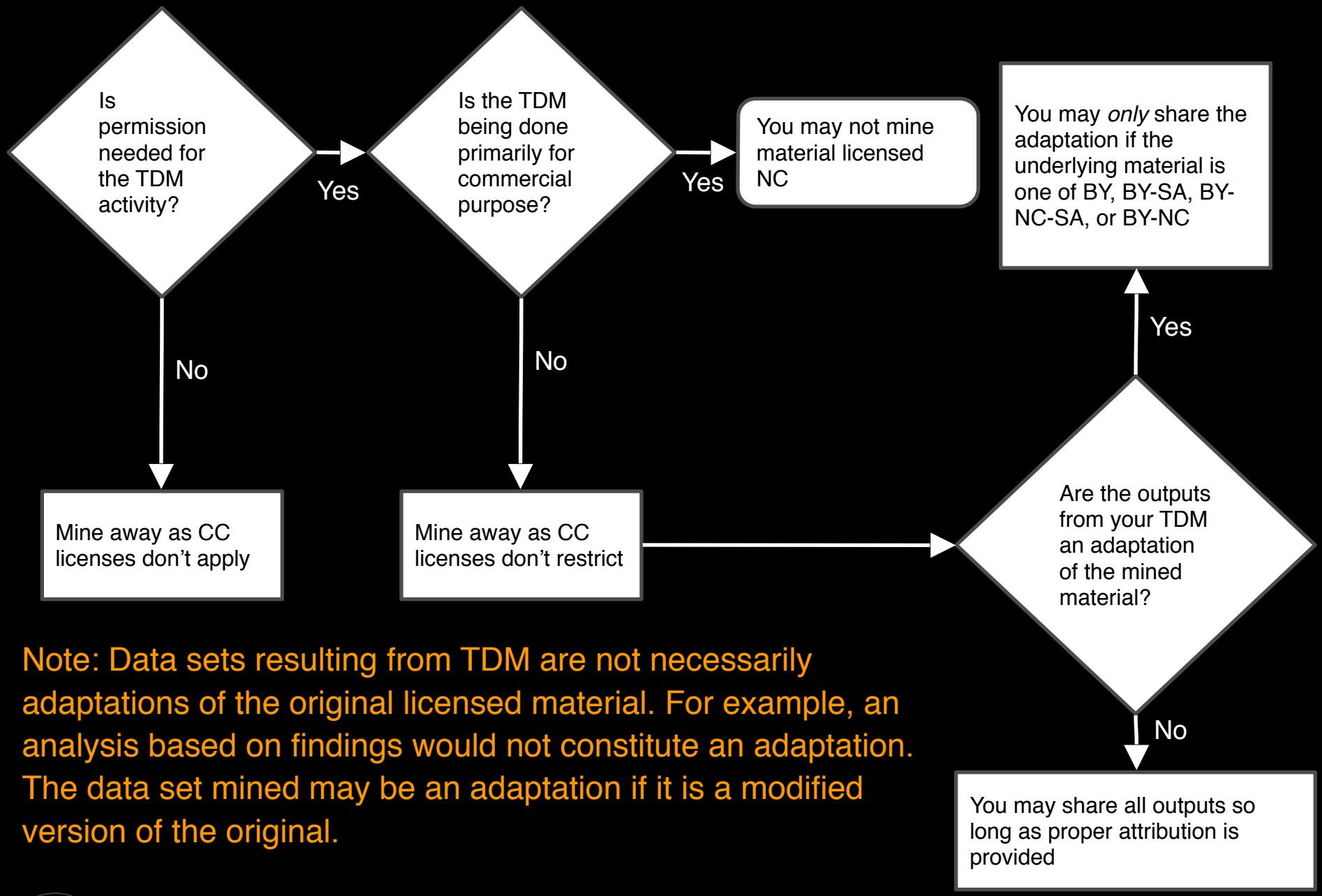


A CC license does not apply to uses such as TDM that qualify as Exceptions and Limitations, and the user does not need to comply with terms and conditions of the license if TDM doesn't implicate copyright or similar rights covered by the CC license.

TDM and CC Licenses



TDM and CC Licenses



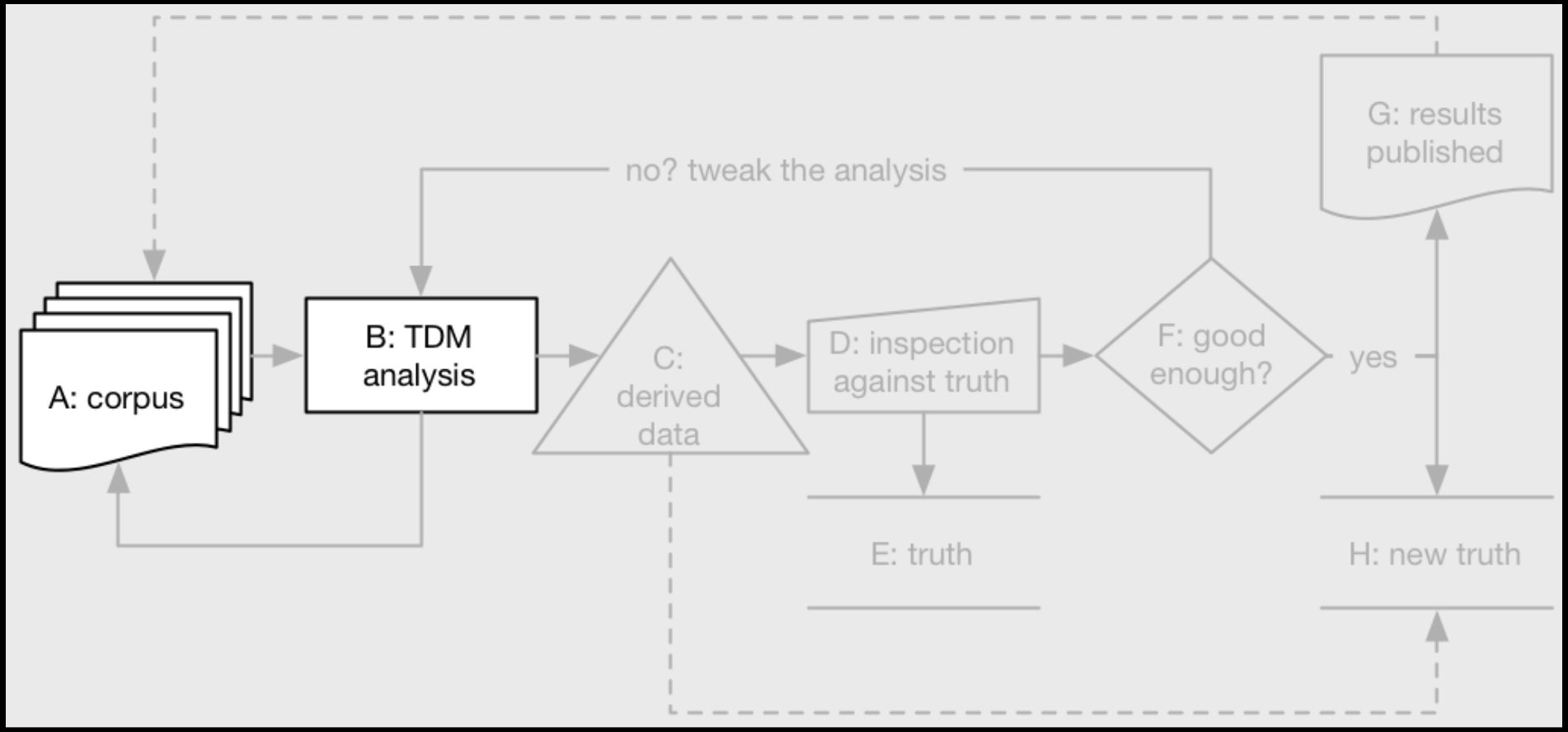
Note: Data sets resulting from TDM are not necessarily adaptations of the original licensed material. For example, an analysis based on findings would not constitute an adaptation. The data set mined may be an adaptation if it is a modified version of the original.

TDM and CC Licenses

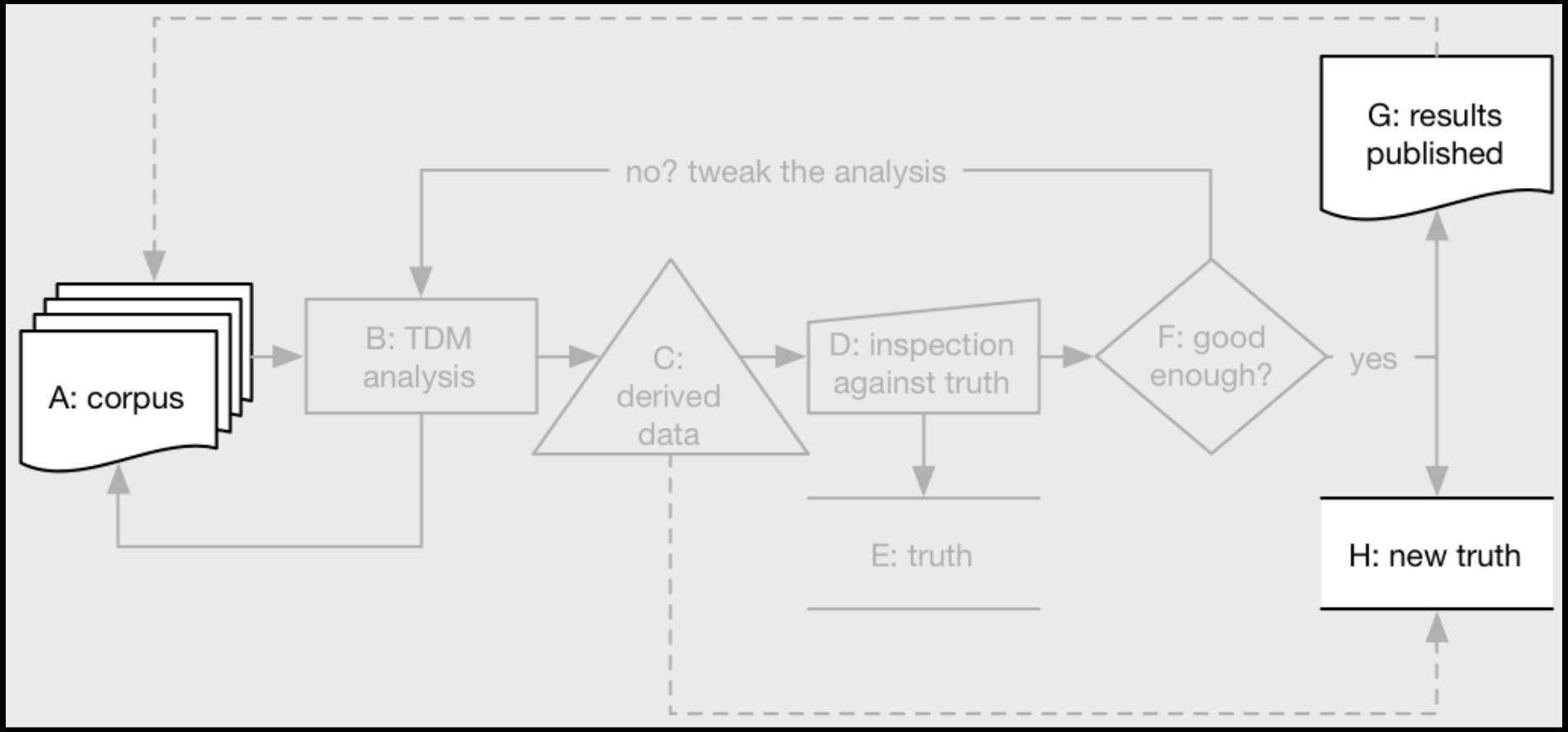
4.0 License	**Permissions Granted** *(✔green = permitted; ✘red = not permitted)*			
	To mine Licensed Material for commercial purposes	To produce Adapted Material	To Share Licensed Material	To Share Adapted Material
BY	✔	✔	✔	✔
BY-SA	✔	✔	✔	✔
BY-NC	✘	✔	✔	✔
BY-NC-SA	✘	✔	✔	✔
BY-ND	✔	✔	✔	✘
BY-NC-ND	✘	✔	✔	✘

Note: The above chart applies only if permission is needed as a matter of copyright and similar rights. If permission is not needed, there is no need to comply with the CC License Terms and Conditions when doing TDM.

TDM and CC Licenses



Any modification in TDM analysis requires running it again on the corpus, a time-consumptive process made easier by a persistent cache of the corpus. Publisher contracts can create hurdles in creating such a persistent cache.

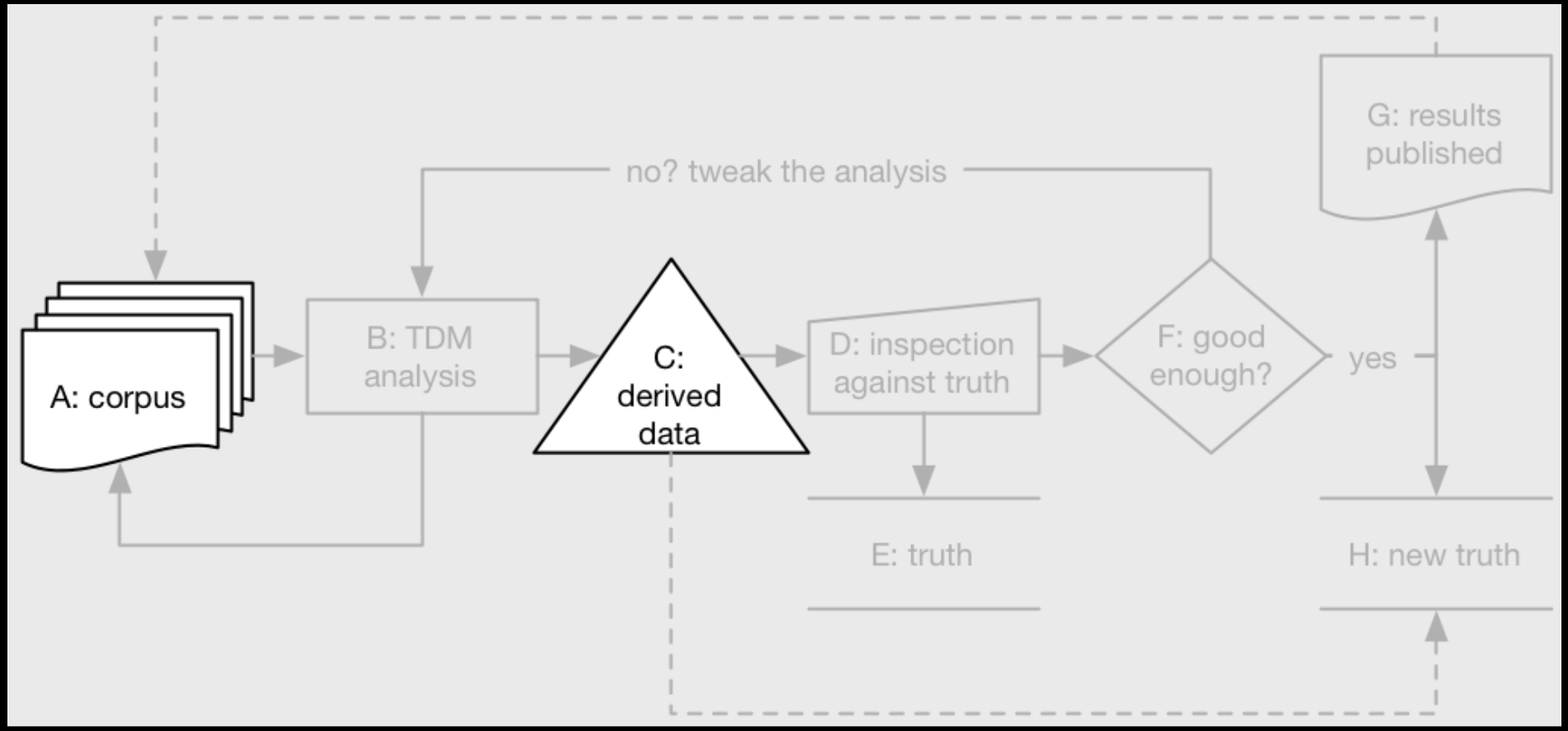


Licenses/contracts on the corpus can affect the license under which the final results and data are published. What would happen if the corpus were made of entities under different kinds of licenses?

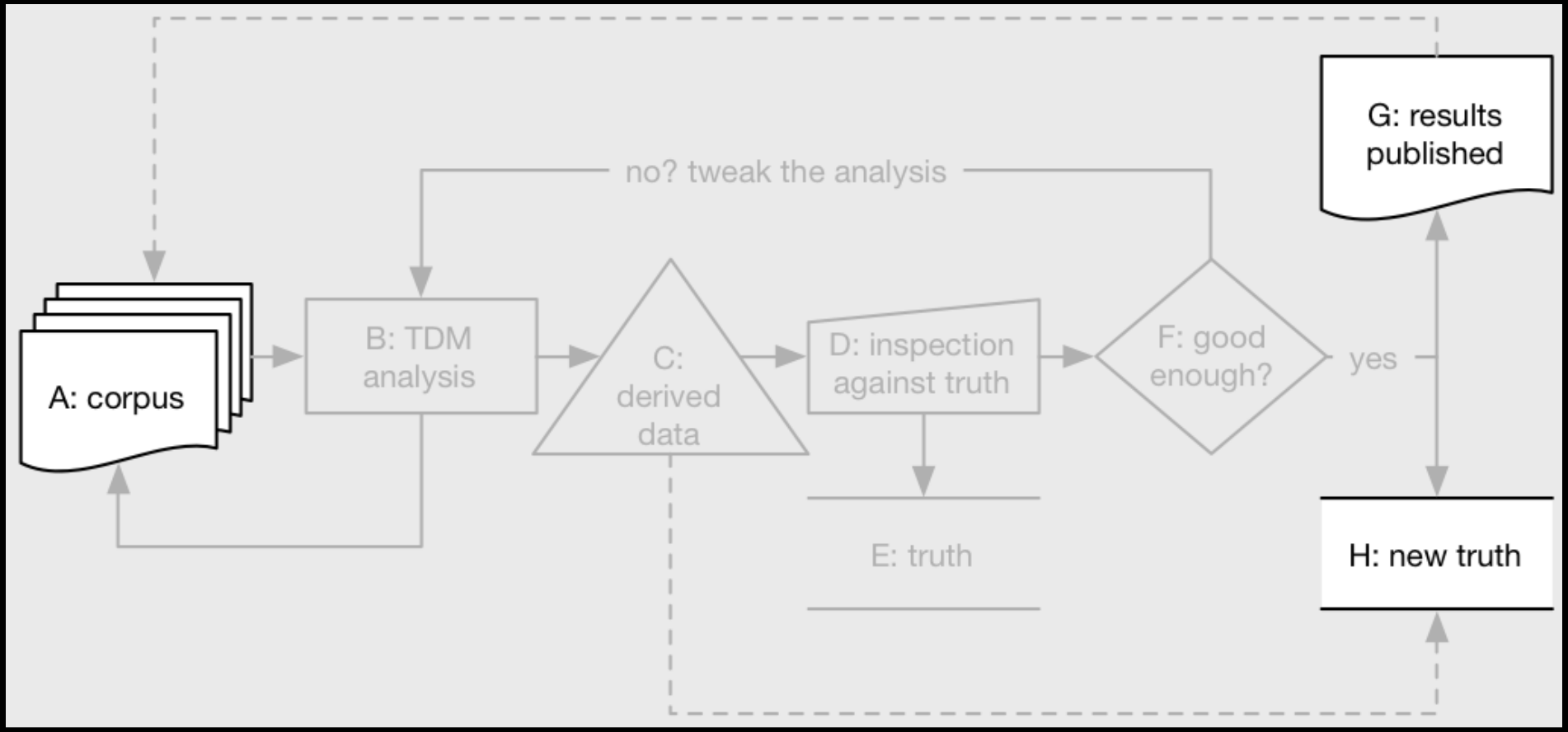
Force11 Data Citation Principles

1. **Importance:** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. **Credit and attribution:** Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. **Evidence:** Where a specific claim rests upon data, the corresponding data citation should be provided.
4. **Unique Identification:** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. **Access:** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. **Persistence:** Metadata describing the data, and unique identifiers should persist, even beyond the lifespan of the data they describe.
7. **Versioning and granularity:** Data citations should facilitate identification and access to different versions and/or subsets of data. Citations should include sufficient detail to verifiably link the citing work to the portion and version of data cited.
8. **Interoperability and flexibility:** Data citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data citation practices across communities.

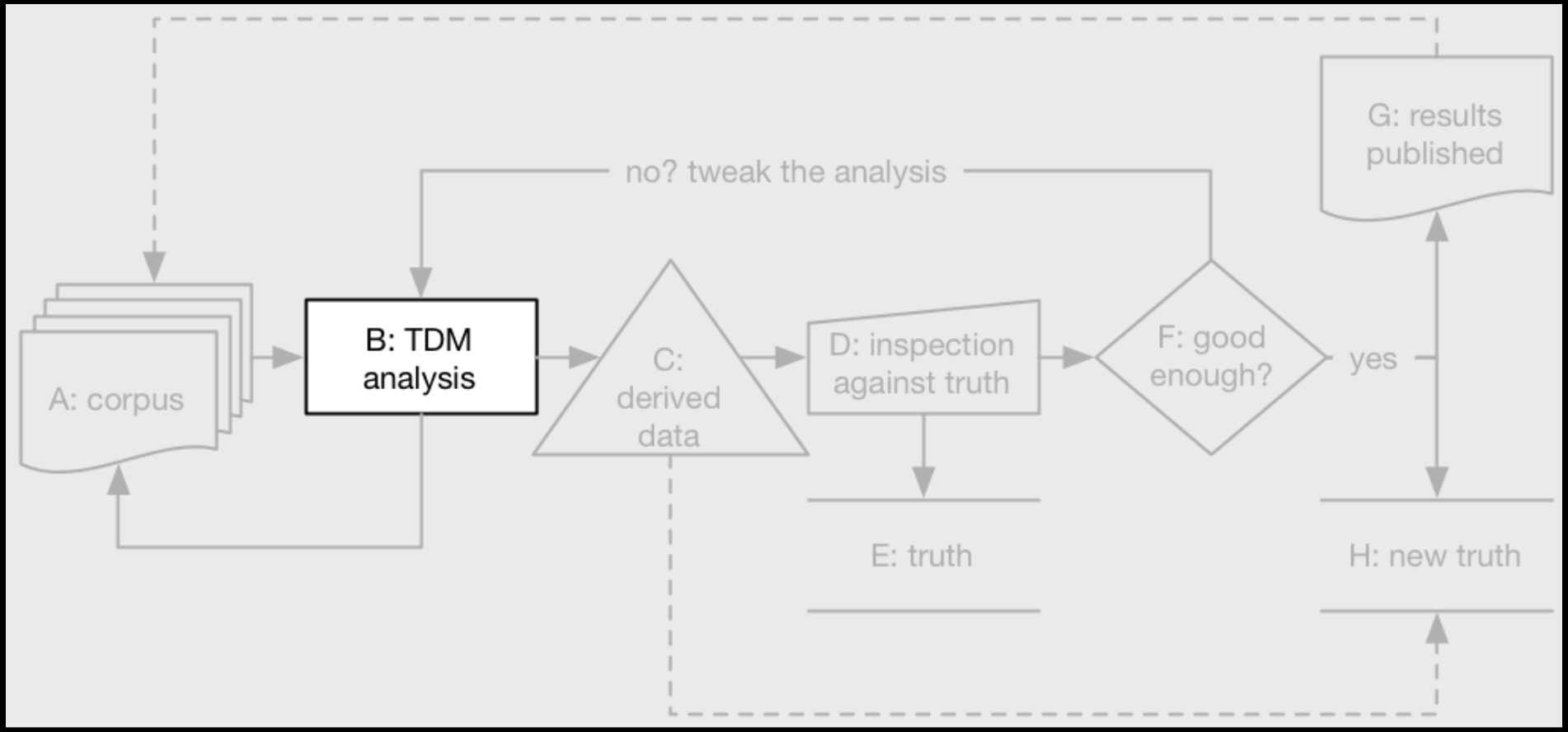
Data Citation and TDM



The Importance Principle

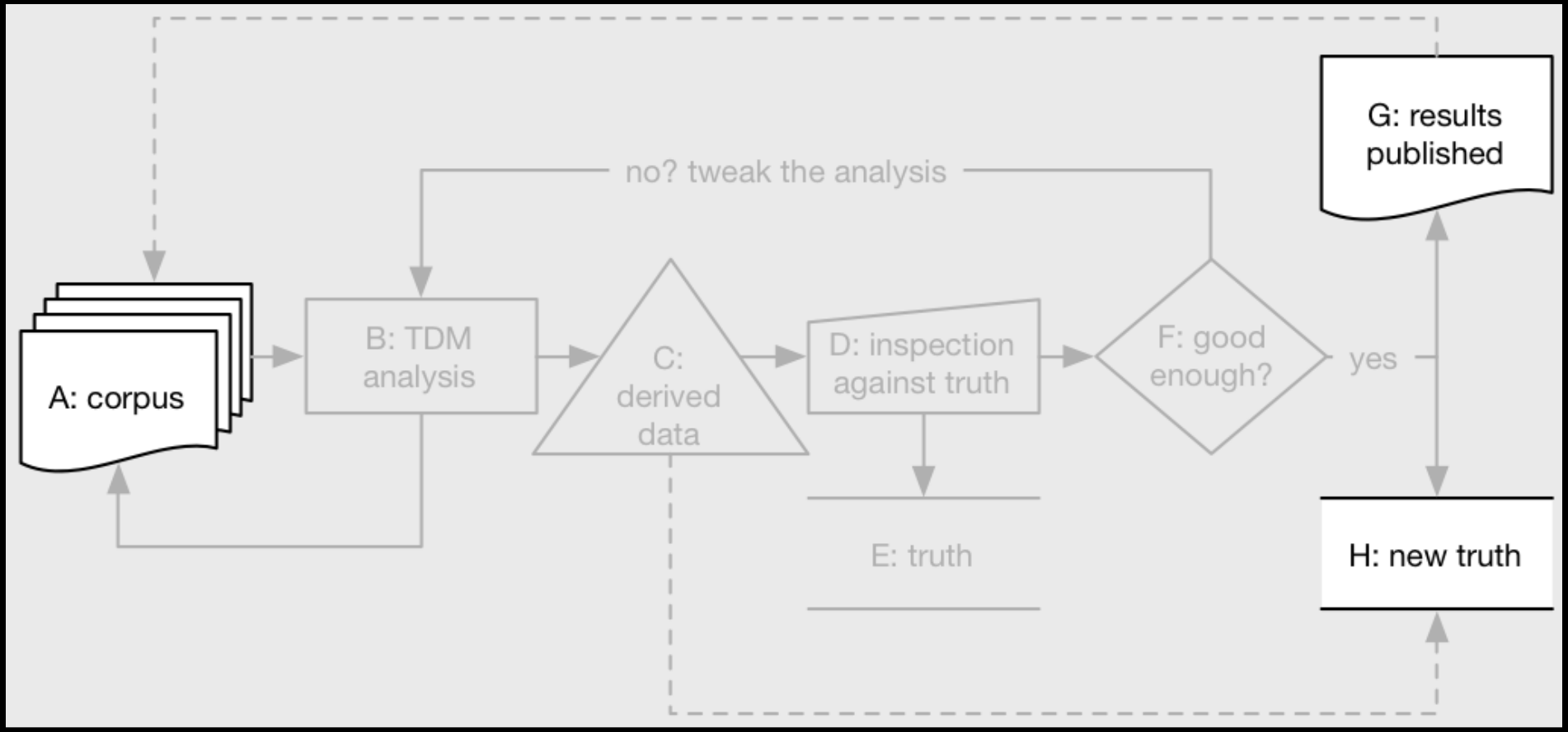


The Credit and Attribution Principle



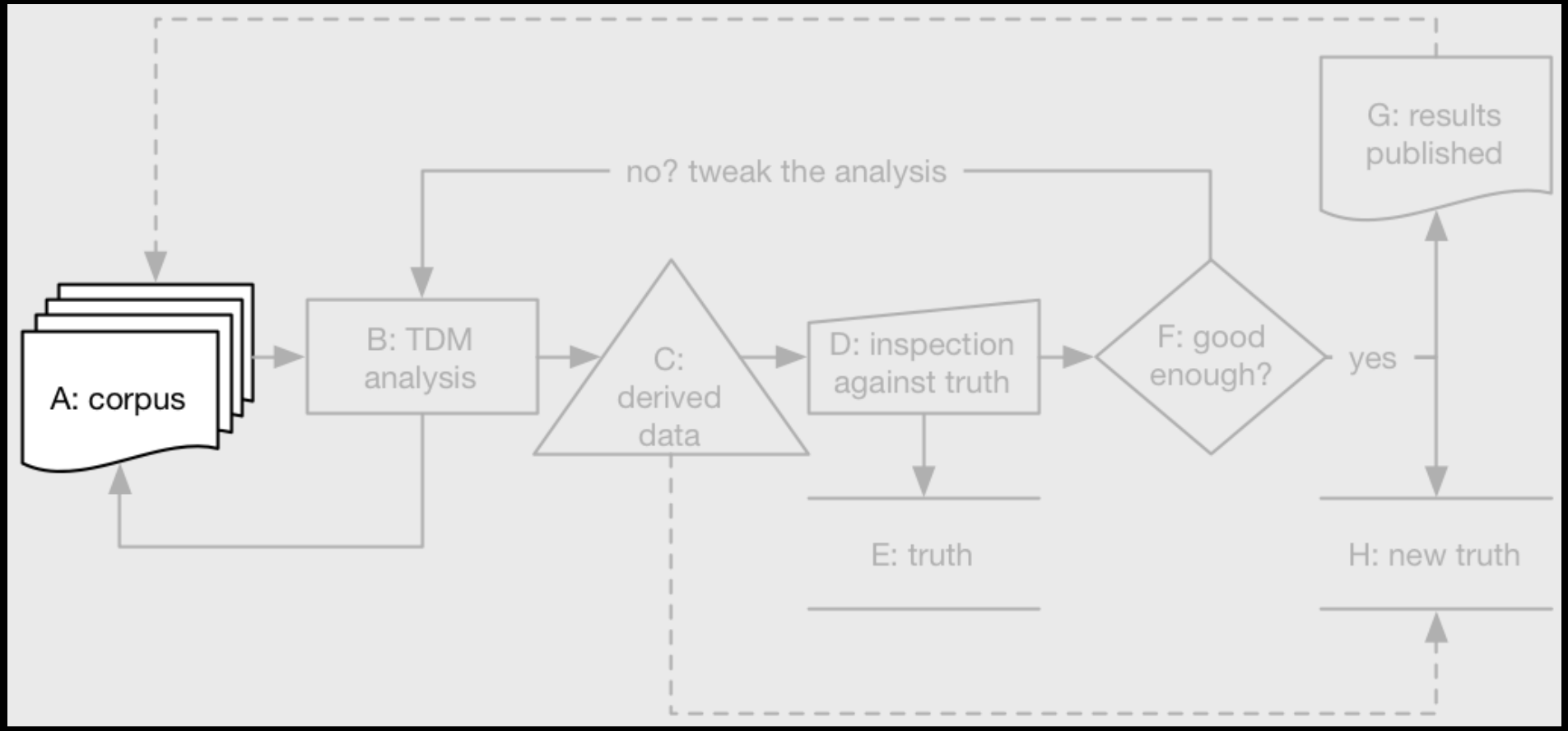
Copyright allows TDM as fair use because of its Transformative nature (in the US). Note: a CC License does not apply to uses (such as TDM) that qualify as Exceptions and Limitations, and the user does not need to comply with terms and conditions of the license if the TDM doesn't implicate copyright

TDM as Fair Use

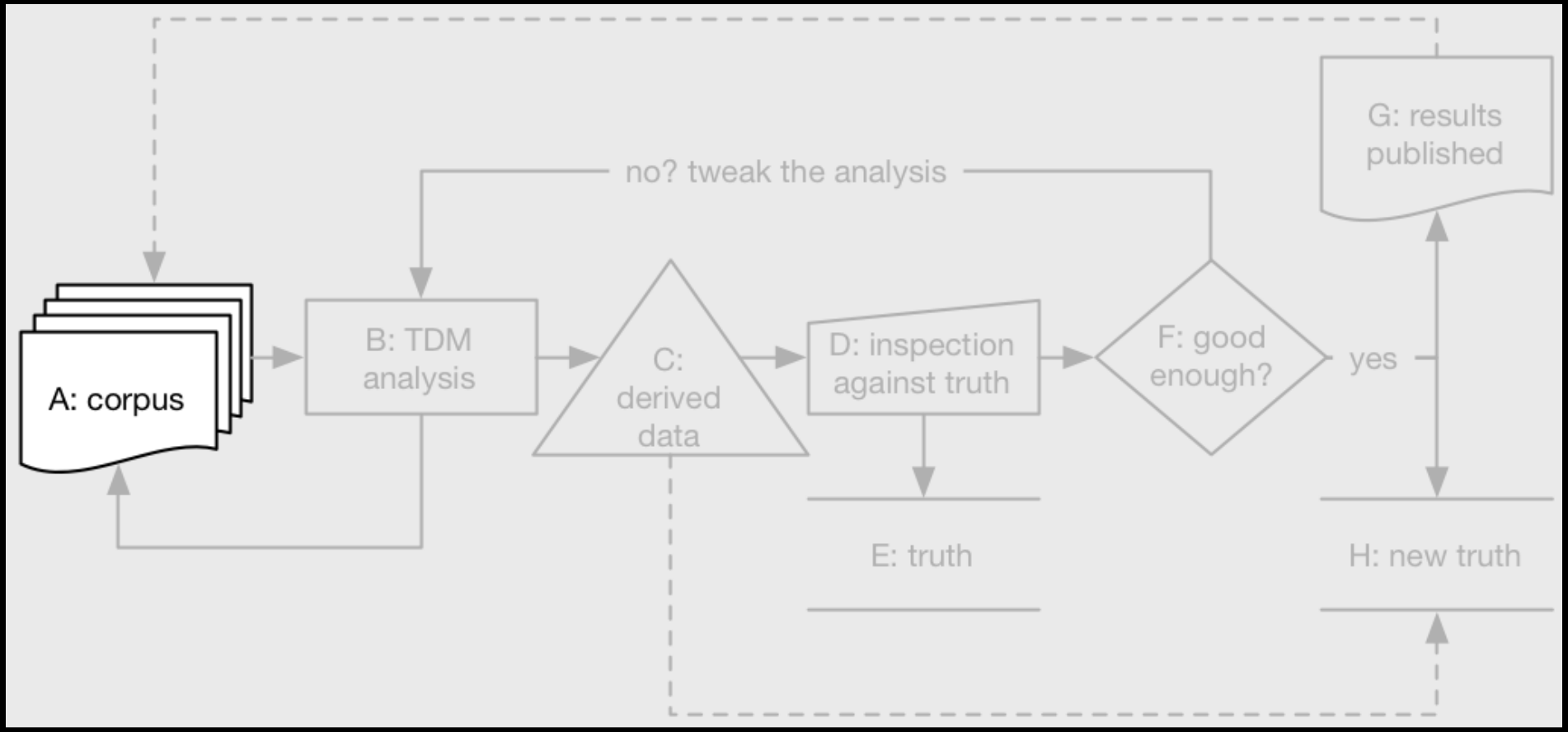


Licenses/contracts on the corpus can affect the license under which the final results and data are published. What would happen if the corpus were made of articles under different kinds of licenses?

TDM and Licensing



Should scholars be able to get back to the corpus?



Should the corpus be open access and be available online for downstream users?

The Access Principle

Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries

LIBER is Europe's largest network of research libraries, with over 400 members.

European Research Organisations Call On Elsevier To Withdraw TDM Policy

Eighteen European research and library organisations, including LIBER, are today calling on Elsevier to withdraw its current policy on text and data mining (TDM)

Our request has been laid out and explained in an open letter to Michiel Kolman, Senior VP Global Academic Relations at the academic publishing company.

We believe that Elsevier's current TDM policy places unnecessary restrictions on researchers. It limits their ability, and their right, to mine content to which they have legal access.

TDM allows researchers to derive information from articles and datasets by seeking patterns in

Core Activities

Scholarly Communication & Research Infrastructures

Reshaping The Research Library

Building Bottom-Up Support

Realising the innovative potential of digital research methods: *a call from the research community.*

1 July 2014

Open Letter to Michiel Kolman, Senior VP Global Academic Relations, Elsevier

On behalf of research community stakeholders, we are calling on Elsevier to withdraw its current policy on text and data mining (TDM).

TDM is a digital research method which enables the analysis of vast and heterogeneous types of content. It has led to new medical and scientific discoveries and is set to be key to increasing the productivity of research and become an established element of research methodologies. Europe is falling behind in the exploitation of TDM because the lack of clarity in the current European copyright framework is disincentivising the uptake of TDM by researchers. In the UK, an exception for TDM has been introduced into legislation. What this means is that TDM will no longer be an activity that is subject to licence in the UK; any researcher will be free to mine content to which their institution has legal access. We see no reason that researchers across Europe and beyond should not have equal rights to mine content to which they have legal access.

Restrictive licences provided by publishers for access to content for the purpose of TDM have the potential to further disadvantage the research community by enforcing strict parameters around

Building Bottom-Up Support

10th July 2014

RE: OPEN LETTER IN RESPONSE TO THE REQUEST FOR ELSEVIER TO WITHDRAW ITS TEXT AND DATA MINING POLICY

Dear Colleagues,

I am writing in regard to your open letter asking Elsevier to withdraw its text and data mining policy. We would like to reassure librarians that we haven't introduced our TDM policy to undermine your calls for copyright exceptions, but rather as the next natural step in the evolution of our TDM services, which have been available since 2006 and have evolved continually over this time. We understand that librarians will continue to lobby for exceptions, and while we disagree on whether these are necessary, would call on all stakeholders to agree that it is important to provide researchers with practical, workable TDM services now no matter the legal framework of the country in which they are based.

We would also appreciate the opportunity to elaborate further here on why we think our policy is fit for the purposes of the research community as well as Elsevier. I would like to first make clear that our

Puneet Kishor
Manager, Science and Data Policy
punkish@creativecommons.org