Definitions, Formulas, and Simulated Examples for Plagiarism Detection with FAIR Metrics

Adam Craig, Adarsh Ambati, Shiladitya Dutta,
Arush Mehrotra, S. Koby Taswell, and Carl Taswell

Brain Health Alliance, Ladera Ranch, California, USA

ASIS&T 82nd Annual Meeting, Melbourne, Australia
20 October 2019

Outline

- Part I: Problem — Plagiarism
- Part II: Solution — FAIR Metrics
- Part III: Application — PORTAL-DOORS Project

Part I: Problem — Plagiarism

Ethics, Integrity, and Character Matter!

- “Many people say that it is the intellect which makes a great scientist. They are wrong: it is character.” – Albert Einstein
- Quoted in “Integrity in Scientific Research: Creating an Environment That Promotes Responsible Conduct”, 2002, US National Academies Press, available at www.ncbi.nlm.nih.gov/books/NBK208712
- For more information and resources on ethics and ethical standards in scholarly research and publishing, refer to the work of COPE, the Committee on Publishing Ethics, at publicationethics.org

Questions in Search of Answers

- Science in politics? And politics in science?
- How to differentiate Information, Misinformation, Disinformation?
- How to know and appreciate the difference between *fake news* and *real news*? What about the problem of *deep fakes*? Truth, lies and misleading deception? Scientific fraud?
- What constitutes *plagiarism of ideas* in scholarly research and publishing? How should *censorship of ideas* be defined?
- **Can we develop metrics, algorithms and software tools to detect and prevent both plagiarism and censorship while also promoting good citation practices?**

A Turn to the Dark Side: Plagiarism

- Plagiarism worsening in scale and scope: Now more sophisticated, defiant of the COPE standards, and brazenly public out in the open
- New forms of plagiarism: especially new kinds of plagiarism of ideas with *idea laundering by exclusive cliques* in scholarly publishing analogous to *money laundering by mafia gangs* in illegal commerce
- Plagiarism further worsened by intentional refusal to cite, and persistent refusal to correct (a benefit-of-the-doubt assumed unintentional?) omission of citation, where both refusals constitute explicit violations of the COPE standards of ethical publishing, and also represent admission of the *consciousness of guilt of plagiarism*
- Many definitions of plagiarism for which two essential criteria exist: 1) the theft of ideas, creative content and/or intellectual property, and 2) the misrepresentation of novel authorship falsely claimed by the thieves who refuse to cite the original authors of the work with acknowledgment of the true creators who first published the content

A Turn to the Dark Side: Censorship

- Impact of *plagiarism* has been worsened by *censorship* of attempts to counter and correct the plagiarism
- Plagiarism supported by journal editors' censorship, ie, their refusal to publish commentary and letters to the editor for objective factual statements intended to inform the journal's readers about the original publications that were plagiarized
- Plagiarism supported by peer review censorship in which responsive papers countering the plagiarism are rejected with the false assertion of non-relevance to the conference or journal — even though an objective computerized analysis confirms explicit matching relevance
- Censorship defined here as an illogical contradictory interpretation of the advertised policies governing the manuscript submission process including the meanings of the key words and key phrases in the lists for the conference topics or journal scope

Return to the Light Side?

- How do we expose plagiarism and censorship?
- Repositories such as pubpeer.com offer a possible approach to publishing responses presumably not subject to censorship by peers of peers who wish to comment in response to published papers that have violated the COPE ethical standards
- As an example, see Taswell's brief commentary at pubpeer.com/publications/F0481960C5C5A98F9CB1FF108E11D0 informing readers about the original Taswell papers that were paraphrased without citing by the Wilkinson et al authors of the FAIR principles published in Nature Scientific Data

Part II: Solution — FAIR Metrics

Questions in Search of Answers

- **Can we develop metrics, algorithms and software tools to detect and prevent both plagiarism and censorship while also promoting good citation practices?**

Concepts for Our FAIR Metrics

- Simple 2x2 table analysis of statements
- Claim can be old or new; claim can also be valid or invalid
- Results in 4 kinds of claims: Quoted, Misquoted, Novel, Plagiarized

Claim	Valid	Invalid
Old	Quoted	Misquoted
New	Novel	Plagiarized

- An *invalid new claim* may exist in the presence of failure to search the literature, paraphrasing without citing (and with persistent refusal to correct omission of citation), plagiarism of ideas, or verbatim plagiarism of words and images
- Is failure to search, refusal to cite, and/or refusal to correct omission of citation acceptable in the current era of COPE standards with computerized search of internet-accessible, web-enabled databases?

Notation for Our FAIR Metrics

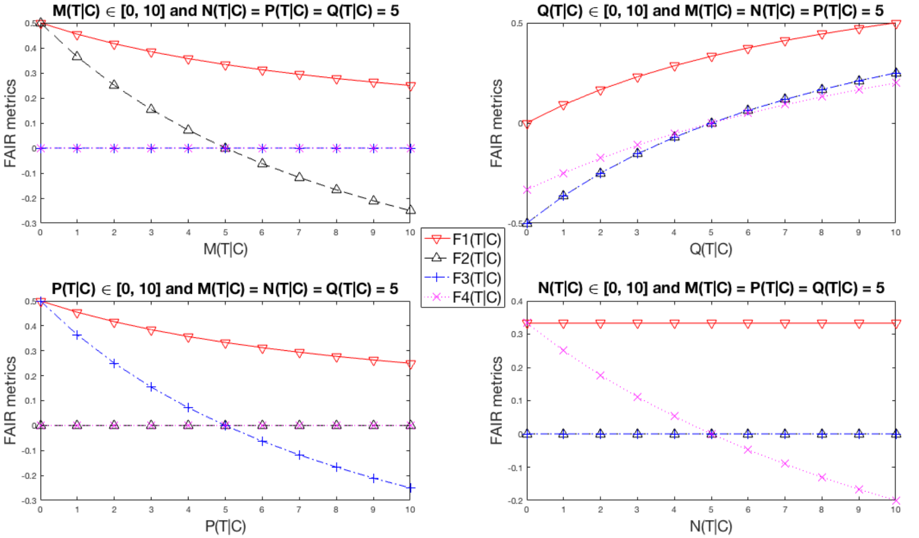
Symbol	Definition
C	set C of statements in a Control paper
$G(A)$	function G operates on set A
$G(A \mid B)$	function G operates on set A given set B)
$M(T \mid C)$	number M of Misquoted (incorrectly cited) statements
$N(T \mid C)$	number N of Novel (uncited) statements
$K(C)$	number K of Known statements
$P(T \mid C)$	number P of Plagiarized (uncited) statements
$Q(T \mid C)$	number Q of Quoted (correctly cited) statements
$R(T \mid C)$	number R of Reported statements
$S(T \mid C)$	number S of Similar statements
T	set T of statements in a Test paper

Formulas for Our FAIR Metrics

Symbol		Formula
$F_1(T\|C)$	$=$	$Q(T\|C)/S(T\|C)$
$F_2(T\|C)$	$=$	$[Q(T\|C) - M(T\|C)]/S(T\|C)$
$F_3(T\|C)$	$=$	$[Q(T\|C) - P(T\|C)]/S(T\|C)$
$F_4(T\|C)$	$=$	$[Q(T\|C) - N(T\|C)]/R(T\|C)$
$S(T\|C)$	$=$	$M(T\|C) + Q(T\|C) + P(T\|C) \leq K(C)$
$R(T\|C)$	$=$	$M(T\|C) + Q(T\|C) + P(T\|C) + N(T\|C) \geq K(C)$

Figure: Formulas for FAIR metrics with condition $0 < S(T|C) \leq K(C) \leq R(T|C)$

Simulated Examples of FAIR Metrics



Current AI-Based Method for Plagiarism Detection

- Using NLP, extraction of RDF triples corresponding to most relevant statements from both the Test T paper and the Control C comparison collection of papers
- Using ML, classification of RDF triples from T in comparison with C for the 4 categories of statements as either Quoted Q, Misquoted M, Novel N, or Plagiarized P
- Automated tally of counts Q, M, N, and P corresponding to the statement counts for Test in comparison with Control
- Automated calculation of FAIR metrics F_1, F_2, F_3, and F_4
- Current work in progress to automate this entire AI-based process for estimating values of the FAIR metrics intended for use with the promotion of fair citation and the detection/prevention of plagiarism

Future AI-Based Approaches for Plagiarism Detection

- Compare performance of the automated AI-based approach with a human-expert-based approach for the analysis of the FAIR metrics
- Enhance the FAIR metric formulas with weighting factors derived from problem-oriented collections of literature for each community of authors publishing in a particular field of scholarly research
- Account for commonality of author citations in reference lists of published articles
- Account for commonality of author attendance at conferences inferred from lists of authors in published conference proceedings
- Validate both unweighted and weighted FAIR metrics on collections of articles known to be either plagiarizing or non-plagiarizing

Use of Acronym ‘FAIR’ and Words ‘Fair’ and ‘Metrics’

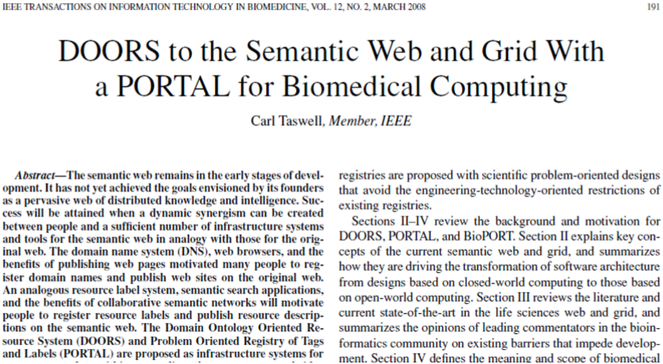
- FAIR principles of Wilkinson et al with acronym ‘FAIR’ for the principles called *Findable, Accessible, Interoperable, Reproducible* are a subset of the PDP and NPDS principles from the PORTAL-DOORS Project paraphrased by Wilkinson et al without citing Taswell
- FAIR metrics of Wilkinson et al are used with the word ‘metrics’ in a manner that is not consistent with its usage in most fields of science
- FAIR metrics of Craig et al are used with acronym ‘FAIR’ for *Fair Acknowledgment of Information Records and Fair Attribution to Indexed Reports* and the word ‘metrics’ in a manner consistent with both the meaning of the word ‘fair’ because it is a recursive acronym, and usage of the word ‘metrics’ with its meaning as a quantitative numerical value for the measure of something

Part III: Application — PORTAL-DOORS Project

DREAM Principles and FAIR Metrics

- In response to the paraphrasing without citing by Wilkinson et al of the Taswell papers from the PORTAL-DOORS Project (PDP), we have created a new name with summarizing phrase for the PDP software design principles and new quantitative analytic methods to evaluate papers for the presence of plagiarism
- DREAM principles with acronym DREAM for *Discoverable Data with Reproducible Results for Equivalent Entities with Accessible Attributes and Manageable Metadata*
- FAIR metrics with acronym FAIR for *Fair Acknowledgment of Information Records and Fair Attribution to Indexed Reports*

Blueprint for the PORTAL-DOORS Project



DOORS to the Semantic Web and Grid With a PORTAL for Biomedical Computing

Carl Taswell, *Member, IEEE*

***Abstract*—The semantic web remains in the early stages of development. It has not yet achieved the goals envisioned by its founders as a pervasive web of distributed knowledge and intelligence. Success will be attained when a dynamic synergism can be created between people and a sufficient number of infrastructure systems and tools for the semantic web in analogy with those for the original web. The domain name system (DNS), web browsers, and the benefits of publishing web pages motivated many people to register domain names and publish web sites on the original web. An analogous resource label system, semantic search applications, and the benefits of collaborative semantic networks will motivate people to register resource labels and publish resource descriptions on the semantic web. The Domain Ontology Oriented Resource System (DOORS) and Problem Oriented Registry of Tags and Labels (PORTAL) are proposed as infrastructure systems for**

registries are proposed with scientific problem-oriented designs that avoid the engineering-technology-oriented restrictions of existing registries.

Sections II–IV review the background and motivation for DOORS, PORTAL, and BioPORT. Section II explains key concepts of the current semantic web and grid, and summarizes how they are driving the transformation of software architecture from designs based on closed-world computing to those based on open-world computing. Section III reviews the literature and current state-of-the-art in the life sciences web and grid, and summarizes the opinions of leading commentators in the bioinformatics community on existing barriers that impede development. Section IV defines the meaning and scope of biomedical

Manuscript received 10/31/2006; online 8/3/2007; print 3/5/2008

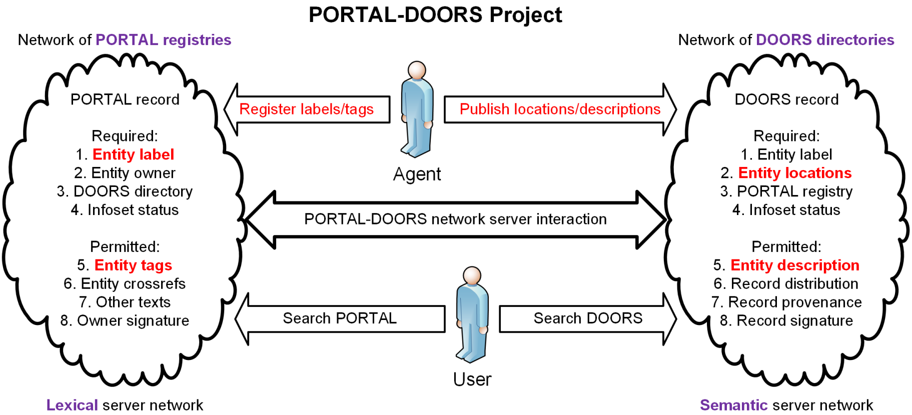
Semantic versus Lexical Information Systems

- A lexical (“dumb”) system is an information system in which words are processed as character strings that have no meaning to the processing agent
- A semantic (“smart”) system is one in which words have defined meaning to the agent processing them with logic-based reasoners
- Semantic search may be efficient, while lexical search inefficient, for the given search task:
    - How many hits returned in response to the search query?
    - Are there too many hits for a person to review?
    - Or if just a few hits, are they relevant?
    - Do the returned hits answer the search question directly or indirectly?
- Semantic information systems can be built with the XML, RDF, OWL, SPARQL stack of technologies for describing and querying resources

PORTAL-DOORS compared to IRIS-DNS

- PORTAL-DOORS for the semantic web has been designed in a manner analogous to IRIS-DNS for the lexical web
- PORTAL (Problem-Oriented Registry of Tags And Labels) is an analogue of IRIS for naming and registering domains
- DOORS (Domain-Ontology Oriented Resource System) is an analogue of DNS for addressing and locating domains
- Using an analogous paradigm with labeled resources instead of named domains, PORTAL-DOORS designed to do for the semantic web what IRIS-DNS does for the lexical web
- PORTAL-DOORS built as a who-what-where diristry, registry, directory network system for identifying, describing, locating and linking things on the internet, web and grid

Schema Design for PORTAL-DOORS



PORTAL-DOORS Project

Network of **PORTAL registries**

PORTAL record

Required:
1. **Entity label**
2. Entity owner
3. DOORS directory
4. Infoset status

Permitted:
5. **Entity tags**
6. Entity crossrefs
7. Other texts
8. Owner signature

Lexical server network

Register labels/tags

Agent

Publish locations/descriptions

PORTAL-DOORS network server interaction

Search PORTAL

User

Search DOORS

Network of **DOORS directories**

DOORS record

Required:
1. Entity label
2. **Entity locations**
3. PORTAL registry
4. Infoset status

Permitted:
5. **Entity description**
6. Record distribution
7. Record provenance
8. Record signature

Semantic server network

Essence of the PORTAL-DOORS Project

- PORTAL-DOORS for the semantic web modeled on the success of IRIS-DNS for the original lexical web
- PORTAL-DOORS designed to address diverse problems: information tsunami (find the needle in the haystack), informatics tower of babel (harmonization for interoperability), cybersilos in scientific discourse, search engine consolidation with monopolies, lexical to semantic transition barriers, fake news in social media and fraud in science
- PORTAL-DOORS operates as a distributed diristry, registry, directory network system of metadata and data repositories
- Semantic search tools and applications to support
  - Translational medical research for drug development, precision medicine, pharmacogenomic molecular imaging, and complex information systems to study gene-brain-behavior relationships
  - Automated meta-analyses of published literature for synthesis of confirmatory and/or contradictory results from clinical trials

Paraphrasing without Citing

- Wilkinson et al 2016 Nature Scientific Data "FAIR guiding principles for scientific data management and stewardship"
- Compared with Taswell 2008 IEEE TITB and Taswell 2010 Future Internet papers on the PORTAL-DOORS Project
- Item-by-item comparison and analysis did not find any novel idea or concept in Wilkinson et al "FAIR principles" that can be described as fundamentally new and/or different from the content, principles, analysis, and discussion of the PORTAL-DOORS Project by Taswell
- All scores tallied by different analysts as human experts with the Craig et al FAIR metrics on the Wilkinson et al FAIR principles paper resulted in values suspicious for absence of fairness
- Recall that the Craig et al FAIR metrics family F_1, F_2, F_3, and F_4 are all defined such that increasing values correspond to increasing fairness, and decreasing values correspond to alerts for possible absence of the different kinds of fairness

Conclusion

- Our PDP and NPDS principles originally published as the foundation for the PORTAL-DOORS Project have been renamed the DREAM principles by us in response to the Wilkinson et al co-authors who unfairly renamed them the FAIR principles.
- Our FAIR metrics, supported by NLP and AI, have been designed to monitor adherence to fair standards of citation in scholarly research and publishing, and to detect and help prevent plagiarism.
- Social engineering, with appropriate incentives and disincentives, remains as important as software engineering for a solution to the continuing problem of plagiarism.
- Consistent non-contradictory use of acronyms with the meanings of the words implied by the acronyms will help address some of the social engineering aspects of the plagiarism problem.

For More Information

- Tom Lehrer 1953 Lobachevsky (with lyrics)
  www.youtube.com/watch?v=gXlfXirQF3A
- DREAM Principles and FAIR Metrics from the PORTAL-DOORS Project for the Semantic Web – *Presented June 2019, 11th IEEE ECAI*
  portaldoors.org/pub/docs/ECAI2019DREAMFAIR0612.pdf
- Managing Scientific Literature with Software from the PORTAL-DOORS Project – *Presented September 2019, 15th IEEE eScience*
  portaldoors.org/pub/docs/BCDC2019PdpDemo0817.pdf
- Definitions, Formulas, and Simulated Examples for Plagiarism Detection with FAIR Metrics – *Presented October 2019, 82nd ASIS&T*
  portaldoors.org/pub/docs/ASIST2019FairMetrics0611.pdf
- DREAM Principles from the PORTAL-DOORS Project and NPDS Cyberinfrastructure – *Submitted, under peer review*
- PDP software demo video available at
  portaldoors.org/pub/mp4/PdpDemoVideo20190924.zip

Contact Info

- www.PORTALDOORS.org
- www.BrainHealthAlliance.org
- ctaswell@BrainHealthAlliance.org
- We welcome collaborators interested in promoting ethics and preventing plagiarism.