Orphan Works As Grist For The Data Mill

Matthew Sag

Associate Professor, Loyola University Chicago School of Law

Paper available available at http://ssrn.com/abstract=2038889).

Three Faces of Library Digitization

- **Preservation**
- **Data production and analysis**
  - Searching books, testing search algorithms, computational linguistics, automated translation, natural language processing, macro-analysis of text
- A platform for **display and distribution** of individual works

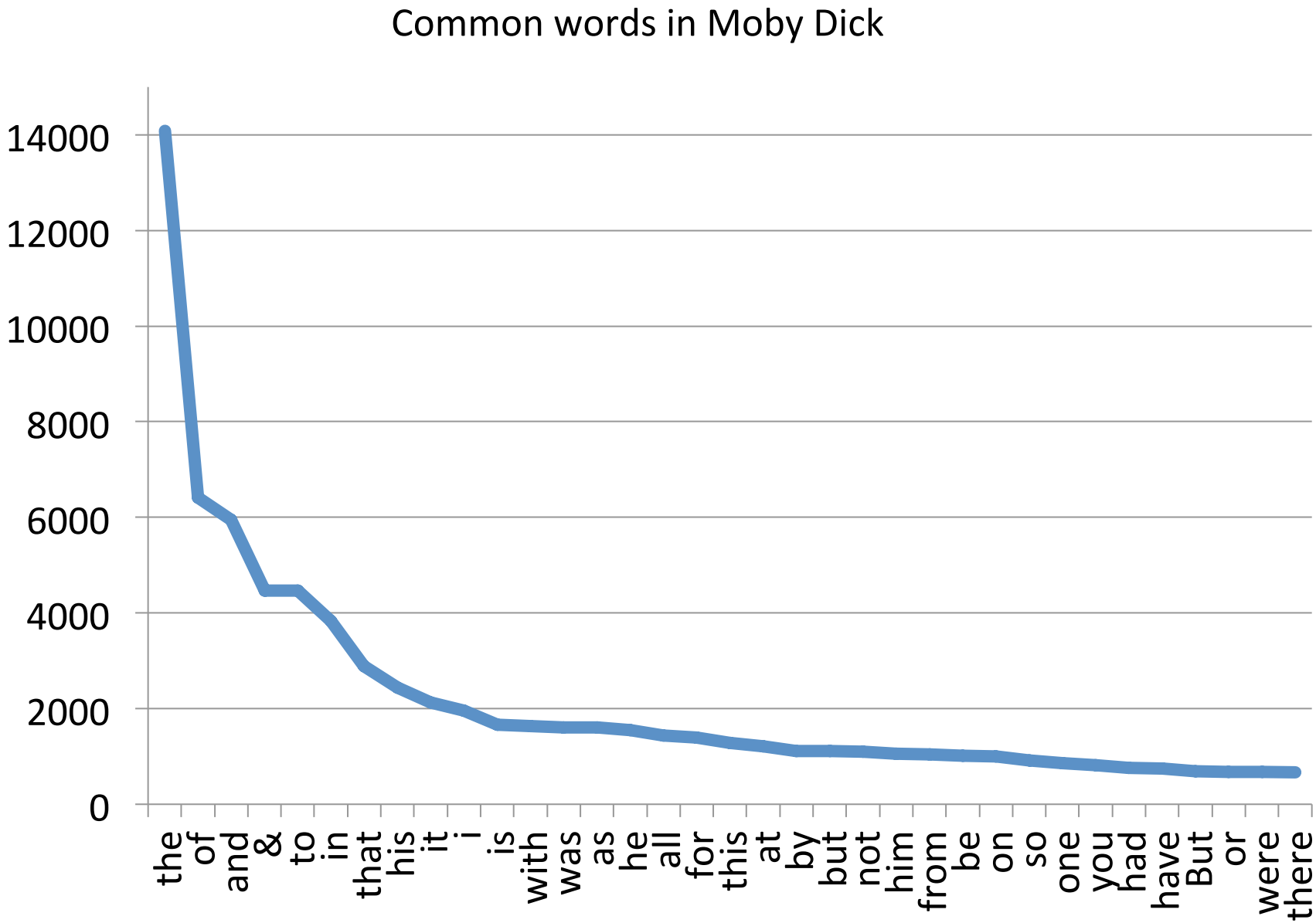
Library digitization and orphan works

- Key Question:
  - Does copying for a ~~non-consumptive~~ **nonexpressive** use implicate the rights of the copyright owner?

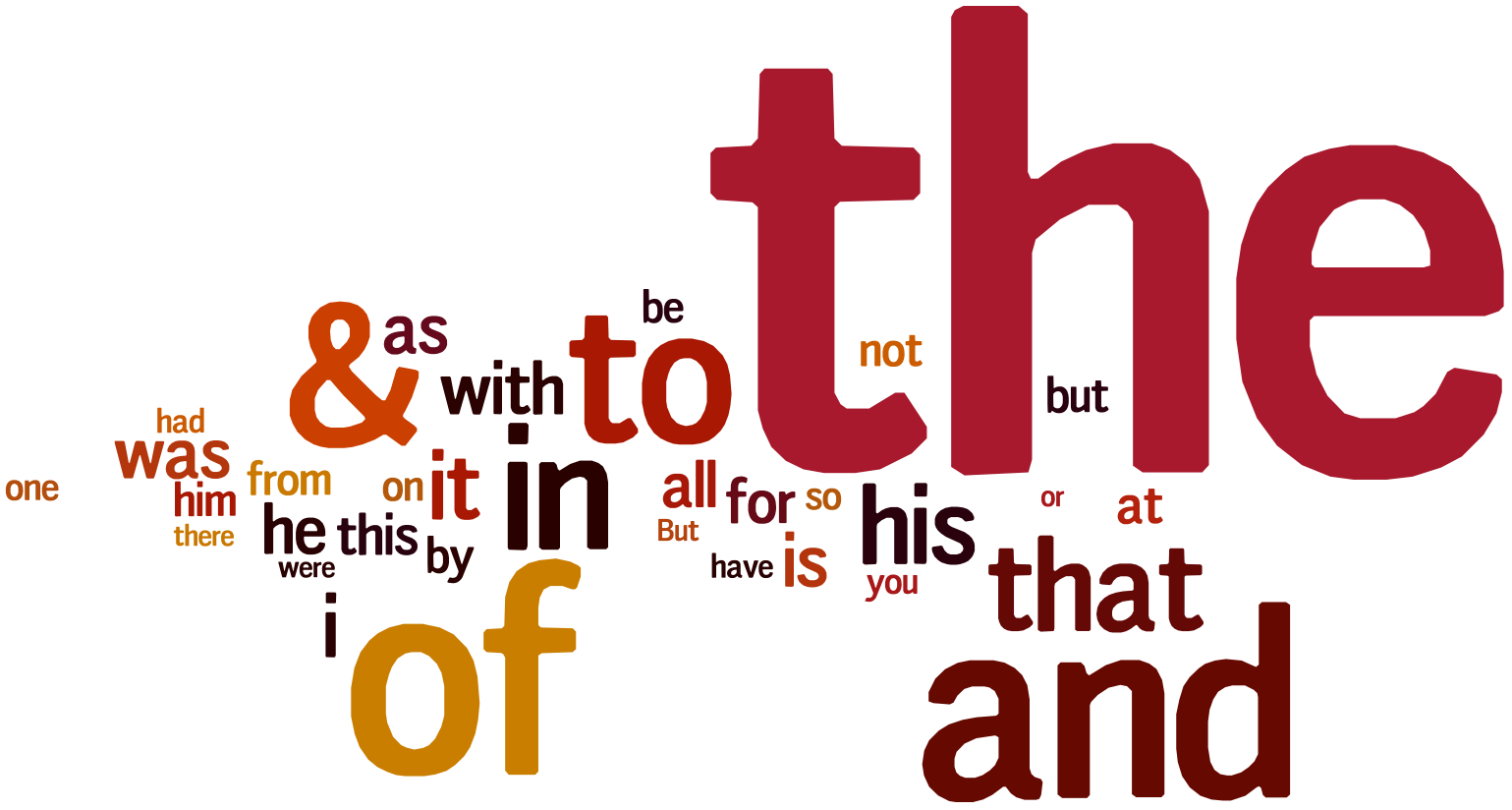
- Note:
  - Orphan works explains why we care, but the orphan status of these works is not directly relevant to the primary question.

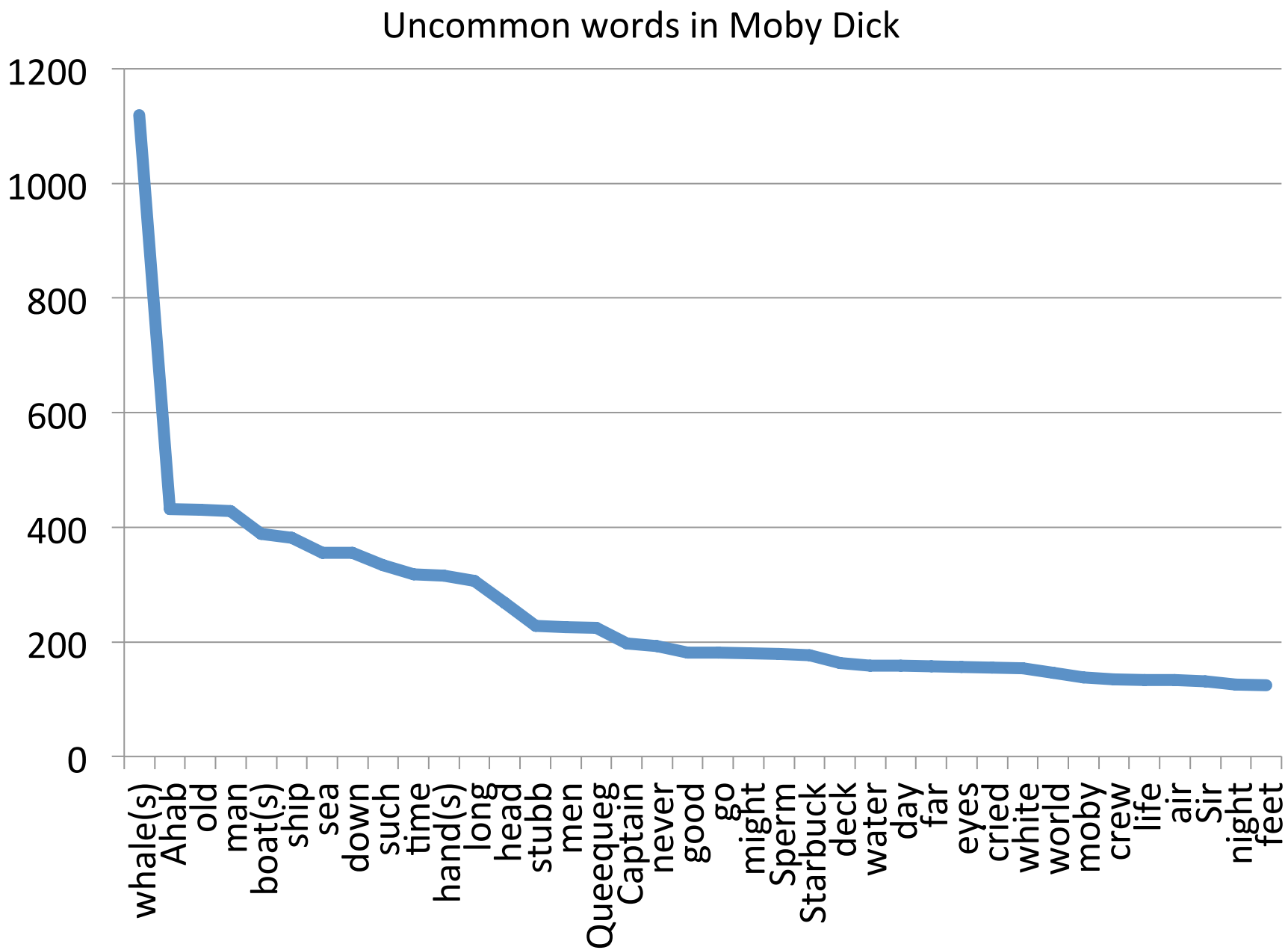
Thought Experiment

- Brian is a savant with total recall
- Moby Dick has its copyright restored
  - (Perpetual Copyright Act of 2014??)
- Brian produces a frequency table


Common words in Moby Dick
14000
12000
10000
8000
6000
4000
2000
0
the
of
and
&
to
in
that
his
it
i
is
with
was
as
he
all
for
this
at
by
but
not
him
from
be
on
so
one
you
had
have
But
or
were
there

Common words in Moby Dick



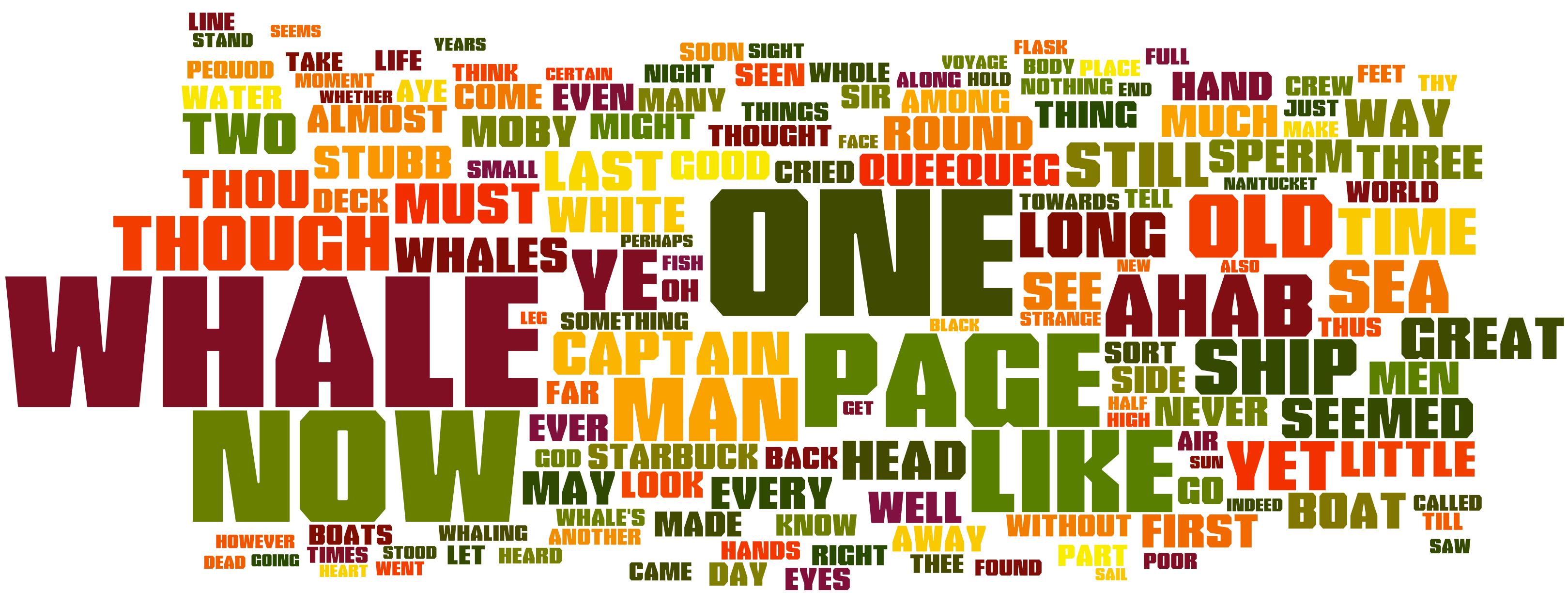

Uncommon words in Moby Dick
1200
1000
800
600
400
200
0
whale(s)
Ahab
old
man
boat(s)
ship
sea
down
such
time
hand(s)
long
head
stubb
men
Queequeg
Captain
never
good
go
might
Sperm
Starbuck
deck
water
day
far
eyes
cried
white
world
moby
crew
life
air
Sir
night
feet

Uncommon words in Moby Dick

Meta Data – a restatement of the obvious

- **Meta data** (even if its valuable) **does not infringe the rights of the copyright owner.**
  - Idea-expression distinction
  - Merger
  - Substantially similarity –>
  - Originality –> –>

Substantially Similarity



Substantially Similarity

Call me Ishmael. Some years ago - never mind how long precisely - having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen, and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off - then, I account it high time to get to sea as soon as I can. This is my substitute for pistol and ball. With a philosophical flourish Cato throws himself upon his sword; I quietly take to the ship. There is nothing surprising in this. If they but knew it, almost all men in their degree, some time or other, cherish very nearly the same feelings towards the ocean with me. There now is your insular city of the Manhattoes, belted round by wharves as Indian isles by coral reefs - commerce surrounds it with her surf. Right and left, the streets take you waterward. Its extreme down-town is the battery, where that noble mole is washed by waves, and cooled by breezes, which a few hours previous were out of sight of land. Look at the crowds of water-gazers there.

Circumambulate the city of a dreamy Sabbath afternoon. Go from Corlears Hook to Coenties Slip, and from thence, by Whitehall northward. What do you see? - Posted like silent sentinels all around the town, stand thousands upon thousands of mortal men fixed in ocean reveries. Some leaning against the spiles; some seated upon the pier-heads; some looking over the bulwarks of ships from China; some high aloft in the rigging, as if striving to get a still better seaward peep. But these are all landsmen; of week days pent up in lath and plaster - tied to counters, nailed to benches, clinched to desks. How then is this? Are the green fields gone? What do they here?

But look! here come more crowds, pacing straight for the water, and seemingly bound for a dive. Strange! Nothing will content them but the extremest limit of the land; loitering under the shady lee of yonder warehouses will not suffice. No. They must get just as nigh the water as they possibly can without falling in. And there they stand - miles of them - leagues. Inlanders all, they come from lanes and alleys, streets and avenues, - north, east, south, and west. Yet here they all unite. Tell me, does the magnetic virtue of the needles of the compasses of all those ships attract them thither?

Originality

[1] **"Goblin-made armour does not require cleaning, simple girl. Goblins' silver repels mundane dirt, imbibing only that which strengthens it."** (J.K. Rowling, Deathly Hallows)

[2] **"… goblin-made armor does not require cleaning, because goblins' silver repels mundane dirt, imbibing only that which strengthens it, such as basilisk venom."** (Harry Potter Lexicon)

[3] **Other than 'Goblin', none of the words in [1] are repeated.** (Matthew Sag)

[4] **There is a high level of similarity between [1] and [2]**(anti-plagiarism software)

Producing Meta Data – Not quite so obvious

- **Hard to argue that a reading machine (e.g. Google Book Search) does not 'reproduce the work' in a 'copy', even if no one reads it.**

- **The distinction between expressive and nonexpressive *works* is well recognized. The same distinction should generally be made in relation to potential *acts* of infringement.**
  - Copying for purely nonexpressive purposes, such as the automated extraction of data, should not be regarded as infringing.

Statutory rights of the author are limited to the communication of original expression to the public

- Consider
  - Threshold of substantial similarity is defined in reference to the perspective of the ordinary observer (with some filtering of facts, ideas, etc.).
  - Intermediate copying does not infringe (screen-play cases), is fair use (reverse engineering cases)

Implications

- **Automated reproduction for nonexpressive uses** (such as search engines, plagiarism detection, and macro-literary analysis) **does not communicate the author's original expression to the public**
  - No expressive substitution, no infringement

Application to Fair Use

- **(1) purpose and character:** Like transformative uses, a nonexpressive use poses no risk of expressive substitution
- (2) nature of the work … "not much use"
- **(3) Amount and Substantiality:** Like transformative uses, because there is no expressive substitution in a nonexpressive use, the amount of copying is qualitatively insignificant.
- **(4) Market effect:** Like transformative uses, a nonexpressive use poses no risk of expressive substitution, thus no cognizable market effect.

In Summary


copyright
nonexpressive
expression
library
use
original
works
work
digitization
Copyright
fair
transformative
public
rights
expressive
copying
law
uses
search
protection
distinction
software
courts
Google
owner's
computer
cases
author's
Act
also
ideas
substitution
book
must
Dick
many
held
three
even
analysis
author
Section
case
made
communication
Nonexpressive
ASA
infringing
article
court
way
making
Library
third
reproduction
available
digital
authors
orphan
market
facts
may
text
libraries
first
make
used
fact
generally
functional
similarity
creation
limited
question
purposes
display
merely
new
words
intermediate
objects
engines
simply
Even
general
relation
character
engineering
reading
particular
apply
information
factor
engine
Guild
parody
different
certain
reproduce
protected
context
infringement
whether
books
data
exclusive
Use
distribution
example
idea-expression
right
issue
subject
potential
Courts
access
nature
much
substantial
Internet
word
copy
significant
automated
extent
principle
issues
copies
value
process
entire
purpose
however
frequency
Sony
owners
Moby
matter
doctrine
effect
clearly
author's
uses
one
copyrighted
owner
substitute
based
interest
Court
Figure
elements