Who Wrote This?
Creator Metadata Quality on Academia.Edu

by

Zachary Schoenberger

A thesis submitted in partial fulfillment of the requirements for the degrees of

Master of Arts in Humanities Computing
and
Master of Library and Information Studies

Humanities Computing/Library and Information Studies
University of Alberta

Abstract

Academic social networking services (SNSs) such as ResearchGate.com or Academia.Edu have recently experienced a surge in popularity (Ortega, 2016). Existing research into academic SNSs have focused on population parameters and social networking usage patterns. Currently, no research has been conducted on the quality of bibliographic metadata on academic SNSs. Bibliographic metadata functions to support user tasks, including finding, identifying, selecting, and obtaining information resources. “Creator” metadata, which describes resource authorship, helps users find and identify digital works in a repository. Additionally, academic researchers rely on author attribution for their professional promotion and prestige, and they are accustomed to scholarly environments which implement standards that support accurate author attribution. This study therefore examines “creator” metadata for University of Alberta publications posted on Academia.Edu, and compares these with publisher created records of the same titles. Metadata quality is assessed through the measurement of completeness, consistency, and accuracy. The study reveals that Academia.Edu “creator” metadata is significantly incomplete compared to publisher metadata, and the frequency of incomplete records increases in proportion to the size of the author cohort. This incompleteness is evidence of poor metadata quality on Academia.Edu. Academia.Edu “creator” metadata is, however, much more consistent than publisher metadata. Finally, accuracy is found to be an inadequate determiner of metadata quality, as the presence of user generated metadata calls into question the conceptual stability of “authenticity” and “authority,” upon which a measure of accuracy depends. This study of metadata quality therefore reveals the complexity and contradiction that underlies this topic. In terms of completeness, Academia.Edu metadata is poor in quality. In terms of consistency, Academia.Edu metadata excels in quality. Finally, the study recommends further investigation into the definition of authority in relation to user-contributed metadata.

Acknowledgement

Thank you to my supervisory committee – Dr. Ali Shiri, Dr. Maureen Engel, and Dr. Tami Oliphant – for discussing, refining, and supporting my ideas throughout this process. Thank you to my parents for their support and guidance throughout my life. To my spouse and best friend, Keren, thank you for your patience and endurance throughout this process, for your never-failing support in all things, for being an inspiration, and, most of all, for your love.

Table of Contents

1. Tables and Figures

1.1 List of Tables

1.2 List of Figures

2. Introduction

In recent years, academic social networking services (SNSs) have exploded in popularity (Ortega, 2016). Websites such as ResearchGate.Com or Academia.Edu, and software such as Mendeley and Zotero, have leveraged the concept of a “social web”[1] to serve academic communities. Academics have demonstrated their affinity for these services; Academia.Edu, for example, boasted 11 million users in 2011 and 49 million in 2017 (Waybackmachine.org). According to a survey conducted by Van Noorden (2014), Academia.Edu is used for (in order of importance): 1) passively waiting for peers to contact them, 2) discover jobs, 3) actively discover peers, and, 4) posting works. Academic dissemination and collaboration is a social activity. Academics engage in social networking through traditional mechanisms such as conferences, workshops, special interest groups, and partnerships across departments and institutions; thus, online social networking services represent the virtual extension of this collaborative nature. Moreover, as a consequence of the distributed nature of online social networking, these services have supported otherwise unlikely international research collaborations, acting in a small way to counter the Western bias inherent in contemporary academia (Van Noorden, 2014). The growth and popularity of academic SNSs is indicative of their likelihood to persist – in some way or another – into the future.

Academic SNSs are different from disciplinary repositories such as Arxiv.Org or from institutional repositories such as University of Pennsylvania’s Scholar Sphere, as they are privately owned, run by individuals rather than collectives, and they piggyback on the popularity of the social networking phenomenon. Academic SNSs, institutional repositories, and disciplinary repositories are, however, similar in that they support dissemination, they traffic in research publications, and they cater

[1] “Social Web” references the use of the World Wide Web to facilitate peer to peer relationships through online social interaction. The social Web is part of “Web 2.0”, where user participation and interaction is at the centre of technological innovation (Halpin & Tuffield, 2010).

to an academic audience. To add trust to digital repositories, developers implement commonly followed standards. Users are likely not aware of the efforts service providers take to uphold these standards. Users do not, however, need to know, as they have come to trust digital library content and take for granted the underlying mechanisms that lead to this trust.

This thesis is concerned with metadata standards in particular. Popular websites such as Academia.Edu do not purport to use metadata standards and the quality of bibliographic metadata on these websites is presently unknown. The goal, therefore, is to examine the quality of "creator" metadata – a particularly important kind of metadata for discovery and identification of works in a scholarly repository environment – and compare this metadata to paired publisher metadata. The intended outcome is to help users of these services understand the trust that can be placed in the metadata they are provided, and to create a context for further research into the quality and trustworthiness of academic SNS metadata. Finally, websites academic SNSs make their resources discoverable and highly available on the web; poor quality metadata threatens the quality of information search and retrieval on search engines, including on the popular Google Scholar portal.

Current research in the area of academic SNSs has largely examined usage patterns and population characteristics of users (Almousa, 2011; Menendez et al. 2012; Thelwall & Kousha, 2014). Almousa (2011) conducted one of the earliest studies of Academia.Edu, comparing activity patterns in different academic ranks. Menendez et al. (2012) conducted a study similar to Almousa (2011), while expanding on the population size and number of measures used; findings were largely similar to Almousa. Thelwall and Kousha (2014) conducted a similar study, examining usage patterns on Academia.edu, including an examination of the Academia.Edu alt-metric score. Most recently, Ortega (2016) has synthesized research on academic SNS usage patterns. Researchers have also discussed the ethical dilemmas around using academic SNSs. Kathleen Fitzpatrick (2015), for example, argues that popular academic SNSs like Academia.Edu compete for users, thus detracting from more legitimate solutions such as institutional repositories. Meanwhile, no studies have examined metadata quality on academic SNSs.

Metadata quality in general is reasonably well studied, but Park (2009) considers the topic underexplored. Metadata quality correlates with the overall effectiveness of information environments, particularly information retrieval (National Information Standards Organization, 2007; Beall, 2006). Good metadata supports common bibliographic user tasks as defined by the Functional Requirements of Bibliographic Records (FRBR), including finding, identifying, selecting, and obtaining information objects. Metadata quality assessment involves the measurement of metadata values or metadata semantics, and prevailing criteria for measuring metadata quality attempt to take into account the functional role that metadata should play. Bruce and Hillmann (2004) developed the most widely accepted set of criteria for examining metadata quality. Of these criteria, Park (2009) determined the criteria of “completeness,” “consistency,” and “accuracy” to be the most commonly used criteria in quality assessment studies. These are the criteria used by this study for examining metadata on Academia.Edu as they are commonly used in quality assessments, they can be examined in the context of “creator” metadata, and past studies can help provide measures for assessing these criteria.

“Creator” is one of the most commonly implemented metadata elements in digital repositories, it belongs to nearly all bibliographic metadata standards, and it is notoriously difficult to implement, as personal names are inherently complex (Cwiok, 2010; Phelps, 2012; Windnagel, 2014). Moreover, the “creator” metadata element is used to indicate the author of a work and is therefore fundamental to the identification and discovery of digital objects. Professional prestige is achieved in large part through recognition and measurement of dissemination activities. Author attribution is therefore a critical part of academic systems of promotion and prestige. Most importantly, poor metadata on the web negatively impacts our ability to trust the web as a reliable source of information.

The thesis is divided into the following sections: 1. Tables and figures; 2. Introduction to the issues and argument; 3. A review of the current literature, thus situating this thesis within its intellectual context; 4. Statement of the study goals; 5. Statement of research questions; 6. Description of methods; 7. Presentation of results; 8. Discussion of findings; 9. Conclusion; 10. References; and, 11. Appendix.

3. Literature Review

3.1 Introduction

Access to academic scholarship rests on accurate and reliable author attribution, which in turn depends on consistent adherence to metadata quality standards (Chapman, 2009; Park, 2009). Researchers, however, have yet to examine “creator” metadata quality on popular academic social networking websites such as Academia.Edu. Academia.Edu allows researchers to post and share publications in an open repository and follow research publications within a public network of peers. Shared research can be indexed by search engines, cited by researchers, measured by analytic services, ingested by automated aggregators, consumed by citation management tools, or end up in a myriad of other places across the web. At the same time, automated processes can be sensitive to name ambiguity or poor quality “creator” metadata (Strotmann and Zhao, 2012; Torvik and Smalheiser, 2009). Finally, metadata exists to support user functions such as finding, identifying, selecting, and obtaining desired materials. Poor “creator” metadata affects the ability of users to perform these core tasks, as “creator” metadata represents a primary access point for resource discovery and use (National Information Standards Organization, 2007).

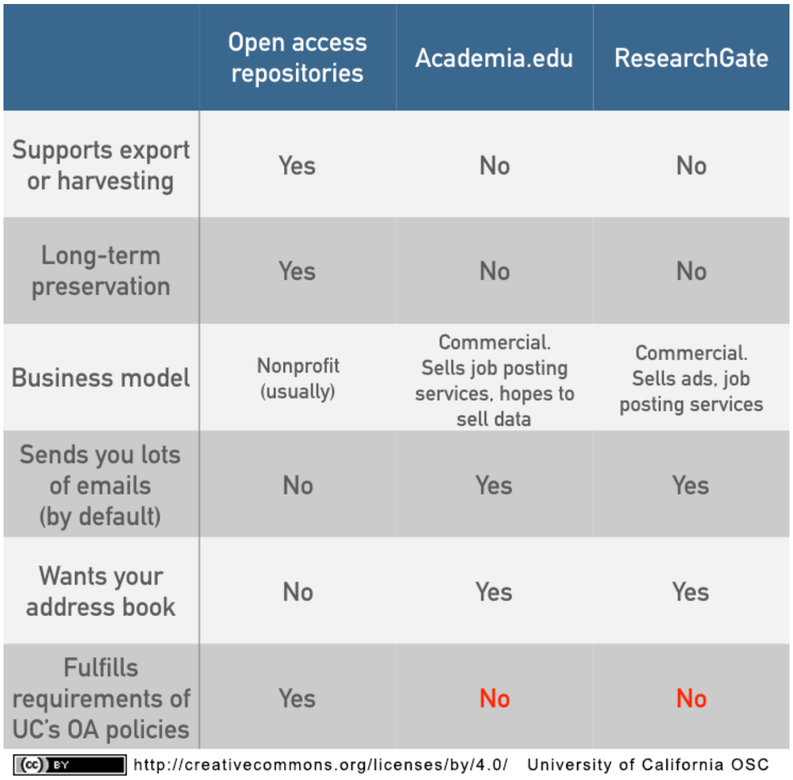
The following literature review addresses the intellectual context underlying this study. Section 1 presents a brief introduction to the context. Section 2 examines the increase in coauthored research and the corresponding rise of academic social networking services (SNSs). The benefits of addressing metadata quality on Academia.Edu are discussed briefly, followed by a look at recent criticism around academic SNSs. Finally, section 2 examines studies of usage patterns and the user population on Academia.Edu. Section 3 provides a broad introduction to the concept of metadata. Section 4 explores technologies and concepts in the world of metadata standards and authority control, and how these can be utilized to create better digital repositories through high quality metadata. Section 5 introduces the functional role of metadata for supporting user needs and their expectations. Section 6 considers “creator” metadata specifically and the functions it supports. Additionally, section 6 examines issues around personal name authority control, identity management solutions, and other tools that may be useful for supporting high quality “creator” metadata in a digital repository. Section 7 examines the current state of metadata quality assessment, focusing on definitions of metadata quality, the impact of metadata quality on digital information environments, and criteria for measuring quality focusing on metadata completeness, consistency, and accuracy. Finally, section 8 presents metadata quality assessment methods as demonstrated in past studies.

3.2 Overview of Academic Social Networking Services

Over the past few decades, research norms have progressively drifted away from single-authorship toward collaborative, co-authored publication. Wuchty, Jones and Uzzi (2007) state that a science and engineering collaborative publication is “currently 6.3 times more likely than a solo-authored paper to receive at least 1,000 citations” (para. 22). “Bluntly stated,” says Blaise Cronin, “teams trump soloists when it comes to scientific output and impact” (2012, p. 22). This shift toward co-author dominance indicates the importance of co-author attribution today. Moreover, it helps explain the recent rise in popularity of academic social networking services (SNSs) such as Academia.Edu, ResearchGate.com, Zenodo, or Mendeley, which can potentially enhance collaboration through social networking.

Academic SNSs are typically privately owned, for-profit social networking services catering to academic users. Academia.Edu was launched in 2008 and has enjoyed considerable success (Almousa, 2011). Academia.Edu is a web-based platform where academic users can post and discover research publications, and can engage with other users through following, recommending, bookmarking, or sharing publications. Broadly speaking, Academia.Edu is a combination of a professional social networking service (i.e. LinkedIn) and a digital scholarly repository (i.e. arxiv.org). Academia.Edu adds additional functionality through push notification services, crowd-sourced question and answer forums, job advertisements, and usage analytics. The "About" section of Academia.Edu claims "Academics use Academia.edu to share their research, monitor deep analytics around the impact of their research, and track the research of academics they follow" (Academia.Edu, 2017). In 2014, Academia.Edu made co-author tagging possible, stating "When your co-author uploads a paper and tags you, you immediately gain access to the analytics for that paper. Co-authors share the views, downloads, and bookmarks of works they've been added to" (The Academia.Edu Team, 2014). Analytics are therefore directly impacted by author metadata quality. Moreover, Academia.Edu caters to users whose livelihoods depend on recognition for research activities; Academia.Edu must therefore support quality "creator" metadata that conforms to these users' expectations.

The University of California San Diego library website has described differences between open access scholarly repositories and academic SNSs, contrasting standards of openness:



	Open access repositories	Academia.edu	ResearchGate
Supports export or harvesting	Yes	No	No
Long-term preservation	Yes	No	No
Business model	Nonprofit (usually)	Commercial. Sells job posting services, hopes to sell data	Commercial. Sells ads, job posting services
Sends you lots of emails (by default)	No	Yes	Yes
Wants your address book	No	Yes	Yes
Fulfills requirements of UC's OA policies	Yes	No	No

Figure 1. Open Access Comparison (University of California OSC, 2015)

Figure 1 presents a definition of “open access” according to the University of California Office of Scholarly Communications, drawing comparison between academic social networking services (specifically ResearchGate.com and Academia.Edu) and open access repositories. “Open access” services meet community standards for long-term preservation and automated harvesting of research materials; in contrast, academic SNSs, according to University of California, do not. Academic SNSs are typically for-profit, and use aggressive recruitment methods. Open source repositories are not-for-profit and are comparatively passive in user recruitment. Finally, open source repositories are designed to meet funding agreements requiring researchers to deposit government funded research in open access repositories. Quality metadata serves the needs of open access repositories by meeting the requirements of external harvesters such as Open Access Initiative (OAI) and long-term perseveration of materials. Metadata achieves these goals by following community-based – often internationally recognized – metadata standards. Academia.Edu is not an open access repository; specifically because of this, Academia.Edu is not required to follow metadata standards, nor does it purport to do so. Academia.Edu is chosen as an object of study because it does not follow standards, and yet, it is a massively popular service affecting the reputations of researchers and the quality of scholarly information on the web. Existing research into academic SNSs only describes user usage patterns and population dynamics, ignoring the importance of quality metadata in services oriented to academic researchers. Although researchers may already be aware that Academia.Edu has low quality metadata, the extent of this quality is not yet determined. Academic users have made Academia.Edu a popular service; with better knowledge of metadata quality, users will be better informed of the consequences of their participation in the service, and will be better equipped to proceed with caution.

In rebuke of Academia.Edu, Kathleen Fitzpatrick (2015) cites several issues with the platform. First, Fitzpatrick takes issue with the website's misrepresentation as an education-affiliated entity through the use of the "edu" top-level domain. The company obtained the domain prior to rules limiting its use to education-affiliated organizations, which Academia.Edu – a commercial enterprise – is not. Next, Fitzpatrick criticizes the company's business model, which is based in part on proceeds from data mining. Fitzpatrick quotes Gary Hall (2015): "Academia.edu has a parasitical relationship to the public education system, in that these academics are labouring for it for free to help build its privately-owned for-profit platform by providing the aggregated input, data and attention value." On the one hand, academics want to participate in these services, fearing missed opportunities to connect with potential collaborators and to disseminate research more widely, which is a fundamental goal of academic research; on the other hand, Academia.Edu profits from academic research, which in turn is dependent on tax-payer financing. Competition from Academia.Edu, Fitzpatrick argues, hinders the advancement of legitimate sharing platforms such as institutional or disciplinary repositories.

Institutional repositories have received enormous funding and have flourished as a result, but have struggled to attract significant user participation (Marsh, 2015). Fitzpatrick attributes this struggle to the institutionally siloed nature of these repositories, which has, in turn, limited the ability of their users to reach global audiences. Disciplinary repositories such as Arxiv.org have enjoyed more success than their institutional counterparts, which Fitzpatrick attributes to a focused attention on user needs within specific communities. Fitzpatrick has responded to this situation through her instrumental involvement in the Humanities Commons project, which seeks to mitigate these issues through a multi-disciplinary repository focused on user needs and social networking. Whether Humanities Commons is a successful answer to Academia.Edu is a recommended area for further study.

Fitzpatrick, in her critique of Academia.Edu, does not explain the difference between its "parasitical" relationship with publicly funded research and the clearly analogous relationship to the "big" publishers of academic journals and the researchers whose success depends on them, or, for that matter, to private monetary investment in research more generally. One might surmise that Fitzpatrick is against any private intervention in scholarly communications. This study both recognizes the problematic nature of Academia.Edu, which requires academics to freely labour in support of the company's corporate earnings, and also acknowledges that private business can infuse new ideas and innovation into the scholarly communications ecosystem. For the profit they stand to earn, Academia.Edu must provide users with services in-kind. Researchers could gain a best-in class platform for dissemination; without accurate author attribution through high quality "creator" metadata, Academia.Edu is letting down the expectations of its constituency. Academic users may reconsider using websites such as Academia.Edu if they discover author attribution is not consistent with their expectations. Such awareness can only be established through examination of the metadata.

Researchers have studied academic SNSs from a variety of perspectives. Almousa (2011) studied Academia.Edu user data, examining users affiliated with two Arts disciplines and two Science/technical disciplines, with an interest in user population characteristics and patterns of usage. Disciplinary affiliation was skewed toward Arts.[2] Post-doctoral users were most active in content distribution (submitting research) and in relationship building (social networking), regardless of disciplinary affiliation. Faculty showed similar research activity but noticeably less relationship activity. Additionally, graduate student behavior was considered "close" to faculty and post-doctoral users, but demonstrated less activity overall. Almousa provides a description of populations and usage patterns, but does not examine metadata quality.

Similar to Almousa's study, Thelwall and Kousha (2014) examined usage on Academia.Edu, grouping users by academic rank, disciplinary-affiliation, and gender. Profile views were compared between groups. Faculty garnered more views than students. In Philosophy and History, gender did not influence views; in Computer Science and Law, users identifying as female received more views than users identifying as male. Finally, Thelwall and Kousha examined correlation between traditional bibliometric measures and Academia.Edu custom usage analytics within the Philosophy group of users, finding no significant evidence of correlation. Bibliometric researchers have generally found little correlation between traditional bibliometric measures and social networking alternative measures such as hit counts, downloads, bookmarks, and recommendations; instead, alternative metrics are commonly considered complementary to traditional measures of impact, offering alternative perspectives through which to examine a researcher's activities (Ortega, 2015). Thelwall and Kousha confirm this view of alternative metrics in their study of Academia.Edu custom analytics.

Menendez et al. (2012) examined data on Academia.Edu to understand how researchers

[2] Ortega (2016) compared academic SNSs and confirmed the notion that humanities and social sciences disciplines vastly outnumber other disciplines on Academia.Edu.

represent themselves (i.e. information practices). The researchers collected approximately 30,000 profiles representing 8 different disciplinary affiliations, examining variables such as academic position and research interests, personal information such as picture or description, and engagement patterns such as publications and interactions with other users. The sample population was composed of "graduate students (49 %), faculty members (36 %), independent researchers (9 %), and post-docs (6 %)." The researchers performed a post-hoc analysis of content submission grouped by disciplinary affiliation: "faculty members contributed significantly more than post-docs, who contributed more than graduate students and independent researchers." The exact distribution, however, is not provided. The researchers concluded that participation on the social networking site closely mirrored the reality of academic hierarchies, with faculty participation exceeding student participation, faculty appearing more open to sharing personal information, and institutional prestige and country of origin significantly affecting patterns of publication and interaction.

3.3 Overview of Metadata

As evident in the preceding discussion, research into academic SNSs has focused on the social networking aspect of these services, particularly examining patterns of usage. No research currently exists which investigates bibliographic metadata on academic SNSs. Metadata as a concept can be associated with a variety of processes and technologies. The term *metadata* was first coined in 1969 for describing "data about data" (Gill, 2016). Within the context of libraries and information science, documentation of "data about data" has existed through traditional cataloging long before electronic information or the web. The term *metadata*, however, has risen in usage alongside the popularity of the web, thus reflecting its common association with electronic information environments (Greenberg, 2005). Metadata, however, does not necessarily refer *exclusively* to electronic data, although it is discussed in this context in this study (Gilliland, 2016). Metadata is more clearly described as structured data about data, meaning that its creation and implementation is purposefully designed to support specific functions or applications (Greenberg, 2005).

Gilliland (2016) applies functional categories to different metadata, including “technical,” “descriptive,” “preservation,” “use,” and “administrative.” Descriptive metadata is “used to identify, authenticate, and describe collections and related trusted information resources” (n.p.). The present study focuses on descriptive metadata, as it is integral for core functions such as identifying and authenticating objects, and is the category to which “creator” metadata belongs. Descriptive metadata supports a variety of functions in the digital environment. For example, “date” and “creator” metadata helps ensure authority and provenance, while “title” and “subject” metadata may assist finding and identifying an object through search and faceted browsing. The purpose of a particular kind of metadata, especially “creator” metadata, can be multiple and overlapping.

3.4 Standards and Authority Control

The advent of the internet and the World Wide Web has shifted scholarly communications away from paper-based formats onto electronic networks. As part of this transformation, information professionals have worked to standardize bibliographic metadata in part to improve discoverability of electronic resources (Greenberg, 2005). Among other results, improved metadata leads to more readily discoverable resources, which in turn contributes to effective scholarly dissemination. Failure to meet set standards impacts the discoverability of scholarship, and the public investment in the academic process sees less return.

Information professionals often try to reach consensus to help manage the complex needs of their community; to meet these needs, communities create metadata standards to which information professionals agree to adhere. Standards help support interoperability across technologies, bridge vocabularies, and provide common ground on which to build future technologies (Woodley, 2016). Standards such as the popular Dublin Core[3] are formalized through international agencies, including the International Standards Organization (ISO), National Information Standards Organization (NISO) [4], and W3C[5]. Dublin Core is therefore able to support cross-discipline interoperability for digital asset description, and is perhaps the most widely adopted standard among digital libraries (Gill, 2016; National Information Standards Organization, 2007; Phelps, 2012; Windnagel, 2014). This picture of metadata standardization only touches the surface of existing frameworks. Information professionals use metadata registries such as *MetadataRegistry.Org* or the *Marine Metadata Interoperability (MMI) Ontology Registry* for publishing and documenting schemas or ontologies; consequently, these registries help support sharing, reuse, and consensus-building. Similarly, description is made more uniform through controlled vocabularies and authority files published by trusted organizations such as Library of Congress (LOC), Getty, or Virtual International Authority File (VIAF) (Woodley, 2016). Enormous efforts are in place to encourage uniform description of personal and corporate names through use of authority files (American Library Association (ALA), 2010). LOC authority files, however, only describe book authors; thus, applying uniform personal names at the article level is a daunting task, and is made worse due to sheer scale of article-level publishing (Elliot, 2010). Despite this problem, all of these initiatives coalesce to provide repositories with the tools and resources

[3] Dublin Core is maintained by the Dublin Core Metadata Initiative. It is a widely-used standard for resource description. Its design is meant to be domain-agnostic, interoperable, and easily extended (DCMI, 2017). In contrast with RDA, which centres on bibliographic description, Dublin Core offers a descriptive framework that may easily be applied to electronic resources outside the scope of heritage institutions (i.e. web resources).

[4] National Information Standards Organization (NISO) was founded in 1929. NISO is accredited by the American National Standards Institute (ANSI). NISO “identifies, develops, maintains, and publishes technical standards to manage information in today's continually changing digital environment. NISO standards apply to both traditional and new technologies and to information across it's whole lifecycle, from creation through documentation, use, repurposing, storage, metadata, and preservation” (National Information Standards Organization, 2017).

[5] W3C is a an international body responsible for creating and maintaining web-specific standards and practices (W3C, 2017).

necessary for creating quality metadata. Successfully applying these tools helps meet the needs and expectations of academic users.

Standards and frameworks have been built to support the specific needs of academic users. NISO created its "Six Principles for Good Metadata" (National Information Standards Organization, 2007) as part of a framework for guiding digital repository developers in the development of "good repositories." "Good" metadata (1) conforms to community standards; (2) supports interoperability; (3) uses authority control and content standards; (4) includes a clear statement of the conditions and terms of use; (5) supports long-term management, curation, and preservation; and (6) should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification. NISO principles for good metadata encourage practices that support the semantic (i.e. descriptive) role of metadata (Park, 2009). NISO principles are therefore useful for creating and maintaining digital collections in ways that enhance finding, identifying, selecting, and obtaining curated objects. This thesis embraces the position that academic social websites should be held to the same (or similar) curatorial rigors as reputable scholarly repositories. This view is based on the shared user population for which these services are designed and the common type of digital object (research materials) which are stored and transmitted by these services. This present study critiques Academia.Edu through the lens of the sixth NISO criteria in particular, as author attribution is critical to authentic and authoritative representation.

Finally, user-contributed metadata challenges traditional notions of standardization and authority in cataloging: "Among the advantages of [user-contributed metadata] is that individual web communities such as affinity groups or hobbyists may be able to create metadata that addresses their specific needs and vocabularies in ways that information professionals who apply metadata standards designed to cater to a wide range of audiences cannot" (Gilliland, 2016, n.p.). Greenberg et al. (2005)

found authors of academic publications demonstrate significant interest in the creation of their own metadata. Moreover, Greenberg found these users could "create good quality metadata when working with the Dublin Core" and "in some cases, of better quality than a metadata professional can produce" (n.p.). Greenberg stipulated that submission design, including textual guidance, contributes to better quality metadata at the point of creation. Although users share in the responsibility for the quality of metadata on Academia.Edu, quality metadata is supported through quality submission processes (Greenberg, 2005). The present study therefore considers how user-contributed "creator" metadata compares to publisher metadata as a precursor to a future study of the Academia.Edu submission process and its role in creating metadata; however, metadata quality that results from the submission process needs to be examined and understood before studying the process itself.

3.5 The Role of Metadata in Supporting Bibliographic Functions

The present study aims to place its study of "creator" metadata quality on Academia.Edu within a functional perspective; this refers to the functional role of metadata in supporting common bibliographic tasks performed by users. These tasks are described by the Functional Requirements for Bibliographic Records (FRBR) cataloging framework (Tillett, 2004). FRBR enumerates the four common tasks conducted by library users and which bibliographic records serve: to *find, identify, select,* and *obtain* materials. Users "find" a particular material by performing a search; search terms are matched against attributes or relationships tied to entities; and, search candidates are returned. Users "identify" objects from the initial search candidates, again through metadata attributes or relationships present in the bibliographic record. The "identification" user task also occurs when users cite materials, and descriptive metadata plays a vital role in supporting the accuracy of citations. Users then "select" a specific manifestation based on this identification. Finally, a user "obtains" an object (i.e. gaining access and using the material) (Tillett, 2004). Metadata standards based on FRBR (i.e. BIBFRAME, RDA, FRAD) intentionally support these core user tasks: finding, identifying, selecting, and obtaining a resource. A functional examination of metadata quality considers how metadata supports these bibliographic functions, especially – in the case of "creator" metadata – finding and identifying. Functional approaches to metadata quality assessment are strongly supported in the literature and are represented by the quality assessment criteria utilized in this study (Bruce & Hillmann, 2004; Lei et al., 2006; Park, 2009; Oochoa & Duval, 2009).

Park (2009) argues for a functional perspective when investigating metadata quality: "Functional requirements can be established by defining both the internal requirements related to the needs of end-users in a local setting and by defining external requirements related to disclosed and exposed local metadata relating to external service providers" (p.214). What are the local needs of users and the external requirements of Academia.Edu? In other words, what is the effect of the present research findings on the Academia.Edu user population and the web more generally? Academic users are likely accustomed to searching and finding research materials through trustworthy services such as library catalogs, disciplinary repositories, and institutional repositories. These websites are known to follow standards of practice that lead to more complete, consistent, and accurate metadata. Users visit Academia.Edu to perform basic user tasks such as discovery and use of digital objects. The simple fact is that users are not able to discover objects fully without complete and accurate metadata. Finally, academic SNSs are made highly visible on the web through Google search indexing, through proactive email campaigns, and by their sheer popularity among researchers.

3.6 Creator Metadata

"Creator" metadata is a commonly recorded metadata element and a critical access point supporting the discovery and identification of digital assets:

> For musical, film, and art works, title, creator, genre, and performance information are typically recorded. For archival papers and records, details of their creation and relationships among them are most important. Information about the creators of these works and their lives is also commonly recorded as metadata in cultural heritage organizations. (Riley, 2017, p. 5)

Not only is the "creator" element commonly recorded across genres and cultural institutions, it is part of a tradition in cataloguing rules. Cwiok (2005) elaborates on the role of "creator" in AACR2:

> AACR2 defines a "personal author" as "the person chiefly responsible for creation of the intellectual or artistic content of a work" (AACR2 2002). The AACR2 concept of main entry illustrates the importance of the role of personal author in bibliographic description. AACR2 defines main entry as "the complete catalogue of an item, presented in the form by which the entity is to be uniformly identified and cited" (AACR2 2002). This essentially means that the entire literary unit may be attributed to a single authoritative entity. Therefore, the objective is to provide access to all works emanating from a particular entity under the appropriate personal name or corporate name. (p. 109)

Creator as a "single authoritative entity" is a concept that has shifted since AACR2[6] was first published. The definition of authorship for digital content has became more complex and fluid, which is reflected in the importance of the Contributor element in Dublin core, which includes Creator (i.e.

[6] Anglo-American Cataloging Rules (AACR) was first released in 1967 and updated (AACR2) in 1978. Anglo-American Cataloging Rules "are designed for use in the construction of catalogues and other lists in general libraries of all sizes. The rules cover the description of, and the provision of access points for, all library materials commonly collected at the present time" (Anglo-American Cataloging Rules, n.d.)

author) as a sub-element of Contributor; this arrangement allows for the expression of complex and fluid expressions of authorship (i.e. webmasters, editors, audio/video technicians, etc.) (Cwiok, 2010). The importance of "creator" metadata is made evident in Resource Description and Access (RDA) [7], where personal name attribute is considered a core element. Under RDA, a personal name element can consist of "preferred names" (9.2.2), "variant names" (9.2), "other identifying attributes" (9.3-9.18), "constructing authorized access points representing persons," and "constructing variant access points representing persons" (9.19.2). "Creator" has long been considered a fundamental metadata element; as such, Dublin Core established Creator – "an entity primarily responsible for making the content of the resource" – as a one of its 15 original elements (DCMI, 2012). Moreover, personal, corporate, and family name metadata fields are present in a majority of descriptive metadata standards, and are among the most commonly implemented metadata in digital library schema (National Information Standards Organization, 2007; Phelps, 2012; Windnagel, 2014). Thus, quality "creator" metadata should be included in any digital scholarly repository.

In a survey of name disambiguation research, Elliot (2010) explains the importance of quality "creator" metadata for information systems:

> Databases and search features must be able to determine whether the person who wrote article A also wrote article B. Searchers may want to call up all items written or created by a particular person. Researchers may need to determine exactly who wrote an article in order to… contact that author to propose future collaboration or ask follow questions about the data. (p. 1)

Similarly, Walker and Armstrong (2014) identify author name control as "critical" to the discovery of

[7] Resource Description and Access (RDA) is an international cataloging standard released in 2010. It is the successor to AACR2 and widely accepted as the current cataloging standard. RDA follows the Functional Requirements of Bibliographic Records, which frames the role of metadata as subservient to the core user tasks (Joint Steering Committee for Development of RDA, 2014).

scholarship in institutional repositories (IRs) and for providing "a more efficient and successful discovery experience for the end user" (p. 9). Disambiguation does not assist with completeness of “creator” metadata. Name disambiguation may however help create more accurate and consistent “creator” metadata.

Walker and Armstrong (2014) found no single satisfactory approach to name authority control in IRs. Lack of agreement on appropriate standards leaves administrators in the position of creating individualized approaches to name control. Inconsistency of author name construction (i.e. "Jane J. Doe" compared to "Jane J Doe" compared to "Jane Doe") in particular contributes to poor user experiences, particularly at the point of asset discovery. The problem is exacerbated by the increasing number of IRs and the lack of standardized authority control. The RDA authority record file is cited by Walker and Armstrong as a model from which to adapt a new framework for disambiguating names in IRs. An RDA authority file is created for active authors and contains unique information such as dates of birth or death to help with disambiguation. RDA authority files, however, currently only apply to book authors and are used typically in online public access catalogs. Walker and Armstrong meanwhile argue that Dublin Core and OAI guidance for “creator” metadata is inadequate for answering the outstanding problems with author naming; in other words, more work needs to be done to solve challenges facing author name disambiguation.

The ORCID initiative is one such area of work, and is a particularly remarkable attempt at offering controlled naming for individual researchers (orcid.org). ORCID assigns a unique identifier to a researcher with which to be identified in a publication. The researcher is responsible for populating a profile that defines unique characteristics about the researcher. ORCID is a potential solution to managing name ambiguity in digital library environments such as Academia.Edu. Critics of ORCID argue the initiative can only be successful if it is broadly accepted and uniformly implemented across systems (Salo, 2009). Walker and Armstrong (2014), in contrast, recommend identity-management tools such as ORCID; they suggest repositories implement ORCID as part of international collaboration to tackle “creator” metadata issues. Additionally, IRs also address “creator” metadata issues through manual and automated remediation projects. Whether tools like ORCID can significantly improve personal name disambiguation is yet to be determined.

Salo (2009) also enumerates the possibilities and obstacles for author name control within institutional repositories. These include external name authority records such as RDA authority files, as well as local name authority records similar to RDA record files but designed using locally acquired researcher data and locally stored authority records. In 2009, platforms such as DSpace and Fedora did not natively support RDA authority record metadata (Chapman et al, 2009; Salo, 2009). Salo considers identity-management tools similar to ORCID as untenable as they depend on international agreement and the active buy-in of researchers across institutions. Current studies, however, indicate existing systems with highly extensible platforms have begun to include local authority records and support for identity management tools such as ORCID (Rosenzweig & Schnitzer, 2015; Baessa et al., 2015; Johnson & Newman, 2014; Thomas, Chen & Clement, 2015). Many of these systems are maintained by organizations as part of the ORCID adoption and integration program (ORCID, 2017). Some institutions have published limited documentation of these projects, although their effect on author metadata quality is undetermined (Rosenzweig & Schnitzer, 2015; Baessa et al, 2015; Johnson & Newman, 2014; Thomas, Chen & Clement, 2015). Thus, identifying outcomes for author metadata quality in IR implementations of ORCID is recommended for further investigation.

3.7 Metadata Quality Assessment and Criteria

Metadata quality is an under-examined area of study (Moen, Stewart & McClures, 1997; Hillmann, 2008; Park, 2009); despite this, a variety of definitions of metadata quality exist. Defining measures of quality – particularly for completeness, accuracy, and consistency – will help determine the questions and methods posed in this study; however, definitions and measures of metadata quality remain complex and up for debate. The study of metadata quality assessment began in the late 1990s, most notably through Moen, Stewart and McClures' (1997) study of highly heterogeneous government (GILS), comparing metadata in traditional library cataloging to those in networked electronic resources. Traditional cataloging is premised on "rule based creation" and "guidance by experts." Metadata, in contrast, "are volatile and distributed," and "not only are rules absent (at the Anglo-American Cataloguing Rules level of detail), there is no consensus that they should be created" (n.p.). As a result, the user guides the development of metadata schemes: "Networked resources are highly heterogeneous, and various metadata schemes appear to reflect attributes assigned in a *de facto* fashion by different user communities" (n.p). Metadata quality assessment therefore should be adaptive and flexible according to needs of the specific user population.

Bruce and Hillmann (2004) presented a "systematic, domain- and method-independent discussion of quality indicators" in a digital library context (p. 1). Bruce and Hillmann cite bibliographic user tasks as a basis for assessing metadata quality, echoing the concerns of Moen, Stewart and McClures: "inevitably, quality is passed downstream from creator, to aggregator, to user" (p. 4). Bruce and Hillmann formulated seven areas for measuring digital library metadata, and a multitude of criteria and indicators by which they may be assessed. These criteria are intentionally "abstract" and "domain-independent," thereby providing the flexibility to assess different metadata schemes. Marc (2016) condenses Bruce and Hillmann's quality criteria and compliance indicators in the following table (pp. 21-22):

Quality Measure	**Quality Criteria**	**Compliance indicators**
Completeness	Does the element set completely describe the objects?	Application profile; documentation
	Are all relevant elements used for each object?	Visual view*; sample
Provenance	Who is responsible for creating, extracting, or transforming the metadata?	OAI server info†; File info, TEI Header‡
	How was the metadata created or extracted?	OAI Provenance; colophon or file description
	What transformations have been done on the data since its creation?	OAI About
Accuracy	Have accepted methods been used for creation or extraction?	OAI About; documentation
	What has been done to ensure valid values and structure?	OAI About; visual view; sample; knowledge of source provider practices; documentation for “creator” provided metadata; known-item search tests
	Are default values appropriate, and have they been appropriately used?	Known-item search tests; visual view
Conformance to Expectations	Does metadata describe what it claims to?	Visual view; external documentation; high ratio of populated elements per record
	Are controlled vocabularies aligned with audience characteristics and understanding of the objects?	Visual view, sample, documentation; expert review
	Are compromises documented and in line with community expectations?	Documentation; user assessment studies
Logical consistency and coherence	Is data in elements consistent throughout?	Visual view
	How does it compare with other data within the community?	Research or knowledge of other community data; documentation
Timeliness	Is metadata regularly updated as the resources change?	Sample or date sort of administrative information
	Are controlled vocabularies updated when relevant?	Test against known changes in relevant vocabularies
Accessibility	Is an appropriate element set for audience and community being used?	Research or knowledge of other community data; documentation
	Is it affordable to use and maintain?	Experience of other implementers; evidence of licensing or other costs.
	Does it permit further value-adds?	Standard format; extensible schema

* “visual view” means the process of evaluating metadata using visual graphical analysis tools, as described in the Dushay and Hillmann [25].

† Open Archives Initiative (home page)

‡ Text Encoding Initiative (home page), http://www.tei-c.org/ (accessed 28 July 2003)

Figure 2. Quality criteria adapted and synthesized by Marc (2016)

Bruce and Hillmann’s “compliance indicators” are heavily reliant on the presence of metadata standards, such as OAI documentation and application profiles. Over the years, a variety of metadata quality assessments have adapted these criteria to study different kinds of metadata in a variety of digital library contexts (Marc, 2016). Examination using Bruce and Hillmann’s criteria has yet to be conducted on academic SNSs. Some adaptation is necessary for academic SNSs, which do not use traditional metadata standards; however, these adaptations can be made from a study of methods in past studies.

Park’s (2009) overview of metadata quality assessment argues that the most common criteria for assessing metadata quality are completeness, accuracy and consistency. According to Park, these criteria are designed to support bibliographic function, or “fitness for purpose.” Park elaborates: "Quality metadata reflect the degree to which the metadata in question perform the core bibliographic functions of discovery, use, provenance, currency, authentication, and administration” (p. 224). Functional requirements of a repository are defined in part by the needs of the users and in part by “external requirements”:

> Functional requirements can be established by defining both the internal requirements related to the needs of end-users in a local setting and by defining external requirements related to disclosed and exposed local metadata relating to external service providers such as the Open Archives Initiative (OAI) (p. 214)

In other words, if the user’s experience requires faceted browsing by year, “date” metadata should be included and formatted appropriately to support this function. An empirical analysis of completeness and accuracy ought to consider the needs of academic users, and to the external requirements for exposed metadata in academic digital repositories. NISO framework for good repositories helps determine user needs in digital scholarly communications environments.

3.7.1 Assessment Criteria: Completeness

Metadata completeness refers to the extent that available and relevant elements are applied to an asset. According to Bruce and Hillmann, “the element set used should describe the target objects as completely as economically feasible” (p. 5). Stvilia and Gasser (2008) echo this perspective of completeness through their discussion of metadata costs and metadata value. A trade-off exists between value derived from quality metadata and the resources – the time, technology, or human labour – required to supply them. Consequently, the present study asserts that a minimum standard for complete metadata in an online academic research sharing environment such as Academia.Edu includes complete application of “creator” metadata. Bruce and Hillmann add “the element set should be applied to the target object population as completely as possible; it does little good to prescribe a particular element set if most of the elements are never used, or if their use cannot be relied upon across the entire collection” (5). In other words, Bruce and Hillmann insist upon reliable application of an element across the collection. Park (2009) reiterates a similar, if deductive, notion of completeness: “full access capacity to individual local objects and connection to the parent local collection(s)” (p. 219). The present study therefore asks: does Academia.Edu provide full access capacity to shared works through complete “creator” metadata?

3.7.2 Assessment Criteria: Consistency

Consistency is described by Park (2009) as one of the three most commonly implemented metadata quality criteria. Consistency can be considered in different ways:

> Conceptual/semantic consistency entails the degree to which the same data values or elements are used for delivering similar concepts in the description of a resource. On the other hand, structural consistency concerns the extent to which the same structure or format is used for

> presenting similar data attributes and elements of a resource.44 For instance, the different formats of encoding the date element (e.g., YYYY-MM-DD or MM-DD-YYYY) may bring forth inconsistency on the structural level. (Park, 2009, p. 221).

The present research focuses on structural consistency. The study considers how names are structured differently across records (eg. “name, name” versus “name, initial”, etc.) within each collection. If the user can trust one name structure (ie. “name, name”), it can be easier to find or identify a specific author (i.e. “Dick, Tracy” versus “Tracy, Dick”). On the other hand, rigid naming structures can exclude naming constructions from non-western cultures, such as multiple last names, or meaningful suffixes and prefixes.

3.7.3 Assessment Criteria: Accuracy

The accuracy of metadata refers to correctness or basis in fact, which includes “the elimination of orthographical errors, conformance to expression of personal names and place names, [and] use of standard abbreviations” (Bruce and Hillmann, 2004, p. 220). The literature is generally in agreement on this definition (Park 2009). Accuracy is considered a measure of quality that greatly affects electronic information services (Park, 2009; Beall, 2006). Beall (2006) studied correctness of date elements and the effect of inaccuracy on search. It is important to note Beall’s findings, that accurate metadata improves both precision and recall. In contrast, inaccurate metadata leads to increased error in information retrieval. In this respect, the accuracy dimension is fundamental to establishing overall metadata quality and, therefore, reliable retrieval of information.

Research on accuracy, however, lacks a critical stance toward notions of “correctness” or “facts” – the measurement of which are difficult to achieve due their inherent relativity. In large heterogenous collections, verifying accuracy can be a nearly impossible task, especially when authoritative records may not exist (Bruce and Hillmann, 2004). Personal names are particularly complex metadata when assessing accuracy. Bruce and Hillmann's notion of "conformance to expression of personal names," for example, is culturally relative. W3C lists eight common divergences in name construction which appear across cultural and linguistic domains (Ishida, 2011). Construction types include: name order (given name may precede family name or vice versa); appended suffixes or prefixes denoting identity characteristics (gender or patrilineal relationships); multiple name parts (maternal and /paternal family names, generational names); middle initials included or excluded; and, multiple variations of these constructions. RDA also makes recommendations for the implementation of culturally dependent personal name constructions (Section F). RDA specifies instructions for the Arabic alphabet (F.1), Burmese and Karen names (F.2), Chinese names containing a non-Chinese given name (F.3), Icelandic names (F.4), Indic names (F.5), Indonesian names (F.6), Malay names (F.7), Roman names (F.8), Romanian names containing a patronymic (F.9) and Thai names (F.10). These standards reveal the dynamic nature of name expressions and the issues with assessing metadata accuracy according to "conformance to expression," which depends on the cultural context to which a name belongs. Personal name construction is therefore dynamic and does not fit easily within the existing framework for measuring metadata accuracy. The present study therefore performs a qualitative assessment, utilizing a content analysis of author names through a deductive approach, including identification of patterns based on a process of tagging and categorization, which is informed in part by the W3C list of personal name constructions.

3.8 Quality Assessment Methods: Past studies and measures

A recent metadata quality assessment on the Google Books digitization project revealed 36% of sampled resources to contain errors, which included the 'creator' element as a source of issue (James & Weiss, 2012). James and Weiss (2012) cite the average high quality digital library as possessing metadata with 1% to 12% rates of error. If reputable information organizations like Google produce less than satisfactory metadata, we should expect even more problematic metadata occurring in less intentional metadata creation environments like Academia.Edu.

The present study requires both qualitative and quantitative methods for assessing the quality of "creator" metadata. Assessment of quality can address semantic structure (format), syntactic structure (schema), and data values (Hillmann, 2008). Academia.Edu, as an academic social networking site, does not have a stated purpose of being interoperable or harvestable; therefore, format and schema have been largely overlooked. This study focuses on data values for Academia.Edu metadata because author name quality is largely related to the presence, accuracy, and consistency of data values. Further research should consider the extent to which Academia.Edu users expect complete, accurate, and consistent author metadata. The present study, however, aims to determine the extent to which Academia.Edu meets these criteria.

The methodology of the present study is determined by a variety of studies. A common approach involves measurement using a scoring method. Hughes (2004) performed a metadata quality assessment on OAI compliant repository metadata, implementing a scoring mechanism to compare elements within records, awarding points based on set criteria. For example, one point was awarded for implementation of a controlled vocabulary; a full point was awarded for a record with no absent fields; and, two tenths of a point were removed for absent fields. Scores were weighted based on the number of terms in a record and the number of records in a sample. This is a common method in metadata quality assessment. Because the present study only focuses on one specific element, it is less important to weight scoring for comparison purposes. In fact, Hughes' study is ineffective for the present purposes because it does not consider the completeness of multi-value elements, which is integral in a study of quality "creator" metadata.

Goovaerts and Leinders (2012) apply a statistical approach for measuring completeness of randomly sampled objects in a discipline-specific OAI-compliant repository. Their measurement technique was adapted from Ochoa and Duval (2009) and M.A. Sicilia et al (2005). M.A. Sicilia et al. argue for a definition of completeness that requires machine-readability as a feature of metadata elements. In contrast with Ochoa and Duval, who assessed entire records, Goovaerts and Leinders focused on specific, required elements. Two random samples (n=100; n=300) were drawn and assessed for occurrence of specified elements, using the first sample as a basis to establish a confidence interval upon which to draw the next sample. The current study replicates the sample sizes used in Goovaerts and Leinders. The metric for completeness involved counting the number of fields in each metadata record that contained a non-null value. Multi-valued fields were considered complete if at least one instance existed. Fields were also weighted, with a higher weighting attributed to fields with greater perceived relevance. This model is effective for a holistic assessment of metadata fields; however, the present study requires an approach that measures the completeness of a specific multivalued element by considering more than one instance. Moreover, weighted evaluation is unnecessary in an assessment of a single element. It is important to note that Goovearts and Leinders assessed metadata for completeness (i.e. existence of values) but not for consistency or accuracy (i.e. correctness of values). The present study – by comparing with publisher records – aims to establish that "creator" metadata is not only complete, but also consistent and accurate. The attention to only one element diminishes the need for complex scoring, and instead allows for simply counting occurrences of correct instances and comparing these counts between publisher and Academia.Edu at the record level.

Assessment of personal name construction in Academia.Edu requires a qualitative approach. Differences between personal name constructions can be counted and statistical measures utilized; however, while counts are useful and they will be taken, a content analysis is arguably as effective, allowing for discovery of patterns in name construction through the analytical process. As discussed earlier, W3C (2011) provides a description of different name constructions; however, we may find through content analysis that these differences in name construction do not fully describe differences between Academia.Edu and publisher records (or, conversely, that not all construction types are reflected). This study turns to qualitative content analysis, which is a method for obtaining meaning from textual data through a process of textual analysis (Hsieh & Shannon, 2005). The analysis can be performed in three different ways: conventional, directed, or summative. In a conventional approach, coding is established through exploration of the text. In a directed approach, codes are guided by a theory or research. In summative content analysis, counts of keywords or content are taken and compared, and accompanied by an interpretation of the text (Hsieh & Shannon, 2005). The W3C list of name constructions provides a basis for a directed approach, considering cultural motivations for differences between metadata values in Academia.Edu and publisher records.

Zavalina (2013) performed a qualitative content analysis of collection metadata records, comparing free-text description to controlled vocabulary to determine if controlled vocabularies could describe digital collections more comprehensively than free-text descriptions. Content analysis involved comparing controlled terms to free-text description within a record, assessing “complementarity” of relationships. Although complementarity is not a useful measure for the present study, the present study uses this example of content analysis as evidence that such an approach is effective.

Rousidis et al. (2014) examined “creator” metadata quality in a Dryad research repository utilizing a mixed method content analysis assessment technique. Specifically, the authors were interested in “date,” “type,” and “creator” metadata quality issues affecting re-usability (especially in the Semantic Web). The authors discovered “problems” with “creator” metadata through manual examination of element values. “Problems” were not based on a rubric of known name issues, but were determined by synthesizing Dublin Core standards with inductive analysis. The naming issues included additional names, missing or added initials, non-English names, “miswritten” names, misused punctuation in initials, miscellaneous issues like irrelevant text, non-standard use of spacing, and, most seriously, the ambiguity of common author names such as “John Adams” or “Sarah Smith” (7-8). These kinds of errors were discovered in 8.71% of all “creator” instances. Rousidis et al. (2014) demonstrate how metadata values can be analyzed and tagged for errors based on a roughly defined set of expectations, and in part as a result of patterns that emerge from the analysis.

3.9 Conclusion

Although metadata is not a new concept, the process of assessing metadata quality in large heterogeneous collections is a relatively nascent area of study. High quality digital libraries adhere to metadata standards such as Dublin Core, and implement frameworks created by organizations such as NISO or Library of Congress. These frameworks support interoperability, long-term information preservation, and effective retrieval and discovery of assets. Researchers can assess metadata quality based on a variety of criteria such as completeness, accuracy, and consistency. Researchers agree, however, that metadata quality is relative to contextual factors, particularly the needs and expectations of the user community and any external requirements. Academic SNSs do not claim to uphold standards of interoperability or long-term preservation; this contrasts with institutional or disciplinary repositories. In other words, we should not expect implementation of interoperable schemas, exposure of metadata through RESTful interfaces or an OAI harvester, use of application profiles or metadata registries, or similar affordances commonly made by high quality digital libraries. In this way, academic SNSs do not adhere to the same external requirements as other academic digital repositories, nor should they be held to this standard.

Academic SNSs do, however, serve an academic user base, and they store and transmit academic research objects; in this regard, academic SNSs should meet the expectations held by the academic community. Academic communities are accustomed to digital library environments that generally use standardized (i.e. Dublin Core) metadata. Academic communities rely on bibliographic functions of metadata, which support finding, identifying, selecting and obtaining digital objects. Quality "creator" metadata is necessary for supporting the discovery (finding and identifying) of digital objects. The academic community is also a professional community, and as such, it is reliant on author attribution as a cornerstone of its profession. Author attribution is integral to professional prestige and promotion in academic institutions. Academia.Edu therefore must provide quality "creator" metadata to meet these needs and expectations.

Quality "creator" metadata may be measured by a variety of criteria. This study aims to determine quality as it supports user functions. For this purpose, the most suitable criteria are completeness, consistency, and accuracy as defined by Bruce and Hillmann (2004) and Park (2009). Completeness is indicated by the extent to which the element is applied across records (does it appear at least once in every record). It also considers reliability of implementation. Reliable implementation of "creator" metadata leads to full access potential. The present study therefore compares Academia.Edu records to publisher records to assess the completeness of the "creator" element. The "creator" element is a multiple-value element; previous studies tend to measure completeness of multiple value elements by determining that one instance of the element is implemented. The present study is therefore novel for its measurement of multiple-value completeness beyond a single instance.

Consistency is another important criterion. Structural consistency is determined by comparing values *within a sample*. This contrasts with accuracy, which compares values *across samples*. For example, given "date" values in a sample derived from one collection, criteria would assess the extent to which these "date" values adhere to common formatting (i.e. yyyy/dd/mm vs mm/dd/yyyy). *Semantic* consistency would compare values similarly, while assessing semantics (i.e. year created vs year published). The present study assesses structural consistency, focusing on naming constructions and spelling in the "creator" value. W3C provides a rubric of naming constructions; this rubric will assist the identification of name construction patterns.

Accuracy is this study's final criteria for metadata evaluation. Accuracy typically involves assessment of correctness or factuality, or orthographic errors. Accuracy is measured by comparing values *across* samples. For example, a "date" element in one collection is compared to the same element from a paired record in another collection. This study considers accuracy as valuable to the study of metadata quality, but is necessarily critical of its definition. In metadata applications, accuracy has a real effect on the quality of information retrieval (Beall, 2006); on the other hand, in the case of user-contributed "creator" metadata, the meaning of "accuracy" is challenged. By providing a personal name that differs from the publisher record, users may challenge the authority of the publisher. Although two paired metadata records appear orthographically different, neither one can be judged "incorrect." Instead, this study considers cultural context of personal names and the unique kinds of provisions made by user-generated metadata.

4. Research Goals

An effective user experience and the efficiency of networked information requires quality "creator" metadata. "Creator" metadata supports user tasks, particularly discovery and identification of materials in an electronic environment (FRBR, 2008). Incomplete metadata affects reliability of access and identification of information objects, or what Park (2009) describes as "full access capacity to objects"; inconsistent metadata causes ambiguity and an incoherent user experience; and inaccurate metadata – orthographical or syntactic errors – are demonstrated to hinder information retrieval (Beall, 2006).

Academic social networking services (SNSs) are one kind of electronic information environment which utilizes bibliographic metadata to support bibliographic user tasks. SNSs have received attention from researchers studying social networking and user behavior, but no study has examined metadata quality for these websites (Almoussa, 2011; Mendendez et al. 2012; Ortega, 2015; Thelwall & Kousha, 2014). A debate over the value of academic SNSs is also underway (Fitzpatrick, 2015), and this debate lacks an empirical perspective considering the quality of author attribution on websites such as Academia.Edu. In making a choice whether or not to use academic SNSs, users should be informed whether these websites support the tasks they need to perform. The present study uses the criteria of completeness, consistency, and accuracy, as defined by Bruce and Hillmann (2004) and Park (2009), as the basis for measuring quality in Academia.Edu metadata. These criteria are chosen because they are demonstrated by the literature to be among the most utilized measures to determine metadata quality (Park, 2009), and because they are common to most metadata quality frameworks, including Bruce and Hillmann (2004) and Stvilia et al. (2007). The study aims to determine if author metadata values on Academia.Edu are similar to publisher values in these three ways. If Acadamia.edu metadata is significantly dissimilar from matched publisher metadata in one or more of these criteria, the metadata presents a significant problem and further research into the causes of metadata quality problems on Academia.Edu is warranted.

5. Research Questions

Metadata quality research for academic social networking sites is necessary to help guide academics in making an informed decision whether or not to utilize these services. Moreover, as these websites increase in popularity, these websites risk spreading lower quality metadata across the web, particularly as they are indexed by search engines such as Google Scholar. Academic systems of prestige and promotion depend on accurate author attribution and, thus, poor author attribution practices works to diminish efforts of academic researchers. The following study asks:

> **Are "creator" metadata for University of Alberta Academia.Edu research materials more or less complete, consistent, and accurate than "creator" metadata in publisher records of the same titles? A problem in one or more of the three criteria – completeness, consistency, and accuracy – indicates a potential problem with "creator" metadata quality generally.**

> **Hypothesis (H1):** No. "Creator" metadata for Academia.Edu materials is of lower quality than publisher records based on issues in at least one of three criteria. This is supported by the general consensus that metadata quality is achieved through the implementation of common metadata standards and the use of naming authorities (National Information Standards Organization, 2007), and that without such standards we should expect poor quality metadata.

> **Null Hypothesis (H0):** Yes. "Creator" metadata for Academia.Edu materials is of equal or better quality than publisher records based on all three criteria. Greenberg et al. (2005) demonstrated that a submission workflow that supports quality metadata (instructions, metadata

application profiles, and user-friendly design) can assist users in creating better quality metadata than, in some cases, trained catalogers (i.e. publishers).

The above research question is subdivided into the following questions:

a) **Are University of Alberta Academia.Edu "creator" metadata values recorded completely when compared to publisher records of the same title?** Park (2009) describes completeness as a measure of full access capacity. Completeness therefore measures the reliable application of an element across the collection and within records. Bruce and Hillmann (2004) and Stvilia et al. (2007) consider completeness an important measure for determining metadata quality, and Park (2009) recognizes it as one of three most commonly implemented criteria in studies of metadata quality.

b) **Does an increase or decrease in author attribution (completeness) correlate significantly with author cohort size (as determined by publisher record "creator" element count)?** Determining correlations with cohort size will help contextualize completeness measures.

c) **Are University of Alberta Academia.Edu "creator" metadata values recorded consistently?** Structural consistency describes coherence across the collection, and therefore measures the extent of structural deviation in values across records (i.e. name, name vs. name, initial etc.). Bruce and Hillmann (2004) and Stvilia et al. (2007) consider consistency an important measure for determining metadata quality, and Park (2009) recognizes it as one of three most commonly implemented criteria in studies of metadata quality.

d) **Are University of Alberta Academia.Edu "creator" metadata values recorded accurately when compared to publisher records of the same title?** Accuracy is measured by comparing orthographic and structural differences between values, across samples (comparing Academia.Edu to publisher metadata at the record level). User-generated metadata, however, provides users with the opportunity to challenge authoritative values, thereby destabilizing a sense of authenticity; consequently, although Academia.edu records may be orthographically or structurally different from publisher records, we cannot conclude that one or the other is inaccurate; we may only conclude they do not match. Bruce and Hillmann (2004) and Stvilia et al. (2007) consider accuracy an important measure for determining metadata quality, and Park (2009) recognizes it as one of three most commonly implemented criteria in studies of metadata quality.

6. Methodology

6.1 Overview

The following section describes the methods employed to conduct this study. The aim of the study is to provide evidence of metadata quality on Academia.Edu relative to publisher metadata, focusing on the "creator" element, using completeness, consistency, and accuracy as assessment criteria. The study was conducted in two stages: pilot and main study. The pilot study involved the collection of a corpus of user data (n=5278)[8] and associated works data (n=19019) from Academia.Edu, the simple random sampling of sample data (n=81), the collection of paired records from OCLC Worldcat[9] (termed publisher records herein) (n=81), and measures for comparing "creator" metadata quality in Academia.Edu records compared to paired publisher records. The main study conducts stratified random sampling based on the disciplinary distribution obtained in the pilot sample to acquire a significantly large sample dataset (n=302), obtains matched publisher records, and applies measures adapted from the pilot study. The overall objective of the study is to answer each research question stated in the previous section. The section includes: 1. An overview of data collection and data sampling procedures for Academia.Edu and publisher records; 2. An overview of the underlying University of Alberta Academia.Edu population and of the sample population; and, 3. The analytical procedures for measuring and comparing the quality of Academia.Edu records with publisher records given the three criteria: completeness, consistency, and accuracy. Completeness is confirmed through

[8] Users were identified and selected based on their profile affiliation to University of Alberta.

[9] The Online Computer Library Cooperative (OCLC) operates the Worldcat union catalog product. Worldcat is a database product maintained by thousands of partner organizations around the world. Information professionals contribute to the creation and verification of resource descriptions, making Worldcat one of the most definitive resources for library catalogers. This study describes Worldcat records as "publisher records" to differentiate between Academia.Edu and the original source of the publication (i.e. the "publisher"), although it is more than likely records have been described and verified by a variety of library professionals from different organizations, according to AAC2R2, but more likely RDA standards. This study also compares with Worldcat sources because the library community traditionally assumes the authority of Worldcat descriptions, making Worldcat records an excellent representation of ideal description practices.

statistical hypothesis testing, and correlations with author cohort size are also observed. Consistency and accuracy are confirmed qualitatively through content analysis, using tagging and inductive methods of categorization and comparison.

6.2 Data Collection

The Academia.Edu population consists of article-level metadata drawn from University of Alberta affiliated Academia.Edu users. University of Alberta users were programmatically identified through the University of Alberta affiliation page of Academia.Edu (Ualberta.Academia.Edu), where affiliated user profiles are listed. The underlying HTML for each user profile (e.g. "http://ualberta.academia.edu/NathanielNelsonFitzpatrick") contains an embedded "user object" (appendix: figure 11), which is a JavaScript object notation (JSON[10]) formatted data container composed of pertinent metadata such as user name, user id, profile page url, university affiliation, department affiliation, institutional rank, and research keywords (appendix: figure 11). User objects were harvested in April 2015 using IPython (Jupyter) Notebook software[11], the Python 3.4 programming language, and with particular assistance from JSON, Requests, and BeautifulSoup Python libraries[12]. Data was stored on a secure local machine, and indexed by user id. User objects provided the departmental and research affiliations for each user for later statistical analysis. User

[10] "JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999." (JSON.ORG)

[11] Jupyter Notebook is a web-browser based Python development environment that allows for easy documentation of code and code re-use. It was ideal for documenting and testing code throughout the collection process.

[12] Python was chosen for the abundance of web scraping toolkits available (i.e. BeautifulSoup), its simple HTTP interfaces (i.e. Requests), and the simplicity of parsing JSON and CSV in Python code. BeautifulSoup is particularly helpful as a tool for targeting and extracting desired elements in a HTTP document. Requests is particularly useful for executing iterative code to obtain data from multiple sites.

objects also provided a definitive list of University of Alberta affiliated user IDs from which I could obtain publication materials, herein referred to as "publication objects."

A publication object (appendix: figure 12) is a JSON formatted container composed of metadata such as title, object id, owner (user) id, format (if the object was uploaded and not only linked), date uploaded (if uploaded), publication source (i.e. journal of publication), type (i.e. paper, book chapter, conference proceedings, etc.), and co-author names and user ids. Publication objects were obtained using the same method for obtaining user objects, with one exception. Unlike profile objects, which are attached to user profiles, publication objects were obtained by supplying an application programming interface (API)[13] with a user identifier as the parameter. For example:

https://ualberta.academia.edu/v0/users/id/details?subdomain_param=api

In turn, the API returned a JSON object containing multiple publication objects. These objects were parsed into individual files using the BeautifulSoup and JSON Python packages. These publication objects were saved locally, indexed by publication identifier and nested within each user folder (marked by user identifier). While all data was retained, only type, owner, title, and co-authors are included in this study.

Metadata was manipulated or mapped for easier statistical processing. Specifically, user academic ranks were mapped to broader categories; for example, “Associate Professor” and “Chair” were both mapped to “faculty.” Additionally, similar spellings between values were algorithmically clustered and consolidated in Open Refine using key collision (fingerprint, metaphone3, 1-gram fingerprint) and nearest neighbour (PPM: radius 1.0/2.0/3.0, block 6) methods. This mapping is also provided in the population description for increased transparency. This data manipulation was

[13] An application programming interface (API) allowed me to access data via the web, to receive it in JSON object format, and to conduct this process iteratively with the help of a list of user IDs.

conducted to summarize academic rank data in a way that was useful for analysis and visualization. Author metadata, in contrast, was not manipulated in any way, as it forms the core of this study's empirical analysis.

6.3 Sampling

The Academia.Edu objects included in this study were limited to a population subset of University of Alberta users. As a University of Alberta student, I am familiar with the departmental affiliations with which University of Alberta users identify, and am therefore able to more effectively and accurately assign disciplinary categories to departmental affiliations for later statistical grouping. Additionally, the proximate population allows for follow-up research, such as qualitative interviews or surveys if warranted.

The population has been further tailored to include only uploads tagged as "papers" in the publication object. The narrowing of the population creates fewer confounding variables; for example, conference proceedings are often submitted by Academia.Edu users, but not indexed by OCLC Worldcat (the aggregation service used by this study). Reducing the population to "papers" in this way created a more reliable population for use in the study.

Sampling began with the generation of a list of University of Alberta "departments" with which the total population of University of Alberta user profiles self-identify. This list was distilled from user profile data using Open Refine software. Department data was extracted from each profile and the entire dataset was faceted by these department values. Each department was mapped to a specific disciplinary category. Disciplinary categories were drawn from a standard vocabulary published by the

National Center for Educational Statistics, called the Classification of Instructional Programs (CIP)[14]. CIP is used to describe instructional programs through a hierarchical classification system. Each Academia.Edu department was matched to a CIP identifier based on my own judgement. To reduce the number of disciplinary categories, CIP categories were mapped to their parent classification. For example, the Academia.Edu department, "Biochemistry," was mapped first to a child category, "Biochemistry, Biophysics, and Molecular Biology (26.02)," and, finally, to the CIP parent category, "Biological and Biomedical Sciences (26)."

The study was structured in two phases: pilot and main phase. The pilot phase (n=81) was conducted to develop and test appropriate methods and provide initial results based on randomly sampled data. Data was sampled from the underlying population of publication objects described in the previous section. Publication IDs from the population were loaded into and R session and selected randomly using the sample function in base R.[15] The pilot sample was created from a random sample technique, and the main study from a stratified random sample (n=301), the distributions for which are showcased in forthcoming population section. The disciplinary affiliation identified in the distribution of the pilot sample provides a basis for the stratified random sample taken in the main phase; this allowed the elimination of less frequently occurring disciplines and ensured comparison between the pilot analysis and the main study analysis if it became necessary. Statistical tests of power and effect confirmed the main study sample size was large enough for generating significant results. There are no differences between the methodologies applied in the pilot and the main phase; however, the statistical

[14] CIP is used by institutions and governments to describe and maintain titles for instructional programs. It was first created in 1980 in the United States by the National Centre for Education Statistics (NCES) (Statistics Canada, 2017). The NCES has maintained CIP since then. Canada maintains its own CIP classification system, which parallels the NCES classification closely. For the purposes of this study, the differences between the U.S. and Canadian CIP are negligible.

[15] R is a functional programming language with built-in tools for statistical analysis and visualization. Packages can be added to the base R functionality to extend and enhance its functionality.

results for the main study will be presented. The pilot was only conducted to determine valid methods and to establish a stratified sampling distribution.

The R base package sample function was used to obtain an initial simple random sample for the pilot study. The sample size was initially set at 100, but was reduced to 81 due to irreconcilable issues in the sample, such as titles which could not be located in OCLC. The stratified sample for the main study required use of an R Apply loop on a sample function, taking arguments for discipline names and stratification values, then assigning the appropriate sample proportion for each disciplinary affiliation, and returning a list of samples, one for each discipline.

6.4 Publisher Matching

The website Import.io[16] was used to obtain data from OCLC Worldcat at the http://www.worldcat.org website. Import.io requests HTML from a webpage and stores structured data in comma separated value (CSV) format according to specifications provided by the user. One Worldcat record was matched to one Academia.Edu sample record based on positive title match. A list of URLs was provided to the Import.io service accompanied by an Academia.Edu object identifier to maintain referential integrity in the Academia.Edu sample. Each URL consisted of a base query pattern paired with a title drawn from the Acadmia.Edu sample; for example:

> http://www.worldcat.org/search?q=ti%3AComparing+Indigenous+and+Western+
> Approaches+to+Autism"&qt=results_page

Import.io returned a structured CSV of results for each search, and the top result was chosen from each result page. If a title and author was not obtained from the Import.io request, the sample was obtained manually from the Worldcat page. Open Refine was used to join Academia.Edu records and publisher

[16] Import.io is a website for iterative harvesting of unstructured or semi-structured data from the web.

records based on the shared Academia.Edu object identifier. If no match could be made, the publication identifier was redrawn from the disciplinary population with the sampled identifiers removed. Publication object metadata was then combined with relevant user metadata, such as Academia.Edu disciplinary affiliation and owner position (academic rank) for later analysis. Finally, every sample was spot-checked to ensure integrity.

6.5 Population

6.5.1 Underlying Population of Academia.Edu Users: Publication Frequency

The entire population consists of 5,278 University of Alberta Academia.Edu users. A small cohort publishes frequently (fewer than 500 users publish more than 25 works) and a much larger group of users are inactive (over 3500 have submitted no works) (figure 5). Figure 3 indicates the publication frequency for the 5,278 University of Alberta users:

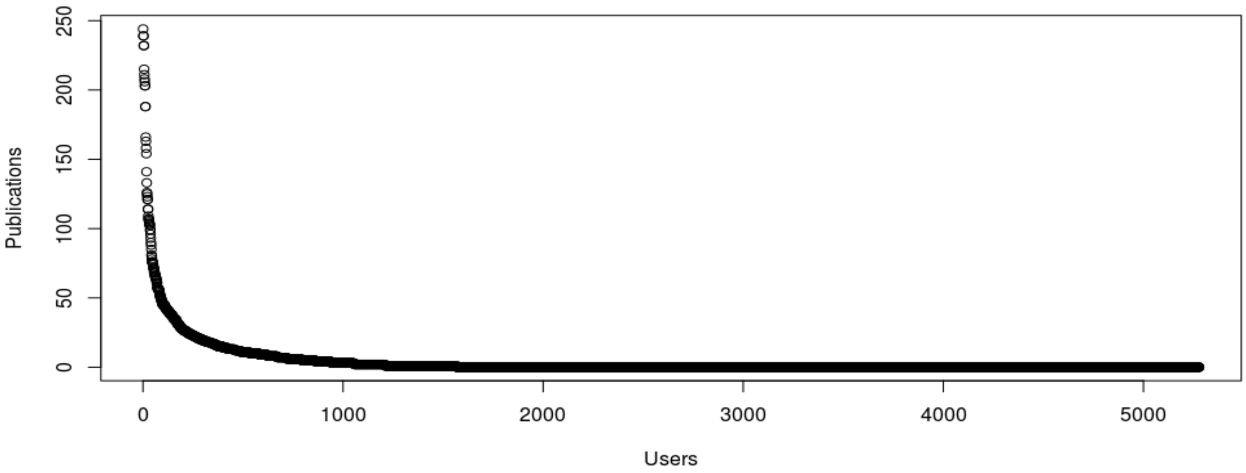


Figure 3. Publication frequency by Academia.Edu users. Note: publications are only those of document type “paper”.

Descriptive statistics for department affiliation and position were obtained through a combination of the Open Refine software and the R programming language, utilizing base, Plyr, DPlyr, Car, Psych, GGVis, and GGPlot2. Frequency of authorship by department was calculated in R. Because publications are drawn from active users, author metadata reflects only a cross-section of active users.

6.5.2 Underlying Population of Academia.Edu Users: Academic Ranks

Users identify themselves with 143 unique academic ranks (many of which are varied spellings of duplicate values). The original ranks were consolidated and mapped to “Student” (graduate and undergraduate), “Faculty” (of any rank, including Librarians, directors, deans, etc.), “Researcher” (e.g. Post-docs, Research Assistants, and laboratory technologists), “Department Member” (a frequently used category in Academia.Edu), “Alumni” (frequently used) and, “Other” (positions that do not fit easily into the above categories, especially administrative positions) (figure 4).

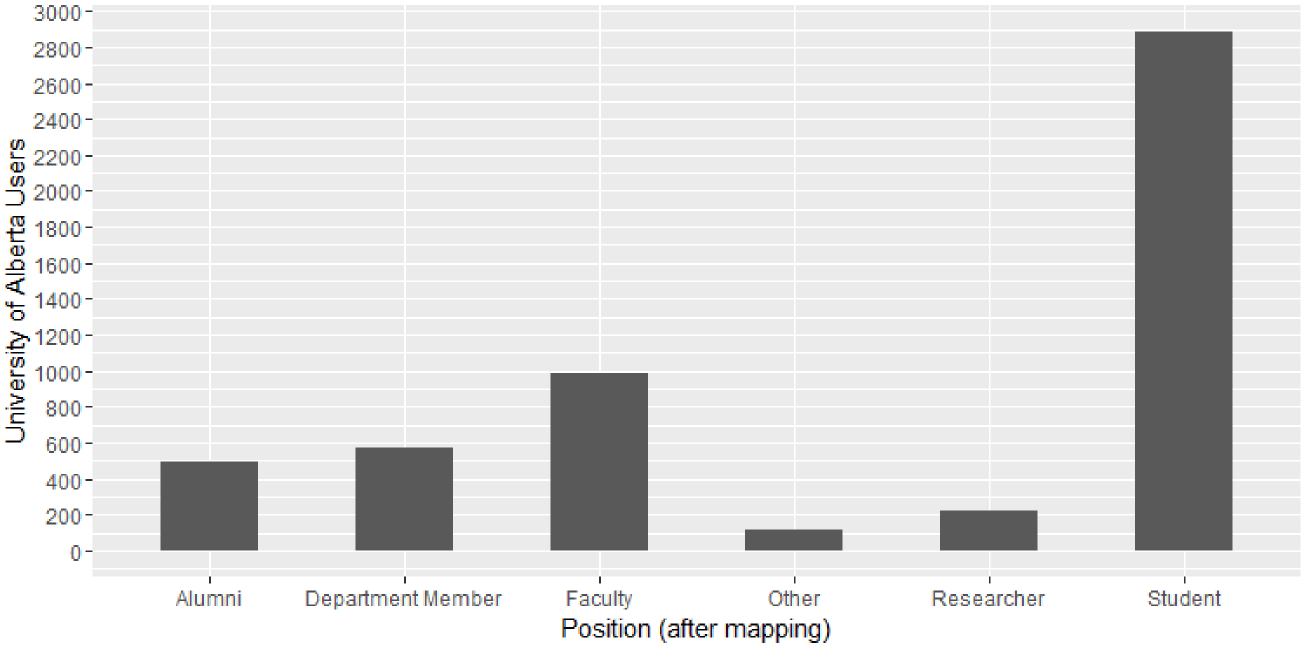


Figure 4. Users affiliated with (mapped, or derived) academic ranks (n=5278).

Raw counts of users by academic rank are shown in a bar plot in figure 4. “Students” are the largest group of users (active or inactive), followed by “Faculty,” “Department Member,” “Alumni,” “Researcher,” and “Other.”

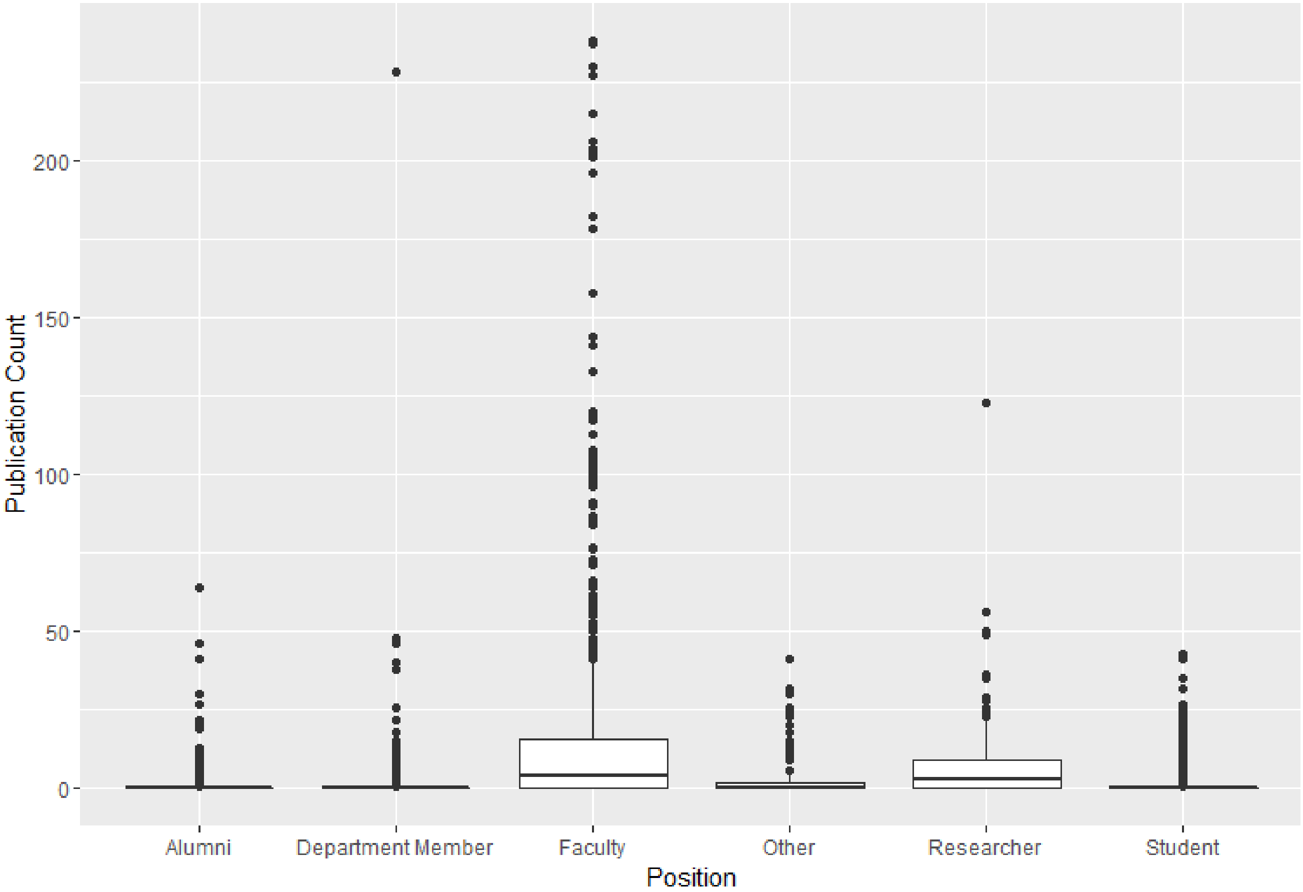


Figure 5. Compares publication frequencies by position in Academia.Edu. Note: publication counts (n=22073) are for all University of Alberta affiliated Academia.Edu users and publications.

Figure 5 shows publication contribution frequencies by position (n=22073). Despite consisting of fewer overall users, faculty far exceed students in publication contributions. Faculty are followed by the “Researcher” category, followed distantly by the remaining categories. The remaining categories demonstrate minimal publication contribution, and are thus mainly representative of the inactive portion of Academia.Edu users. Moreover, this finding has larger implications for the study, as the sample of works taken from Academia.Edu represents mainly faculty and researchers, neglecting inactive users (i.e. students). This may deceive the reader into assuming Academia.Edu is dominated by faculty; in fact, a larger proportion of Academia.edu users are students, and this population is essentially overlooked in this study.

6.5.3 Underlying Population of Academia.Edu Users: Disciplinary Affiliations

From the University of Alberta user population (n=5278), users identify with 289 unique academic departments (many, however, share similar spellings, indicating duplication). These 289 departments were mapped to 65 Classification of Instructional Programs (CIP) definitions. The 65 CIP definitions were further reduced to 25 parent classification categories used in both the pilot and main study. Below is an arbitrarily ordered list of the 25 parent classifications represented in the entire population:

AGRICULTURE, AGRICULTURE OPERATIONS, AND RELATED SCIENCES.
AREA, ETHNIC, CULTURAL, GENDER, AND GROUP STUDIES.
BIOLOGICAL AND BIOMEDICAL SCIENCES.
BUSINESS, MANAGEMENT, MARKETING, AND RELATED SUPPORT SERVICES.
COMMUNICATION, JOURNALISM, AND RELATED PROGRAMS.
COMPUTER AND INFORMATION SCIENCES AND SUPPORT SERVICES.
EDUCATION.
ENGINEERING.
ENGLISH LANGUAGE AND LITERATURE/LETTERS.
FAMILY AND CONSUMER SCIENCES/HUMAN SCIENCES.
FOREIGN LANGUAGES, LITERATURES, AND LINGUISTICS.
HEALTH PROFESSIONS AND RELATED PROGRAMS.
HISTORY.
LEGAL PROFESSIONS AND STUDIES.
LEISURE AND RECREATIONAL ACTIVITIES.
LIBERAL ARTS AND SCIENCES, GENERAL STUDIES AND HUMANITIES.
LIBRARY SCIENCE.
MATHEMATICS AND STATISTICS.
MULTI/INTERDISCIPLINARY STUDIES.
NATURAL RESOURCES AND CONSERVATION.
PHILOSOPHY AND RELIGIOUS STUDIES.
PHYSICAL SCIENCES.
PSYCHOLOGY.
SOCIAL SCIENCES.

VISUAL AND PERFORMING ARTS

Table 1. Disciplinary affiliations.

These 25 disciplines vary in contributions to Academia.Edu. The highest contributors (above 1000 works) include health, biology, engineering, physical sciences, computer and information sciences, and education:

Discipline	**Publications**
HEALTH PROFESSIONS AND RELATED PROGRAMS	4203
BIOLOGICAL AND BIOMEDICAL SCIENCES	2000
ENGINEERING	1817
PHYSICAL SCIENCES	1515
COMPUTER AND INFORMATION SCIENCES AND SUPPORT SERVICES	1397
EDUCATION	1023
FOREIGN LANGUAGES, LITERATURES, AND LINGUISTICS	953
SOCIAL SCIENCES	863
AGRICULTURE, AGRICULTURE OPERATIONS, AND RELATED SCIENCES	824
NATURAL RESOURCES AND CONSERVATION	613
BUSINESS, MANAGEMENT, MARKETING, AND RELATED SUPPORT SERVICES	500
LIBRARY SCIENCE	443
PSYCHOLOGY	428
ENGLISH LANGUAGE AND LITERATURE/LETTERS	401
LIBERAL ARTS AND SCIENCES, GENERAL STUDIES AND HUMANITIES	323
MULTI/INTERDISCIPLINARY STUDIES	302
PHILOSOPHY AND RELIGIOUS STUDIES	294
LEISURE AND RECREATIONAL ACTIVITIES	215
AREA, ETHNIC, CULTURAL, GENDER, AND GROUP STUDIES	187
COMMUNICATION, JOURNALISM, AND RELATED PROGRAMS	94
MATHEMATICS AND STATISTICS	66
FAMILY AND CONSUMER SCIENCES/HUMAN SCIENCES	65
LEGAL PROFESSIONS AND STUDIES	23
HISTORY	6
VISUAL AND PERFORMING ARTS	5

Table 2. Raw counts of publications by disciplinary category on Academia.Edu. Note: Only for publications of document type “paper” (n = 18560)

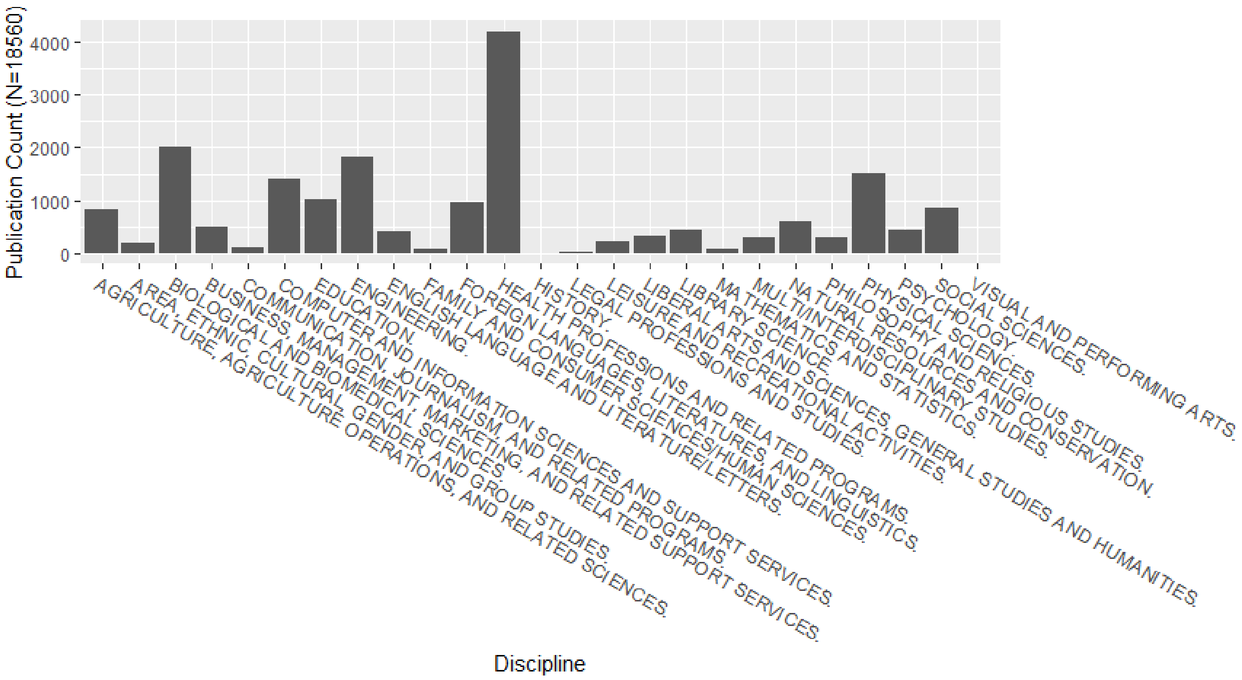


Figure 6. Publication frequencies by disciplinary category.
Note: Only for publications of document type “paper” (n = 18560)

Evidently, more works are shared on Academia.Edu from *Health Professions and Related Programs*, *Physical Sciences*, and *Biological and Biomedical Sciences* than from other disciplines (Table 2; Figure 6). The stratified sample taken in the main study is designed to sample faithful representations from different disciplines. After simple random sampling in the pilot phase, only 16 of the total 25 disciplines were drawn from the population:

Discipline	**Pilot Sample Sizes** (Simple Random Sample) (n=81)	**Pilot** (Relative Frequency) (n=81)	**Main Study Sample Sizes** (based on pilot distribution) (n=302)
AGRICULTURE, AGRICULTURE OPERATIONS, AND RELATED SCIENCES.	4	0.05	15
AREA, ETHNIC, CULTURAL, GENDER, AND GROUP STUDIES.	1	0.01	4
BIOLOGICAL AND BIOMEDICAL SCIENCES.	10	0.12	37
BUSINESS, MANAGEMENT, MARKETING, AND RELATED SUPPORT SERVICES.	1	0.01	4
COMMUNICATION, JOURNALISM, AND RELATED PROGRAMS.	1	0.01	4
COMPUTER AND INFORMATION SCIENCES AND SUPPORT SERVICES.	4	0.05	15
EDUCATION.	4	0.05	15
ENGINEERING.	10	0.12	37
ENGLISH LANGUAGE AND LITERATURE/LETTERS.	1	0.01	4
FOREIGN LANGUAGES, LITERATURES, AND LINGUISTICS.	5	0.06	19
HEALTH PROFESSIONS AND RELATED PROGRAMS.	21	0.26	78
LIBRARY SCIENCE.	3	0.04	11
NATURAL RESOURCES AND CONSERVATION.	2	0.02	7
PHILOSOPHY AND RELIGIOUS STUDIES.	2	0.02	7
PHYSICAL SCIENCES.	8	0.1	30
SOCIAL SCIENCES.	4	0.05	15

Table 3. Pilot and main study sample sizes

The main study sample is therefore limited to the 16 disciplines, which are more common to the Academia.Edu corpus.

6.6 Analytical Procedures

6.6.1 Overview

A content analysis methodology guided this research project. Content analysis involves the quantitative or qualitative examination of a corpus of textual data (Krippendorff, 2004). A simplified method of quantitative analysis was employed for assessing completeness through a process of counting occurrences and applying inferential statistical tests. In contrast, consistency and accuracy required an interpretive method of analysis, where tagging was performed through an inductive process of categorization based in part on W3 naming conventions, as well as patterns that emerged from the analytical process; categories were then described for statistical occurrences.

<u>Variables: Dependent</u>

Author Name - a metadata element describing the owner or co-author as defined in a record

Author Count - the sum number of authors indicated in a single author metadata element.

<u>Variables: Independent</u>

Source - the source of the record (Publisher or Academia.Edu)

Position (Academia.Edu only) - the academic rank of an owner of a work

Department (Academia.Edu only)- the Academia.Edu department to which the owner of a work belongs

Discipline (Academia.Edu only) - the disciplinary classification of a user, used for classifying works for whom the user is responsible, derived using the Classification of Instructional Programs (CIP) vocabulary, and correlated with department variable.

6.6.2 Completeness

6.6.2.1 Overview

Measurements were performed between May and August 2016 on data acquired in April 2016. Author metadata completeness was measured by acquiring the number of authors belonging to an author metadata element. The process of measurement included obtaining a count of Academia.Edu authors (author count) for both Academia.Edu (the owner and any co-authors) and publisher records. Each title was recorded twice, once for each source. Academia.Edu splits author names by default; whereas Worldcat author names are split by semi-colon separators (i.e. "Bob Smith; Yulduz Foreezi") and must be further split into separate columns with the help of Open Refine. Author counts were obtained by recording the number of authors belonging to one record. Author counts were recorded with the title of the work, the Academia.Edu disciplinary affiliation of the work's owner, the Academia.Edu academic rank of the work's owner, and the source of the count (Academia.Edu or Publisher). This CSV file was then imported into R for statistical analysis and visualization.

6.6.2.2 Statistical Testing

The Wilcoxon Rank Sum Test was chosen for hypothesis testing because the subjects of study – the metadata records – were independently sampled, as each record originates from a unique population (i.e. each group of records was created independently from the other). The Wilcoxon Rank Sum Test (also called the Mann-Whitney U test) is a non-parametric hypothesis test suited to comparing two separate (heterogenous) non-normal populations: "The goal consists of comparing the central tendencies of the two samples, to test whether the locations of the respective populations are equal or not" (Bonnini, 2014, n.p.). Thus, a positive test indicates a difference, or a "shift", in the location of central tendencies when comparing the two samples (i.e. the two samples are likely to originate from significantly different underlying populations). This test was chosen because the records are not, strictly speaking, pair matched, because the original creation of the records in each of the two sample populations (Academia.Edu and publisher records) are entirely distinct.[17] To test this assumption, a Levene test of homogeneity[18] was performed, revealing the two populations are significantly non-normal (are strongly skewed) and heterogeneous (have unequal variance); thus, the nonparametric Wilcoxon Rank Sum Test was chosen. Testing effect size on non-parametric data is a complicated task. Therefore, Cohens D statistic was utilized to test the effect size between samples, despite it being a parametric test, thus it should be considered a limitation of the test.

Significantly different distribution means demonstrate that Academia.Edu users over or under attribute authorship in metadata records compared to publisher records. Descriptive data and visual analysis of trends will help determine if authorship is under or over attributed in Academia.Edu records. Moreover, by testing author cohort sizes and disciplinary affiliation, this study will determine whether size of cohorts has an appreciable affect on under or over attribution.

The R Statistical Software was used to perform statistical analyses. The specific descriptive and inferential methods used have been discussed in the above sections; results will be discussed in the following results section. Summary statistics, including tendency, dispersion, and shape, of main sample data will be provided. A Levene test of heterogeneity was conducted to determine if the two samples (publisher compared to Academia.Edu) share equal variances. Initial exploration of the data

[17] This contrasts with a pair matched sample, which would be drawn from the same population; for example, in a hypothetical test of a new drug, two eight year old male children are sampled from a population, one is part of the experimental group and the other is part of the control, but they are said to be pair matched because they are drawn from the same population and share "matched" characteristics. In the current study, the two samples are drawn from separate populations (one is from publishers, and one is from Academia.Edu), which are clearly different, and thus "independent," populations.

[18]

indicates non-normal distributions and unequal or heteroscedastic populations. A non-parametric test was therefore required for significance testing.

6.6.3 Consistency

Consistency (also described as comparability or coherence) was measured by examining conceptual and structural differences between the same element across a corpus of records (Bruce and Hillmann, 2004; Stvilia et al., 2007). Conceptual consistency refers specifically to the extent to which an element describes similar concepts in a resource. Structural consistency refers to the extent to which an element utilizes a common format (i.e. mm-dd-yyyy vs. dd-mm-yy). Conceptual inconsistency necessarily results in structural inconsistency as two conceptually different expressions will logically reveal different structures (i.e a personal name is structured differently from a publisher name). Structural inconsistency, however, does not necessarily result in conceptual inconsistency (i.e. both mm-dd-yyyy and dd-mm-yy are conceptually consistent as they each represent temporal data).

Consistency was measured by assessing the extent of variation in name construction (i.e. name order) within a sample (i.e. within the Academia.Edu sample and within the publisher sample). Academia.Edu does not specify a standard; therefore, we must determine consistency through an internal comparison of values to arrive at a measure of coherence. Consistency was revealed within a collection by measuring the range of structural expressions across all “creator” elements in each record:

1) Match work owner name in Academia.Edu records to corresponding names in publisher records. Disregard records that cannot be matched in this way. This ensures that each set of records is evenly and fairly compared. In 7 of 302 instances, the owner’s name did not appear in the publisher metadata; these records were removed from the sample population, thus when calculating a relative count of name variants, 295 was the denominator.

2) Split element values by token (i.e. by space between tokens).
3) Inductively identify each token type and code each element according to the order of these token types. The token types include ‘name’ (i.e. a whole name), ‘hyphenated’ (i.e. a hyphenated name), ‘initial’ (i.e. one letter presumably indicating an initial), and a credential (i.e. a commonly recognized abbreviation or salutation such as ‘PhD’ or ‘Dr.’). The different permutations of these token types were identified and counted for the Academia.Edu publication owner name and its paired value in the publisher record (n=295).
4) Count the occurrences of each code for each sample.
5) Determine the more consistent set of records by identifying the set with the highest concentration of values adhering to a single format, and the spread of token types across the sample: how many type combinations exist in a sample; and are type combinations evenly spread across the sample, or concentrated in one or two combinations?

6.6.4 Accuracy

Orthographic errors are the most common form of accuracy problem found in metadata (Bruce and Hillmann, 2004; Beall, 2006; Stvilia et al., 2007). Other variations could include construction differences (i.e. name order). Thus, the measure for accuracy was determined by observing shifts in orthography and structure. Whereas consistency measures the comparability of values within a collection, accuracy measures the comparability of matched values between collections. The process involved:

1) Owner name was matched to its counterpart in the publisher record (same as the consistency process). In 7 of 302 instances, the owner’s name did not appear in the publisher metadata; these records were removed from the sample population, thus when calculating a relative count of name variants, 295 was the denominator.

2) Paired values were normalized for case and punctuation. Open Refine was used to check paired values for exact matches. Values with variation in spelling or construction (n=200) were separated from exact matches.

3) Name changes fall into three major categories: initial inclusion or exclusion, name changes (adding, dropping, or changing names), and orthography (i.e. spelling).

4) Acadmia.Edu owner names were compared with their publisher counterpart and the following seven variant types were identified (examples show change from publisher to Academia.Edu):

Publisher Value	**Academia.Edu Value**	**Type of change**
craig o. heinke	craig heinke	Middle initial included/excluded
francis pelletier	francis jeffry pelletier	Different name parts (added middle name)
poveda c.	cesar poveda	Different name parts (extended first name)
russell greiner	russ greiner	Orthography (Russell > Russ)
ehud ben zvi	ehud benzvi	Different name parts (conjoined last name)
andrе́ p grace	andre grace	Orthography (special character)
r yousefi moghaddam	nima yousefi	Different name parts (name change)

Table 4. Examples of orthographic and structural differences between publisher and academia.edu “creator” values. Types are adapted from W3 recommendations on name construction (Ishida, 2011).

5) The 200 non-matching values were compared qualitatively and categorized using the above tagging scheme. The results are visualized in the results section.

The two data sets were not compared to determine which source is more accurate; rather, the two were compared to determine the extent to which they match in orthography and structure. A strong mismatch was evident from the outset, as 200 of 295 (68%) – a majority – of values do not match. The results section will reveal in what ways the values vary (i.e. how do users represent their name different from the publisher?). By exploring how names vary, this section lends a qualitative perspective on the ways that user-contributed metadata shapes “creator” metadata.

7. Results

7.1 Completeness: Overall

	Academia	Publisher
n	302	302
mean	2.48	4.27
sd	2.85	3.87
median	1.00	3.00
min	1.00	1.00
max	25.00	50.00
range	24.00	49.00
skew	4.11	6.08
kurtosis	23.48	63.71
se	0.16	0.22

Table 5. Overall comparison of main study group summary statistics.

As evident in table 5, the sample mean for “creator” attribution is smaller (almost half) in Academia.Edu records (2.48) compared to publisher records (4.27), as are sample medians (1 and 3, respectively). The means of the underlying population appear different; thus, a trend of author under-attribution is strongly evident in Academia.Edu records. Incomplete “creator” metadata affects user experience in search and browse, effectiveness of search, and negatively impacts the reputation of researchers whose professional advancement is dependent on receiving recognition for research contributions. These issues will be examined further in the discussion.

7.2 Completeness: Hypothesis Testing

All hypothesis testing was conducted with the Wilcoxon Rank Sum (unpaired) test at a 95% level of confidence (α=.05)[19]. The pilot study (n=81) Wilcoxon Rank sum test revealed a p-value of

[19] The Wilcoxon Rank Sum test assumes independence between two or more observations. It tests whether a shift in the central tendency has occurred when comparing two groups. A Wilcoxon Rank Sum test poses the null hypothesis: “the distributions of two populations are equal.” Therefore, a significant result reveals the sample population distributions are not equal (i.e. they do not come from the same underlying population) (Fay & Proschan, 2010). In the context of the present

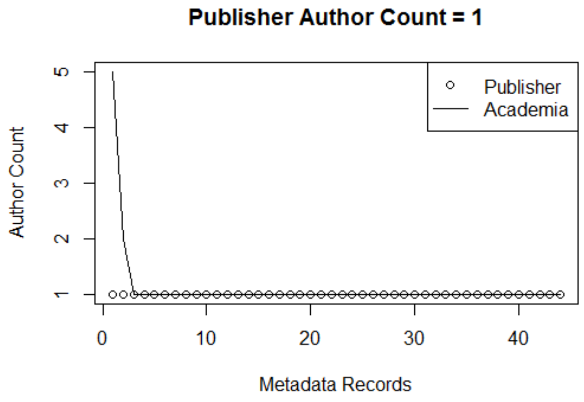
5.058e-09, thus indicating a significant difference between Academia.Edu author counts and publisher author counts. This provides initial evidence for incomplete author metadata records in the pilot metadata. The same test was performed on the main study between groups (n=302) and revealed a p-value of less than 2.2e-16. This study therefore rejects the null hypothesis that no difference in distribution means exists between the two groups, and embraces the research hypothesis that a shift in means exist between the two. Thus, we can conclude that author metadata in Academia.Edu is either over or under-represented. As evident in the summary statistics (table 5), particularly in the author-cohort size comparison (table 6), the trend is strongly toward under-attribution of authorship in Academia.Edu records. In the following section we test author cohort size. Finally, we examine the qualities that differentiate author naming in Academia.Edu from publisher metadata at the record-specific level.

7.3 Completeness: Cohort Size

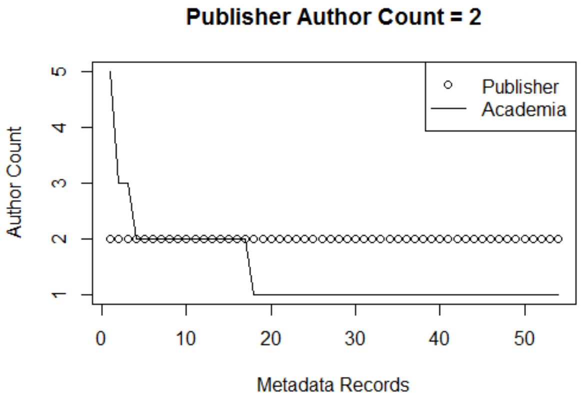
Summary statistics were calculated for “creator” element count in Academia.Edu records when grouped by author cohort size. Author cohort sizes were determined by the matching publisher record author metadata count.

study, a significant result indicates author counts are likely to have been drawn from different populations (i.e. different author counts).

Academia.Edu Summary: Publisher Author Count = 1	
n	44
mean	1.11
sd	0.62
median	1
min	1
max	5
range	4

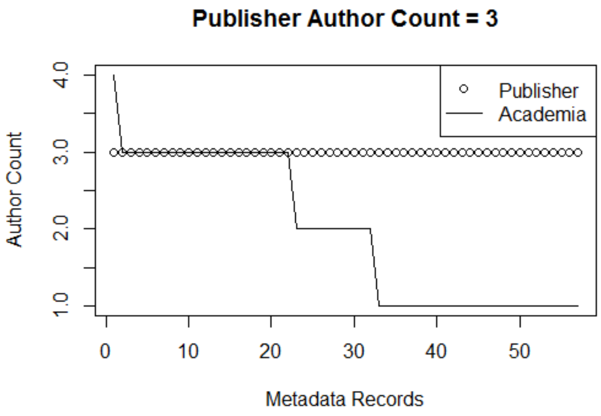


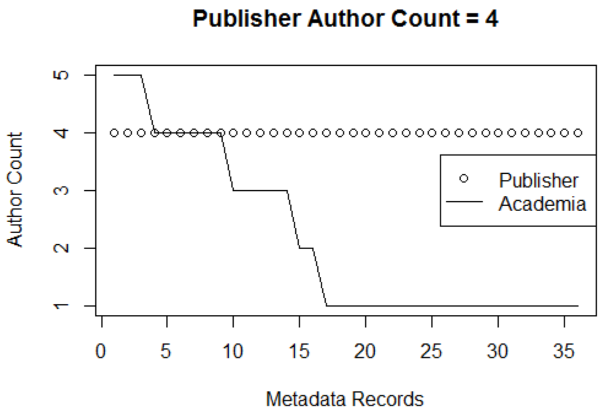
Academia.Edu Summary: Publisher Author Count = 2	
n	54
mean	1.41
sd	0.74
median	1
min	1
max	5
range	4



Academia.Edu Summary: Publisher Author Count = 3	
n	57
mean	1.96
sd	0.94
median	2
min	1
max	4
range	3

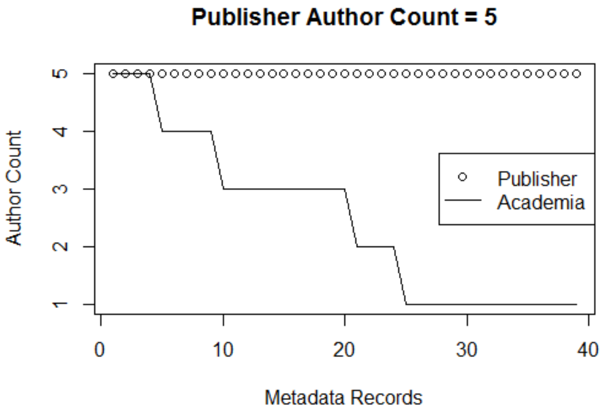
Academia.Edu Summary: Publisher Author Count = 4	
n	36
mean	2.17
sd	1.46
median	1
min	1
max	5
range	4

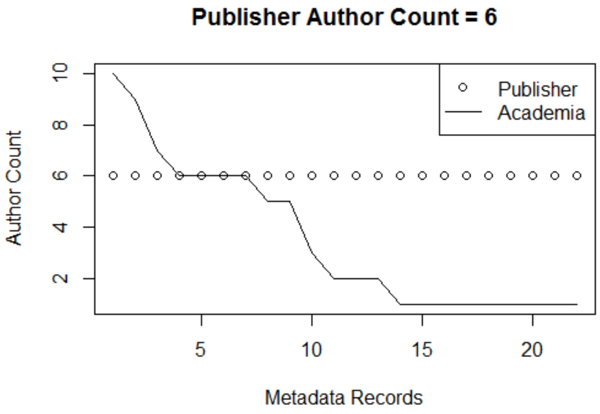

Publisher Author Count = 3


Publisher Author Count = 4

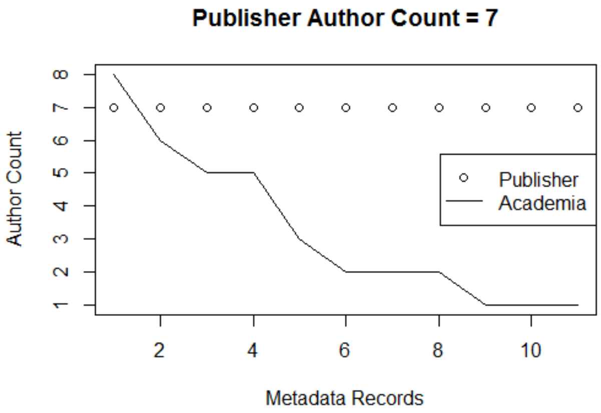
Academia.Edu Summary: Publisher Author Count = 5	
n	39
mean	2.46
sd	1.39
median	3
min	1
max	5
range	4

Academia.Edu Summary: Publisher Author Count = 6	
n	22
mean	3.55
sd	2.91
median	2
min	1
max	10
range	9

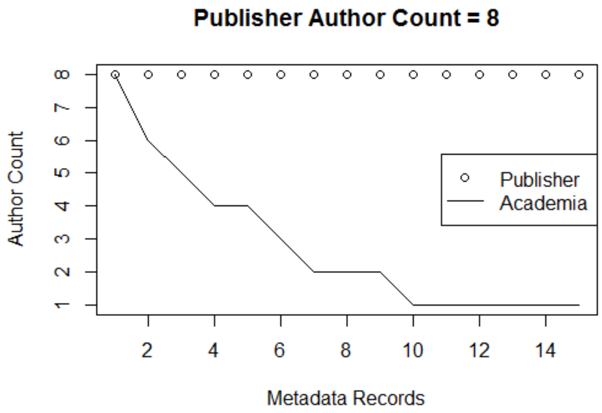

Publisher Author Count = 5


Publisher Author Count = 6

Academia.Edu Summary: Publisher Author Count = 7	
n	11
mean	3.27
sd	2.37
median	2
min	1
max	8
range	7



Academia.Edu Summary: Publisher Author Count = 8	
n	15
mean	2.80
sd	2.18
median	2
min	1
max	8
range	7



Academia.Edu Summary: Publisher Author Count = 9	
n	12
mean	4
sd	3.10
median	3
min	1
max	9
range	8

Academia.Edu Summary: Publisher Author Count >= 10	
n	12
mean	11.25
sd	7.65
median	10.50
min	1
max	25
range	24

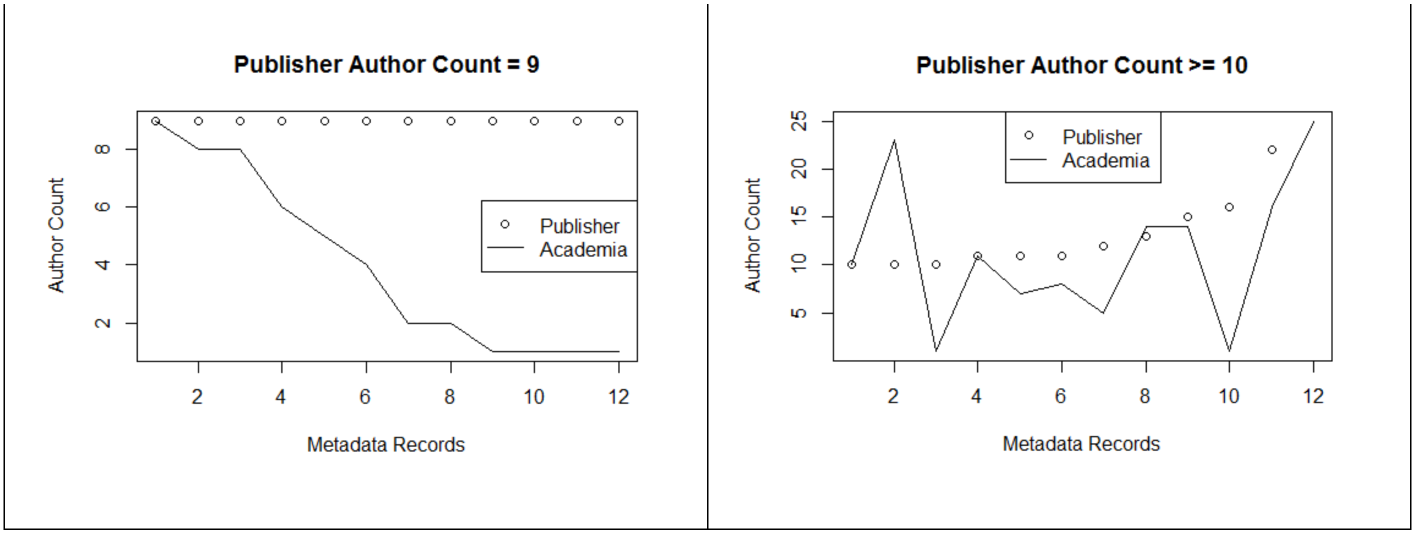


Figure7. Tables containing central tendency measures for academia.edu “creator” metadata grouped by author cohort size (main study sample). Charts compare record-level Academia.edu authorship attribution to publisher attribution, grouped by publisher author count.

Academia.Edu and publisher records are compared in figure 7 at the record level, plotted one author cohort size at a time. Works authored by one author (Figure 7: Academia.Edu Summary: Publisher Author Count = 1) are generally attributed accurately in Academia.Edu, except for two works that give additional attribution to authors (observations one and two). In contrast, approximately one quarter of two-author (Figure 7: Academia.Edu Summary: Publisher Author Count = 2) (14 of 54 records are accurate) and two thirds of three-author (Figure 7: Academia.Edu Summary: Publisher Author Count = 3) (21 of 57 records are accurate) publications under-attribute authorship in Academia.Edu. One sixth of four-author (Figure 7: Academia.Edu Summary: Publisher Author Count = 4) publications (6 of 36 records are accurate) are under-attributed. Five (Figure 7: Academia.Edu Summary: Publisher Author Count = 5) and six author cohorts (Figure 7: Academia.Edu Summary: Publisher Author Count = 6) demonstrate even greater under-attribution (4 of 39 and 4 of 22, respectively, demonstrate accurate author attribution). Seven (Figure 7: Academia.Edu Summary: Publisher Author Count = 7), eight (Figure 7: Academia.Edu Summary: Publisher Author Count = 8), and nine (Figure 7: Academia.Edu Summary: Publisher Author Count = 9) author cohorts reveal almost zero accuracy (0 of 11, 1 of 15, and 1 of 12, respectively, demonstrate accurate author attribution).

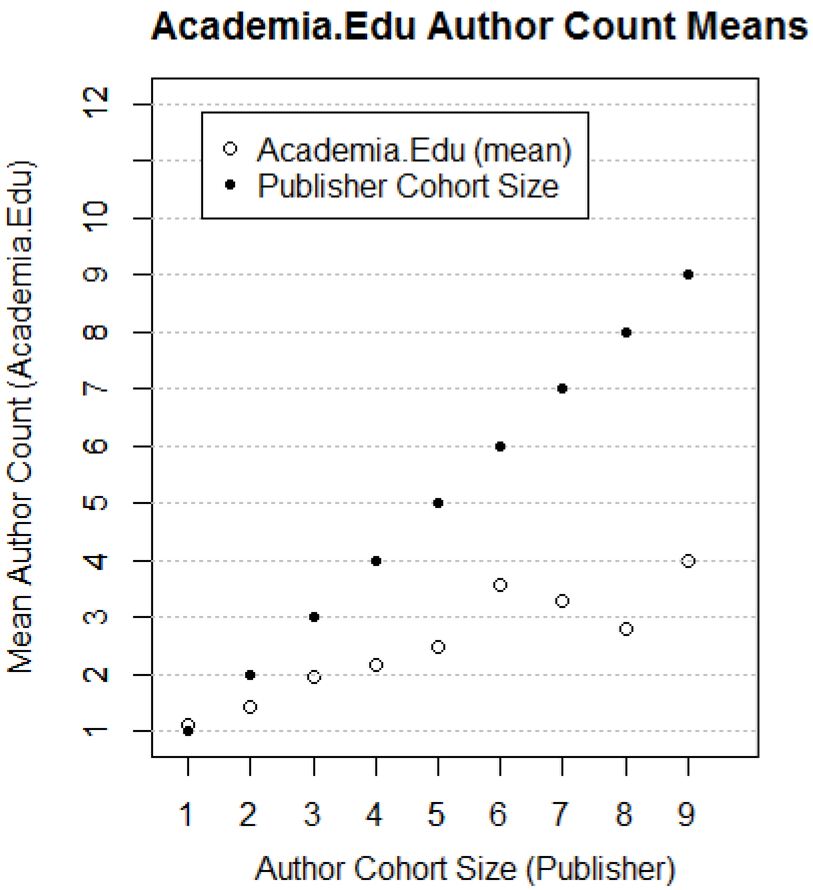


Figure 8. Mean author attribution in Academia.Edu records (group means by publisher cohort size). Note: group '10>' is trimmed.

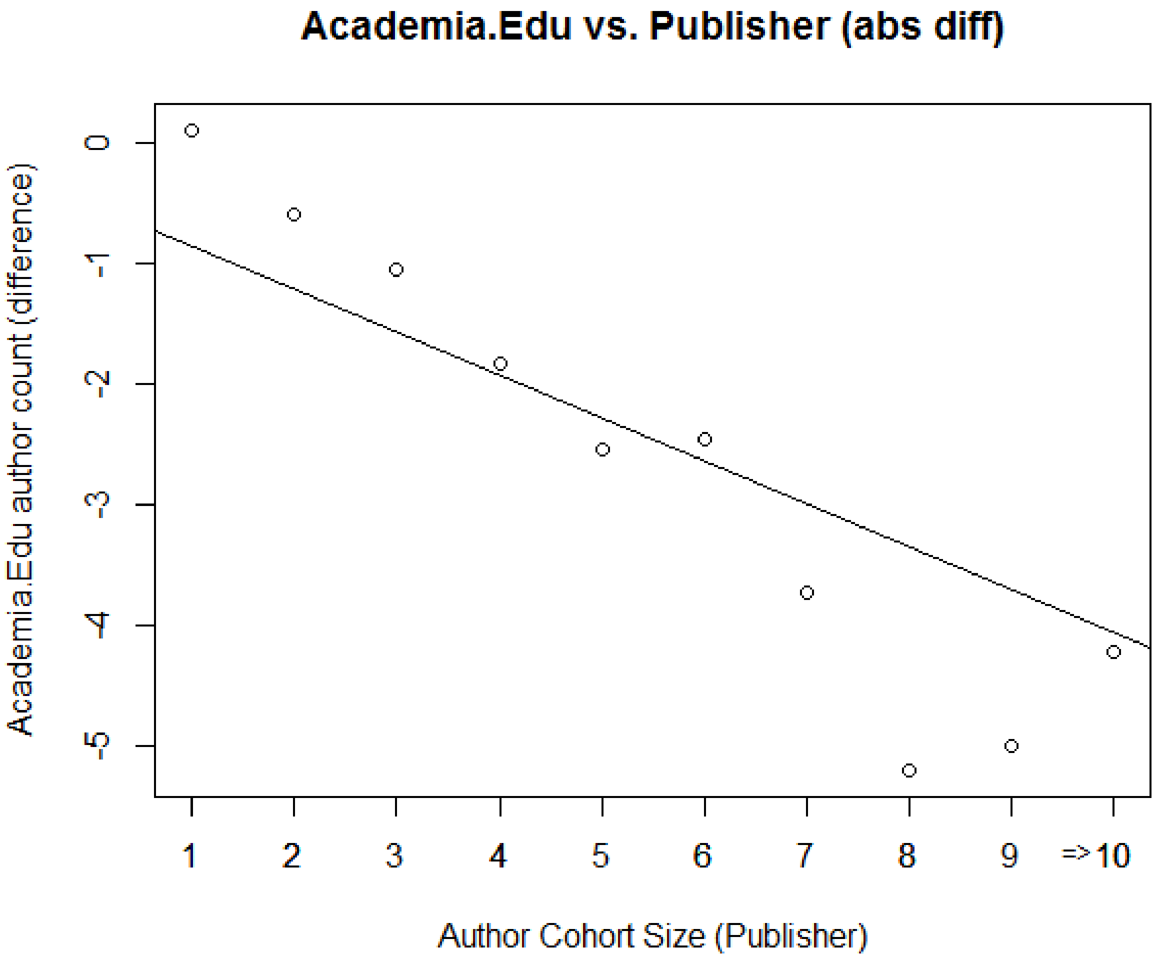


Figure 9. Difference between publisher cohort size and mean author attribution in Academia.Edu records.

As author cohort size increases, so does the trend of incomplete author attribution. Above 10 authors, the tendency to under-attribute authorship becomes less predictable, but overall it maintains the trend of increasing under-attribution. Figures 8 and 9 reveal the difference between the publisher cohort size and Academia.Edu mean cohort size. One author cohorts reveal an inversion of the common trend; instead, in this group, mean attribution rates actually exceed publisher attribution. This is because one author cohorts are easiest to attribute fully, while a small number of publications attribute authors otherwise unrecognized by the publisher. Overall, however, the pattern revealed is a general tendency toward increasing under-attribution as cohort size increases, as evinced by the generalized linear model in figure 9. The statistical significance of these patterns is determined in the next section through hypothesis and effect testing.

7.4 Completeness: Cohort Size Hypothesis Testing

Cohort Size	n	P Value	Effect
1	88	0.16	0.25997
2	108	2.24E-11	1.132226
3	114	1.91E-11	1.550266
4	72	3.73E-08	1.771168
5	78	5.65E-14	2.577901
6	44	4.95E-04	1.194132
7	22	5.64E-04	2.223864
8	30	2.25E-06	3.376745
9	24	3.55E-05	2.277867
>=10	24	0.30	0.484009

Table 6. Wilcoxon Rank Sum test p value and Cohens D effect size arranged by cohort size (determined by publisher author metadata). Main study.

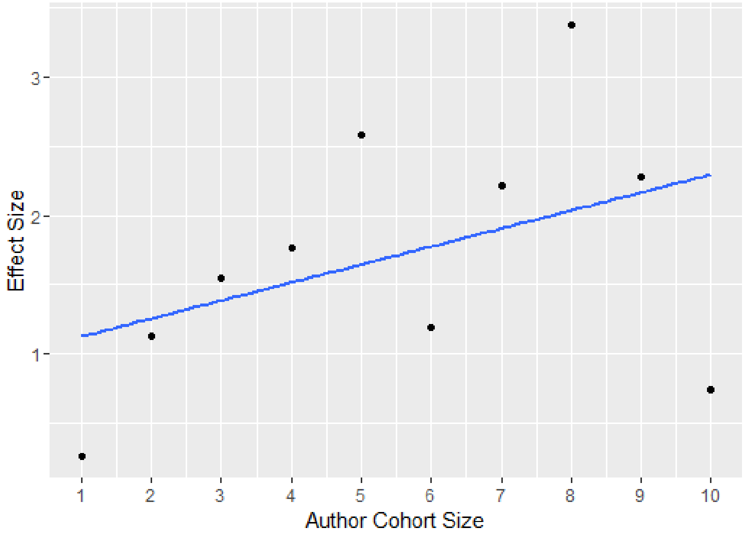


Figure 10. Cohens d-statistic plotted by author cohort size, including generalized linear model.

Through the calculation of significance using Wilcoxon Rank Sum tests and Cohen's d-statistic for effect size (table 6), it is determined that author cohort size is directly correlated to author metadata completion (figure 10). A direct correlation between increased author cohort sizes and increases in effect size is indicated. Only single author cohorts and cohorts larger than 10 demonstrate statistically insignificant differences, and this is reflected in effect size as well. Complete author metadata among single author cohorts is an expected outcome, as only one author requires attribution, and this is provided by default on Academia.Edu. Author cohorts consisting of 2 to 9 authors demonstrate significantly incomplete "creator" metadata, and present an upward trend in effect size as cohort size increases; the largest cohorts (>=10), however, demonstrate slightly better rates of attribution on Academia.Edu with hypothesis testing generating insignificant results despite an adequate sample size (n=24). We can conclude that author cohort size has a tremendous effect on "creator" metadata completion, especially as author cohort size increases. Further examination of author attribution among extremely large cohorts (>10) could yield interesting findings.

7.5 Consistency

		Name Construction Types				
Code	**Example**	**Token 1**	**Token 2**	**Token 3**	**Token 4**	**Token 5**
nn	Matthew Gushta	name	name	--------------	------------	--------
nh	Joyce Magill-Evans	name	hyphenated	--------------	------------	--------
in	I Filanovsky	initial	name	--------------	------------	--------
hn	Brett-Maclean, Pamela	hyphenated	name	--------------	------------	--------
nnn	Kaysi Eastlick Kushner	name	name	name	------------	--------
nin	Suzanne M Kresta	name	initial	name	------------	--------
inn	H. Dean Cluff	initial	name	name	------------	--------
cnnn	Dr. Prakash Chandra Mondal	credential	name	name	name	--------
nnnn	Colleen Cassady St Clair	name	name	name	name	--------
niin	Manohara P J Senaratne	name	initial	initial	name	--------
niinn	Kazi Md. Shammi Tunvir	name	initial	initial	name	name
nhc	Martin Ferguson-Pell, PhD	name	hyphenated	credential	------------	--------
n	Ghosh	name	-------------	--------------	------------	--------
ih	K Rodriquez-Capote	initial	hyphenated	--------------	------------	--------
ni	Muehlenbachs K.	name	initial	--------------	------------	--------
nnh	Sandra Jean Garvie-Lok	name	name	hyphenated	------------	--------
hin	Lori-Ann R Sacrey	hyphenated	initial	name	------------	--------
nih	Alvaro R Osornio-Vargas	name	initial	hyphenated	------------	--------
nni	Salway, Sarah M	name	name	initial	------------	--------
nii	Lagravere, M. O.	name	initial	initial	------------	--------
iin	F M Christensen	initial	initial	name	------------	--------
iih	AO El-Kadi	initial	initial	hyphenated	------------	--------
niih	Ayman O S El-Kadi	name	initial	initial	hyphenated	--------
iiih	A O S El-Kadi	initial	initial	initial	hyphenated	--------
iiin	H L M Nye	initial	initial	initial	name	--------
niii	Schmiegelow F.K.A.	name	initial	initial	initial	--------

Table 7. Construction types (n=26) identified in Academia.Edu owner names and paired publisher names.

Code	Academia.Edu	Publisher
nn	256	97
nh	15	4
in	3	56
hn	2	1
nnn	4	12
nin	8	50
inn	1	2
cnnn	2	0
nnnn	1	1
niin	1	2
niinn	1	0
nhc	0	1
n	0	1
ih	0	1
ni	0	15
nnh	0	1
hin	0	1
nih	0	3
nni	0	4
nii	0	5
iin	0	32
iih	0	1
niih	0	1
iiih	0	1
iiin	0	1
niii	0	1

Table 8. Counts of name construction types (n=26). n = ‘name’, i = ‘initial’, h = ‘hyphenated’, c= ‘credential’

	Academia.Edu	Publisher
n	11.00	24.00
mean	26.73	12.25
sd	76.16	23.70
median	2.00	1.50
min	1.00	1.00
max	256.00	97.00
range	255.00	96.00
skew	2.45	2.29
kurtosis	4.47	4.61
se	22.96	4.84

Table 9. Comparison of name construction types.

Personal names in Academia.Edu user records and publisher records reveal a wide variety of name construction types (table 7; n=26). As evinced by table 7, personal names range from single token to five token schemas, and demonstrate a range of permutations. Academia.Edu author names demonstrate greater consistency as a majority of values (~87%) adhere to the “name name” construction type (n = 257) (table 8; table 9). In contrast, publisher values are spread across several construction types: ‘name name’ (n=97), ‘initial name’ (n=56), and ‘name initial name’ (n=50) (table 8; table 9). Additionally, there are more than twice as many outlier name construction types; for example, Academia.Edu contains 4 different construction types where n=1, in contrast with publisher values, which reveal 12 different construction types where n=1 (table 8). Publisher values are incoherent and inconsistent compared to Academia.Edu values, which are evidently far more coherent and consistent.

The consistency of personal names on Academia.Edu is indicative of an effective submission process. Consistent naming provides many benefits, including an improved user experience, especially when browsing facets by author, as well as more reliable identification of authors by name. This reliability is also likely to improve attempts to disambiguate between similar names, as consistent naming makes differentiating parts of a name less difficult. Finally, publisher name inconsistency indicates a serious threat to the authority of publisher values. Publisher efforts to follow common standards for personal names are evidently failing. These issues will be examined at greater depth in the discussion.

7.6 Accuracy

A qualitative examination of author metadata revealed differences between author names in Academia.Edu records compared to publisher records. These differences are herein referred to as variants.

Type of change	Publisher Example	Academia.Edu Example	Count
First name extended	W Makis	William Makis	102
First name included	Ghosh	Sunita Ghosh	1
First name changed	R Yousefi Moghaddam	Nima Yousefi	1
Middle initial included	T Wellicome	Troy I Wellicome	2
Middle initial excluded	Matthew D Benson	Matthew Benson	99
Middle name included	K Tunvir	Kazi MD Shammi Tunvir	1
Middle name excluded	Rhoda Suubi Muliira	Rhoda Muliira	6
Middle name extended	E H Tuna	Emine Hande Tuna	4
Last name changed	Sami S Botros	Samy Soliman	3
Last name hyphenated	Brett-Maclean Pamela	Pamela Brett-Maclean	3
Last name extended	De Nicola Z	Nicola de Zanche	2
Typography (spelling)	Russell Greiner	Russ Greiner	3
Accented character included	Lagravere M O	Manuel Lagravère	1
Accented character excluded	André P Grace	Andre Grace	2
Credentials included	Prakash Chandra Mondal	Dr. Prakash Chandra Mondal	2
Credentials excluded	Martin Ferguson-Pell PhD	Martin Ferguson-Pell	1
Name order change	Sookram Sunil	Sunil Sookram	43

Table 10. Main study sample subpopulation (n=38) of name variants with qualitative tagging (see table 8 for categorical examples). All “creator” values originate from Academia.Edu document owners.

Letter case was not factored in tagging of name variants. In the main sample (n=302), 7 records had to be removed because the Academia.Edu owner name did not appear in the publisher metadata (n=295). After comparing Academia.Edu author metadata to publisher metadata, 200 (68% of sample) Academia records demonstrate name variations compared to the publisher record. Name changes fall into three major categories: initial inclusion or exclusion, name changes (adding, dropping, or changing names), and spelling. The overwhelming majority of changes occur through the extension of the first name from initial into the full name (102 cases), the exclusion of a middle initial (99 cases), and a change in name order (43); all other differences tend to be outliers. Academia.Edu values compared to publisher values demonstrate common use of the full middle name. Outlier cases include the addition or removal of diacritical marks, the conjunction of names, and the use of different names entirely. Although outliers, they represent cases where users have defined themselves in interesting ways; in particular, users may add or remove diacritical marks to assert cultural-linguistic identity, or conjoin or change names to indicate changes in marital status. In summary, the trend in Academia.Edu is variation from the publisher norms, rather than reproducing publisher author metadata.

68% of Academia.Edu personal names demonstrate some variation from the publisher records. In other words, only 32% of “creator” values are the same as publisher records. In the process of describing their name, the majority of Academia.Edu users self-represent differently from how they are represented by publishers. Determining if this measure of “accuracy” is effective as a measure of metadata quality depends on whether we assign authority to the publisher record, or to the user-generated metadata. On the one hand, the publisher is a traditional source of authority, and they are expected to follow metadata standards. Because federated search will commonly obtain the resource from the publisher, federated search will reproduce publisher metadata. On the other hand, the user may be considered an alternative source of authority, as they are in a position to self-represent. This problem is not easily resolved, although it will be addressed in the discussion.

8. Discussion

8.1 “Creator” Metadata Quality

This study compared “creator” metadata for University of Alberta users of Academia.Edu with “creator” metadata from matched publisher records obtained through the OCLC Worldcat aggregation service. The study examined metadata according to three criteria: completeness, consistency, and accuracy. The discussion section answers each research question and provides analysis in the context of current literature. Foremost, the study asked:

> **Are “creator” metadata for University of Alberta Academia.Edu research materials more or less complete, consistent, and accurate than “creator” metadata in publisher records of the same titles? A problem in one or more of the three criteria – completeness, consistency, and accuracy – indicates a potential problem with “creator” metadata quality generally.**

Results were consistent with the research hypothesis that “creator metadata for Academia.Edu materials are of lower quality than publisher records based on at least one of three criteria.” The study results identify completeness and accuracy of metadata as strongly deviating from publisher record “creator” metadata. The consistency of Academia.Edu “creator” metadata, however, is far superior to publisher records. “Creator” metadata on Academia.Edu are therefore highly incomplete, but at the same time very consistent. Accuracy is not a “cut and dried” measure of quality, as its determination depends on stable definitions of “authority” and “authenticity;” on Academia.Edu, the resource creator is also the metadata creator, thereby imbuing Academia.Edu metadata with an ere of authenticity that necessarily challenges the authority of the publisher. Moreover, the inconsistency of publisher metadata indicates fundamental failures in publishers’ efforts to standardize personal name constructions, thus further frustrating our ability to assign authority to publisher records. If any firm conclusion can be reached, it is that metadata quality is an intensely complex area of study with many internal contradictions. Arriving at a summative judgement on the quality of Academia.Edu metadata is therefore difficult, and the study instead embraces the conclusion that, above all else, a study of user-generated metadata on Academia.Edu reveals the subjectivity of concepts such as "authenticity" and "authority."

8.2 Completeness

8.2.1 Overall Completeness

With respect to the overall completeness of "creator" metadata, the study asked:

> Are University of Alberta Academia.Edu "creator" metadata values recorded completely when compared to publisher records of the same title? Park (2009) describes completeness as a measure of full access capacity. Completeness therefore measures the reliable use of an element across the collection and within records. Bruce and Hillmann (2004) and Stvilia et al. (2007) consider completeness an important measure for determining metadata quality, and Park (2009) recognizes it as one of three most commonly implemented criteria in studies of metadata quality.

Academia.Edu "creator" metadata is significantly incomplete. The mean number of authors per record for Academia.Edu compared to the mean number of authors per publisher record is 2.48 and 4.27, respectively. Completeness criteria involve the measure of a record's "access capacity" (Park, 2009). "Creator" metadata represents a primary access point within information systems for supporting search and browse. Without complete "creator" metadata, the record is at least partially incapable of supporting its functional capacity to help users find and identify materials. Academia.Edu is at risk of breaking users' trust, as they may hold and expectation for complete "creator" metadata through experience with other services that uphold bibliographic standards, such as institutional and disciplinary repositories, and online catalogs. Further research could investigate the extent to which users of academic web-based services expect complete “creator” metadata, to help place the current findings in the context of user expectations.

Incomplete metadata pose dire consequences for users. Metadata functions to support user tasks, including finding and identifying materials. Completeness of metadata is one of the most important measures of quality, and “creator” is one of the most important elements for establishing access to an item. Consequently, incomplete metadata may lead to greater difficulty finding and reliably identifying research materials. Academic authors also depend on author attribution for providing professional credibility. Incomplete “creator” metadata in academic SNSs poses a moderate, but existing, threat to a researcher’s reputation. Incomplete “creator” metadata therefore risks diminishing user trust in the platform.

A cursory examination of Google Scholar reveals extensive indexing of Academia.Edu research. Incomplete “creator” metadata therefore also potentially decreases the effectiveness of search services on the web by decreasing precision and recall on author-based searches. If Academia.Edu materials are aggregated by indexing services, or if Academia.Edu decides it wants to make research harvestable, the incompleteness of “creator” metadata will pose an even greater risk to accurate scholarly information exchange on the web.

8.2.2 Completeness and Author Cohort Size

With respect to “creator” metadata completeness and author cohort size, the study asked:

> Does an increase or decrease in author attribution (completeness) correlate significantly with author cohort size (as determined by publisher record “creator” element count)? Determining correlations with cohort size will help contextualize completeness measures.

Academia.Edu "creator" metadata revealed significant under-attribution in author cohorts ranging from 2 to 9 authors. Hypothesis testing of author counts from Academia.Edu and publisher records revealed significant differences in cohort sizes 2 through 9. Testing using the Wilcoxon Rank Sum test revealed, with 95% confidence, the samples originate from underlying populations with different distributions. Within cohort sizes ranging from 2 through 9, the effect size of under-attribution increased in a linear trend. As cohort size increases, so does the likelihood of leaving authors off the record. Cohorts of 10 or more authors, however, deviate from this trend with respect to the Wilcoxon Rank Sum test, and demonstrate lower effect size compared to all other cohort sizes. Academic SNSs are evidently not supporting high quality "creator" metadata when completeness is considered. Repositories, whether academic SNSs or otherwise, should strive to create submission processes that encourage better author attribution at the time of ingest. As a consequence of incomplete metadata, users risk receiving diminished recognition for research contributions. As cohort size increases, the number of coauthors missing from the record also increases, thereby affecting more researchers in the process. Due to the majority of authorship being conducted in collaboration, higher under-attribution in larger cohorts may lead to negative consequences for the majority of researchers (Cronin, 2012).

8.3 Consistency

Regarding "creator" metadata consistency, the study asked:

> Are University of Alberta Academia.Edu "creator" metadata values recorded consistently?

Structural consistency describes coherence across the collection, and therefore measures the extent of structural deviation in values across records (i.e. name, name vs. name, initial etc.). Bruce and Hillmann (2004) and Stvilia et al. (2007) consider consistency an important measure

> for determining metadata quality, and Park (2009) recognizes it as one of three most commonly implemented criteria in studies of metadata quality.

Measures of consistency reveal that Academia.Edu provides strongly consistent “creator” metadata across records. Approximately 87% of Academia.Edu values adhere to one name construction (name name). Moreover, Academia.Edu demonstrated 11 different name construction types. This contrasted markedly with publisher values, which did not reveal a strong concentration in any kind of construction type; values were instead spread across 24 different name construction types. The reason for this is likely two fold: 1) Academia.Edu requires users to create their own names in a highly structured submission process, which likely elicits common name construction types (‘name’ ‘name’), and 2) publisher values originate from a gambit of different journals, each of which maintain its own naming construction standards, thus leading to a lack of standardization across publishers. The inconsistency in name construction exacerbates problems with name ambiguity, which is known to negatively effective information retrieval (Walker & Armstrong, 2014). Inconsistent metadata creates uncertainty for the user, making identification and discovery more difficult. Academia.Edu metadata, however, excels in this particular criteria.

Developers of repositories can look to Academia.Edu to determine how greater consistency can be achieved in resource description, for example through a study of submission processes. Additionally, publishers should be weary of how their inconsistent practices are affecting the networked information environment. Hillman describes the importance of consistency:

> The quality of “searchability” nicely illustrates the value of consistency. Users expect to be able to search collections of similar objects using similar criteria, and increasingly they expect search results and indicative indexes to have similar structures and appearance (7).

“Similar structures and appearances” is exactly where publisher “creator” metadata is lacking, and Academia.Edu “creator” metadata excels. The uniform experience in Academia.Edu meets expectations for a reliable experience. On the other hand, highly consistent naming may not be entirely positive. Consistency in personal names may, for example, reveals a potentially rigid, highly structured submission process on Academia.Edu, which could be limiting the range of personal name construction types expressed in different cultures (Ishida, 2011; RDA Toolkit). Consistency in this particular study could, however, also be biased by a culturally homogenous faculty at the University of Alberta. In contrast, publisher names may be from research publications whose authorship represents a more culturally diverse population. In other words, there is strong evidence for consistent personal naming on Academia.Edu, but a variety of potential causes for this consistency exist, including a potentially biased sample.

Publisher inconsistencies, however, are pronounced, and OCLC Worldcat is representative of a broad range of publishing sources. Inconsistencies undoubtedly find their way into federated services such as OCLC Worldcat or a library catalogue. When a user of these services seeks a particular “creator” name, but is unable to rely on consistent naming structures, ambiguity is bound to result. Personal name ambiguity already plagues information systems, as personal names are typically not unique, but rather shared by many researchers within the same field or even institution (Salo, 2009; Walker & Armstrong, 2014); “creator” metadata inconsistency exacerbates this problem. For example, determining “A. Lee” from “Lee, A.” only complicates personal name disambiguation when the user is unsure if naming structures are consistently implemented (i.e. “first name, last name” or “last name, first name”). Inconsistencies lead to redundant browsing facets and confusing results in search. A solution for publishers is to enter personal names according to a standard and agreed upon format (i.e.

RDA), to reuse personal names using tools such as VIAF or ORCID, and by reusing names from existing records.

8.4 Accuracy

In regards to "creator" metadata accuracy, the study asked:

> Are University of Alberta Academia.Edu "creator" metadata values recorded accurately when compared to publisher records of the same title? Accuracy is measured by comparing orthographic and structural differences between values. User-generated metadata, however, provides users with the opportunity to challenge authoritative values, thereby destabilizing the notion of authoritativeness; consequently, although Academia.edu records may be orthographically different from publisher records, we cannot conclude that one or the other is inaccurate; we may only conclude they do not match in quality. Bruce and Hillmann (2004) and Stvilia et al. (2007) consider accuracy an important measure for determining metadata quality, and Park (2009) recognizes it as one of three most commonly implemented criteria in studies of metadata quality.

68% of names on Academia.Edu demonstrated structural or orthographic differences compared to their matched publisher record. Academia.Edu, simply stated, deviates from the publisher values. An authentic object is expected to mimic metadata from the original manifestation of the item (i.e. publisher metadata) and, publisher values are traditionally considered an authoritative source because they are indexed and reproduced by reputable services such as OCLC Worldcat. Thus, according to this perspective, Academia.Edu fails to reflect the authority of the object's source, the publisher, and therefore provides untrustworthy "creator" metadata.

Variation in orthography and structure could potentially lead to ambiguity with regard to author attribution. Ambiguity is a common, yet serious issue in author naming; for example, from the present study, I had to determine if "Sami S Botros" and "Sami Soliman" were the same person. Significant manual research was required to determine the two are indeed the same. Ideally, accuracy across repositories is most desirable; in practice, however, this is not a likely scenario due to resource constraints. The question, then, is whether we should even strive to reproduce a single authority and, consequently, place less emphasis on the importance of accuracy as a quality assessment criteria.

Traditionally, a publisher resource would be considered a metadata authority; Academic SNSs, however, can be perceived as an alternative authority, as Academic SNSs are one of few services that require and succeed at obtaining user-contributed metadata. As a provider of user-contributed metadata, Academic SNSs present an opportunity to reconsider our notion of authenticity and even the application of accuracy as a criteria for assessing metadata quality. Greenberg et al. (2005) found creators were likely to create better metadata than information professionals, given a high quality submission process. Changes in orthography and structure could therefore be indicative of users' efforts to represent themselves *more* accurately. Accuracy is therefore not a simple measure of quality and it does not, in the present context, provide a satisfactory conclusion on "creator" metadata.

8.5 Limitations

The sample is limited by the selection of University of Alberta affiliated researchers, thus biasing representation. This is somewhat mitigated by the university being one of the larger research institutions in Canada and thus reasonably representative of research activities among larger North American institutions. Additionally, limitations have been introduced by the restriction of publication objects tagged as "papers." This particular metadata are user controlled and thus somewhat arbitrarily assigned and not always accurate, therefore leading to some immeasurable uncertainty in the population. It is acknowledged that many of the sampled publications will be either monographs, book chapters, conference proceedings, or other types of materials, thus introducing variation between object types. Selection of CIP identifiers for disciplinary categories is also biased by the researcher's subjective judgement. Finally, the sample only represents a snapshot of Academia.Edu records from a specific point in time (April 2016). Changes in the population parameters undoubtedly will have occurred since the data was collected.

8.6 Recommendations

This study recommends that researchers actively assess whether their author contributions are being accurately portrayed in academic SNSs, and whether they are appropriately assigning attribution in the description of their own publications. Additionally, highly consistent “creator” metadata on Academia.Edu is indicative of, among other indications, a quality submission process. Developers and UX designers may be interested to further study the submission process that leads to consistency on Academia.Edu. Lower completeness in Academia.Edu metadata may also be, in part, a result of certain attitudes held by researchers. Gaining insight into attitudes toward coauthor attribution and toward use of academic SNSs is therefore an important area of future study. Finally, the findings that resulted from this study have shed light on the complexity of notions such as authenticity and authority. A large body of research exists around unpacking the meaning of these terms, and, due to the scope of this study, this research has not been discussed. Further study of consistency and accuracy in user-generated metadata as it relates to notions of authenticity and authority may yield an interesting and informative discussion.

9. Conclusion

Academic SNSs and competing scholarly communications services such as institutional repositories are similar in that they support dissemination, they traffic in research publications, and they cater to an academic audience. Users of these services should, therefore, expect some level of standardization among the different services. Metadata standardization is a basic requirement for creating “good” repositories (NISO, 2007); thus, users should expect metadata standardization from academic SNSs such as Academia.Edu. This study has revealed the affect of non-standardized metadata on Academia.Edu. Academia.Edu “creator” metadata is notably incomplete, and this is a serious issue that needs to be addressed before users place their trust in the service. Incomplete metadata is also affecting larger cohorts more than smaller cohorts, and is, therefore, likely affecting disciplines that tend to publish more in larger cohorts. Failure to attribute authorship likely hurts early career academics the worst, as the impact of attribution on their professional promotion is greater, and therefore failure to be attributed is a greater issue in their case. As social media continues to play a larger role in professional self-representation, failure to attribute will continue to affect academic users of these services.

Academia.Edu features user-generated metadata, and the extent to which it is “low quality” is partially determined by how much emphasis we place on the authority of the publisher. NISO (2007) recommends good objects in repositories be authoritative and demonstrate authenticity; as we have learned from “creator” generated metadata, however, determining authority and authenticity is a matter of perspective. Poor consistency between publisher values provides grounds for challenging the authority of publisher records. At the same time, user-generated metadata in Academia.Edu has led to significantly incomplete records, thus damaging the credibility of Academia.Edu metadata.

This study argued that failure in at least one quality criterion would indicate a problem with metadata quality. Incomplete "creator" metadata is enough to satisfy this threshold and deem Academia.Edu to have some quality issues. Academia.Edu metadata, however, should not be dismissed entirely. The user generated model should be valued by repository developers, who may appreciate the rich expression that is available through user-contributed metadata. Moreover, user-contributed metadata is not resource intensive. The human resources costs related to metadata creation have been cited by many as a deterrent to creating good quality metadata (Stvilia & Gasser, 2008) . On the other hand, "rich expression" also risks inaccuracy and incomplete entries, as this study has identified. Academic users rely on author attribution to support their livelihood, and the discovery of objects is largely dependent on the completeness of metadata. Thus, the current study recommends Academia.Edu take steps to improve "creator" metadata, and users of the service be forewarned that "creator" metadata quality may be less than adequate for their needs. Without addressing issues in metadata quality, Academia.Edu is failing to meet its obligation as an academic service. Academics depend on accurate attribution for their research contributions, and online dissemination helps achieve this goal.

Among all other conclusions, the results of this study provide strongest evidence for the sheer complexity of metadata quality assessment. Authority and authenticity are not simple concepts to define, and how we determine these definitions has a direct impact on the outcomes of quality assessments. This study therefore lends support to Park (2009), who warns that metadata quality is an understudied field. At the same time, it is the position of standards organizations such as NISO that quality metadata exist in support of interoperable, user-friendly repositories. Thus, further study of metadata quality indicators, especially consistency and accuracy, is recommended.

References

Academia.Edu (2014). Introducing Co-authored papers. Retrieved January 20, 2017 from https://medium.com/@academia/introducing-co-authored-papers-6f74361c1104#.j3zpqddbo

Almousa, O. (2011). Users' classification and usage-pattern identification in academic social networks. In *2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies* (AEECT) (1–6).

Anglo-American Cataloguing Rules (n.d.). About. *AACR2.* Retrieved from http://www.aacr2.org/about.html

Baessa, M., Lery, T., Grenz, D., & Vijayakumar, J. K. (2015). Connecting the pieces: Using ORCIDs to improve research impact and repositories. *F1000Research*, *4*. https://doi.org/10.12688/f1000research.6502.1

Beall J. (2006). Metadata and data quality problems in the digital library. *Journal of Digital Information*, 6(3).

Birrell, D., Dunsire, G., & Menzies, K. (2010). Match Point: Duplication and the Scholarly Record: The Online Catalogue and Repository Interoperability Study (OCRIS), and Its Findings on Duplication and Authority Control in OPACs and IRs. *Cataloging & Classification Quarterly*, 48(5), 377–402. http://doi.org/10.1080/01639371003738723

Bonnini, S. (2014). *Nonparametric hypothesis testing: rank and permutation methods with applications in R.* Chichester, U.K.: Wiley.

Bruce TR, Hillmann, D.I. (2004) "The continuum of metadata quality: Defining, expressing, exploiting." In: Hillmann DI, Westbrooks, EL., (Eds). *Metadata in Practice*. Chicago, IL: American Library Association.

Chapman, J. W., Reynolds, D., & Shreeves, S. A. (2009). Repository Metadata: Approaches and Challenges. *Cataloging & Classification Quarterly*, 47(3-4), 309–325. http://doi.org/10.1080/01639370902735020

Cronin, B. (2012). Collaboration in Art and in Science: Approaches to Attribution, Authorship, and Acknowledgment. *Information & Culture: A Journal of History*, (1), 18.

DCMI (2017). Specifications. *Dublin Core Metadata Initiative*. Retrieved from http://dublincore.org/specifications/

Elliott, S. (2010). Survey of Author Name Disambiguation: 2004 to 2010. *Library Philosophy & Practice*, 1–10.

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. http://doi.org/10.1214/09-SS051

Fitzpatrick, K. (2015). "Academia. Not Edu." Planned Obsolescence. Retrieved from http://www.plannedobsolescence.net/academia-not-edu/

Goovaerts, M., & Leinders, D. (2012). Metadata quality evaluation of a repository based on a sample technique. In *Metadata and Semantics Research*, (181–189).

Greenberg, J. (2005). Understanding Metadata and Metadata Schemes. Retrieved form http://www.ils.unc.edu/mrc/pdf/greenberg05understanding.pdf

Halpin, H., Tuffield, M. (2010) A Standards-based, Open and Privacy-aware Social Web: W3C Social Web Incubator Group Report. W3C Incubator Group Report. Retrieved from: https://www.w3.org/2005/Incubator/socialweb/XGR-socialweb-20101206/

Hillmann, DI. (2008) Metadata quality: From evaluation to augmentation. *Cataloguing & Classification Quarterly*, 46(1). 15.

Hillmann, D. I., & Bruce, T. R. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. *ALA Editions*. Retrieved from http://ecommons.cornell.edu/handle/1813/7895

Hsieh H.F., Shannon S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15 (9) 1277-88.

Hughes, B. (2004). Metadata quality evaluation: Experience from the open language archives community. *Digital Libraries: International Collaboration and Cross-Fertilization.* Springer. 320–329.

Ishida, R. (2011). "QA: Personal Names" in www.W3.org. Retrieved from https://www.w3.org/International/questions/qa-personal-names

James, R., & Weiss, A. (2012). An Assessment of Google Books' Metadata. *Journal Of Library Metadata*, *12*(1), 15-22. doi:10.1080/19386389.2012.652566

Johnson, R., & Newman, L. (2014). Extending the Hydra Head to Create a Pluggable, Extensible Architecture: Diving into the Technology of Hydramata. Retrieved from http://www.doria.fi/handle/10024/97598

Joint Steering Committee for Development of RDA (2014). Background. *RDA: Resource Description and Access*. Retrieved from http://www.rda-jsc.org/rda.html#background

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology.* 2nd ed. Thousand Oaks, Calif.: Sage.

Lei Y, Sabou, M., Lopez, V., Zhu, J., Uren, V., Motta, E. (2006). An infrastructure for acquiring high quality semantic metadata. *3rd European Semantic Web Conference: Budva, Montenegro*. 11-4.

Marc, D.T. (2016). Assessing Metadata Quality and Terminology Coverage of a Federally Sponsored Health Data Repository. (Dissertation)

Marsh, R. M. (2015). The role of Institutional Repositories in Developing the Communication of Scholarly Research. *OCLC Systems & Services: International Digital Library Perspectives.*

Moulaison, H. L. (2015). The expansion of the personal name authority record under Resource Description and Access Current status and quality considerations. *IFLA Journal*, 41(1), 13–24. http://doi.org/10.1177/0340035215570044

National Information Standards Organization. (2007). A Framework of Guidance for Building Good Digital Collections. 3rd Edition. Retrieved from http://www.niso.org/publications/rp/framework3.pdf

National Information Standards Organization (2017). About. *National Information Standards Organization*. Retrieved from http://www.niso.org/about/

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, (2-3), 67.

ORCID (2017). About. Retrieved from www.Orcid.org

Ibid. (2017) Adoption and Integration Program. Retrieved from http://orcid.org/content/adoption-and-integration-program

Ortega, J. L. (2015). Disciplinary differences in the use of academic social networking sites. *Online Information Review,* 39(4), 520–536. http://doi.org/10.1108/OIR-03-2015-0093

Ortega, J. Luis. (2016) *Social network sites for scientists : a quantitative survey.* Cambridge, MA: Chandos.

Ovadia, S. (2014). ResearchGate and Academia.edu: Academic Social Networks. *Behavioral & Social Sciences Librarian*, 33(3), 165–169. http://doi.org/10.1080/01639269.2014.934093

Park, J.-R. (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly*, 47(3-4), 213–228. http://doi.org/10.1080/

Park, J.-R., & Tosaka, Y. (2010). Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly*, 48(8), 696–715. http://doi.org/10.1080/01639374.2010.508711

Phelps, T. E. (2012) An Evaluation of Metadata and Dublin Core Use in Web-based Resources. *Libri* 62(4), 326–35.

Resource Description and Access (2017). Resource Description and Access Toolkit. Retrieved from http://www.rdatoolkit.org/

Riley, J. (2017) Understanding Metadata: What is Metadata and What is it For? Retrieved from http://www.niso.org/publications/press/understanding_metadata

Rosenzweig, M., & Schnitzer, A. E. (2015). An initiative to address name ambiguity Implementing ORCID at a large academic institution. *College & Research Libraries News*, *76*(5), 260–264.

Rousidis, D., Garoufallou, E., Balatsoukas, P., & Sicilia, M.-A. (2014). Data Quality Issues and Content Analysis for Research Data Repositories: The Case of Dryad. In *Let's Put Data to Use: Digital Scholarship for the Next Generation, 18th International Conference on Electronic Publishing, Thessaloniki, Greece*. Retrieved from http://elpub.scix.net/data/works/att/106_elpub2014.content.pdf

Salo, D. (2009). Name Authority Control in Institutional Repositories. *Cataloging & Classification Quarterly*, 47(3-4), 249–261. http://doi.org/10.1080/01639370902737232

Sicilia, M.A., Garcia, E., Pages, C., Martinez, J.J., Gutierrez, J.M. (2005). Complete metadata records in learning object repositories: some evidence and requirements. *International Journal of Learning Technol.* 1(4), 411–424

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, 43(1), 1–43. http://doi.org/10.1002/aris.2009.1440430113

Statistics Canada (2017). Background. *Classification of Instructional Programs (CIP) Canada 2016*. Retrieved from http://www.statcan.gc.ca/eng/subjects/standard/cip/2016/introduction#bg

Strotmann, A. & Zhao, D. (2012) Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, 63(9), 1820-1833 http://dx.doi.org/10.1002/asi.22695

Stvilia, B., & Gasser, L. (2008). Value-based metadata quality assessment. *Library & Information Science Research*, 30(1), 67–74. http://doi.org/10.1016/j.lisr.2007.06.006

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720–1733. http://doi.org/10.1002/asi.20652

Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., Han, J. (2011) Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. *ASONAM*, 121–128.

Thelwall, M., & Kousha, K. (2015). ResearchGate: Disseminating, communicating, and measuring Scholarship? *Journal of the Association for Information Science and Technology*, 66(5), 876–889. http://doi.org/10.1002/asi.23236

Thomas, W. J., Chen, B., & Clement, G. (2015). ORCID Identifiers: Planned and Potential Uses by Associations, Publishers, and Librarians. *The Serials Librarian*, *68*(1–4), 332–341. https://doi.org/10.1080/0361526X.2015.1017713

Torvik, V.I., & Smalheiser, N.R. (2009). Author name disambiguation in Medline. *ACM Transactions on Knowledge Discovery in Data*, 3(3), 11.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature*, *512*(7513), 126-129.

W3C (2017). Mission. *W3C*. Retrieved from https://www.w3.org/Consortium/mission

Walker, L. A., & Armstrong, M. (2014). "I cannot tell what the dickens his name is": Name Disambiguation in Institutional Repositories. *Journal of Librarianship & Scholarly Communication*, 2(2), 1–10. http://doi.org/10.7710/2162-3309.1095

Windnagel, A. (2014). The Usage of Simple Dublin Core Metadata in Digital Math and Science Repositories. *Journal of Library Metadata* 14(2), 77–102.

Appendix

```
{
        "id": 18448,
         "first_name": "Nathaniel",
         "last_name": "Nelson-Fitzpatrick",
         "page_name": "NathanielNelsonFitzpatrick",
         "domain_name": "ualberta",
         "created_at": "2008-11-24T04:20:48.103-08:00",
         "display_name": "Nathaniel Nelson-Fitzpatrick",
         "url": "http://ualberta.academia.edu/NathanielNelsonFitzpatrick",
         "photo": "https://0.academia- photos.com/18448/80311/88153/s65_nathaniel.nelson-
         fitzpatrick.jpg",
         "department": {
              "id": 9046,
              "name": "Electrical and Computer Engineering",
              "url": "http://ualberta.academia.edu/Departments/Electrical_and_Computer_
              Engineering",
              "university": {
                     "id": 234,
                     "name": "University of Alberta",
                     "url": "http://ualberta.academia.edu/"
              }
        },
        "position": "Graduate Student",
        "position_id": 3,
        "interests": [
              {
                  "id": 17733,
                  "name": "Nanotechnology",
                  "url": "http://www.academia.edu/People/Nanotechnology"
              },
              {
                  "id": 4758,
                  "name": "Electronics",
                  "url": "http://www.academia.edu/People/Electronics"
              },
              {
                  "id": 59,
                  “name": "Polymer Engineering",
                  "url": "http://www.academia.edu/People/Polymer_Engineering"
              }
        ]
}
```

Figure 11. A JSON formatted Academia.Edu “user object” obtained directly from a user’s public profile.

```
{
    "attachments": [
          {
           "title": "",
           "file_name": "2012_Saffih_IEEE_NEWCAS.pdf",
           "download_url":
           "https://www.academia.edu/attachments/30364568/download_file",
           "file_type": "pdf",
           "id": 30364568
          }
        ],
        "document_type": "other",
        "urls": [
          {
           "url": "http://dx.doi.org/10.1109/NEWCAS.2012.6329024",
           "id": 427378
          }
        ],
        "research_interests": [
          {
           "url": "https://www.academia.edu/People/Electrical_Engineering",
           "name": "Electrical Engineering",
           "id": 49
          },
          {
           "url": "https://www.academia.edu/People/Computer_Vision",
           "name": "Computer Vision",
           "id": 854
          },
          {
           "url": "https://www.academia.edu/People/Nanofabrication",
           "name": "Nanofabrication",
           "id": 8702
          },
        ],
        "current_user_can_edit": null,
        "coauthors_can_edit": true,
        "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic CMOS
        image sensors",
        "co_author_tags": [
          {
           "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic
           CMOS image sensors",
           "display_order": 0,
  "tagging_user_id": 952713,
  "name": "Faycal  Saffih",
  "tagged_user_id": 1264954,
  "co_author_invite_id": null,
  "affiliation": "University of Guelph",
  "work_id": 2318878,
  "id": 3802299,
  "email": "f***h@gmail.com"
},
{
  "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic
  CMOS image sensors",
  "display_order": 4194304,
  "tagging_user_id": 952713,
  "name": "R. Evoy",
  "tagged_user_id": 35703559,
  "co_author_invite_id": 326374,
  "work_id": 2318878,
  "id": 3802300,
  "email": "e***y@ece.ualberta.ca"
},
{
  "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic
  CMOS image sensors",
  "display_order": 6291456,
  "tagging_user_id": 952713,
  "name": "Nathaniel Nelson-Fitzpatrick",
  "tagged_user_id": 18448,
  "co_author_invite_id": null,
  "affiliation": "University of Alberta",
  "work_id": 2318878,
  "id": 3802301,
  "email": "n***n@ualberta.ca"
},
{
  "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic
  CMOS image sensors",
  "display_order": 7340032,
  "tagging_user_id": 952713,
  "name": "Nathan Fitzpatrick",
  "tagged_user_id": null,
  "co_author_invite_id": 4045760,
  "work_id": 2318878,
  "id": 17653887,
  "email": "n***k@uwaterloo.ca"
},
    {
      "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic
      CMOS image sensors",
      "display_order": 7864320,
      "tagging_user_id": 952713,
      "name": "Amro M Elshurafa",
      "tagged_user_id": 39602001,
      "co_author_invite_id": null,
      "work_id": 2318878,
      "id": 17653888,
      "email": "a***a@gmail.com"
    },
    {
      "title": "Fabrication of CMOS-compatible nanopillars for smart bio-mimetic
      CMOS image sensors",
      "display_order": 8126464,
      "tagging_user_id": 952713,
      "name": "Stephane Evoy",
      "tagged_user_id": null,
      "co_author_invite_id": 2756729,
      "work_id": 2318878,
      "id": 17653904,
      "email": "s***y@ualberta.ca"
    }
  ],
  "internal_url": "https://www.academia.edu/2318878/Fabrication_of_CMOS-
  compatible_nanopillars_for_smart_bio-mimetic_CMOS_image_sensors",
  "current_user_is_owner": null,
  "metadata": {
  "abstract": "In this paper, nanopillars with heights of 1\u03bcm to 5\u03bcm and
  widths of 250nm to 500nm have been fabricated with a near room temperature
  etching process. The nanopillars were achieved with a continuous deep reactive ion
  etching technique and utilizing PMMA (polymethylmethacrylate) and Chromium as
  masking layers. As opposed to the conventional Bosch process, the usage of the
  unswitched deep reactive ion etching technique resulted in nanopillars with smooth
  sidewalls with a measured surface roughness of less than 40nm. Moreover, undercut
  was nonexistent in the nanopillars. The proposed fabrication method achieves etch
  rates four times faster when compared to the state-of-the-art, leading to higher
  throughput and more vertical side walls. The fabrication of the nanopillars was
  carried out keeping the CMOS process in mind to ultimately obtain a CMOS-
  compatible process. This work serves as an initial step in the ultimate objective of
  integrating photo-sensors based on these nanopillars seamlessly along with the
  controlling transistors to build a complete bio-inspired smart CMOS image sensor
  on the same wafer.",
  "publication_name": "Proceedings of NEWCAS (New Circuits and Systems
  Conference)",
            "publication_date": {
              "year": 2012,
              "errors": {

              },
              "day": 11,
              "month": 10
            },
            "publisher": "IEEE",
            "more_info": "Silicon Etching, High Aspect Ratio, Unswitched Bosch Etching,
                    Mixed-Mode Etching, Pseudo Bosch Etching, Silicon Nanopillars"
          },
          "preview_url": null,
          "owner_id": 952713,
          "id": 2318878,
          "created_at": "2012-12-21T15:31:08.239-08:00"
}
```

Figure 12. A JSON formatted Academia.Edu "publication object." Data was extracted using the Google Refine Expression Language (GREL) in Open Refine[20] and exported as CSV.

[20] Open Refine is ideally suited to the task of organizing and transforming metadata. Open Refine imports and exports JSON, CSV, and many other formats, and allowed me to "join" different scraped datasets using common user ids and publication ids.