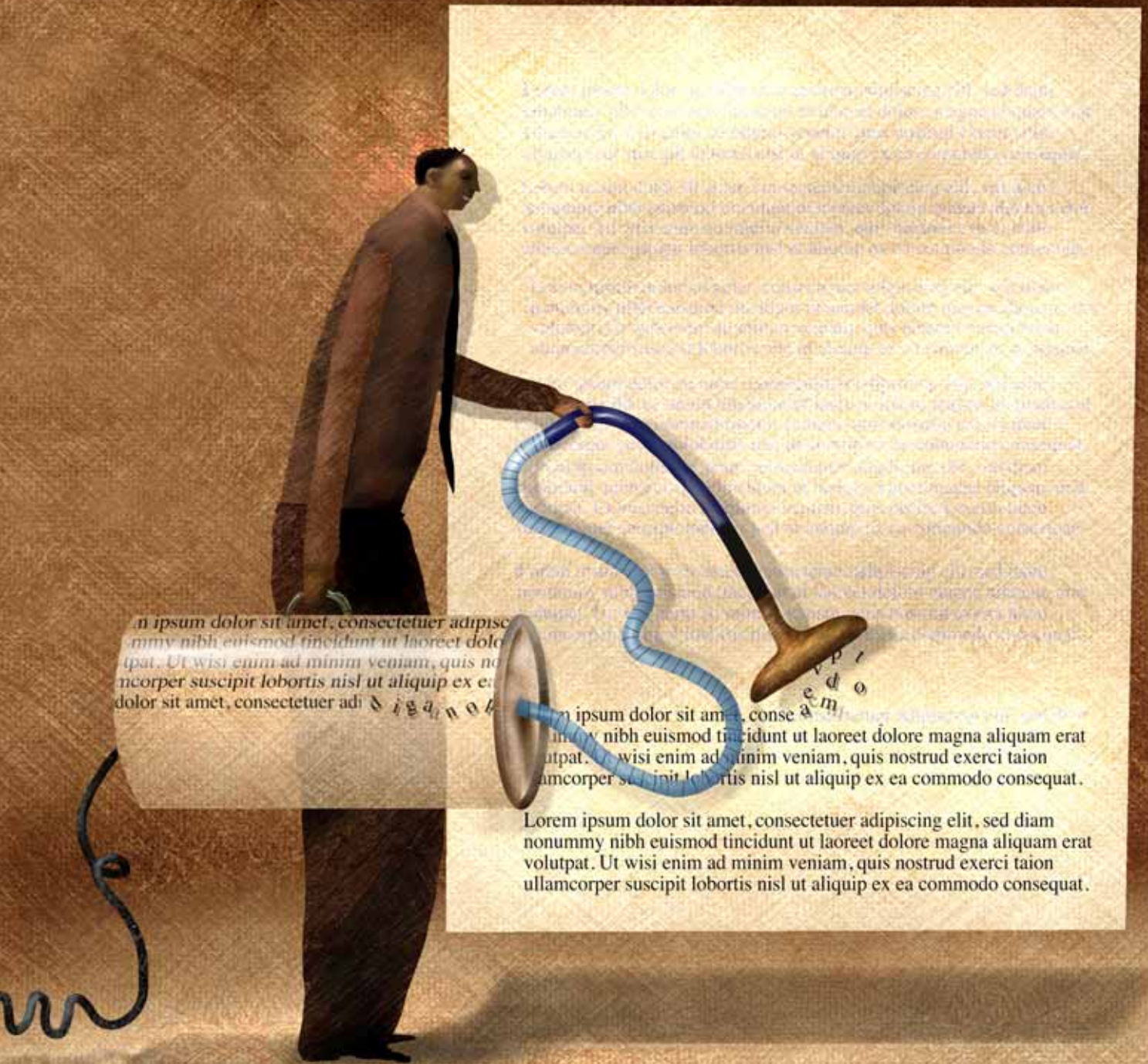


# Plagiarism: A New Look At An Old Problem

John M. Barrie, Ph.D.  
Turnitin and iThenticate - Founder



## **Profession prior to plagiarism**

- Undergraduate: U.C. Berkeley, Rhetoric and Neurobiology
- Doctorate: U.C. Berkeley — Biophysics Multidisciplinary Graduate Group, Neurobiology
  - Dissertation: Theoretical and computational electro-neurophysiology - Spatiotemporal dynamics of the neocortical EEG (aka, the physiology of perception)

## How did I get here?

- Concurrent research regarding technologies to extend traditional educational models
  - Created class websites (1994-2000)
  - Goals:
    - Expose undergraduates to peer review process
    - Allow students to share information in ways not possible without the Internet
  - Results published as a 1996 study in *Science*:
    - Acquisition of core course ideas was facilitated
    - Students acquired Internet skills unrelated to the class
    - Faculty interacted with the class website < 15 hrs/term
  - Unexpected observation: There were significant levels of digital plagiarism and collusion among the Berkeley students
  - 1996 prediction: Intellectual property theft (plagiarism) would become an enormous problem for educators and for Academia

**Turnitin has thousands of institutional clients in >50 countries representing over 9,000,000 students and we receive >20,000 student papers per day**



**Hundreds of universities in the UK (JISC)**

**Cornell University  
UCLA, UC San Diego, UC Davis, UC Santa Cruz, etc.  
Georgetown University  
Duke University  
California Institute of Technology  
Colgate University  
Rice University  
Boston University  
California State University System  
Georgia State University System  
Michigan Tech  
Houston Unified School District  
The University of Maryland System  
Rochester Institute of Technology  
Rutgers University System  
US Military Academy, West Point  
Trinity College  
Swarthmore College  
The Citadel  
The University of Western Ontario  
Manakau Institute of Technology, New Zealand**

**ithenticate**

**A service of iParadigms**

## Aside

- We index the works from some of the largest aggregators (going back five years)
- Students are referencing (or not) information from the Internet in amounts 1000 times greater than information from the more traditional sources that we index

# **Unattributed use of another person's ideas seems to be a general problem in our society**

- Many politicians and corporate leaders rarely write their own speeches, policy positions or books
- Some researchers routinely take credit for work done with their grant money
- Some famous journalists have research teams write their articles
- Many legal opinions written by distinguished judges and justices are really written by their judicial clerks
- Plagiarism in academia is rampant (according to the largest study of plagiarism, more than 40% of students have admitted to plagiarizing from the Internet)
- We find that about 30% of the more than 20,000 papers we receive each day are less than original

# Similarities between publishing and academia

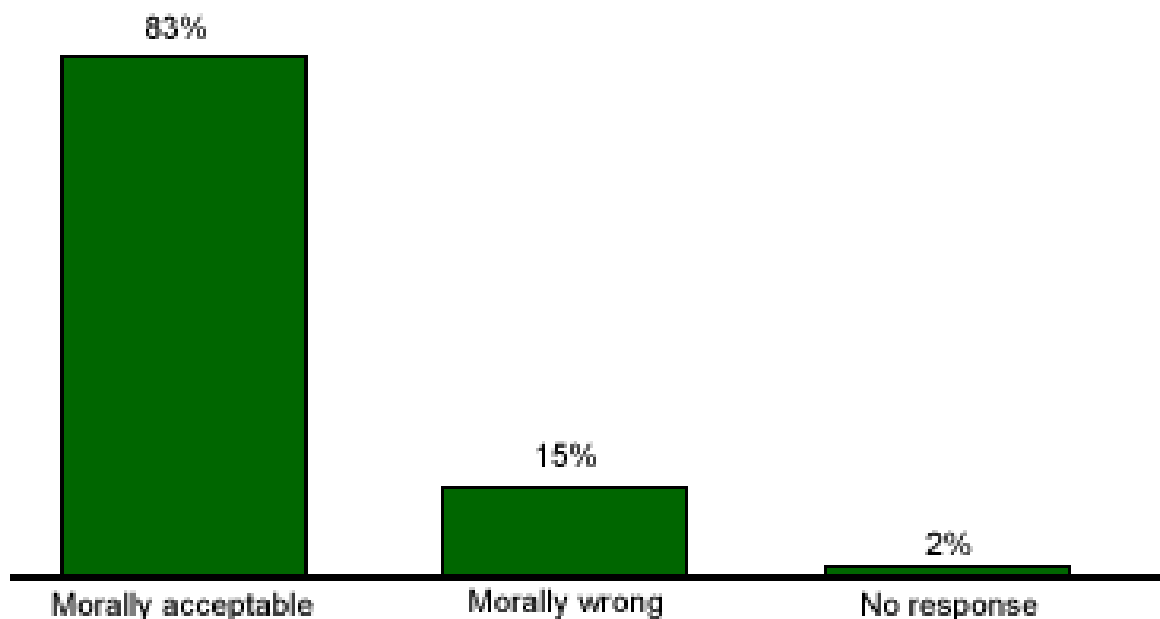
- Both institutions consider plagiarism to be a capital crime
- Both institutions suffer sporadic plagiarism scandals brought to light mainly by chance
- Both institutions rely on a code of honor and integrity to deter plagiarism and fraud
- Both institutions typically react to plagiarism scandals by placing the offending party's head on a stick for others to see to further deter unethical activity
- In both cases, each institution is being reactive instead of proactive and their digital plagiarism problem continues... in spite of all efforts to deter

# Plagiarism in the media and in academia is a digital problem

- Attitudes among students (your future authors and researchers) regarding **digital** intellectual property theft are enlightening. From a September 30, 2003 Gallup Poll:

## Moral Acceptability of Downloading Music for Free

*Next, I'm going to read you a list of issues. Regardless of whether or not you think it should be legal, for each one, please tell me whether you personally believe that in general it is morally acceptable or morally wrong. How about downloading music from the Internet for free?*



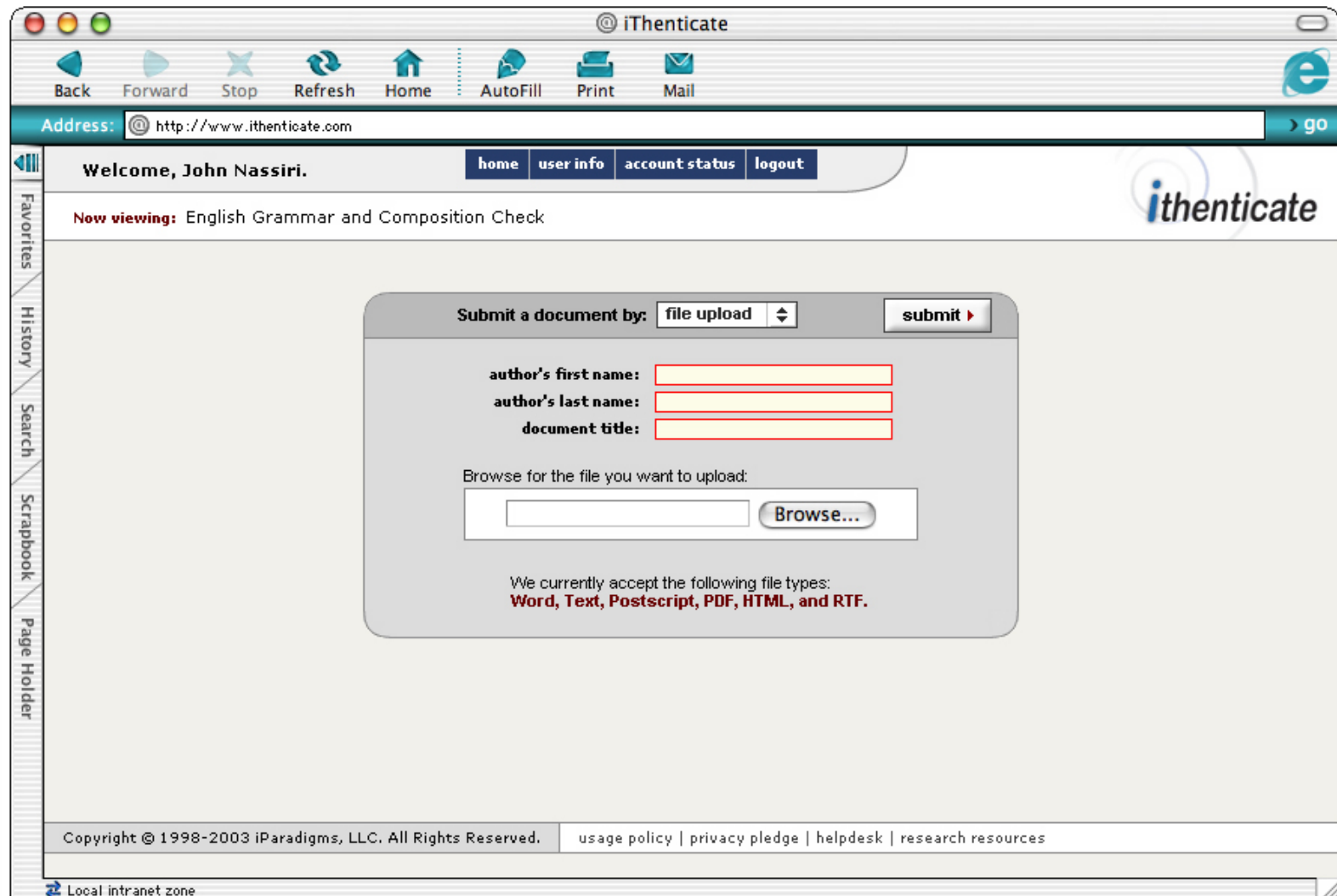
- Only a digital solution can address this digital problem



## **Publishers have some additional problems that academics grading papers may not have**

- Potential for lawsuits from an increasingly litigious society
- Immediate detriment to reputation
- Loss of revenue from down-stream intellectual property theft
- Could involve award-winning journalists, Pulitzer Prize winning authors or university presidents (such as Richard Judd, the President of Central Connecticut State University)

# Step One: The digital manuscript is submitted over the Internet and to our computers



# Finding a needle in a haystack



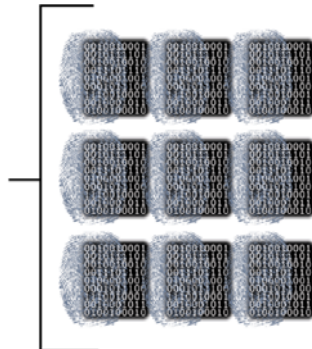
**1. Manuscript or article submitted to iThenticate**



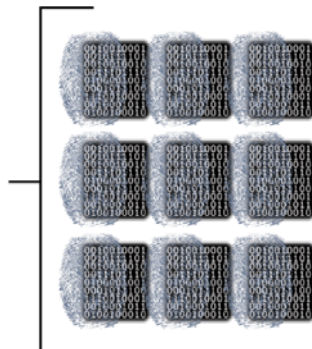
**2. Computer transforms manuscript into a digital fingerprint: a very long string of numbers**

**3**

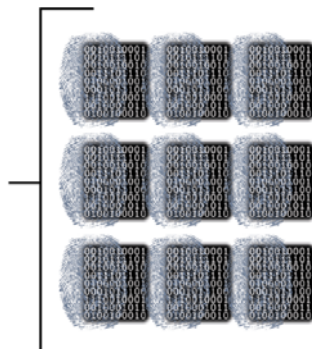
**Copy of Internet**



**Electronic Books**



**Journals / Periodicals**



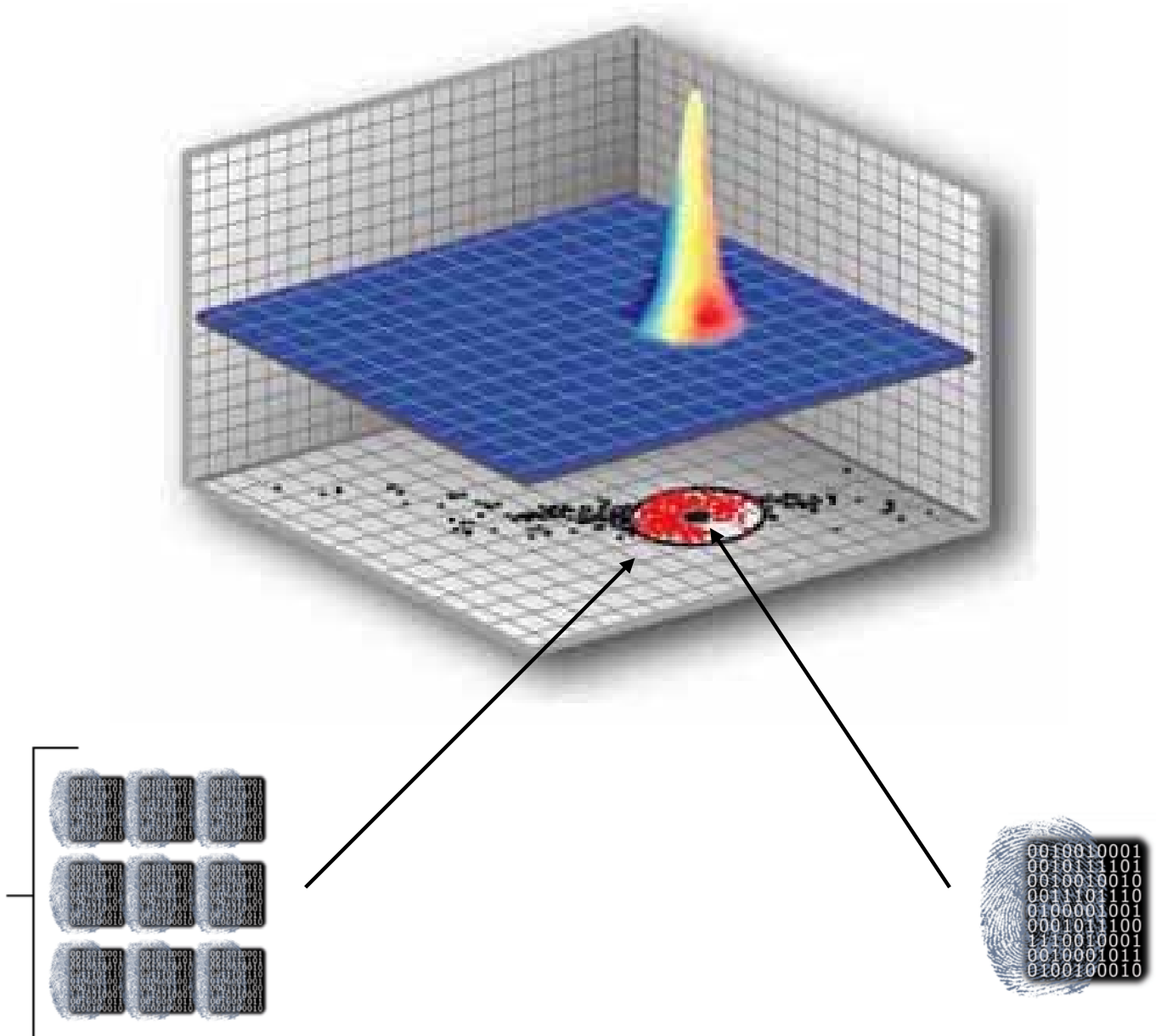
**4**

**Extract matching documents**



How can we compare a manuscript or article to **BILLIONS** of documents in any reasonable amount of time?

- We re-map the digital fingerprint of the manuscript or article into a high dimensional space and test for clustering



# Generating the Originality Report

Entire process: < 10 seconds



Matching passages  
from 3+ billion  
Internet web pages:  
downloaded at a rate of  
40 million pages/day



Matching passages  
from thousands of  
digital books



Matching passages  
from tens of millions  
of periodical articles



5

Compare matching  
passages to original  
manuscript or article

6

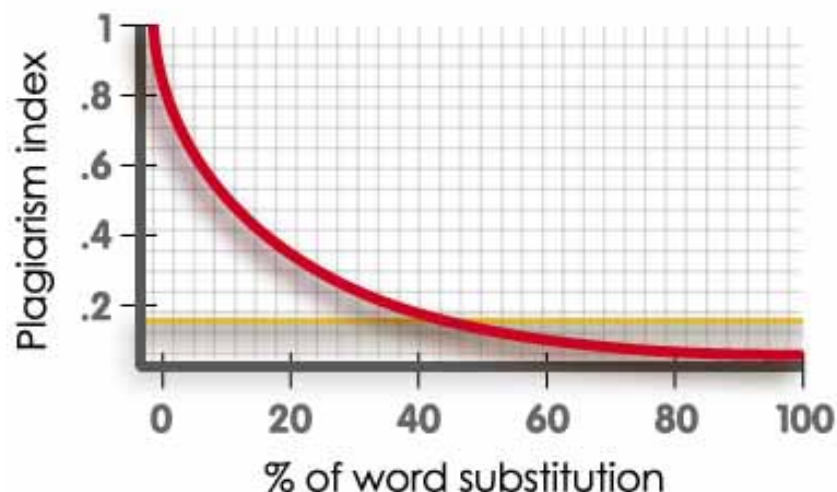
# Detection of word substitution or alteration

## MACBETH MANUSCRIPT FROM THE INTERNET (INTRO PARAGRAPH)

Macbeth is presented as a mature man of definitely established character, successful in certain fields of activity and enjoying an enviable reputation. We must not conclude, there, that all his volitions and actions are predictable; Macbeth's character, like any other man's at a given moment, is what is being made out of potentialities plus environment, and no one, not even Macbeth himself, can know all his inordinate self-love whose actions are discovered to be-and no doubt have been for a long time-determined mainly by an inordinate desire for some temporal or mutable good.

## SAME MANUSCRIPT WITH **MODIFIED** WORDS

Macbeth is **shown** as **an empowered** man of **well-established** character, **prosperous** in **several** fields of **life** and enjoying an **esteemed** reputation. We **mustn't** conclude, **therefore**, that all **of** his volitions and actions **will be foreseeable** ; Macbeth's **essence** , like **most** other **men** at **any** given **time** , is **what's** being **created** out of potentialities **and his** environment, and no one, not even Macbeth himself, can **discern** all his **immoderate** self-love whose **behaviors** are **found** to be-and **without** doubt have been for **some** time-determined **primarily** by an **extreme** desire for **a** temporal or **changeable** good.



# Detection of sentence or paragraph addition

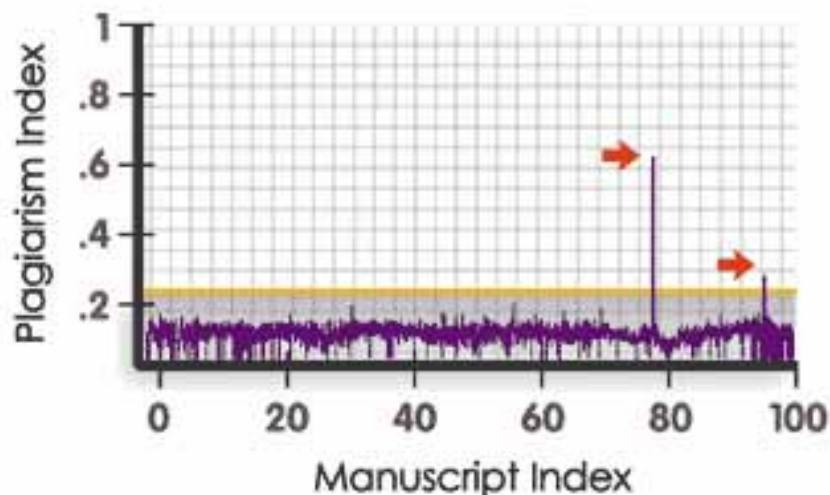
## PAPER A MACBETH INTERNET DERIVED PAPER (INTRO PARAGRAPH)

Macbeth is presented as a mature man of definitely established character, successful in certain fields of activity and enjoying an enviable reputation. We must not conclude, there, that all his volitions and actions are predictable; Macbeth's character, like any other man's at a given moment, is what is being made out of potentialities plus environment, and no one, not even Macbeth himself, can know all his inordinate self-love whose actions are discovered to be-and no doubt have been for a long time-determined mainly by an inordinate desire for some temporal or mutable good.

## PAPER A + B

## MACBETH MODIFIED TEST PAPER WITH COMBINED **ADDED CONTENT**

Shakespeare's famous play, Macbeth, is one of his great tragedies based around the classic theme of the hero's fatal flaw. Macbeth is presented as a mature man of definitely established character, successful in certain fields of activity and enjoying an enviable reputation. **Yet, like any man, he is human, and thus in possession of flaw and foibles, hidden that they may be from public eye, and hinted at by foreshadow only by the author.** We must not conclude, there, that all his volitions and actions are predictable; Macbeth's character, like any other man's at a given moment, is what is being made out of potentialities plus environment, and no one, not even Macbeth himself, can know all his inordinate self-love whose actions are discovered to be-and no doubt have been for a long time- determined mainly by an inordinate desire for some temporal or mutable good. **This desire being so strong under certain circumstances as to override all others, even, as is usually the case in tragedy, the ultimate desire of self-preservation.**



# Suspect manuscripts are highlighted

© iThenticate

Back Forward Stop Refresh Home AutoFill Print Mail

Address: @ http://www.ithenticate.com go

Welcome, John Nassiri. home user info account status logout

ithenticate

Now viewing: English Grammar and Composition Check

page: [ 1 ]

Results for: English Grammar and Composition Check

show these documents: low % ↔ high %

show: new

delete move to...

<input type="checkbox"/>	author	title	report	links	file	doc id	date
<input type="checkbox"/>	Mickelson, Mallory	<a href="#">Chapters 1-3</a>			.txt	1180910	04-15-03
<input type="checkbox"/>	Mickelson, Mallory	<a href="#">Chapters 4-6</a>			.txt	1180910	04-15-03
<input type="checkbox"/>	Dudson, Frank	<a href="#">Chapters 7-11</a>			.txt	1180910	04-15-03

Copyright © 1998-2003 iParadigms, LLC. All Rights Reserved. usage policy | privacy pledge | helpdesk | research resources

Local intranet zone



# Every instance of matching text is underlined and color-coded indicating the possible source

author: Doe, John    title: Topics in Neurobiology    paper ID: 46657    submitted: 03-21-03 12:15 PM

similarity index: ■ (68% matching text)    version: #1 (03-05-03)    [side-by-side version](#)

Sources:	link	reanalyze and exclude	match
Internet	<a href="http://www.t2.net/schask/ssmanual7.html">http://www.t2.net/schask/ssmanual7.html</a>	<input type="checkbox"/>	34%
Internet	<a href="http://www.schoolsucks.com/text/paper.cfm">http://www.schoolsucks.com/text/paper.cfm</a>	<input type="checkbox"/>	13%
Internet	<a href="http://www.psy.plym.ac.uk/year2/schizophrenia/mann/htm">http://www.psy.plym.ac.uk/year2/schizophrenia/mann/htm</a>	<input type="checkbox"/>	10%
Internet	<a href="http://www.salmon.psy.plym.ac.uk/year2/schizo/htm.html">http://www.salmon.psy.plym.ac.uk/year2/schizo/htm.html</a>	<input type="checkbox"/>	8%
Database	<a href="#">Submitted to Art Center College of Design on 2002-11-03</a>	<input type="checkbox"/>	3%

**Report text:**

It is hard to comprehend the pain caused by a chronic disease like schizophrenia without personal experience-- either as a victim or by having a relationship with a victim. In actuality no individual is immune from schizophrenia, it affects individuals, families, communities, and society as a whole. My aunt's struggle with schizophrenia for twenty years has been challenging for our family, and with a lot of courage she shared her story with me.

It has been twenty five years since I first became mentally ill. As I approach forty-five, I find myself still struggling with the same symptoms, still crippled by the same fears and paranoia. I am haunted by an evasive picture of what my life could have been, whom I might have become, what I might have accomplished. My schizophrenia is a sad realization, a painful reality that I live with everyday. I probably inherited a predisposition to mental illness; my uncle was diagnosed as having dementia praecox, an earlier term for schizophrenia. In my last year in high school, I began to experience personality changes. I did not realize the significance of the changes at the time, and I think others denied them-but looking back I can see that they were the earliest signs of the illness. I became increasingly withdrawn and sullen. I felt alienated and lonely and hated everyone. I even considered suicide. I felt as if there was a huge gap between me and the rest of the world; everybody seemed distant from me.

Schizophrenia is defined as: a group of psychoses characterized by confused and disconnected thoughts, emotions, and perceptions. Schizophrenia is a brain disorder, which is identified by specific concrete symptoms. Schizophrenia is not a split personality, or multi-personality, and it has been proven that schizophrenia is not caused by childhood trauma, bad parenting, or poverty. Schizophrenia is marked by extreme thought disorder, and is usually treatable with medication. Given proper support, many people with schizophrenia can learn how to deal

# Individual sources can be directly compared to the original manuscript

The screenshot shows a web browser window titled "iThenticate" with the address bar displaying "http://www.ithenticate.com". The main content area is titled "iThenticate Originality Report" and includes a version dropdown set to "#1 (03-05-03 21:12)". A table of report details is shown:

<b>author:</b> Doe, John
<b>title:</b> Schizophrenia: An Ongoing Struggle
<b>submitted:</b> 03-23-03 21:12 PDT
<b>paper ID:</b> 578390
<b>similarity index:</b> <span style="display: inline-block; width: 10px; height: 10px; background-color: yellow; border: 1px solid black;"></span> (68% matching text)

Below the table are links for "print version" and "help". To the right, a "match" bar shows percentages: 34%, 13%, 10%, 8%, and 3%. The "url" field contains a long URL from CNN, and the "info" field states: "Live Internet link (This is a current link on the Internet, and can be accessed at the above URL)".

The "Report text:" section contains two paragraphs of text. The first paragraph discusses the difficulty of understanding schizophrenia without personal experience. The second paragraph is a personal narrative starting with "It has been twenty five years since I first became mentally ill..." and includes several phrases underlined in red, which correspond to the underlined text in the "Source:" section.

The "Source:" section displays the title "My Two Faces" by Elizabeth Nasstan, AP Wire Services. The text below the title is a copy of the original manuscript, with the same red underlines as the report text. A vertical navigation menu on the left of the source text includes links for "events", "home", "medcenter", and "questions".

At the bottom left, a status bar indicates "Local intranet zone".

# Publishers have an additional problem

As of mid-2001 there were more than 7,200 different pirated books on the Internet

The screenshot shows a web browser window titled "iThenticate". The address bar displays "http://www.ithenticate.com". The page content includes the following information:

**author:** Rosenquist, Marjorie    **title:** Publisher's Internet Survey Chap 3-5  
**similarity index:** ■ (99% matching text)    **doc ID:** 46657

source	link	match
Internet	<a href="http://www.t2.net/schask/ssmanual7.html">http://www.t2.net/schask/ssmanual7.html</a>	96%
Proquest	<a href="http://www.schoolsucks.com/text/paper.cfm">http://www.schoolsucks.com/text/paper.cfm</a>	91%
Internet	<a href="http://www.psy.plym.ac.uk/year2/schizophrenia/mann/ftr">http://www.psy.plym.ac.uk/year2/schizophrenia/mann/ftr</a>	89%
Internet	<a href="http://www.salmon.psy.plym.ac.uk/year2/schizo/htm.html">http://www.salmon.psy.plym.ac.uk/year2/schizo/htm.html</a>	77%
Internet	<a href="http://www.t2.net/schask/ssmanual7.html">http://www.t2.net/schask/ssmanual7.html</a>	52%
Proquest	<a href="http://www.schoolsucks.com/text/paper.cfm">http://www.schoolsucks.com/text/paper.cfm</a>	34%
Internet	<a href="http://www.psy.plym.ac.uk/year2/schizo">http://www.psy.plym.ac.uk/year2/schizo</a>	33%
Internet	<a href="http://www.salmon.psy.plym.ac.uk/year2/schizo/h">http://www.salmon.psy.plym.ac.uk/year2/schizo/h</a>	31%
Internet	<a href="http://www.t2.net/schask/ssmanual7.html">http://www.t2.net/schask/ssmanual7.html</a>	12%
Proquest	<a href="http://www.schoolsucks.com/text/paper.cfm">http://www.schoolsucks.com/text/paper.cfm</a>	5%
Internet	<a href="http://www.psy.plym.ac.uk/year2/schizophrenia/mann/ftr">http://www.psy.plym.ac.uk/year2/schizophrenia/mann/ftr</a>	5%
Internet	<a href="http://www.salmon.psy.plym.ac.uk/year2/schiz">http://www.salmon.psy.plym.ac.uk/year2/schiz</a>	4%
Internet	<a href="http://www.t2.net/schask/ssmanual7.html">http://www.t2.net/schask/ssmanual7.html</a>	4%
Internet	<a href="http://www.schoolsucks.com/te">http://www.schoolsucks.com/te</a>	1%
Internet	<a href="http://www.psy.plym.ac.uk/year2/schizophrenia/mann/ftr">http://www.psy.plym.ac.uk/year2/schizophrenia/mann/ftr</a>	1%
Internet	<a href="http://www.salmon.psy.plym.ac.uk/year2/schiz">http://www.salmon.psy.plym.ac.uk/year2/schiz</a>	1%
Proquest	<a href="#">Submitted to Art Center College of Design on 2002-11-03</a>	1%

Local intranet zone

The right side of the browser window displays text from the document being searched, including the following paragraphs:

plagiarists [had to find appropriate works from a limited pool of resources, usually a nearby library, and copy them by hand. Since these resources were almost always professionally written, the risk of detection was very high.](#)

The Internet now makes it easy to find thousands of relevant sources in seconds, and in the space of a few minutes plagiarists can find, copy, and paste together an entire term paper or essay. Because much of the material online is produced by other students, it is often difficult or impossible for educators to identify plagiarism based on expectations of student-level work.

[Even when an instructor does suspect plagiarism, the sheer size of the Internet seems to work in the plagiarist's favor. Search engines can be useful for tracking down suspect passages, but even they have their limitations, given the number, variety, and password-protected nature of many websites. Even where search engines do prove useful, manually searching the Internet for matches of hundreds of student papers can be a formidable task.](#)

Additionally, the seemingly "public" nature of online content blurs the distinction between publicly and privately owned information. Electronic resources, by nature easily reproducible, are not perceived as "intellectual property" in the same way that their material counterparts are. Just as file transfer programs such as Napster make it easy to trade copyrighted music files most people would never think to steal in physical form, [the Internet makes plagiarism easy for students who might have thought twice about copying from a book or published article.](#)

Perhaps the greatest resources for would-be plagiarists are the hundreds of online paper-mills, or "cheatsites", that exist solely for the purpose of providing students with quick-fix homework and term-paper solutions. Many of these services contain hundreds of thousands of papers on a wide variety of topics, and some even offer customized papers for an additional fee. The fact that many of these sites have become profitable ventures (complete with paid

# A sample from La Nazione (5 Feb 2002)

## A random article regarding Bill Clinton

author: Anon Anon title: clinton paper ID: 12905385 submitted: 03-19-04 4:11 PM PST

similarity index:  (99% matching text) version: # 1 (03-19-04)  [side-by-side version](#)

### Sources:

	link	match
Internet	<a href="http://lanazione.it/art/2002/02/04/2988053">http://lanazione.it/art/2002/02/04/2988053</a>	99%

### Report text:

La trappola  
"Uccidere Clinton al torneo di golf"  
Ecco il piano diabolico di Al Qaeda

NEW YORK, 5 FEBBRAIO 2002 - Bill Clinton ama il golf ma questa passione avrebbe potuto costargli molto cara: secondo documenti e video scoperti in Afghanistan, Al Qaida aveva progettato di ucciderlo ed uno degli scenari previsti era appunto quello di colpire mentre l'ex presidente partecipava a un torneo.

Le forze americane in Afghanistan hanno recentemente trovato un campo di addestramento degli uomini di Osama bin Laden in una località chiamata Shomali. E qui, secondo l'agenzia americana Upi, che avrebbero rinvenuto il materiale riguardante Bill Clinton.

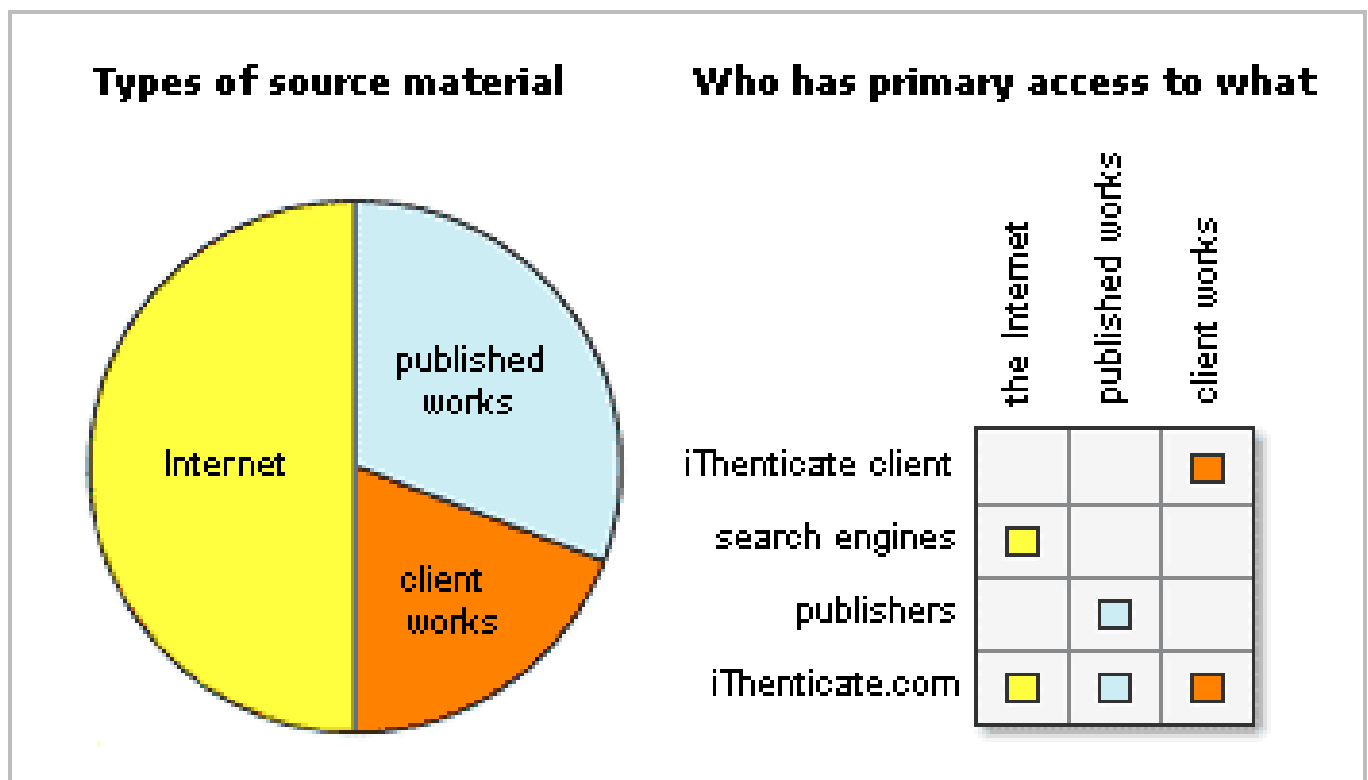
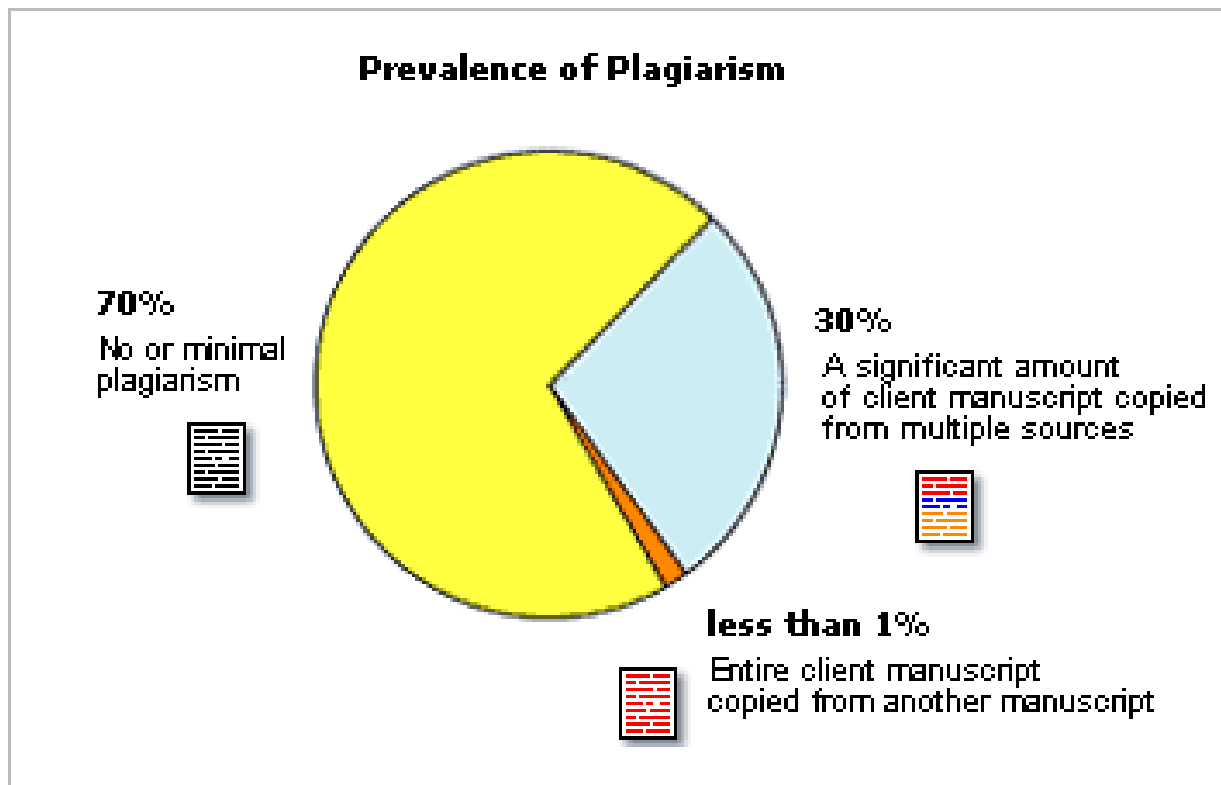
C'erano varie ipotesi allo studio per ucciderlo. Una di queste era appunto quella di organizzare un attentato su un campo da golf. Quando l'ex presidente giocava, secondo Al Qaida, il livello di protezione non era mai elevato.

In un video trovato nella base di Shomali, viene simulato appunto un attentato durante un torneo di golf, con un gruppo di terroristi che spara all'impazzata contro alcuni giocatori.

<Si tratta di un video che probabilmente faceva parte di un programma di addestramento per un attentato contro Clinton. Anche se in questo caso non vi è un riferimento specifico, tutti sanno che a Clinton piace il golf e che nei fine settimana spesso andava a giocare>, ha detto alla Upi Keith Indema, un consulente americano del governo provvisorio afgano.

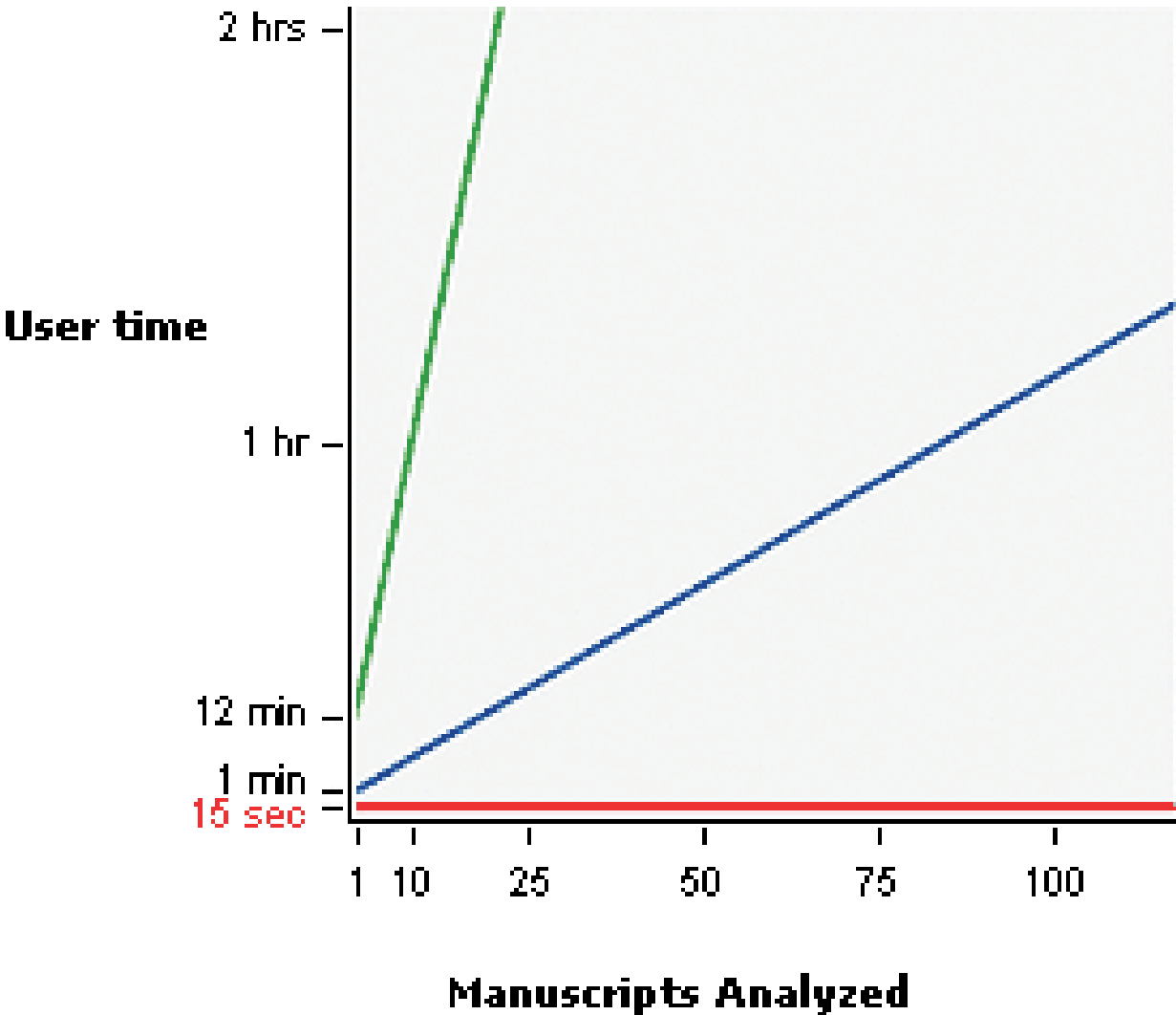
Secondo l'agenzia, in alcuni dei documenti trovati a Shomali vi sono vari riferimenti diretti all'ex presidente. <Abbiamo cercato di uccidere Clinton ma non ci siamo riusciti>, si legge ad esempio su un quaderno scritto in arabo. Altre annotazioni, sempre in arabo, elencano i vari sistemi usati dai servizi segreti americani per garantire l'incolumità del presidente. Vengono esaminati anche i loro punti deboli.

# Why not just use Google to search for source material?



# Searching the Internet by hand is a massive time investment

User search time



- search time using search engine (multi-source plagiarism papers)
- search time using search engine (single-source plagiarism papers)
- search time using **iThenticate.com** (ALL papers)

## **Final Thought**

- **We have always viewed Turnitin as a tool (not a complete solution) to stop academic plagiarism. Similarly, we are interested in working with colleagues such as yourselves to see how iThenticate can contribute to improvements in the best practices of publishing. We see iThenticate as a tool for pre-publishing due diligence and as a tool to detect post-publishing intellectual property theft.**

**Thank you.**





# Intellectual Property (IP) — Copyright

The purpose of Copyright Law is “to promote the Progress of Science and useful Arts....”

*U.S. Constitution, Art. I § 8, cl. 8.*

“... the aim of copyright is to give an author an exclusive right sufficient to create an incentive to produce, but not so great a right as to undermine the public domain.” These rights must be for a limited term and they must “promote the progress of science.”

Lawrence Lessig, *The Future of Ideas* (New York: Random House, 2001), 98.

# Copyright Law — Fair Use

“the fair use of a copyrighted work ... for purposes such as criticism, comment, news reporting, teaching, scholarship, or research, is not an infringement of copyright.”

*17 U.S.C.A. § 107.*

## Copyright Law — Legal Opinion

“... the Company’s activity may be fairly characterized as ‘criticism’ as that term is used in the preamble of Section 107 of the Copyright Act. According to Webster’s Dictionary (2d Ed., 1996), criticism means ‘fault finding or censure’ or ‘the act of judging the merits of something.’ By this definition, the Company, by investigating the integrity of an author’s work, is engaged in a form of criticism: the Company is judging the merits of the author’s work. As a result, we conclude that the Company’s ‘criticism’ of written works constitutes the type of activity that the courts have traditionally characterized, and the legislature has recognized, as ‘fair use’ of copyrighted material. Our opinion is further supported by a closer look at whether the Company has sufficiently transformed the original author’s work. Certainly, the Company’s use involves a complete transformation of the raw material when the fingerprint is created. Further, the purpose of the Company’s fingerprint creation and analysis is to identify potential plagiarism, which has absolutely nothing to do with the purpose of the original work.”

Foley & Lardner, *Legal Opinion*, 2003.

## Copyright Law — Legal Opinion

“...the Company’s [originality] report, by identifying potential plagiarists, provides new insights and understandings about the original. We believe that the identification of plagiarists is the type of activity that the fair use doctrine is intended to protect for the ‘enrichment of society.’”

Foley & Lardner, *Legal Opinion*, 2003.

## Copyright Law — Legal Opinion

“In short, for the same reason delineated above with regard to our plagiarism analysis, we are of the opinion that storing a copy of an author’s work in a database to be used solely for the purpose of comparing the work to other works constitutes a ‘fair use’ of the work. In view of the foregoing, it is our opinion that under Copyright Law, both types of use by the company of written works, despite the lack of express consent of the author, fall within the ambit of the fair use doctrine and, accordingly, the Company should not be liable for the claims of copyright infringement ... a court would find in favor of a defense of fair use.”

Foley & Lardner, *Legal Opinion*, 2003.

# Copyright Law — Example of Fair Use

- Leslie A. Kelly was a photographer who posted his pictures on a website
- Arriba owned a search engine for images
  - They copied ALL of Kelly's pictures without permission
  - They copied the ENTIRE picture
  - They transformed the pictures into thumbnails
  - They stored the thumbnails in a database
  - They are a commercial venture and profited from their service
  - They did not harm the market value of Kelly's work
- **The Court found that Arriba was making a fair use of Kelly's pictures.**

*Kelly v. Arriba Soft Corp., 9th Cir., No. 00-55521,  
2/6/02*