Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Ανάλυση του φαινομένου της 'φθοράς' των συνδέσμων στις διαδικτυακές αναφορές για την επιστημονική βιβλιογραφία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΠΟΣΤΟΛΟΣ Σ. ΓΑΡΟΣ

Επιβλέπων: Παναγιώτης Τσανάκας
Καθηγητής

Αθήνα, 8 Οκτωβρίου 2024

National Technical University of Athens
School of Electrical and Computer Engineering

Division of Computer Science and Technology

Analysis of the 'link decay' phenomenon in web references for scientific literature

DIPLOMA THESIS

of

APOSTOLOS S. GAROS

Supervisor: Panagiotis Tsanakas
Professor

Issued by the three-member committee of inquiry in 8th of October 2024

.....	.....	.....
Panagiotis Tsanakas	Dimitrios Soudris	Diomidis Spinellis
Professor NTUA	Professor NTUA	Professor AUEB

Athens, October 2024

.....
Apostolos S. Garos
Graduate of Electrical and Computer Engineering, NTUA
October 2024

Περίληψη

Η σταθερότητα των διευθύνσεων URL στις ακαδημαϊκές δημοσιεύσεις είναι ζωτικής σημασίας για τη διατήρηση της ακεραιότητας και της προσβασιμότητας του επιστημονικού έργου. Ωστόσο, η φθορά των διευθύνσεων URL - το φαινόμενο κατά το οποίο οι σύνδεσμοι καθίστανται ανενεργοί ή ανακατευθύνονται - αποτελεί σημαντική απειλή για την αξιοπιστία της έρευνας σε διάφορους επιστημονικούς κλάδους. Η παρούσα μελέτη διερευνά τη διαφύλαξη των διευθύνσεων URL στην επιστημονική βιβλιογραφία, με ιδιαίτερη έμφαση στις δημοσιεύσεις σχετικά με τη μηχανική λογισμικού (software engineering). Πραγματοποιώντας εκτεταμένη ανάλυση σχεδόν 50.000 ακαδημαϊκών εργασιών, η έρευνα αυτή αποκαλύπτει μια μέση ημιζωή (half-life) URL 9.68 ετών και ένα συνολικό ποσοστό αδράνειας URL 31,22%. Η μελέτη εξετάζει παράγοντες όπως οι πλατφόρμες φιλοξενίας, το μέσο δημοσίευσης και τα χαρακτηριστικά του domain, αποκαλύπτοντας μια σύνθετη συσχέτιση με τη σταθερότητα της διεύθυνσης URL, η οποία δείχνει ότι ορισμένα περιβάλλοντα ευνοούν περισσότερο τη μακροπρόθεσμη διαφύλαξη της διεύθυνσης από άλλα. Τα ευρήματα αυτά υπογραμμίζουν την πολυπλοκότητα της επίτευξης μακροπρόθεσμης ψηφιακής διατήρησης και υπογραμμίζουν την ανάγκη για βελτιωμένες πρακτικές και συνεργατικές προσπάθειες για τη διασφάλιση της προσβασιμότητας των διαδικτυακών πόρων στην ακαδημαϊκή έρευνα. Τα αποτελέσματα αυτής της έρευνας εκτείνονται πέρα από τη μηχανική λογισμικού, προσφέροντας ιδέες σχετικές με όλους τους τομείς που εξαρτώνται από τη μακροβιότητα των ψηφιακών αναφορών.

Λέξεις κλειδιά: Φθορά URL, Ψηφιακή διαφύλαξη, Αξιοπιστία URL στις ακαδημαϊκές δημοσιεύσεις, Δημοσιεύσεις στη μηχανική λογισμικού, Πλατφόρμες αρχειοθέτησης και διαφύλαξη URL, Ακεραιότητα επιστημονικής επικοινωνίας.

Abstract

The stability of URLs in academic publications is crucial for preserving the integrity and accessibility of scholarly work. However, URL rot — the phenomenon where links become inactive or redirect — poses a significant threat to the reliability of research across various disciplines. This study delves into the persistence of URLs in scholarly literature, with a particular focus on software engineering publications. By conducting an extensive analysis of nearly 50,000 academic papers, this research reveals an average URL half-life of 9.68 years and an overall URL inactivity rate of 31.22%. The study examines factors such as hosting platforms, publication venues, and domain characteristics, uncovering a varied correlation with URL stability, which indicates that some environments are more conducive to long-term URL preservation than others. These findings underscore the complexity of achieving long-term digital preservation and emphasize the necessity for improved strategies and collaborative efforts to safeguard the accessibility of online resources in academic research. The implications of this research extend beyond software engineering, offering insights relevant to all fields that depend on the longevity of digital references.

Keywords and Phrases: URL rot, Digital preservation, URL stability in academic publications, Link decay in scholarly literature, Software engineering references, Archival platforms and URL persistence, Scholarly communication integrity

Contents

List of Figures

Εκτεταμένη Περίληψη

Επισκόπηση

Η ραγδαία ανάπτυξη του ψηφιακού τοπίου τις τελευταίες δεκαετίες έχει επιφέρει σημαντικές αλλαγές στον τρόπο διάδοσης και πρόσβασης στην ακαδημαϊκή γνώση. Οι επιστήμονες και οι ερευνητές βασίζονται ολοένα και περισσότερο σε διαδικτυακούς πόρους, όπως αποθετήρια δεδομένων, λογισμικά ανοιχτού κώδικα, ηλεκτρονικά περιοδικά και ιστοσελίδες, για να ενισχύσουν τις δημοσιεύσεις τους και να προσφέρουν πρόσθετο περιεχόμενο στους αναγνώστες. Σε αυτό το πλαίσιο, οι σύνδεσμοι URL (Uniform Resource Locators) διαδραματίζουν καίριο ρόλο, παρέχοντας άμεση πρόσβαση σε αυτούς τους ψηφιακούς πόρους. Η αξιοπιστία και η διαθεσιμότητα αυτών των συνδέσμων είναι καθοριστικής σημασίας για τη συνεχή πρόσβαση και χρήση των παραπομπών που υποστηρίζουν τις επιστημονικές εργασίες.

Ωστόσο, η φθορά των συνδέσμων, γνωστή ως link decay ή URL rot, αποτελεί μια σοβαρή πρόκληση στην ακεραιότητα της επιστημονικής έρευνας. Με τον όρο φθορά συνδέσμων αναφερόμαστε στην κατάσταση κατά την οποία οι σύνδεσμοι που οδηγούσαν κάποτε σε συγκεκριμένους ψηφιακούς πόρους παύουν να λειτουργούν ή οδηγούν σε περιεχόμενο που δεν είναι πλέον διαθέσιμο. Οι αλλαγές στην αρχιτεκτονική του διαδικτύου, η μετακίνηση περιεχομένου χωρίς σωστή ανακατεύθυνση, η λήξη των ονομάτων τομέα και η διαγραφή ιστοσελίδων είναι μερικές από τις κύριες αιτίες αυτού του φαινομένου. Αυτή η δυσλειτουργία μπορεί να επηρεάσει βαθιά την αναπαραγωγιμότητα και τη διαχρονική αξιοπιστία της ακαδημαϊκής βιβλιογραφίας, καθώς καταστρέφει τη δυνατότητα των αναγνωστών να επαληθεύσουν τις πηγές ή να έχουν πρόσβαση στο πρωτότυπο υλικό.

Είναι σημαντικό να αναγνωριστεί ότι η φθορά των συνδέσμων δεν αποτελεί νέο πρόβλημα. Από τις πρώτες μέρες του διαδικτύου, μελέτες όπως αυτές του Koehler [14] έδειξαν ότι ένα σημαντικό ποσοστό των συνδέσμων που παρατίθενται σε ακαδημαϊκές δημοσιεύσεις καθίσταται ανενεργό μέσα σε λίγα χρόνια από τη δημοσίευση. Αυτό το φαινόμενο παραμένει διαχρονικό και, παρά τις τεχνολογικές βελτιώσεις, εξακολουθεί να αποτελεί ένα επίμονο πρόβλημα. Ιδιαίτερα στον τομέα της μηχανικής λογισμικού, όπου οι σύνδεσμοι προς ψηφιακά εργαλεία και αποθετήρια κώδικα αποτελούν ουσιαστικό μέρος της έρευνας, η φθορά των URL μπορεί να επηρεάσει δραματικά την ικανότητα των ερευνητών να επαληθεύσουν και να επαναλάβουν τις μελέτες.

Σε αυτό το πλαίσιο, η παρούσα εργασία εξετάζει τη φθορά των συνδέσμων στη βιβλιογραφία της μηχανικής λογισμικού, προσφέροντας μια συστηματική ανάλυση της σταθερότητας των URL, αναγνωρίζοντας τους παράγοντες που επηρεάζουν την επιμονή τους και προτείνοντας στρατηγικές για τη μείωση των αρνητικών επιπτώσεων της αποδόμησης των ψηφιακών αναφορών. Οι αναφορές αυτές όχι μόνο διευκολύνουν την πρόσβαση σε σημαντικά δεδομένα και κώδικα, αλλά επίσης διασφαλίζουν την ακεραιότητα της επιστημονικής διαδικασίας.

Κίνητρο της Έρευνας

Το κίνητρο πίσω από την παρούσα έρευνα βασίζεται στην αυξανόμενη ανησυχία για τη διαρκή σταθερότητα των διαδικτυακών αναφορών και τον αντίκτυπο της φθοράς των URL στην ακαδημαϊκή κοινότητα. Σε έναν κόσμο όπου η επιστημονική πρόοδος βασίζεται όλο και περισσότερο στη χρήση ψηφιακών πόρων, είναι κρίσιμο να διασφαλιστεί ότι οι σύνδεσμοι αυτοί παραμένουν λειτουργικοί με την πάροδο του χρόνου. Ειδικά στον τομέα της μηχανικής λογισμικού, όπου οι διαδικτυακοί πόροι, όπως τα αποθετήρια κώδικα, τα συστήματα συνεχούς ενσωμάτωσης και οι τεχνικές τεκμηριώσεις, είναι άρρηκτα συνδεδεμένα με την ερευνητική διαδικασία, η φθορά των URL μπορεί να επιφέρει σοβαρές συνέπειες.

Η αυξανόμενη εξάρτηση από ψηφιακές πλατφόρμες για την αποθήκευση και την ανταλλαγή δεδομένων έχει μετατρέψει τα URL σε θεμελιώδη στοιχεία της επιστημονικής βιβλιογραφίας. Ωστόσο, όταν οι σύνδεσμοι αυτοί παύουν να λειτουργούν, η πρόσβαση στο αρχικό περιεχόμενο καθίσταται δύσκολη ή αδύνατη. Αυτό όχι μόνο υπονομεύει την αξιοπιστία της έρευνας αλλά δημιουργεί και ένα χάσμα στην αναπαραγωγιμότητα των αποτελεσμάτων, κάτι που είναι ζωτικής σημασίας για την επιστημονική μέθοδο.

Παρόλο που έχουν προταθεί λύσεις, όπως η χρήση μόνιμων ταυτοποιητών (DOI), τα οποία διασφαλίζουν την διατήρηση της προσβασιμότητας των ψηφιακών αναφορών, η φθορά των URL παραμένει διαδεδομένο πρόβλημα. Η έλλειψη αναλυτικών μελετών που να εξετάζουν τη σταθερότητα των URL στη βιβλιογραφία της μηχανικής λογισμικού καθιστά αναγκαία την ανάπτυξη εμπεριστατωμένων στρατηγικών που να αντιμετωπίζουν το πρόβλημα σε βάθος. Η επιμονή αυτών των αναφορών αποτελεί ακρογωνιαίο λίθο για τη διατήρηση της επιστημονικής συνέχειας, και, ως εκ τούτου, απαιτείται μια βαθύτερη κατανόηση των παραγόντων που οδηγούν στη φθορά τους.

Η παρούσα έρευνα προέκυψε ως απάντηση σε αυτήν την ανάγκη, επιχειρώντας να γεφυρώσει το υπάρχον κενό στη βιβλιογραφία και να προσφέρει πρακτικές λύσεις για τη βελτίωση της διατήρησης των ψηφιακών αναφορών. Μέσω της ανάλυσης χιλιάδων ακαδημαϊκών εργασιών και εκατοντάδων χιλιάδων URL, αυτή η μελέτη επιδιώκει να ρίξει φως στο μέγεθος του προβλήματος και να αναδείξει τις προκλήσεις που παραμένουν.

Στόχοι της Έρευνας

Ο πρωταρχικός στόχος αυτής της εργασίας είναι η συστηματική διερεύνηση της διαφύλαξης των URL στις επιστημονικές εργασίες μηχανικής λογισμικού και η αναγνώριση των βασικών παραγόντων που επηρεάζουν τη σταθερότητά τους. Αυτή η διερεύνηση πραγματοποιείται μέσω της ανάλυσης σχεδόν 50.000 ακαδημαϊκών εργασιών, από τις οποίες εξάγονται και εξετάζονται πάνω από 130.000 URL. Η μελέτη επιδιώκει:

- **Ποσοτικοποίηση της φθοράς των URL:** Καθορισμός του βαθμού αποδόμησης των URL στη βιβλιογραφία της μηχανικής λογισμικού μέσω του υπολογισμού της μέσης ημιζωής των URL και του συνολικού ποσοστού ανενεργών URL.

- **Ανάλυση Παράγοντων Επιρροής:** Εξέταση διάφορων παραγόντων που μπορεί να επηρεάζουν τη σταθερότητα των URL, όπως ο τύπος των πλατφορμών φιλοξενίας, τα χαρακτηριστικά των δημοσιεύσεων και τα ειδικά χαρακτηριστικά των τομέων. Στόχος είναι η αναγνώριση μοτίβων που μπορούν να εξηγήσουν τις διαφοροποιήσεις στην επιμονή των URL.

- **Αξιολόγηση της Συσχέτισης με Παραδοσιακούς Δείκτες:** Διερεύνηση της σχέσης μεταξύ της σταθερότητας των URL και των παραδοσιακών ακαδημαϊκών δεικτών όπως ο δείκτης απήχησης των περιοδικών (journal impact factor) και η κατάταξη CORE, για να διαπιστωθεί εάν αυτοί οι δείκτες σχετίζονται με την πιθανότητα επιμονής των URL.

- **Αναγνώριση Αναδυόμενων Προκλήσεων:** Ανάδειξη νέων εμποδίων στην επαλήθευση των URL, όπως η επίδραση των μηχανισμών προστασίας τομέων όπως το Cloudflare, που δυσκολεύουν την ακριβή αξιολόγηση της διαθεσιμότητας των URL.

Σχετική Βιβλιογραφία

Η φθορά των συνδέσμων (URL rot) έχει μελετηθεί μερικώς τα τελευταία χρόνια, με μελέτες να εξετάζουν τη σταθερότητα και τη μακροχρόνια επιμονή των συνδέσμων που χρησιμοποιούνται σε ακαδημαϊκές δημοσιεύσεις. Οι πρώτες έρευνες που αναδείκνυαν το ζήτημα πραγματοποιήθηκαν στα τέλη της δεκαετίας του 1990. Η μελέτη του Koehler [14], μία από τις πλέον επιδραστικές, ανέδειξε ότι η ημιζωή των URL, δηλαδή ο χρόνος που απαιτείται για να γίνει ανενεργό το 50% των συνδέσμων, κυμαινόταν στα δύο χρόνια. Η έρευνα αυτή προσέλκυσε σημαντική προσοχή στην ακαδημαϊκή κοινότητα και ενέπνευσε περαιτέρω έρευνες σχετικά με τη φθορά των συνδέσμων.

Στη μελέτη τους, οι Lawrence et al. [17] ανέλυσαν πάνω από 67 577 συνδέσμους από περίπου 100 826 άρθρα στον τομέα της επιστήμης υπολογιστών και βρήκαν ότι το ποσοστό ανενεργών συνδέσμων έφτανε έως και το 53% για δημοσιεύσεις που εκδόθηκαν το 1994. Μετά από προσπάθειες διόρθωσης των συνδέσμων, το ποσοστό αυτό μειώθηκε στο 2,9%, υπογραμμίζοντας τη σημασία της ενεργής διαχείρισης και της τακτικής επαλήθευσης των URL.

Παρόμοια, ο Spinellis [31] ανέλυσε 4224 συνδέσμους από δημοσιεύσεις στους τομείς της επιστήμης υπολογιστών, χωρίς να διορθώσει πιθανά λάθη σύνταξης ή γραμματικά λάθη στους συνδέσμους. Η μελέτη κατέδειξε ότι η μέση ημιζωή των συνδέσμων ήταν τέσσερα έτη, υπογραμμίζοντας τις προκλήσεις που υπάρχουν στη διατήρηση της μακροπρόθεσμης σταθερότητας των συνδέσμων.

Μελέτες όπως του Wren [38] στον τομέα της βιοϊατρικής και του Wagner [36] στη διαχείριση της υγείας έδειξαν ότι η φθορά των URL δεν περιορίζεται σε έναν τομέα, αλλά είναι διαδεδομένη σε διάφορους επιστημονικούς κλάδους. Το πρόβλημα της φθοράς των συνδέσμων αγγίζει περίπου το 22% των αναφορών στη βιοϊατρική και φτάνει το 49,3% στη διαχείριση της υγείας, γεγονός που αποδεικνύει την ευρύτητα του ζητήματος.

Σημαντικές προόδους στην αντιμετώπιση της φθοράς των συνδέσμων έχουν γίνει μέσω ψηφιακών υπηρεσιών όπως το LOCKSS (Lots of Copies Keep Stuff Safe) και το CLOCKSS (Controlled LOCKSS), οι οποίες δημιουργούν πολλαπλά αντίγραφα ψηφιακού περιεχομένου σε διαφορετικές τοποθεσίες, διασφαλίζοντας την πρόσβαση ακόμα και όταν κάποιος σύνδεσμος καταστεί ανενεργός [5]. Αυτά τα συστήματα διαδραματίζουν κρίσιμο ρόλο στη μακροπρόθεσμη διατήρηση των ψηφιακών πόρων και στη μείωση των επιπτώσεων της φθοράς των URL.

Η έρευνά μας βασίζεται σε αυτές τις προηγούμενες μελέτες, εστιάζοντας στις ιδιαιτερότητες της φθοράς των συνδέσμων στον τομέα της μηχανικής λογισμικού, όπου οι παραπομπές σε αποθετήρια κώδικα και τεχνικά έγγραφα είναι κρίσιμες για την επαλήθευση και την επαναληψιμότητα των ερευνών. Επιπλέον, εξετάζουμε τις στρατηγικές ψηφιακής διατήρησης που εφαρμόζονται για να διασφαλίσουν την μακροπρόθεσμη πρόσβαση στους ψηφιακούς πόρους, και διερευνούμε τις οικονομικές πτυχές της βιωσιμότητας αυτών των στρατηγικών.

Μεθοδολογία

Η μεθοδολογία της παρούσας εργασίας βασίζεται σε μια πολυεπίπεδη προσέγγιση, η οποία περιλαμβάνει τη συλλογή, επεξεργασία, ανάλυση και επαλήθευση δεδομένων, με στόχο τη διερεύνηση της επιμονής των URL στις επιστημονικές δημοσιεύσεις στον τομέα της μηχανικής λογισμικού. Το σχέδιο της έρευνας είναι διαμορφωμένο έτσι ώστε να διασφαλίσει την ακεραιότητα, την επαναληψιμότητα και τη διαφάνεια των ευρημάτων.

Συλλογή Δεδομένων

Η συλλογή δεδομένων ξεκινά από την καθιερωμένη βιβλιογραφική βάση DBLP (Digital Bibliography & Library Project), η οποία παρέχει ένα ευρύ φάσμα βιβλιογραφικών πληροφοριών για τις επιστημονικές δημοσιεύσεις στον τομέα της πληροφορικής. Το DBLP αποτελεί μια αξιόπιστη και ενημερωμένη πηγή, με πάνω από πέντε εκατομμύρια καταχωρίσεις που καλύπτουν δημοσιεύσεις από επιστημονικά περιοδικά και συνέδρια.

Για τους σκοπούς αυτής της έρευνας, τα δεδομένα προήλθαν από το DBLP με τη μορφή του αρχείου `dblp.xml`, το οποίο περιλαμβάνει μεταδεδομένα σχετικά με τις δημοσιεύσεις, όπως τους συγγραφείς, τους τίτλους, τους χώρους δημοσίευσης και, κυρίως, τους ηλεκτρονικούς συνδέσμους (ee links) προς το πλήρες κείμενο των εργασιών. Εστιάζουμε στις δημοσιεύσεις που αφορούν τη μηχανική λογισμικού, λαμβάνοντας υπόψη συνέδρια και περιοδικά με υψηλή επιστημονική επίδραση, καλύπτοντας χρονολογίες από το 1971 έως το 2023.

Συνολικά, επιλέχθηκαν εργασίες μηχανικής λογισμικού για περαιτέρω ανάλυση. Τα PDF αυτών των εργασιών συλλέχθηκαν χρησιμοποιώντας το εργαλείο doi2pdf, το οποίο αυτοματοποιεί τη λήψη ακαδημαϊκών εργασιών μέσω των αναγνωριστικών DOI (Digital Object Identifier). Η προσαρμοσμένη έκδοση του εργαλείου χρησιμοποιεί εναλλακτικές πηγές σε περίπτωση αποτυχίας λήψης από την κύρια πηγή, ενώ χρησιμοποιεί την πλατφόρμα arXiv για προδημοσιεύσεις, όταν αυτό είναι αναγκαίο. Αυτή η προσέγγιση διασφάλισε υψηλό ποσοστό επιτυχίας, με τη λήψη αρχείων PDF, καλύπτοντας το 89.3% των επιλεγμένων εργασιών.

Τα αρχεία PDF υποβλήθηκαν σε επεξεργασία με τη χρήση της βιβλιοθήκης Grobid (GeneRation Of BIbliographic Data), μια μηχανική εκμάθηση για την εξαγωγή και ανάλυση βιβλιογραφικών δεδομένων σε μορφή XML. Το Grobid μετατρέπει τα PDF σε XML, διατηρώντας τη δομή των εγγράφων, που είναι ουσιαστική για τις επόμενες αναλύσεις. Η επιλογή του Grobid στηρίζεται στην υψηλή του απόδοση στην εξαγωγή περιεχομένου από ακαδημαϊκά έγγραφα, όπως αποδεικνύεται από τις εκτενείς αξιολογήσεις. Επιτύχαμε υψηλό ποσοστό επιτυχίας στη μετατροπή, με αρχεία XML, που αντιπροσωπεύουν το 97.1% των ληφθέντων αρχείων PDF.

Η εξαγωγή των ηλεκτρονικών συνδέσμων πραγματοποιήθηκε μέσω ενός προσαρμοσμένου Python σκριπτ, το οποίο χρησιμοποιεί κανονικές εκφράσεις για τον εντοπισμό και την εξαγωγή προτύπων URL από τα κείμενα. Δόθηκε ιδιαίτερη προσοχή σε τυπογραφικές ανωμαλίες και μη τυποποιημένες μορφές URL, και με ειδικούς μηχανισμούς επιτεύχθηκε ο καθαρισμός και η κανονικοποίηση των συνδέσμων, διασφαλίζοντας τη συντακτική ορθότητά τους για τις επόμενες φάσεις επαλήθευσης. Με αυτή την προσεκτική μεθοδολογία διασφαλίστηκε η ακεραιότητα του συνόλου δεδομένων URL, θέτοντας μια ισχυρή βάση για τη σωστή επαλήθευση και ανάλυση.

Για την αποθήκευση των δεδομένων, χρησιμοποιήθηκε το PostgreSQL, το οποίο προσφέρει σταθερότητα και υποστήριξη σύνθετων ερωτημάτων. Το σχήμα της βάσης δεδομένων περιλαμβάνει πίνακες για τη διαχείριση των δεδομένων που συλλέχθηκαν, όπως φαίνεται στο Σχήμα 4.2, επιτρέποντας την αποδοτική αναζήτηση και ανάλυση των δεδομένων.

Ανάλυση Δεδομένων

Η ανάλυση δεδομένων επικεντρώνεται στην αυτοματοποιημένη επαλήθευση και αξιολόγηση της διατήρησης των συνδέσμων URL που εξήχθησαν από τις επιλεγμένες δημοσιεύσεις. Χρησιμοποιήθηκαν τεχνικές ανάλυσης συνδέσμων και ανίχνευσης δικτυακών σφαλμάτων για να προσδιοριστεί η κατάσταση κάθε URL, καθώς και αναλύσεις παλινδρόμησης για να διερευνηθούν οι παράγοντες που επηρεάζουν τη διατήρησή τους.

Αρχικά, όλοι οι σύνδεσμοι URL υποβλήθηκαν σε έλεγχο ορθότητας σύμφωνα με τα σχετικά πρότυπα του διαδικτύου (RFC 1738 και διάδοχοι), εξαλείφοντας κακοσχηματισμένα URL που δεν θα μπορούσαν

να είναι ενεργά. Για την αυτοματοποιημένη επαλήθευση των συνδέσμων, χρησιμοποιήθηκε το εργαλείο ανοιχτού κώδικα curl-cffi, το οποίο επέτρεψε την αυτόματη πρόσβαση σε κάθε σύνδεσμο και την παρακολούθηση τυχόν ανακατευθύνσεων HTTP. Οι προσπάθειες αυτές καταγράφηκαν στη βάση δεδομένων με πληροφορίες όπως οι κωδικοί κατάστασης HTTP και τυχόν δικτυακά σφάλματα.

Για την ενίσχυση της αξιοπιστίας της διαδικασίας επαλήθευσης, υιοθετήθηκε ένας μηχανισμός επαναπροσπάθειας για συνδέσμους που δεν ανταποκρίθηκαν εντός 10 δευτερολέπτων. Εφαρμόστηκε καθυστέρηση 45 δευτερολέπτων πριν από κάθε νέα προσπάθεια, με ανώτατο όριο τρεις επαναλήψεις, ώστε να αντιμετωπιστούν προσωρινά ζητήματα δικτύου. Οι σύνδεσμοι που επέστρεψαν απάντηση HTTP 200 μετά από όλες τις ανακατευθύνσεις και επαναλήψεις καταγράφηκαν ως ενεργοί, ενώ οι υπόλοιποι θεωρήθηκαν ανενεργοί.

Προκειμένου να διασφαλιστεί η ακρίβεια της αυτοματοποιημένης επαλήθευσης, πραγματοποιήθηκε δειγματοληπτικός χειροκίνητος έλεγχος 200 συνδέσμων. Κατά τη διάρκεια αυτού του ελέγχου, επαληθεύτηκαν οι σύνδεσμοι που είχαν χαρακτηριστεί ως ενεργοί για να διαπιστωθεί αν οδηγούσαν στο αναμενόμενο περιεχόμενο. Οι έλεγχοι αυτοί αποκάλυψαν ότι το 94% των συνδέσμων ήταν όντως ενεργοί και σχετικοί, επιβεβαιώνοντας την αξιοπιστία της αυτοματοποιημένης διαδικασίας. Ωστόσο, εντοπίστηκε ένα μικρό ποσοστό συνδέσμων που είχαν καταστεί ανενεργοί ή οδηγούσαν σε μη σχετικό περιεχόμενο, υπογραμμίζοντας τη σημασία των περιοδικών χειροκίνητων ελέγχων.

Επιπλέον, πραγματοποιήθηκαν αναλύσεις παλινδρόμησης για να εξεταστούν οι παράγοντες που επηρεάζουν τη διατήρηση των συνδέσμων URL. Χρησιμοποιήθηκε ένα μοντέλο γραμμικής παλινδρόμησης, το οποίο επέτρεψε την ανάλυση της επίδρασης διαφόρων παραμέτρων, όπως οι κατατάξεις των περιοδικών CORE, οι Journal Impact Factors και το μήκος των συνδέσμων URL σε χαρακτήρες και συνιστώσες διαδρομής.

Εργαλεία και Τεχνολογίες

Για την πραγματοποίηση της μελέτης μας, χρησιμοποιήθηκε ένας αριθμός εργαλείων και τεχνολογιών που επέλεξαν με βάση την αξιοπιστία τους και την καταλληλότητά τους για τη διαχείριση των διαφόρων πτυχών της έρευνάς μας. Κάθε εργαλείο έπαιξε κρίσιμο ρόλο στη διασφάλιση της ακρίβειας και της αποδοτικότητας στη συλλογή, επεξεργασία και ανάλυση των δεδομένων.

PostgreSQL: Για την αποθήκευση και διαχείριση των δεδομένων, επιλέχθηκε η βάση δεδομένων PostgreSQL, ένα ισχυρό σύστημα σχεσιακών βάσεων δεδομένων ανοικτού κώδικα. Η PostgreSQL είναι γνωστή για τα προηγμένα χαρακτηριστικά της, όπως η υποστήριξη σύνθετων ερωτημάτων και η εξασφάλιση της ακεραιότητας των δεδομένων. Το ACID-compliant σύστημά της την καθιστά ιδανική για τη διαχείριση μεγάλων συνόλων δεδομένων, όπως αυτά της μελέτης μας, και διευκόλυνε την αποθήκευση των δεδομένων σε δομημένους πίνακες όπως οι `papers`, `urls`, και `venues`.

doi2pdf: Για την αυτοματοποίηση της λήψης των αρχείων PDF, χρησιμοποιήθηκε το εργαλείο doi2pdf, ένα εργαλείο που βασίζεται στη γλώσσα Python και ανακτά ακαδημαϊκές εργασίες χρησιμοποιώντας τα αναγνωριστικά DOI. Η προσαρμοσμένη έκδοση που χρησιμοποιήσαμε περιελάμβανε βελτιώσεις για τη διαχείριση εναλλακτικών πηγών λήψης σε περίπτωση αποτυχίας της κύριας πηγής, καθώς και τη χρήση της πλατφόρμας arXiv για τον εντοπισμό προδημοσιεύσεων.

Grobid: Το Grobid (GeneRation Of BIbliographic Data) χρησιμοποιήθηκε για τη μετατροπή των PDF σε δομημένη μορφή XML. Το Grobid χρησιμοποιεί μηχανική εκμάθηση για την εξαγωγή και ανάλυση του περιεχομένου των ακαδημαϊκών εγγράφων, διατηρώντας σημαντικές πληροφορίες τμηματοποίησης των εγγράφων που ήταν απαραίτητες για την εξαγωγή των URL. Η χρήση του σχήματος TEI (Text Encoding Initiative) διασφαλίζει ότι τα δεδομένα είναι τόσο μηχανικά αναγνώσιμα όσο και κατανοητά από τον άνθρωπο.

Python: Αναπτύχθηκαν διάφορα προσαρμοσμένα Python scripts για την υποστήριξη των διαφόρων σταδίων της μελέτης μας. Τα scripts αυτά αυτοματοποίησαν εργασίες όπως η εξαγωγή και

ανάλυση των URL, ο έλεγχος της συντακτικής ορθότητας των συνδέσμων και η διαχείριση του μηχανισμού επαναπροσπάθειας κατά την επαλήθευση των URL. Η ευκολία χρήσης και οι ισχυρές βιβλιοθήκες της Python την καθιστούν ιδανική επιλογή για την ανάπτυξη τέτοιων εργαλείων.

curl_cffi: Η βιβλιοθήκη curl_cffi χρησιμοποιήθηκε εκτενώς για την αυτοματοποίηση της επαλήθευσης των συνδέσμων URL. Η βιβλιοθήκη αυτή παρέχει διασύνδεση της Python με το εργαλείο curl, υποστηρίζοντας πρωτόκολλα όπως το HTTP και το HTTPS. Το εργαλείο αυτό μας επέτρεψε να ελέγξουμε συστηματικά την κατάσταση των URL, καταγράφοντας τις απαντήσεις και τα σφάλματα δικτύου.

pandas and matplotlib: Για την ανάλυση και απεικόνιση των δεδομένων, χρησιμοποιήθηκαν οι βιβλιοθήκες pandas και matplotlib της Python. Η pandas προσφέρει ισχυρές δομές δεδομένων και λειτουργίες που επέτρεψαν την αποδοτική διαχείριση, καθαρισμό και μετασχηματισμό των δεδομένων. Η matplotlib χρησιμοποιήθηκε για τη δημιουργία διαγραμμάτων που απεικόνισαν τις αναλύσεις μας, παρέχοντας σαφή και κατανοητή παρουσίαση των δεδομένων.

SQLAlchemy: Για τη σύνδεση των Python σεριπτς με τη βάση δεδομένων PostgreSQL, χρησιμοποιήθηκε η βιβλιοθήκη SQLAlchemy. Αυτή η βιβλιοθήκη διευκόλυνε την αλληλεπίδραση με τη βάση δεδομένων μέσω Python κώδικα, επιτρέποντας την αποδοτική εκτέλεση ερωτημάτων και τη διαχείριση των δεδομένων.

Απειλές στην Εγκυρότητα

Η ενότητα αυτή εξετάζει τις πιθανές περιοριστικές παραμέτρους της μελέτης, οι οποίες κατηγοριοποιούνται σε εσωτερική εγκυρότητα, εξωτερική εγκυρότητα, κατασκευαστική εγκυρότητα και αξιοπιστία. Κάθε κατηγορία αφορά διαφορετικές πτυχές του σχεδιασμού και της εκτέλεσης της μελέτης που ενδέχεται να επηρεάσουν τα αποτελέσματα.

Εσωτερική Εγκυρότητα: Η εσωτερική εγκυρότητα σχετίζεται με την ακρίβεια και αξιοπιστία των διαδικασιών εξαγωγής και επαλήθευσης των συνδέσμων URL. Παρά τη χρήση ολοκληρωμένων αυτοματοποιημένων τεχνικών, όπως έλεγχοι σχημάτων, φιλτράρισμα μέσω βάσεων δεδομένων και επαλήθευση κωδικών κατάστασης HTTP, υπάρχει πάντα η πιθανότητα σφαλμάτων κατά την εξαγωγή δεδομένων. Λανθασμένη ερμηνεία συνδέσμων URL λόγω προβλημάτων μορφοποίησης ή εσφαλμένης ανάλυσης και οι εγγενείς περιορισμοί της αυτοματοποιημένης επαλήθευσης (π.χ. αδυναμία επαλήθευσης της συνάφειας του περιεχομένου) θα μπορούσαν να επηρεάσουν τα αποτελέσματα. Για τη μείωση αυτού του κινδύνου, πραγματοποιήθηκε χειροκίνητος έλεγχος ενός σημαντικού υποσυνόλου των ανενεργών συνδέσμων. Παρά την χρησιμότητα αυτού του ελέγχου, η χειροκίνητη επαλήθευση για ενεργούς συνδέσμους κρίθηκε περιττή, καθώς συνδέσεις που δεν είναι προσβάσιμες μηχανικά δεν μπορούν να ελεγχθούν και από τον άνθρωπο. Ωστόσο, η διαδικασία χειροκίνητης επαλήθευσης ενδέχεται να εισαγάγει υποκειμενικές προκαταλήψεις, ιδιαίτερα για συνδέσμους που οδηγούν σε γενικές σελίδες ή έχουν ασαφείς διαδρομές, καθώς βασίζεται στην κρίση για τη συνάφεια του περιεχομένου.

Εξωτερική Εγκυρότητα: Η εξωτερική εγκυρότητα αναφέρεται στην ικανότητα γενίκευσης των ευρημάτων της μελέτης. Η παρούσα μελέτη επικεντρώνεται αποκλειστικά σε δημοσιεύσεις στον τομέα της μηχανικής λογισμικού, όπως αυτές καταχωρίζονται στο DBLP. Αν και το DBLP αποτελεί μια αξιόπιστη βιβλιογραφική βάση, είναι πιθανό να υπάρχουν διαφορετικά μοτίβα σταθερότητας συνδέσμων σε πιο εξειδικευμένα επιστημονικά περιοδικά ή συνέδρια. Ενώ η μελέτη δεν διεκδικεί γενίκευση πέρα από το συγκεκριμένο πλαίσιο της, θα ήταν έκπληξη αν παρατηρούνταν σημαντικά διαφορετικά μοτίβα σε άλλους τομείς της πληροφορικής.

Κατασκευαστική Εγκυρότητα: Η κατασκευαστική εγκυρότητα αφορά τον βαθμό στον οποίο η μελέτη μετρά αυτό που προτίθεται να μετρήσει. Στην παρούσα μελέτη, οι ορισμοί των ¨ενεργών' και ¨ανενεργών' συνδέσμων βασίζονται κυρίως σε αυτοματοποιημένα κριτήρια, όπως οι κωδικοί κατάστασης HTTP και τα δικτυακά σφάλματα. Αυτή η προσέγγιση ενδέχεται να μην καταγράφει πλήρως τις

λεπτές πτυχές της λειτουργικότητας ενός συνδέσμου, ειδικά σε περιπτώσεις όπου οι σύνδεσμοι ανακατευθύνονται σε περιεχόμενο σχετικό με το πεδίο αλλά όχι στο αρχικά αναμενόμενο. Η εξάρτηση από αυτοματοποιημένα εργαλεία για την εξαγωγή και επαλήθευση των συνδέσμων μπορεί να παραβλέψει λεπτομέρειες σχετικά με την ενεργότητα και τη συνάφεια των συνδέσμων.

Αξιοπιστία: Η αξιοπιστία σχετίζεται με τη συνέπεια των αποτελεσμάτων της μελέτης. Η μελέτη αυτή χρησιμοποιεί μια πολυδιάστατη διαδικασία για τη συλλογή δεδομένων (π.χ., βιβλιογραφικές πληροφορίες, εργασίες) και την επαλήθευση της σταθερότητας των συνδέσμων (π.χ., εξαγωγή συνδέσμων, έλεγχοι συνδέσμων). Παρά την απλότητα κάθε βήματος, πάντα υπάρχει η πιθανότητα εμφάνισης σφαλμάτων στην υλοποίηση του συστήματος.

Αποτελέσματα

Σταθερότητα συνδέσμων URL με την πάροδο του χρόνου

Παρουσιάστηκε μια επισκόπηση της σταθερότητας των συνδέσμων URL μεταξύ 1995 και 2023. Παρατηρήθηκε ότι οι σύνδεσμοι των πιο πρόσφατων δημοσιεύσεων διατηρούνται ενεργοί σε μεγαλύτερο ποσοστό σε σύγκριση με παλαιότερες δημοσιεύσεις. Το ποσοστό των ενεργών URL για το 2023 ήταν 83.77%, ενώ για το 1996 ήταν μόλις 15.97%. Ο υπολογισμός του χρόνου ημιζωής των URL, δηλαδή ο χρόνος που απαιτείται για να γίνει ανενεργό το 50% των URL που δημοσιεύθηκαν σε ένα συγκεκριμένο έτος, έδειξε ότι η ημιζωή των URL είναι 9.68 χρόνια.

Σταθερότητα συνδέσμων URL ανάλογα με τον σκοπό φιλοξενίας

Οι σύνδεσμοι που φιλοξενούνται σε πλατφόρμες ελέγχου εκδόσεων (π.χ. GitHub, Bitbucket) ή σε ψηφιακά αρχεία (Zenodo, arXiv) είναι πιο σταθεροί σε σύγκριση με το σύνολο των URL. Παρά τη σχετική σταθερότητα των συνδέσμων αρχείων, η χρήση τους παραμένει περιορισμένη (22.000 σύνδεσμοι αρχείων έναντι 136.000 συνδέσμων ελέγχου εκδόσεων).

Σταθερότητα συνδέσμων URL ανάλογα με τον χώρο δημοσίευσης

Η σταθερότητα των URL παρουσιάζει σημαντική διακύμανση μεταξύ των επιστημονικών συνεδρίων και περιοδικών, με ποσοστά ενεργών συνδέσμων που κυμαίνονται από 54% έως 81%. Παρατηρήθηκε ότι τα κορυφαία επιστημονικά συνέδρια και περιοδικά της μηχανικής λογισμικού (π.χ. FSE, ASE, ICSE) βρίσκονται στην πρώτη μισή του πίνακα σταθερότητας.

Σταθερότητα συνδέσμων URL ανά τμήμα εγγράφου

Οι σύνδεσμοι που εμφανίζονται σε περιλήψεις επιστημονικών άρθρων (abstracts) παρουσιάζουν το υψηλότερο ποσοστό σταθερότητας (80.82%), ενώ οι σύνδεσμοι στις αναφορές (bibliography) παρουσιάζουν το χαμηλότερο ποσοστό (68.01%).

Σταθερότητα συνδέσμων URL ανά πρωτόκολλο (scheme)

Οι σύνδεσμοι URL αποτελούνται από διάφορα τμήματα, με το πρωτόκολλο να αποτελεί το πρώτο στοιχείο κάθε συνδέσμου. Στην ανάλυση μας εξετάζονται δύο κύρια πρωτόκολλα, το http και το https, τα οποία χρησιμοποιούνται σε σύγχρονες διαδικτυακές αναφορές. Το πρωτόκολλο https προσφέρει κρυπτογραφημένη πρόσβαση, προστατεύοντας από επιθέσεις μέσω του πρωτοκόλλου TLS. Παρά τις αρχικές προσδοκίες ότι οι https σύνδεσμοι μπορεί να είναι πιο ευαίσθητοι σε προβλήματα, η ανάλυση έδειξε το αντίθετο.

- http: 58.67%
- https: 81.58%

Η μεγαλύτερη σταθερότητα των https συνδέσμων μπορεί να αποδοθεί στο γεγονός ότι οι https συνδέσμοι είναι νεότεροι, καθώς το πρωτόκολλο αυτό υιοθετήθηκε αργότερα σε σχέση με το http. Επίσης, είναι πιθανό οι συνδέσμοι https να σχετίζονται με καλύτερες πρακτικές αναφοράς που εφαρμόζονται πιο πρόσφατα.

Σταθερότητα συνδέσμων URL ανά ανώτατο τομέα (TLD)

Η ανάλυση των ποσοστών ενεργών συνδέσμων URL ανά ανώτατο τομέα (TLD) αποκάλυψε σημαντικές διαφορές στη σταθερότητα. Οι τομείς που περιλαμβάνουν περισσότερους από 50 συνδέσμους εξετάστηκαν για την αποφυγή στρεβλών αποτελεσμάτων λόγω μικρών δειγμάτων. Τα αποτελέσματα φαίνονται στην ακόλουθη λίστα.

1. .press: 100.0%
2. .dev: 89.44%
3. .blog: 87.13%
4. .cc: 86.5%
5. .to: 85.25%

Οι τομείς .press, .dev, και .blog παρουσιάζουν τα υψηλότερα ποσοστά ενεργών συνδέσμων, ενώ άλλοι τομείς όπως .io και .ly (δεν φαίνονται στη λίστα) παραμένουν σταθεροί αλλά σε χαμηλότερα ποσοστά.

Σταθερότητα συνδέσμων URL ανά τομέα (domain)

Η περαιτέρω ανάλυση επικεντρώθηκε σε συγκεκριμένους τομείς (domains) και τα ποσοστά ανενεργών συνδέσμων. Στην παρακάτω λίστα παρουσιάζονται τα ποσοστά ανενεργών συνδέσμων για τους κύριους τομείς, με βάση τον απόλυτο αριθμό ανενεργών συνδέσμων.

1. nasa.gov: 72.08%
2. cmu.edu: 70.59%
3. mit.edu: 53.23%
4. stanford.edu: 51.32%
5. mozilla.org: 45.54%

Ο τομέας nasa.gov παρουσίασε το υψηλότερο ποσοστό ανενεργών συνδέσμων (72.08%), ενώ το cmu.edu και το mit.edu ακολούθησαν με ποσοστά άνω του 50%.

Σφάλματα HTTP

Η ανάλυση των σφαλμάτων HTTP περιλαμβάνει την κατηγοριοποίηση των συνδέσμων URL με βάση τους κωδικούς κατάστασης HTTP που έλαβαν κατά τον έλεγχο προσβασιμότητας. Παρακάτω παρουσιάζεται η κατανομή αυτών των σφαλμάτων.

- Not Found (404): 74.65%
- Forbidden (403): 18.16%
- Internal Server Error (500): 1.52%
- Service Unavailable (503): 0.91%
- Gone (410): 0.8%
- Other: 3.95%

Η ανάλυση αποκάλυψε ότι τα σφάλματα 404 Not Found ήταν τα πιο κοινά, ακολουθούμενα από σφάλματα 403 Forbidden και 500 Internal Server Error.

Ανάλυση σφαλμάτων δικτύου

Η ανάλυση των σφαλμάτων δικτύου επικεντρώθηκε σε ζητήματα συνδεσιμότητας που αντιμετωπίστηκαν κατά την πρόσβαση σε συνδέσμους URL. Στη μελέτη αυτή, τα σφάλματα δικτύου που καταγράφηκαν κατά τους ελέγχους προσβασιμότητας των συνδέσμων URL είναι συγκεκριμένα σφάλματα UNIX. Αυτά τα σφάλματα προσφέρουν πληροφορίες για τις τεχνικές προκλήσεις που αντιμετωπίζονται κατά την πρόσβαση σε συνδέσμους από συστήματα τύπου UNIX, τα οποία χρησιμοποιήθηκαν στις αυτοματοποιημένες διαδικασίες επαλήθευσης URL σε αυτή την εργασία.

- Couldn't resolve host (6): 62.49%
- Operation timeout (28): 21.4%
- Certificate not verified with known CAs (60): 8.73%
- Failed to connect to host (7): 3.29%
- Unsupported protocol (1): 1.48%
- Other: 2.61%

Τα πιο συνηθισμένα σφάλματα αφορούσαν προβλήματα DNS και συνδέσεις που δεν μπόρεσαν να επιτευχθούν.

Ανάλυση παλινδρόμησης

Η ανάλυση παλινδρόμησης εντόπισε τις σχέσεις μεταξύ πολλαπλών παραγόντων και της διατήρησης των συνδέσμων URL στις ακαδημαϊκές δημοσιεύσεις. Από παρατήρηση των αποτελεσμάτων, δεν φαίνεται να υπάρχει κάποια σημαντική συσχέτιση της επιτυχίας των URLs με τους παράγοντες που εξετάστηκαν.

- Μήκος URL: -0.001

- Αριθμός στοιχείων διαδρομής: -0.015
- Journal Impact Factor: 0.016
- Κατάταξη CORE: 0.028

Συζήτηση

Σύγκριση με Προηγούμενες Μελέτες

Η παρούσα μελέτη, αξιοποιώντας ένα σημαντικό σύνολο δεδομένων, συνεισφέρει στην εξέλιξη της κατανόησης της διατήρησης των συνδέσμων URL στην επιστημονική βιβλιογραφία. Η ανάλυση αυτή διαφοροποιείται από προηγούμενες μελέτες, τόσο λόγω του ευρύτερου συνόλου δεδομένων που χρησιμοποιήθηκε όσο και λόγω της πιο περιεκτικής προσέγγισης που υιοθετήθηκε για την κατάταξη των ενεργών συνδέσμων URL. Η απόφαση να ταξινομηθούν οι σύνδεσμοι ως ενεργοί όταν δεν υπήρξαν σφάλματα ευθυγραμμίζεται με το μέγεθος του συνόλου δεδομένων, καθιστώντας αυτή την προσέγγιση πρακτική και αναγκαία.

Η σύγκριση των ποσοστών προσβασιμότητας των URL ανά έτος αποκαλύπτει σημαντικές προόδους στη διαθεσιμότητα ψηφιακών πόρων. Για παράδειγμα, η μελέτη μας κατέγραψε ποσοστό ενεργών URL 83.77% για το 2023, σε αντίθεση με το 15.97% που καταγράφηκε το 1996.

Επιπλέον, το γενικό ποσοστό ενεργών URL της μελέτης μας, 68.78%, συμφωνεί με προηγούμενες μελέτες, όπως αυτή των Casserly and Bird [3], η οποία ανέφερε ποσοστά μεταξύ 56.4% και 81.4%. Παρόλο που αυτά τα ποσοστά δείχνουν βελτίωση σε σχέση με τις παλαιότερες μελέτες (όπως αυτή του Lawrence [17]), η αύξηση της ημιζωής των URL είναι σημαντική, φτάνοντας τα 9.68 χρόνια, σε σύγκριση με τα 6.5 χρόνια που ανέφερε ο Bansal [1].

Συνολικά, αυτή η μελέτη παρέχει νέες γνώσεις και δείχνει την εξέλιξη των τεχνολογικών προόδων και των πρακτικών αρχειοθέτησης που έχουν συμβάλει στη σταθερότητα των ψηφιακών αναφορών στην ακαδημαϊκή βιβλιογραφία.

Σημασία των Τάσεων Σταθερότητας των URL

Η ανάλυση των τάσεων σταθερότητας των URL ανά έτος προσφέρει πολύτιμες πληροφορίες για τη διαθεσιμότητα ψηφιακών πόρων στη βιβλιογραφία. Το ποσοστό προσβασιμότητας για το 2020, 79.98%, δείχνει βελτίωση σε σχέση με τα προηγούμενα έτη, όπως το 1999, όπου μόλις το 28.88% των συνδέσμων ήταν ενεργοί.

Η εξέλιξη αυτή αντικατοπτρίζει τη γενική τάση προς πιο σταθερές ψηφιακές αναφορές. Αυτή η αλλαγή μπορεί να αποδοθεί σε διάφορους παράγοντες, όπως η ευρεία υιοθέτηση πιο ανθεκτικών τεχνολογιών ιστού, οι βελτιώσεις στην αρχειοθέτηση ψηφιακού περιεχομένου και η μετάβαση από το πρωτόκολλο HTTP στο HTTPS, που προσφέρει αυξημένη ασφάλεια και αξιοπιστία.

Παρά την αυξημένη σταθερότητα των URL, η μεταβλητότητα που παρατηρήθηκε υπογραμμίζει την ανάγκη για συνεχή προσαρμογή των πρακτικών παραπομπής, ώστε να διασφαλίζεται η μακροχρόνια προσβασιμότητα των αναφορών.

Επίδραση του Χώρου Δημοσίευσης και του Τομέα στη Σταθερότητα των URL

Η ανάλυση της σταθερότητας των URL ανάλογα με τον χώρο δημοσίευσης και τον ανώτατο τομέα (TLD) αποκαλύπτει σαφή μοτίβα. Τα ποσοστά ενεργών συνδέσμων URL παρουσίασαν σημαντική

διακύμανση ανά χώρο δημοσίευσης, με το χαμηλότερο ποσοστό να καταγράφεται στο REJ (54.25%) και το υψηλότερο στο ICSME (81.37%).

Επιπλέον, οι διαφορές ανά TLD δείχνουν ότι τομείς όπως .press και .dev παρουσίασαν υψηλότερα ποσοστά ενεργών συνδέσμων (100% και 89.44% αντίστοιχα), γεγονός που μπορεί να αποδοθεί σε διαφορετικές πρακτικές διαχείρισης τομέων και ανανέωσης.

Η κατανόηση αυτών των διαφορών είναι κρίσιμη για τους ερευνητές και τους εκδότες, καθώς αναδεικνύει τη σημασία της επιλογής των τομέων ή των χώρων δημοσίευσης κατά την παραπομπή ψηφιακών πόρων.

Μηχανισμοί Προστασίας Τομέων

Κατά τη διάρκεια της αυτόματης πρόσβασης στους τομείς, παρατηρήθηκε ότι ένα σημαντικό ποσοστό τομέων που κατατάχθηκαν ως ανενεργοί προστατεύονται από υπηρεσίες όπως το Cloudflare. Αυτές οι υπηρεσίες έχουν σχεδιαστεί για την προστασία των ιστοσελίδων από διάφορες απειλές, όπως επιθέσεις DDoS και αυτοματοποιημένη κίνηση.

Η παρουσία αυτών των μηχανισμών προστασίας δυσκολεύει την αξιολόγηση της προσβασιμότητας των URL, καθώς συχνά περιλαμβάνουν διαδικασίες επαλήθευσης όπως CAPTCHA, που εμποδίζουν την αυτόματη πρόσβαση. Αυτό το εύρημα υπογραμμίζει την ανάγκη για βελτίωση των μεθοδολογιών μελέτης της διατήρησης των URL.

Τεχνικές Προκλήσεις στην Προσβασιμότητα των URL

Η μελέτη αποκάλυψε αρκετές τεχνικές προκλήσεις στην προσβασιμότητα των URL. Σφάλματα HTTP όπως το Not Found (404) και το Forbidden (403) αναδείχθηκαν ως κύριοι παράγοντες προβλημάτων, με τα σφάλματα 404 να αντιπροσωπεύουν το 74.65% των περιπτώσεων και τα σφάλματα 403 το 18.16%.

Επιπλέον, σφάλματα δικτύου όπως το Couldn't resolve host (62.49%) και το Operation timeout (21.4%) επιβεβαιώνουν τις προκλήσεις στη σταθερότητα των συνδέσεων.

Η αντιμετώπιση αυτών των προκλήσεων απαιτεί καινοτόμες λύσεις, όπως η ενσωμάτωση εργαλείων αυτόματης επιθεώρησης συνδέσμων και η βελτίωση των πολιτικών πρόσβασης σε ψηφιακούς πόρους.

Συμπεράσματα από την Ανάλυση Παλινδρόμησης

Η ανάλυση παλινδρόμησης που πραγματοποιήθηκε στη μελέτη αυτή προσφέρει πληροφορίες για τους παράγοντες που επηρεάζουν τη διατήρηση των συνδέσμων URL. Παρά τις προσδοκίες, τα αποτελέσματα δεν αποκάλυψαν ισχυρή συσχέτιση μεταξύ των εξεταζόμενων παραμέτρων και της σταθερότητας των συνδέσμων.

Συμπεράσματα

Σύνοψη Ευρημάτων

Η διατήρηση των συνδέσμων URL στην επιστημονική βιβλιογραφία διαμορφώνεται από έναν πολύπλοκο συνδυασμό παραγόντων, όπως οι τεχνολογικές προκλήσεις, οι πολιτικές και πρακτικές των χώρων δημοσίευσης, καθώς και τα χαρακτηριστικά των διαδικτυακών τομέων. Αυτή η μελέτη παρείχε μια ολοκληρωμένη ανάλυση της σταθερότητας των URL στις επιστημονικές δημοσιεύσεις της μηχανικής λογισμικού, αποκαλύπτοντας τόσο τη σημαντική πρόοδο όσο και τις συνεχιζόμενες προκλήσεις στον τομέα της ψηφιακής διατήρησης.

Τα ευρήματα της έρευνάς μας δείχνουν μια αξιοσημείωτη βελτίωση στη διάρκεια ημιζωής των URL, με αύξηση στα 9.68 χρόνια. Αυτή η επέκταση στη μακροβιότητα των URL υποδηλώνει ότι οι προσπάθειες για την ενίσχυση της ανθεκτικότητας των ψηφιακών αναφορών αποδίδουν θετικά αποτελέσματα. Ωστόσο, η δυναμική και συχνά προσωρινή φύση των διαδικτυακών πόρων εξακολουθεί να προκαλεί σημαντικές προκλήσεις, όπως καταδεικνύεται από την προσωρινή αδυναμία πρόσβασης σε ορισμένους συνδέσμους URL. Αυτά τα ευρήματα τονίζουν την ανάγκη για συνεχή παρακολούθηση και προσαρμοστικές στρατηγικές για τη διατήρηση της αξιοπιστίας των ψηφιακών αναφορών στην ακαδημαϊκή επικοινωνία.

Επιπλέον, η ανάλυση παλινδρόμησης αποκάλυψε την απουσία ισχυρών συσχετίσεων μεταξύ της σταθερότητας των URL και παραγόντων όπως ο δείκτης επιρροής των περιοδικών και η κατάταξη CORE. Αυτό το αποτέλεσμα υποδηλώνει ότι η διατήρηση των URL επηρεάζεται πιθανότατα από ένα ευρύτερο σύνολο παραγόντων που δεν καταγράφηκαν πλήρως σε αυτή τη μελέτη. Η πολυπλοκότητα αυτών των σχέσεων υπογραμμίζει τη σημασία της περαιτέρω έρευνας για τον εντοπισμό πρόσθετων παραμέτρων που συμβάλλουν στη μακροβιότητα των URL στις ακαδημαϊκές δημοσιεύσεις.

Μελλοντική Έρευνα

Η μελλοντική έρευνα θα πρέπει να επικεντρωθεί στη διερεύνηση πρόσθετων παραγόντων που επηρεάζουν τη σταθερότητα των URL, όπως τα χαρακτηριστικά των πλατφορμών φιλοξενίας και οι πρακτικές ψηφιακής διατήρησης. Επίσης, απαιτείται βελτίωση των αυτοματοποιημένων εργαλείων επαλήθευσης για την καλύτερη αντιμετώπιση των προστατευτικών μηχανισμών όπως το Cloudflare. Επιπλέον, παρόμοιες μελέτες θα πρέπει να διεξαχθούν σε άλλους επιστημονικούς τομείς για να κατανοηθεί αν οι παρατηρούμενες τάσεις είναι γενικές, καθώς και να αξιολογηθεί η επίδραση των νέων τεχνολογιών ιστού στη μακροβιότητα των ακαδημαϊκών αναφορών.

Τελικές Παρατηρήσεις

Συμπερασματικά, αυτή η μελέτη συνέβαλε σημαντικά στην κατανόηση της σταθερότητας των URL στη βιβλιογραφία της μηχανικής λογισμικού. Τα ευρήματα αναδεικνύουν τόσο την πρόοδο που έχει σημειωθεί όσο και τις προκλήσεις που παραμένουν στη διατήρηση ψηφιακών πόρων. Η παρατηρούμενη αύξηση στη διάρκεια ημιζωής των URL είναι ενθαρρυντική, αλλά τα συνεχιζόμενα ζητήματα προσωρινής αδυναμίας πρόσβασης και η έλλειψη ισχυρών συσχετίσεων στην ανάλυση παλινδρόμησης υποδηλώνουν ότι χρειάζεται περαιτέρω δουλειά.

Καθώς το ψηφιακό περιβάλλον συνεχίζει να αλλάζει, η ακαδημαϊκή κοινότητα πρέπει να παραμένει σε εγρήγορση και να προσαρμόζει τις πρακτικές της. Η συνεργασία μεταξύ ερευνητών, εκδοτών και αρχειοθετών θα είναι απαραίτητη για την ανάπτυξη και εφαρμογή αποτελεσματικών στρατηγικών που θα διασφαλίζουν την ακεραιότητα και μακροβιότητα της ακαδημαϊκής επικοινωνίας. Το μέλλον της ψηφιακής διατήρησης εξαρτάται από την ικανότητά μας να προβλέψουμε και να ανταποκριθούμε στις εξελισσόμενες προκλήσεις της σταθερότητας των URL στον ακαδημαϊκό τομέα.

Chapter 1

Introduction

1.1 Purpose of the Thesis

In the digital age, the accessibility and preservation of online resources have become critical concerns in the realm of academic scholarship. The increasing reliance on web-based references in scientific literature has introduced new challenges in maintaining the integrity and reproducibility of research. URLs serve as gateways to these online resources, yet their transient nature poses a significant threat to the longevity and reliability of scholarly communications.

URL rot, the phenomenon where hyperlinks become inactive or point to different content than originally intended, is a growing issue that undermines the integrity of academic work. As the web evolves, many URLs referenced in scholarly publications decay over time, leading to the loss of access to crucial resources. This decay not only hampers the verification of research findings but also disrupts the continuity of knowledge dissemination. In fields like software engineering, where online repositories, documentation, and digital tools are frequently cited, the persistence of URLs is particularly vital.

Despite the recognition of URL rot as a problem, comprehensive studies specifically addressing the persistence of URLs in software engineering literature have been limited. Previous research has largely focused on broader fields or has not fully accounted for the rapid advancements and specific practices within software engineering. This gap highlights the need for an in-depth investigation into the factors that influence URL stability in this domain, which forms the core motivation for this thesis.

1.2 Problem Statement

The problem at the heart of this research is the instability of URLs in software engineering scholarly works, which threatens the reproducibility and reliability of research in this field. While URL rot has been acknowledged across various academic disciplines, there has been insufficient focus on how it manifests within software engineering, a field heavily reliant on digital resources. Furthermore, there is a lack of understanding regarding the specific factors that contribute to URL decay in this context.

Given the importance of digital continuity in software engineering research, where access to tools, code repositories, and documentation is essential, the instability of URLs presents a significant challenge. This thesis seeks to address this challenge by exploring the extent of URL rot in software engineering literature and identifying the factors that influence URL stability.

1.3 Structure of the Thesis

This thesis is organized into several chapters, each focusing on different aspects of URL persistence in software engineering scholarly works. The structure is as follows:

- **Chapter 1: Introduction** - This chapter introduces the thesis, outlining its purpose, the problem statement, and the structure of the research.
- **Chapter 2: Background** - Provides an overview of the concepts related to URL persistence and link decay, including definitions, key characteristics of links in scholarly publications, and their importance in academic research.
- **Chapter 3: Related Work** - Discusses previous studies on URL rot and digital preservation in academic publishing, highlighting institutional repositories, decentralized preservation systems, web archiving, and other relevant approaches.
- **Chapter 4: Methodology** - Describes the research design, data collection methods, and analytical techniques employed in this study. It includes details on tools and technologies used, such as PostgreSQL, doi2pdf, Grobid, Python, and others, and addresses potential threats to the validity of the research.
- **Chapter 5: Results** - Presents the empirical findings of the research, including analyses of URL stability over time, by hosting purpose, by venue, by document section, and by scheme, domain, and host. This chapter also examines network errors and the outcomes of the regression analysis.
- **Chapter 6: Discussion** - Analyzes the implications of the results, comparing them with previous studies, and discussing the impact of venue and domain on URL stability, domain protection mechanisms, technical challenges in URL accessibility, and insights gained from the regression analysis.
- **Chapter 7: Conclusion** - Summarizes the key findings of the research, discusses the contributions and implications of the study, and suggests directions for future work. This chapter also includes final remarks on the significance of the study and its impact on the field of digital preservation in scholarly communication.

Each chapter builds upon the previous one, culminating in a comprehensive understanding of URL persistence in software engineering literature and offering insights into improving the longevity and reliability of digital references in academic publications.

Chapter 2

Background

2.1 Definition of Link Decay

Link decay, also known as URL rot or link rot, refers to the phenomenon where hyperlinks that were once functional and pointed to specific web resources become inactive or lead to content that is no longer available. This issue arises from several factors, including the deletion of web pages, changes in web server configurations, domain expiration, or the relocation of content without proper redirection. The transient nature of web content means that hyperlinks, which are intended to provide stable references to additional information, often fail to maintain their integrity over time.

In academic and scientific literature, the impact of link decay is particularly pronounced. Researchers frequently cite web resources to support their findings, offer supplementary data, or direct readers to relevant external content. However, when these web references become inaccessible, it undermines the reliability and reproducibility of the research. This can lead to several adverse outcomes, such as the inability to verify sources, the loss of crucial contextual information, and a decrease in the overall credibility of the scholarly work.

The concept of link decay has been recognized since the early days of the internet. Early studies, such as those conducted by Koehler [14], highlighted the rapid rate at which URLs could become inactive, demonstrating that a significant percentage of web references in scholarly papers ceased to function within a few years of publication. Subsequent research has consistently found that link decay is a widespread and persistent problem across various fields of study.

Understanding link decay involves examining the structure of URLs and how changes to their components can lead to decay. A URL consists of several parts: the protocol (e.g., http or https), the domain name, the path, and, optionally, query parameters. Changes to any of these components can render a URL inactive. For example, changes to the domain name, such as when a website is rebranded or acquired, can make all previous URLs pointing to that domain obsolete. Similarly, modifications to the server's directory structure or file names can break the paths specified in the URLs.

Link decay can manifest in different forms. A common outcome is a "404 Not Found" error, indicating that the resource no longer exists at the specified address. In other cases, the URL might redirect to an unrelated page, or it might lead to a generic homepage instead of the specific referenced content. These varying outcomes complicate efforts to address link decay, as different strategies may be required depending on the type and cause of the decay.

To mitigate the effects of link decay, several approaches have been developed. Persistent identifiers, such as Digital Object Identifiers (DOIs), provide stable links to digital content, even if the

location of the content changes. Archival services, such as the Internet Archive's Wayback Machine, capture and preserve snapshots of web pages over time, allowing researchers to access historical versions of decayed links. Additionally, best practices for citing web resources, such as including access dates and using reputable, stable sources, can help reduce the incidence of link decay.

2.2 Key Characteristics of Links in Scholarly Publications

In scholarly publications, links serve as vital tools for connecting readers to additional resources, supporting data, and supplementary materials. The characteristics of these links play a crucial role in determining their longevity and reliability. Understanding these characteristics is essential for developing strategies to mitigate link decay and ensure the integrity of scholarly communication.

2.2.1 Diversity in Type and Purpose

Links in scholarly publications can point to a wide variety of resources, including academic papers, datasets, software repositories, official websites, and multimedia content. The stability of these links varies significantly based on the type of resource they reference. For instance, links to peer-reviewed articles hosted on established digital libraries or repositories tend to be more stable than links to personal web pages or blogs, which are more susceptible to changes and deletions.

2.2.2 Contextual Placement

The context in which links are used within scholarly publications influences their stability and impact. Links can be embedded in different sections of a paper, such as the introduction, methods, results, discussion, and references. Links in the references section, such as DOIs or URLs to digital archives, are essential for verifying sources and accessing cited works. In contrast, links within the body of the text may provide additional context, direct readers to supporting data, or offer further reading. The placement and purpose of links affect their likelihood of decay and the consequences of such decay.

2.2.3 Hosting Domain Stability

The stability of a link is also influenced by the hosting domain. Resources hosted on domains associated with academic institutions, government agencies, and reputable organizations are generally more stable than those hosted on commercial or personal domains. This stability often results from better maintenance practices, longer-term planning for digital preservation, and institutional support for maintaining web resources.

2.2.4 Technical Protocols

Technical characteristics, such as the use of HTTP or HTTPS protocols, play a significant role in the stability of links. HTTPS links, which provide secure, encrypted connections, are increasingly common and are often considered more reliable than HTTP links. However, HTTPS links can still decay if the underlying content is moved or deleted, or if the security certificates expire without renewal.

2.2.5 URL Format and Structure

The format and structure of links can greatly affect their longevity. Short, well-structured URLs are generally more stable than long, complex ones with numerous query parameters. This is because shorter URLs are less likely to be affected by changes in the underlying directory structure or server configurations, making them less prone to decay.

2.2.6 Use of Persistent Identifiers

The use of persistent identifiers, such as Digital Object Identifiers (DOIs), is a critical characteristic of links in scholarly publications. DOIs provide a stable, permanent link to digital content, regardless of changes to the URL where the content is hosted. Managed by registration agencies, DOIs ensure the continuity and accessibility of the linked content. The widespread adoption of DOIs in academic publishing has significantly improved the reliability and stability of links in scholarly works.

2.2.7 Conclusion of Key Characteristics

The key characteristics of links in scholarly publications significantly impact their stability, reliability, and role in academic communication. As highlighted, the diversity in type and purpose, contextual placement, hosting domain stability, technical protocols, URL format and structure, and the use of persistent identifiers all contribute to the overall integrity and longevity of these links. Understanding these factors is crucial for mitigating link decay and maintaining the trustworthiness of scholarly works.

2.3 Importance of Links in Scholarly Publications

Links in scholarly publications are of paramount importance for several reasons. They serve as essential tools for enhancing the credibility, transparency, and reproducibility of academic research. The following points highlight the significance of links in scholarly work and the critical role they play in the academic ecosystem.

2.3.1 Verification and Validation of Sources

Firstly, links enable verification and validation of sources. When researchers cite web resources, they provide readers with direct access to the referenced materials. This allows readers to verify the accuracy of the cited information, assess the quality of the sources, and understand the context in which the references were used. The ability to access original sources is fundamental to the academic practice of building upon previous work and ensuring the integrity of scholarly communication.

2.3.2 Enhancing Transparency and Reproducibility

Secondly, links enhance the transparency and reproducibility of research. In many scientific fields, reproducibility is a cornerstone of research integrity. By providing links to datasets, software repositories, and supplementary materials, authors enable other researchers to replicate their experiments, validate their findings, and build on their work. This transparency fosters a collaborative research environment and contributes to the advancement of knowledge.

2.3.3 Facilitating Knowledge Dissemination

Links also facilitate the dissemination of knowledge. In the digital age, the internet is a primary medium for the distribution and consumption of scholarly content. Hyperlinks enable seamless navigation between related works, allowing researchers to explore a network of interconnected studies. This interconnectedness accelerates the dissemination of ideas and findings, making it easier for researchers to stay informed about the latest developments in their fields.

2.3.4 Increasing Visibility and Impact of Research

Moreover, links contribute to the visibility and impact of research. When scholarly works are linked to from other papers, databases, or online platforms, they become part of a larger academic discourse. This interconnectedness can increase the visibility of a researcher's work, leading to higher citation counts and greater recognition within the academic community. Additionally, links from reputable sources can enhance the perceived credibility and authority of a publication.

2.3.5 Supporting Teaching and Learning

In educational contexts, links are invaluable for teaching and learning. Educators often use scholarly publications to teach students about current research and methodologies. Links to external resources, datasets, and multimedia content provide students with a richer learning experience, allowing them to engage with the material more deeply. By exploring linked content, students can gain a more comprehensive understanding of the subject matter and develop critical thinking skills.

2.3.6 Facilitating Interdisciplinary Research

Links also play a vital role in interdisciplinary research. Modern scientific inquiries often span multiple disciplines, requiring researchers to draw upon a diverse range of sources. Links facilitate this cross-disciplinary integration by providing direct access to relevant works from various fields. This interconnectedness supports innovative research approaches and the synthesis of new ideas across disciplinary boundaries.

2.3.7 Supporting Digital Preservation

Lastly, links support digital preservation and the long-term accessibility of scholarly content. By using persistent identifiers and linking to reputable digital archives, researchers can ensure that their work remains accessible to future generations. Digital preservation initiatives, such as those undertaken by institutional repositories and national libraries, rely on stable links to maintain the scholarly record over time.

In conclusion, links in scholarly publications are indispensable for ensuring the credibility, transparency, and impact of academic research. They enable verification of sources, enhance reproducibility, facilitate knowledge dissemination, and support interdisciplinary collaboration. The strategic use of links, particularly persistent identifiers, can significantly improve the stability and accessibility of scholarly content, safeguarding the integrity of academic communication in the digital age.

2.3.8 Conclusion of Importance of Links in Scholarly Publications

In conclusion, links in scholarly publications are fundamental to the integrity, transparency, and dissemination of academic research. They enable verification of sources, enhance the reproducibility of studies, and facilitate the exchange of knowledge across disciplines. By providing direct access to referenced materials, links support the academic practice of building upon previous work and ensure that research findings are accessible and verifiable. Additionally, links increase the visibility and impact of research by connecting scholarly works within a broader academic network, fostering collaboration and advancing knowledge. Overall, the thoughtful integration of links in scholarly publications is essential for supporting robust and reliable academic communication.

Chapter 3

Related Work

3.1 Previous Studies on URL Rot

The persistence of URLs used in research papers has been a subject of extensive study, with numerous investigations examining their stability and permanence over time. These studies have provided valuable insights into the factors affecting the half-life of URLs and the implications of using them in academic research. In this section, we review and critique the key findings from previous research on URL rot and identify gaps in the literature that our study aims to address.

One of the earliest and most influential studies on URL persistence was conducted by Koehler [14]. Koehler's research measured the half-life of URLs, defined as the time it takes for half of the URLs to become inactive, and found it to be approximately two years. This study highlighted the rapid rate at which URLs can decay, drawing significant attention to the issue within the academic community. A follow-up study by Koehler [15] reinforced these findings, showing that the problem of URL rot persists over time.

Lawrence et al. [17] conducted a comprehensive analysis of 67 577 URLs extracted from 100 826 computer science research articles from the CiteSeer database. Their study revealed that the percentage of invalid URLs varied based on the publication year, peaking at 53% for papers published in 1994. After applying correction techniques such as using search engines and manually fixing URLs, they managed to reduce the number of invalid links to around 2.9%. This study underscored the importance of active management and correction of URLs to mitigate the effects of link decay.

Germain [10] offered a different perspective by imposing stricter criteria for URL persistence. Unlike previous studies that allowed for minimal corrections, Germain only considered links to the referenced files as valid. Her study, which tested 64 web references from 31 academic journal articles, reported a half-life of three years. This approach highlighted the variability in URL persistence depending on the criteria used to define an active link.

Spinellis [31] adopted a rigorous methodology by analyzing 4224 URLs from computer science research papers obtained from IEEE and ACM digital libraries. Spinellis did not correct for syntax or lexicographical errors and classified URLs solely based on their ability to download the referenced resource. The study reported an average URL half-life of four years, emphasizing the challenges in maintaining URL validity over time.

McCown et al. [24] focused on articles published in D-Lib Magazine from July 1995 to August 2004. They extracted 4387 web references from 453 articles and conducted extensive trials over 25 weeks, finding that approximately 30% of the URLs were unreachable. This study illustrated the temporal nature of URL accessibility and the necessity for periodic verification.

Wren [38] conducted an extensive analysis of 1630 URLs found in biomedicine articles using ab-

Table 3.1: Prior studies on URL persistence: characteristics and findings

Authors	Year	Papers	URLs	Research field	Half-life (years)	Inactive URLs (%)
Koehler [14]	1999	-	361	Cross-field	2	14.8%
Lawrence et al. [17]	2000	100 826	67 577	Computer science	-	23–53%
Germain [10]	2000	31	64	Cross-field	3	51.5%
Casserly and Bird [3]	2003	1425	500	Library and information sciences	-	18.6%
Spinellis [31]	2003	2471	4224	Computer science	4	28%
Wren [38]	2004	-	1630	Biomedicine	-	22%
Koehler [15]	2004	-	361	Cross-field	-	66.2%
McCown et al. [24]	2005	453	4387	Cross-field	10	30%
Wren et al. [40]	2006	7337	1113	Dermatology	-	18.3%
Wren [39]	2008	-	7462	Biomedicine	-	20%
Wagner et al. [36]	2009	-	2011	Healthcare management	-	49.3%
Saberi and Abedi [29]	2012	748	3734	Social sciences	8.9	27%
Loan and Shah [21]	2020	221	358	Library and information sciences	-	32.1%
Bansal and Parmar [2]	2020	1564	1724	Cross-field	1.76	43.3%
Bansal [1]	2021	273	1921	Library and information sciences	6.5	23.31%
Escamilla et al. [8]	2023	2 641 041	253 590	Cross-field	-	6.02%
This study	2024	49 268	133 686	Software Engineering	9.68	31.22%

stracts from MEDLINE. The study concluded that about 22% of the URLs were no longer available. A follow-up study by Wren et al. [40] in dermatology journals reported an 18.3% inaccessibility rate. In 2008, Wren [39] reenacted the 2004 study, analyzing 7462 URLs and obtaining similar results. These studies consistently demonstrated the persistent problem of URL rot in biomedical literature.

Wagner et al. [36] tested 2011 URLs in health care management journals published between 2002 and 2004, finding that 49.3% of the URLs were not reachable. The study emphasized that URL decay is a significant issue in health care management literature, reflecting broader trends across various disciplines.

Loan and Shah [21] examined 358 web citations in articles published between 2007 and 2011 in the Journal of Informetrics. They found that 32.12% of the URLs were unavailable. In the same year, Bansal and Parmar [2] analyzed 1724 URLs from papers published in 2015-2016 in the Current Science Journal, reporting a 43.33% inaccessibility rate and an estimated half-life of 1.76 years. Bansal [1] later studied 1921 URLs from the DESIDOC Journal of Library & Information Technology, finding a longer half-life of 6.5 years. These studies indicate some improvement in URL accessibility over time, particularly for more recent publications.

Extending the investigation to software and code references, Escamilla et al. [8] analyzed over 253 590 URLs from more than 2.6 million papers in 2023. They discovered that 93.98% of URLs pointing to Git Hosting Platforms (GHP) remained active, suggesting that URLs related to software repositories are more stable than other types of web references.

3.2 Digital Preservation in Academic Publishing

Digital preservation is a critical strategy in academic publishing, aimed at ensuring the longevity and accessibility of scholarly content in the face of technological changes and the inherent instability of digital formats. Effective digital preservation combats URL rot, a phenomenon where web links become inactive or lead to content that is no longer available, by maintaining the integrity and

accessibility of digital assets over time. This section explores various digital preservation strategies and their relevance to combating URL rot, supported by insights from recent studies.

3.2.1 Institutional Repositories

Definition and Importance

Institutional repositories store and manage digital content created by members of an academic institution, including research papers, datasets, and multimedia materials. They play a vital role in ensuring long-term access to digital assets. By centralizing the preservation efforts, these repositories ensure that valuable academic resources are protected from digital decay and remain accessible to future generations of researchers and students. They also facilitate the dissemination of knowledge by providing open access to the institution's scholarly output, thereby enhancing the visibility and impact of the research conducted within the institution.

Case Studies

Colorado State University Libraries This institution successfully implemented digital archiving practices that include identifying, selecting, packaging, and archiving local digital assets for long-term access and migration. Their approach is comprehensive and involves several key steps:

- **Identification**: The process begins with identifying digital content that needs preservation. This includes research papers, datasets, multimedia materials, and other scholarly outputs created by the university's members.
- **Selection**: Once identified, the materials are carefully selected based on their importance, relevance, and potential for future use. This step ensures that the most valuable content is prioritized for preservation.
- **Packaging**: The selected digital assets are then packaged in a way that ensures their integrity and usability over time. This involves converting files into stable formats and organizing them with appropriate metadata.
- **Archiving**: Finally, the packaged materials are archived in the institutional repository, where they are stored securely and made accessible for long-term use.

Colorado State University Libraries emphasize the need for collaborative approaches to maximize limited resources and ensure sustainability. By involving various stakeholders, including librarians, IT professionals, and researchers, they create a robust framework for digital preservation that leverages collective expertise and shared responsibilities. Their strategy not only addresses immediate preservation needs but also builds a sustainable model for ongoing digital stewardship.

> "The digital preservation strategies at Colorado State University Libraries are designed to ensure the long-term access and usability of digital assets. Through a collaborative approach, the institution effectively manages its digital content and mitigates the risks associated with digital decay." [28]

Academic Libraries in Ghana Libraries in Ghana highlight the necessity for comprehensive digital preservation policies, disaster planning, adequate funding, and staff development to support long-term preservation efforts. The study by Adjei, Mensah, and Amoaful (2019) outlines several critical components for effective digital preservation in Ghanaian academic libraries:

- **Policy Development**: Establishing clear and comprehensive digital preservation policies that provide a mandate and direction for preservation activities.
- **Disaster Planning**: Developing and implementing disaster plans to protect digital content from unforeseen events such as natural disasters or technical failures.
- **Funding**: Securing adequate funding to support ongoing preservation efforts, including the purchase of necessary technology and the training of staff.
- **Staff Development**: Investing in the continuous development of staff skills and knowledge in digital preservation techniques and best practices.

The case of academic libraries in Ghana underscores the importance of a holistic approach to digital preservation, integrating policy, planning, funding, and staff development to ensure the longevity and accessibility of digital assets.

3.2.2 Decentralized Preservation Systems

LOCKSS (Lots of Copies Keep Stuff Safe)

LOCKSS [5] creates multiple copies of digital content across various locations, ensuring that even if one copy becomes corrupted or lost, others remain accessible. This redundancy is crucial in preventing URL rot by maintaining the availability of digital resources despite changes in individual web pages or servers. Developed initially to preserve access to e-journals, the LOCKSS system employs a peer-to-peer network of computers to cache content, thereby ensuring its longevity and accessibility. Each node in the LOCKSS network regularly polls its peers to verify the integrity of its cached content and to repair any corrupted or missing pieces from other copies. This process of continuous auditing and repair underpins the system's robustness.

CLOCKSS (Controlled LOCKSS)

CLOCKSS (Controlled LOCKSS) takes the principles of the LOCKSS system further by incorporating a governance model that involves collaboration between publishers and librarians. This community-governed initiative is designed to ensure that digital content remains accessible and unaltered over time, providing a trustworthy archive for scholarly publications. CLOCKSS operates similarly to LOCKSS but with enhanced controls and oversight, ensuring that preserved content is only accessible when it is no longer available from any publisher. This approach guarantees the integrity and authenticity of archived materials, which is paramount in academic publishing.

CLOCKSS archives [5] are particularly valuable in academic publishing, where the integrity of scholarly records is paramount. By preserving digital content in a controlled environment with collaborative governance, CLOCKSS ensures that scholarly materials remain accessible and credible for future research and reference.

3.2.3 Digital Stewardship Programs

University Examples

Institutions like Colorado State University Libraries and Brigham Young University Libraries integrate planning, policy development, and technological solutions to safeguard digital content. Successful programs include comprehensive ingestion processes, ensuring digital content is correctly archived and retrievable. These programs emphasize the importance of a holistic approach to digital preservation, which includes not only the technical aspects of digital storage and access but also the administrative and policy frameworks that support these efforts [27].

For example, the digital stewardship program at Colorado State University Libraries focuses on a systematic process of digital preservation that includes the identification, selection, and archiving of digital assets. Similarly, Brigham Young University Libraries have developed robust policies and technological infrastructures to support the long-term preservation of their digital collections. These programs demonstrate the effectiveness of integrating various components of digital stewardship to achieve comprehensive preservation goals.

Challenges and Solutions

A study of academic libraries in Ghana highlights the need for comprehensive digital preservation policies, disaster planning, adequate funding, and staff development to support long-term preservation efforts. These libraries face challenges such as limited financial resources, insufficient infrastructure, and a lack of trained personnel. To address these issues, the study recommends the development of clear digital preservation policies, the establishment of disaster recovery plans, the securing of sustainable funding sources, and the continuous professional development of staff [7].

3.2.4 Web Archiving

Strategies and Tools

Web archiving involves systematically collecting and preserving web content to ensure its longevity. Effective web archiving strategies include regular snapshots of web pages and the use of specialized tools to manage and retrieve archived content. Tools such as the Internet Archive's Wayback Machine are instrumental in this process, capturing snapshots of web pages at various points in time, thereby allowing users to access historical versions of websites [23].

Examples and Case Studies

Various international initiatives and case studies demonstrate the effectiveness of web archiving in preserving digital content. The Internet Archive's Wayback Machine, for instance, regularly captures web pages and makes them available to the public, ensuring that even if a website goes offline, its content can still be accessed. Other examples include national web archiving programs, such as those run by the Library of Congress and the British Library, which aim to preserve their respective countries' web heritage [23].

3.2.5 Metadata and Standards

Role of Metadata

Proper metadata is essential for managing and retrieving digital content. Preservation metadata standards help ensure that digital objects remain accessible and understandable over time. Metadata provides critical information about the digital object's creation, structure, and preservation history, which is crucial for its long-term usability.

Current Practices

The adoption of metadata standards like PREMIS (Preservation Metadata: Implementation Strategies) supports the consistent documentation of preservation actions. PREMIS, in particular, provides a comprehensive framework for capturing the information necessary to support the long-term preservation of digital materials. By implementing these standards, institutions can ensure that their digital content remains accessible and understandable over time [37].

3.2.6 Economic Considerations

Cost of Preservation

Financial resources significantly impact the sustainability of digital preservation efforts. Institutions must balance the costs of digital preservation against the potential losses of digital content. The costs associated with digital preservation include technological infrastructure, staffing, and ongoing maintenance. Without adequate funding, even the best-laid preservation plans can falter.

Funding Models

Different funding models and strategies, including endowments and grants, can support ongoing digital preservation activities. Collaborative funding efforts between institutions can also help mitigate costs. For example, some institutions have established endowments specifically for digital preservation, while others have secured grants from governmental and non-governmental organizations to support their efforts. Collaborative funding models, where multiple institutions share the costs and benefits of preservation activities, also offer a viable solution for ensuring the sustainability of digital preservation efforts [16].

3.3 Summary of Findings

The literature review reveals that URL rot, characterized by the inaccessibility of web links over time, is a widespread issue affecting various academic fields. Early studies by Koehler [14, 15] and Lawrence et al. [17] highlight the rapid decay of URLs, with significant percentages becoming inactive within a few years. Subsequent research by Germain [10], Spinellis [31], and others corroborate these findings, emphasizing the need for robust digital preservation strategies.

To combat URL rot, several preservation strategies have been identified. Institutional repositories and decentralized systems like LOCKSS and CLOCKSS play crucial roles in maintaining the accessibility of digital content through redundancy and community governance [5, 28]. Digital stewardship programs and web archiving initiatives further support these efforts by integrating comprehensive planning and technological solutions [27, 23]. This research aims to build on these findings by exploring the specific factors influencing URL rot in software engineering, assessing the

effectiveness of preservation strategies, and investigating economic considerations for sustainable digital preservation.

Chapter 4

Methodology

This chapter delineates the comprehensive methodology employed in this study to investigate the persistence of URLs in scientific publications, specifically within the domain of software engineering. With the increasing prevalence of digital references in academic literature, it is imperative to understand the longevity and reliability of URLs to maintain the integrity of scholarly work. Our methodology integrates both automated and manual approaches to ensure a thorough examination of URL persistence. This chapter provides a detailed account of the processes involved in data collection, extraction, verification, and analysis. By meticulously outlining each step, we aim to facilitate transparency and reproducibility in our research.

4.1 Research Design

The overall research design of this study, visualized in Figure 4.1, is structured to systematically address our research questions regarding the persistence and decay of URLs in scientific publications. Our approach is segmented into four major phases: data collection, data processing, data analysis, and validation. Each phase is designed to build upon the previous one, ensuring a comprehensive examination of URL persistence.

In the data collection phase, we identify and gather relevant research papers from the DBLP database [20], which is a well-regarded repository of bibliographic information on major computer science publications. This phase involves downloading the papers in PDF format and preparing them for further processing.

The data processing phase includes extracting URLs from the downloaded PDFs and parsing them for analysis. This step is crucial as it transforms raw data into a structured format that can be effectively analyzed. We employ various tools and methods to ensure accuracy and efficiency in this phase.

The data analysis phase involves both automated and manual verification of the extracted URLs, as well as regression analysis to identify factors influencing URL decay. Automated verification is conducted using custom scripts and tools to check the current status of each URL. Manual auditing is performed on a sample of URLs to ensure the reliability of the automated process and to provide qualitative insights into URL persistence.

Finally, the validation phase includes manual auditing to verify the results of the automated process and regression analysis, ensuring that the research findings are robust and reliable. This comprehensive approach validates our findings against existing studies.

Each of these phases is detailed in the subsequent sections, providing a clear and thorough

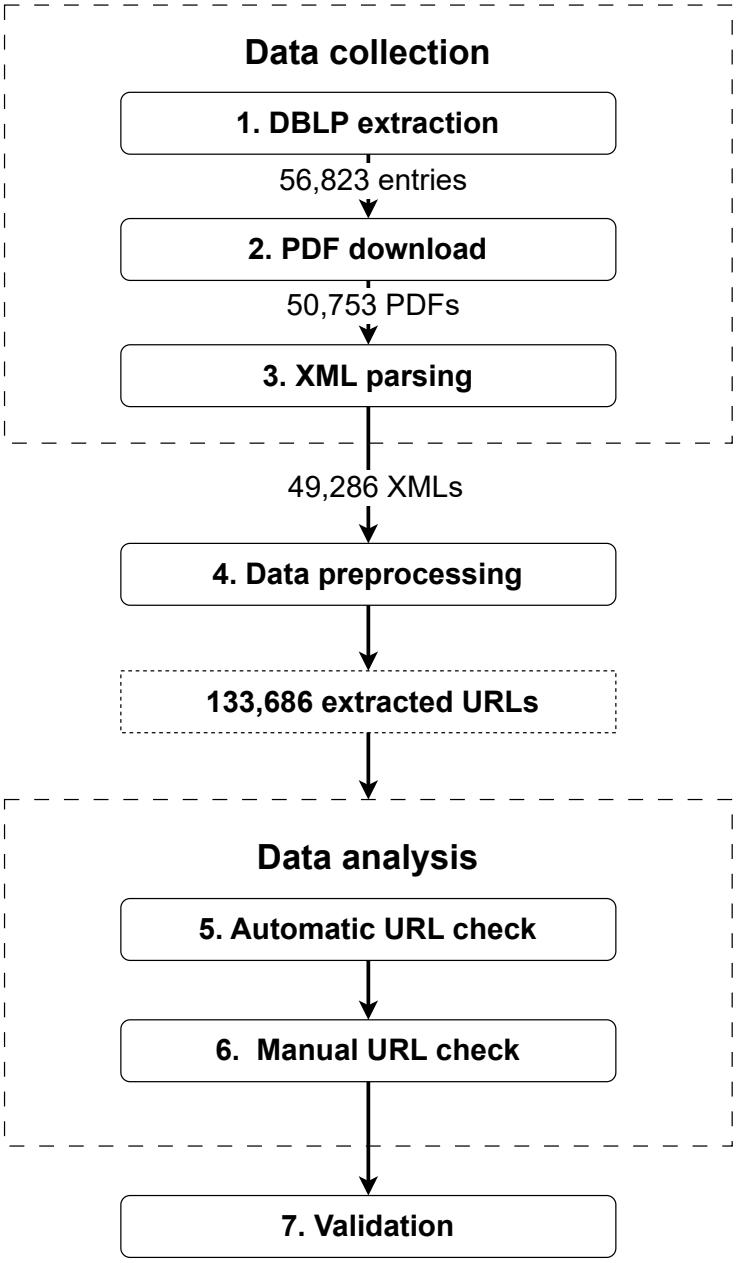

Figure 4.1: Experimental methodology

explanation of the methodologies employed in this study.

4.2 Data Collection

4.2.1 Source of Data: DBLP

DBLP (Digital Bibliography & Library Project) is a highly regarded and comprehensive computer science bibliography that has been in existence since 1993. It offers open access to a vast collection of bibliographic information on major computer science journals, conference proceedings, and other scholarly publications. DBLP is known for its reliability, accuracy, and extensive coverage, which includes over five million publications from various subfields within computer science [20].

The choice of DBLP as the primary data source for this study is motivated by several factors. First, its comprehensive coverage ensures that we capture a broad spectrum of software engineering literature. Second, DBLP is updated regularly, providing current and relevant data essential for analyzing trends in URL persistence. Finally, the structured format of DBLP data, available in XML, facilitates efficient extraction and processing of bibliographic records.

To gather data, we downloaded the most recent DBLP data dump available as of February 1, 2024. This data dump, in the form of a `dblp.xml` file, contains detailed metadata about each publication, including titles, authors, publication venues, and electronic edition ("ee") links. These links often include URLs to the full text of the papers, making DBLP an invaluable resource for our study.

Our focus is on software engineering publications, and we selected relevant venues indexed by DBLP. These venues include top-tier conferences and journals known for their impact and contributions to the field. Table 4.1 lists the selected venues, detailing their acronyms, full names, and types (conference or journal). The publications span from 1971, the earliest year with available records, to 2023, the latest complete year at the time of data collection. In total, we identified 56 823 software engineering papers for further analysis.

4.2.2 PDF Download and Processing

To process the selected papers, we used a customized version of `doi2pdf`, a Python-based tool that automates the retrieval of academic papers based on their DOI identifiers [4]. Our customized version enhances the tool's capabilities by supporting multiple mirrors in case of primary source failures and utilizing arXiv for preprint versions when necessary. This approach ensured a high retrieval rate, successfully downloading 50 753 PDFs, covering 89.3% of the identified papers.

Once downloaded, the PDFs were processed using Grobid (GeneRation Of BIbliographic Data), a machine learning library specialized in extracting and parsing bibliographic data into structured XML [22]. Grobid converts PDFs to XML, preserving essential document sectioning information, which is crucial for subsequent analyses. The choice of Grobid is supported by its superior performance in extracting content from scholarly documents, as demonstrated in comprehensive evaluations [34]. We achieved a high conversion success rate, parsing 49 286 XML files, accounting for 97.1% of the retrieved PDFs.

Table 4.1: Analyzed venues: conferences (C) and journals (J) in the field of software engineering, indexed by DBLP.

Acronym	**Full name**	**Type**
ASE	IEEE/ACM International Conference on Automated Software Engineering	C
ESEM	International Symposium on Empirical Software Engineering and Measurement	C
FASE	Fundamental Approaches to Software Engineering	C
FSE	ACM SIGSOFT Symposium on the Foundations of Software Engineering	C
GPCE	Generative Programming and Component Engineering	C
ICPC	IEEE International Conference on Program Comprehension	C
ICSE	International Conference on Software Engineering	C
ICSM	IEEE International Conference on Software Maintenance	C
ICSME	International Conference on Software Maintenance and Evolution	C
ICST	IEEE International Conference on Software Testing, Verification and Validation	C
ISSTA	International Symposium on Software Testing and Analysis	C
MODELS	International Conference On Model Driven Engineering Languages And Systems	C
MSR	Working Conference on Mining Software Repositories	C
RE	IEEE International Requirements Engineering Conference	C
SANER	IEEE International Conference on Software Analysis, Evolution and Reengineering	C
SCAM	International Working Conference on Source Code Analysis & Manipulation	C
SSBSE	International Symposium on Search Based Software Engineering	C
WCRE	Working Conference on Reverse Engineering	C
ASEJ	Automated Software Engineering	J
ESE	Springer - Empirical Software Engineering	J
IJSEKE	International Journal of Software Engineering and Knowledge Engineering	J
ISSE	Innovations in Systems and Software Engineering	J
IST	Information and Software Technology	J
JSS	Elsevier - Journal of Systems and Software	J
REJ	Requirements Engineering Journal	J
NOTES	ACM SIGSOFT Software Engineering Notes	J
SMR	Journal of Software: Evolution and Process	J
SOSYM	Software and System Modeling	J
SPE	Software: Practice and Experience	J
SQJ	Software Quality Journal	J
STVR	Software Testing, Verification and Reliability	J
SW	IEEE Software	J
TOSEM	ACM - Transactions on Software Engineering Methodology	J
TSE	IEEE - Transactions on Software Engineering	J

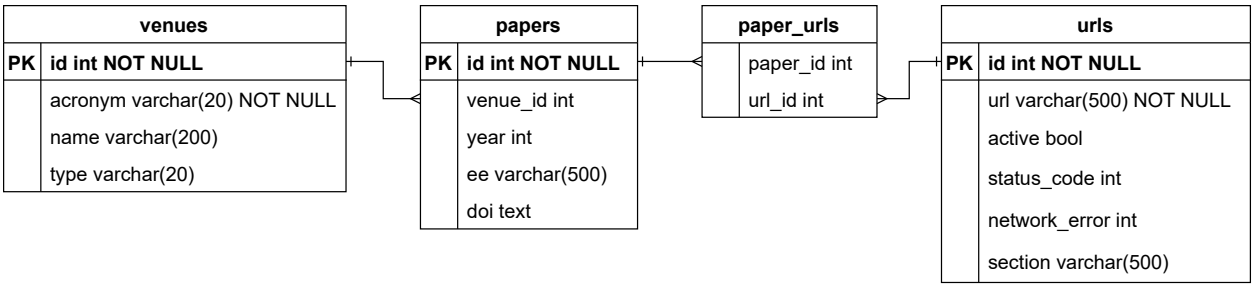

Figure 4.2: Database schema

4.2.3 URL Extraction and Parsing

The extraction of URLs from the XML documents was carried out using a custom Python script. This script employs regular expressions to identify and extract URL patterns within the text. Given that URLs can appear in various formats and contexts, our script was designed to handle common issues such as line breaks and adjacent punctuation marks that can affect URL validity.

Several challenges were encountered during URL extraction, particularly related to typographic anomalies and non-standard URL formats. To address these, our script included mechanisms to clean and standardize the extracted URLs, ensuring they were syntactically correct and ready for subsequent verification phases. This meticulous approach ensured the integrity of our URL dataset, laying a solid foundation for accurate verification and analysis.

By following this detailed methodology for data collection, we ensured that our study is based on a robust and reliable dataset, allowing for meaningful analysis of URL persistence in scientific publications.

4.2.4 Data Storage

For conducting this study, we utilized PostgreSQL as our primary relational database management system. PostgreSQL was chosen for its robustness, support for complex queries, and its wide adoption in both industry and academia.

The database schema, illustrated in Figure 4.2, consists of several key tables, each serving a specific purpose in organizing and managing the data collected for our research. Below is a detailed description of the tables and their roles:

venues

This table stores metadata about the publication venues (such as conferences and journals). The columns include:

- `id`: A unique identifier for each venue.
- `acronym`: The short form or abbreviation of the venue's name.
- `name`: The full name of the venue.
- `type`: Indicates whether the venue is a conference (C) or a journal (J).

papers

This table contains information about the papers analyzed in the study. The columns include:

- `id`: A unique identifier for each paper.
- `venue_id`: A foreign key linking to the `venues` table, indicating the publication venue of the paper.
- `year`: The year in which the paper was published.
- `ee`: The electronic edition link, typically a URL to the paper's online version.
- `doi`: The Digital Object Identifier, providing a persistent link to the paper.

urls

This table holds information about each URL encountered in the papers, along with its verification status. The columns include:

- `id`: A unique identifier for each URL.
- `url`: The actual URL string.
- `active`: A boolean flag indicating whether the URL was found to be active.
- `status_code`: The HTTP status code received when attempting to access the URL.
- `network_error`: Any network error encountered during URL verification.
- `section`: The section of the paper where the URL was found (e.g., abstract, body, references).

paper_urls

This junction table associates URLs with the papers in which they were found. The columns include:

- `paper_id`: A foreign key linking to the `papers` table.
- `url_id`: A foreign key linking to the `urls` table.

This structured schema allows for efficient querying and analysis. For instance, it enables the retrieval of all URLs associated with papers from a specific venue or the determination of URL activity status across different years or paper sections. By organizing the data in this way, we ensure that it is well-structured and easily accessible for comprehensive analysis.

4.3 Data Analysis

4.3.1 Automated URL Verification

The automated URL verification process was designed to efficiently assess the status of a large number of URLs extracted from the selected papers. To achieve this, we employed a combination of URL parsing and network error detection techniques.

Initially, we parsed all extracted URLs to ensure they were syntactically correct according to the relevant Internet standards (RFC 1738 and successors). This step was crucial in filtering out malformed URLs that had no chance of being active, thereby streamlining the subsequent crawling process.

The actual verification of URLs was conducted using the open-source tool `curl-cffi` [18], which allowed us to automate the process of accessing each URL and following any HTTP redirections. We categorized the outcomes of these attempts into various types of errors, including network errors at different protocol levels and erroneous HTTP status codes (anything other than 200). These details were meticulously recorded in the database, specifically in the `urls` table with columns for `status_code` and `network_error`.

To enhance the robustness of our verification process, we incorporated a retry mechanism. For URLs that did not respond within a 10-second window or encountered other errors, we implemented a 45-second wait before retrying, up to three attempts. This approach helped to mitigate the impact of transient network issues, ensuring that temporary inaccessibility did not lead to URLs being incorrectly marked as inactive.

Ultimately, URLs that successfully returned a HTTP 200 response after all redirections and retries were recorded as active, while all others were deemed inactive. This automated verification process was instrumental in providing a comprehensive overview of URL persistence in the collected data.

4.3.2 Manual URL Audit

To complement our automated verification process and ensure its accuracy, we conducted a manual audit of a sample of 200 URLs. This audit focused on URLs that were reported as active by our automated checks. The rationale behind this was that if a URL could not be accessed programmatically, it was unlikely that a human could access it either, thus rendering manual checks on inactive URLs unnecessary.

During the manual audit, we verified if the URLs marked as active indeed led to relevant content. A URL was considered active if it redirected to content that reasonably matched the expected destination. This step was crucial for identifying false positives, where URLs might lead to generic or unrelated content despite being technically active.

Our findings from the manual audit were encouraging, with 94% of the URLs confirmed as active and relevant. This high accuracy rate reinforced the reliability of our automated verification process. However, we did identify a small percentage of URLs that had become inactive between the automated and manual checks, as well as some that led to irrelevant content. These insights helped us refine our methodology and highlighted the importance of periodic manual audits in complementing automated verification processes.

4.3.3 Regression Analysis

As part of our analysis, we conducted several regression analyses to explore potential factors influencing URL persistence. The goal was to identify variables that significantly correlated with the likelihood of URLs remaining active or becoming inactive.

We employed a linear regression model due to its simplicity and ease of interpretability. This model allowed us to isolate and understand the impact of individual variables on URL persistence. Among the variables considered were the CORE venue rankings, journal impact factors, and URL length measured in both characters and path components.

The CORE rankings and journal impact factors were chosen based on the assumption that higher-quality venues might adopt better practices for managing and preserving web references. URL length was considered because longer URLs tend to have more points of potential failure, making them more susceptible to becoming inactive.

The insights gained from these regression analyses were valuable in understanding the dynamics of URL persistence in scholarly publications. They provided a nuanced view of how different factors contribute to URL decay, informing best practices for citing and managing web references in academic research.

4.4 Tools and Technologies Used

In this section, we provide a detailed account of the tools and technologies employed throughout our study. The choice of these tools was driven by their specific features, reliability, and suitability for handling the diverse aspects of our research. Each tool played a critical role in ensuring the accuracy and efficiency of our data collection, processing, and analysis.

4.4.1 PostgreSQL

For data storage and management, we selected PostgreSQL, a powerful, open-source relational database system. PostgreSQL is renowned for its advanced features, such as support for complex queries, robust data integrity, and extensive extensibility. Its ACID-compliant nature ensures reliable transactions, making it ideal for storing and manipulating large datasets, such as those in our study. The relational model facilitated the structured storage of data, with tables such as `papers`, `paper_urls`, `urls`, and `venues`, each serving specific purposes in our database schema [**?**].

4.4.2 doi2pdf

To automate the downloading of PDF documents, we utilized doi2pdf, a Python-based tool that retrieves academic papers using their DOI identifiers. The version we used was customized to enhance its download capabilities. Specifically, we added support for trying multiple mirrors in case of primary source failure and used arXiv to locate preprint versions when necessary. This tool was crucial for efficiently obtaining a vast number of PDFs from our selected venues and years, ensuring a high retrieval rate of the identified papers [4].

4.4.3 Grobid

Grobid, short for GeneRation Of BIbliographic Data, was employed to convert the downloaded PDF files into structured XML format. Grobid excels at extracting and parsing the content of scholarly documents, preserving essential sectioning information, which was vital for our subsequent URL extraction. It relies on machine learning to provide highly accurate text segmentation and metadata extraction. The use of TEI (Text Encoding Initiative) schema by Grobid ensures that the output is both machine-readable and human-understandable, maintaining the rich semantic structure of the documents [22].

4.4.4 Python

Various custom Python scripts were developed to support different stages of our study. These scripts were integral in automating tasks such as URL extraction from XML documents, parsing URLs for syntactic correctness, and managing the retry mechanism during URL verification. Python's extensive libraries and ease of use made it an excellent choice for developing these custom tools, ensuring flexibility and efficiency in handling our data.

4.4.5 curl_cffi

The `curl_cffi` library was used extensively for automated URL verification. This Python interface to the `curl` command-line tool supports a wide range of protocols, including HTTP, HTTPS, and FTP, making it versatile for accessing and testing URLs. By automating the process of URL access and redirection following, `curl_cffi` allowed us to systematically verify the status of each URL, recording outcomes such as HTTP status codes and network errors. Its robustness and reliability were pivotal in conducting the automated checks efficiently [19].

4.4.6 pandas and matplotlib

For the analysis and visualization of our data, we utilized the `pandas` and `matplotlib` libraries in Python. `pandas` is a powerful data manipulation tool that provides data structures and functions needed to manipulate structured data seamlessly. It was used for data cleaning, transformation, and analysis, allowing us to handle the large datasets involved in our study efficiently. `matplotlib` was employed to create the various graphs and visualizations that supported our analysis, providing clear and insightful representations of our data [25, 12].

4.4.7 SQLAlchemy

To facilitate the connection between our Python scripts and the PostgreSQL database, we used SQLAlchemy, a SQL toolkit and Object-Relational Mapping (ORM) library for Python. SQLAlchemy provides a full suite of well-known enterprise-level persistence patterns, designed for efficient and high-performing database access. It allowed us to interact with our PostgreSQL database using Pythonic code, simplifying the process of querying, updating, and managing our data [32].

4.5 Threats to Validity

This section discusses potential limitations of this study, categorized into internal validity, external validity, construct validity, and reliability. Each category addresses different aspects of the study's design and execution that could influence the results.

4.5.1 Internal Validity

Internal validity concerns the accuracy and reliability of the URL extraction and validation processes. Despite employing comprehensive automated techniques, such as scheme checks, database filtering, and status code validation, there remains a potential for data extraction errors. Misinterpretation of URLs due to incorrect parsing or format issues and the inherent limitations of automated validation (e.g., inability to verify content relevance) could affect the results. To mitigate this risk, we conducted a manual audit of a significant subset of the inactive URLs. While

auditing inactive URLs was useful, auditing active URLs was deemed unnecessary as URLs that are not machine accessible cannot be audited by a human either. However, the manual verification process introduces subjective biases, particularly for URLs leading to generic pages or having ambiguous paths, as it relies on intuitive judgment to determine content relevance.

4.5.2 External Validity

External validity pertains to the generalizability of the study's findings. This study focuses specifically on software engineering papers indexed by DBLP. As a researcher in this field, I am confident in DBLP as a robust bibliographic index. Nevertheless, it is possible that niche venues within the field exhibit different URL stability patterns. While this study does not claim generality beyond its stated focus, it would be surprising if significantly different patterns were observed in other niche venues.

4.5.3 Construct Validity

Construct validity addresses the extent to which the study measures what it intends to measure. In this study, the operational definitions of "active" and "inactive" URLs depend heavily on automated criteria such as HTTP status codes and network errors. This approach may not fully capture the nuances of URL functionality, particularly in cases where URLs redirect to content relevant to the domain but not what the original authors intended to reference. The reliance on automated tools for URL extraction and validation might overlook subtle aspects of URL activity and relevance.

4.5.4 Reliability

Reliability concerns the consistency of the study's results. This study employs a multi-step process to gather data (e.g., bibliographic information, papers) and verify URL stability (e.g., parsing, URL extraction, URL checks). Although each step in isolation is straightforward, there is always a possibility that specific bugs crept into our implementation.

Chapter 5

Results

This chapter presents the empirical findings of our study on URL stability in software engineering scholarly works. Our analysis covers temporal trends, the influence of hosting purposes, differences across publication venues, variations by document section, and the impact of URL components such as scheme, domain, and host. Additionally, we examine the types of HTTP and network errors encountered and conduct a regression analysis to understand factors contributing to URL persistence. The results are illustrated with graphs and tables to provide a clear overview of our findings. This examination aims to uncover patterns that can inform practices for ensuring the longevity and accessibility of digital references in academic literature.

In our analyses we excluded 2723 URLs that were malformed during the extraction process.

5.1 URL stability over time

We begin by providing a temporal overview of URL decay, concentrating on the period from 1995 to 2023. Papers from earlier years (1971–1994) included in our analysis referenced fewer than 50 URLs per year on average; these were excluded from the visualizations in this section as they contributed minimally to the analysis.

Figure 5.1 illustrates the evolution of active and inactive URLs over time, alongside the total number of URLs analyzed. The number of URLs referenced in software engineering papers has increased over time, a trend likely correlated with the general rise in the number of papers in the field (not shown). This growth, however, also increases the risk of disruptions to the scholarly record, as URLs are generally less stable than other academic artifacts such as papers.

The percentage of active URLs varies by year, with a noticeable trend showing that URLs in more recent papers are more likely to remain active compared to those in older publications. For example, in 2023, 83.77% of URLs were still active, indicating a relatively high level of accessibility, whereas in 1996, only 15.97% of URLs remained active. A common metric used to assess this trend is the URL half-life—the time required for half of the URLs published in a particular year to become inactive. As shown in Table 3.1, we observe a half-life of 9.68 years, which is on the higher end compared to other studies on URL persistence.

5.2 URL stability by hosting purpose

With increased awareness of the problem of URL instability, the use of archival services for Web references is also increasing over time. The increase in popularity of version control systems—

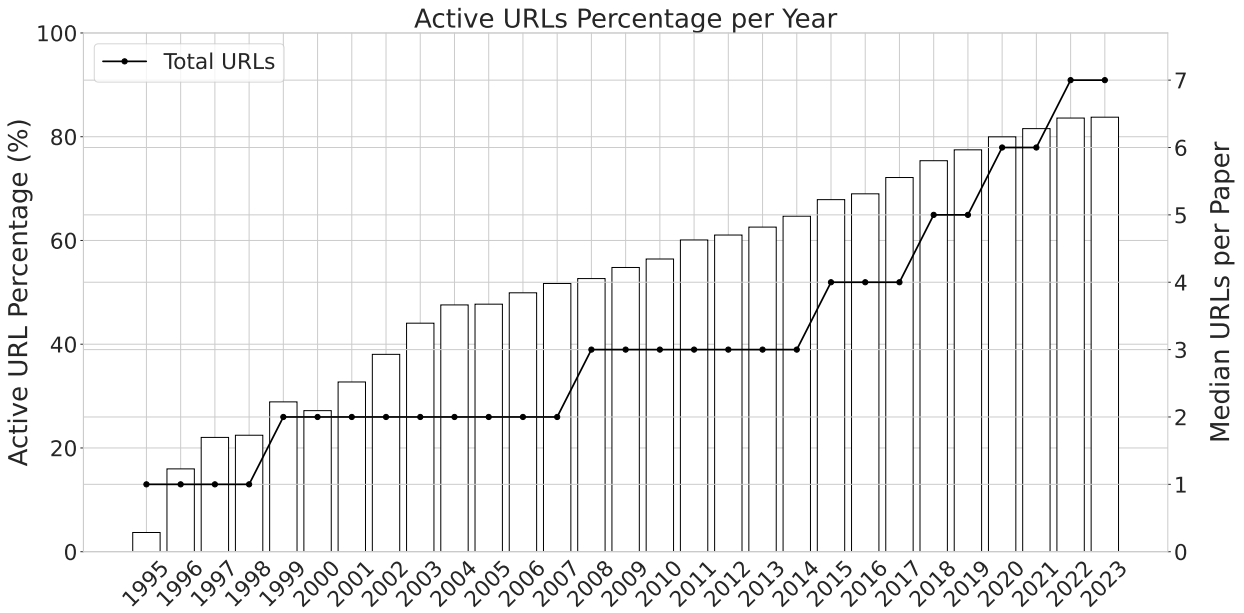

Figure 5.1: Active URLs over time (1995-2023)

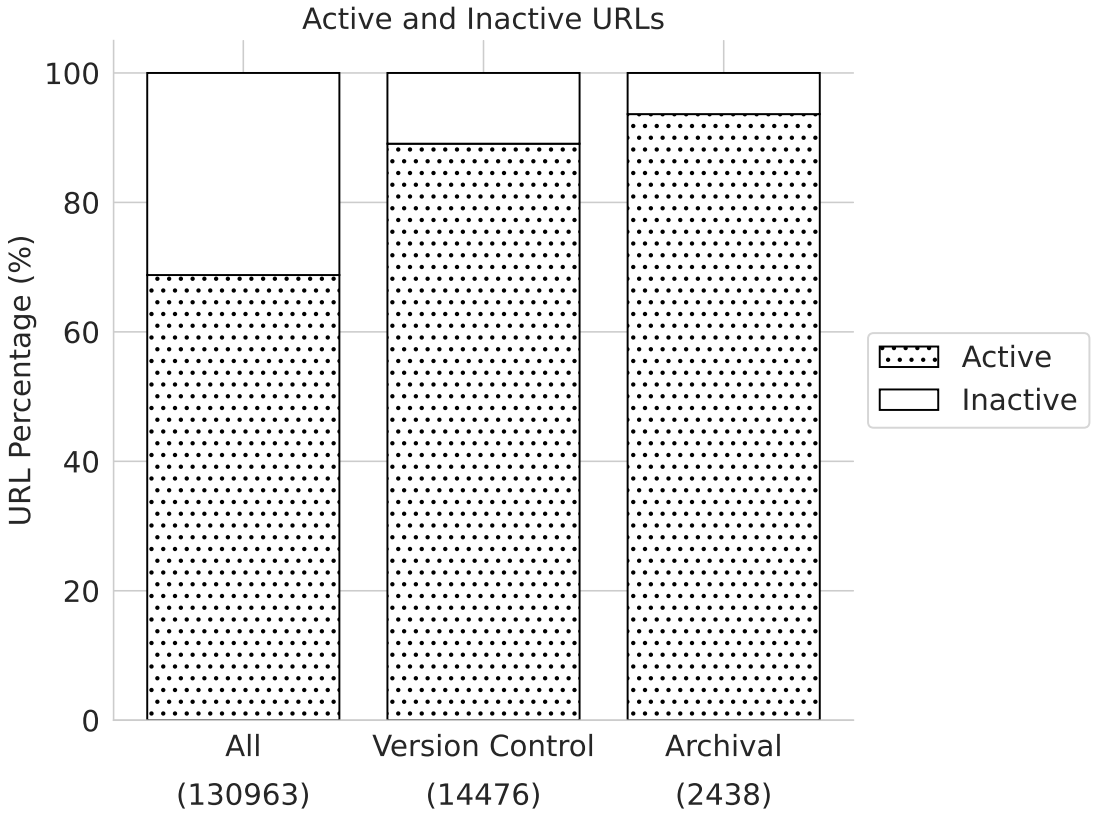

Figure 5.2: Ratio of active URLs by hosting purpose: all, version control systems, archival platforms.

Table 5.1: Publication venues by active URL ratio

Venue	**Active URL** (%)	Venue	**Active URL** (%)
ICSME	81.37%	SOSYM	68.13%
FSE	80.69%	GPCE	66.65%
SANER	80.05%	RE	66.44%
TOSEM	79.92%	SMR	65.73%
MSR	79.79%	SQJ	65.35%
ICPC	79.32%	ASEJ	65.06%
ISSTA	79.31%	IST	64.36%
ASE	77.62%	SPE	64.04%
SCAM	77.34%	JSS	63.29%
ICSE	75.93%	SSBSE	62.11%
ICST	75.33%	ISSE	61.54%
ESEM	74.62%	ICSM	57.49%
ESE	74.33%	NOTES	56.01%
TSE	73.19%	IJSEKE	55.67%
STVR	71.92%	WCRE	55.38%
MODELS	69.76%	SW	55.10%
FASE	69.29%	REJ	54.25%

which are not archival services, but do provide versioning—is also contributing to the kind of URLs scientists use in their papers.

To investigate how these practices contribute to the persistence of URL references in software engineering papers, we classified URLs as: version control URLs, archival URLs, and all URLs, regardless of of special origin. To be classified as a version control, a URL must point to popular version control platforms such as GitHub, Bitbucket, or GitLab. These platforms are typically used for hosting and managing collaborative projects (most often for code, but also for data and websites), involving regular updates and revisions.

Archival URLs, on the other hand, are those that lead to digital archives or repositories with the stated mission of long-term preservation of digital assets. We classify as archival URLs those that point to the following platforms: Zenodo, arXiv, Internet Archive, Software Heritage, and Figshare.

Figure 5.2 shows a breakdown of URL stability by hosting type. Both version control and archival URLs tend to are more stable than the entire URL corpus, and by a significant margin. Also, as one would hope, archival URLs are more stable than VCS URLs, although only by a small margin. The use of archival URLs remain overall scarce though: about 22k URLs for archival versus 136k for version control URLs.

5.3 URL stability by venue

Table 5.1 provides a breakdown of the ratio of active URL by venue (see Table 4.1 for an expansion of the venue acronyms). We could not identify any clear pattern although, anecdotally, the top venues in the field (e.g., FSE, TOSEM, ASE, ICSE, and TSE) are all in the top-half of the ranking by active URL ratio. The spread in the ratios is very high (54–81%) and does not appear to be induced by the conference age.

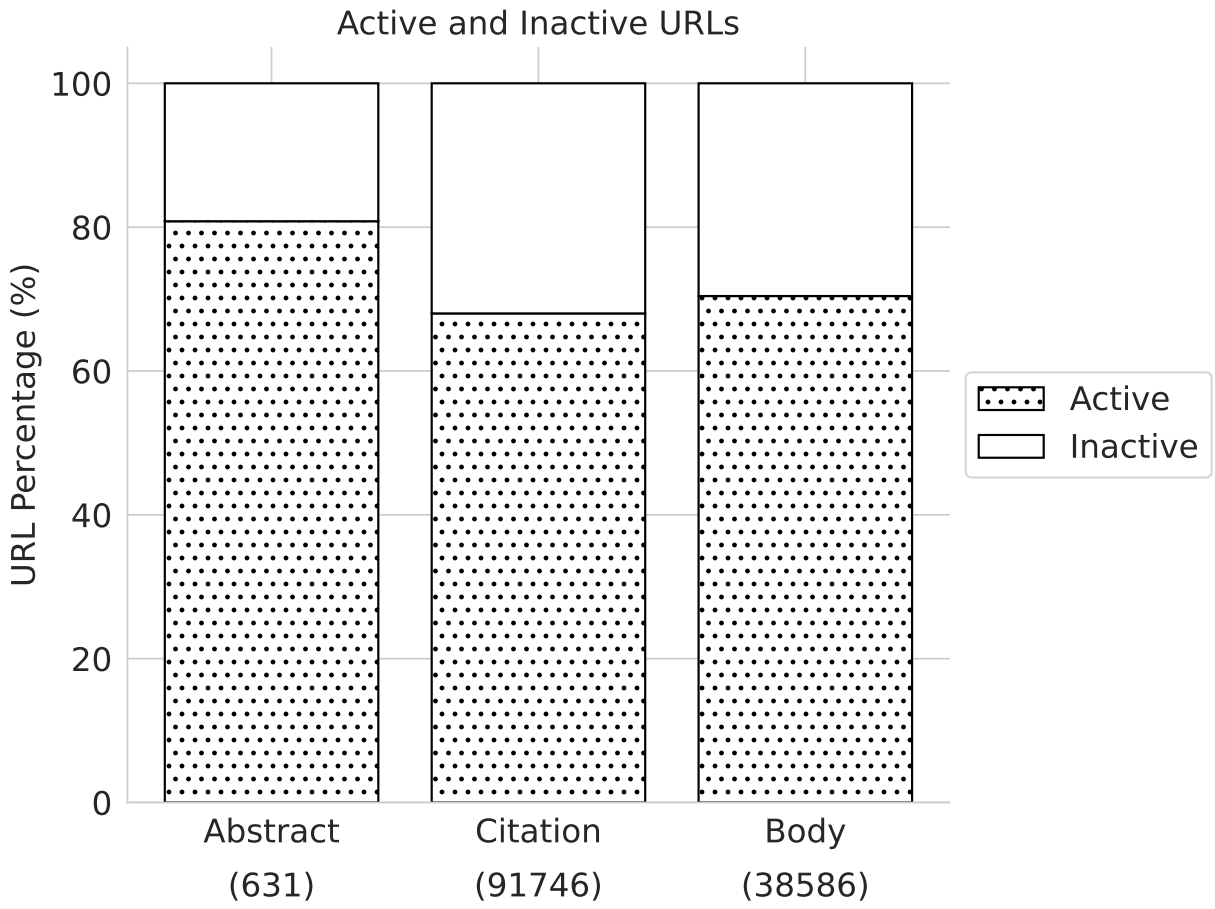

Figure 5.3: Active versus Inactive URLs by section

5.4 URL stability by document section

URLs are used in papers for different purposes and can hence appear in different document sections. We analyzed three classes of document placements, based on the section a URL was found in: abstract, citation ("References" sections), body (everywhere else in the document). A breakdown of active URL ratios by placement is shown in Figure 5.3.

URLs are not used much in abstracts, but when they are they appear to be very stable (80.82% of active ratio). URLs are used the most in citations (e.g., DOI URLs or website for non-paper references), but there they show the worst active ratio: 68.01%. In between, both in terms of frequency of use and active ratio, we find URLs located elsewhere in papers, with an active ratio of 70.42%.

5.5 URL stability by scheme, domain, host

URLs are composed of various parts, the initial ones being URI scheme (e.g., `http`) and host name (e.g., `acm.org`), with the latter being further decomposable into a top level domain (e.g., `.com`) and the rest of it. In this subsection we examine the impact of the various URL parts on

Table 5.2: Top-level URL domains (TLD) by active URL ratio

TLD	**Total URLs**	**Active URL** (%)
`.press`	157	100.0%
`.dev`	161	89.44%
`.blog`	101	87.13%
`.cc`	363	86.5%
`.to`	61	85.25%
`.co`	271	83.76%
`.be`	460	83.7%
`.ai`	98	83.67%
`.io`	3,231	79.54%
`.ly`	1,029	79.4%

URL rot in scientific papers.

URI scheme Two main URI schemes are used these days for Web references: `http` and `https`, the latter denoting Web accesses cryptographically protected against man-in-the-middle attacks via the TLS protocol.

We initially expected `https` URIs to be more fragile than `http` ones, because more "moving parts" can fail with the former: a certificate can expire, client and host can disagree on acceptable cryptographic algorithms, etc. To our surprise, the reverse is true. Active URL percentages by protocol are as follows:

- `http`: 58.67%
- `https`: 81.58%

This can be explained by the fact that, taken a long enough observation period (50 years in our case), `https` URLs tend to be younger (due to protocol adoption delays) and hence more likely to be still active to this day. But it is also possible that `https` URLs correlated with better URL referencing practices that were not followed decades ago.

Top-Level Domains (TLDs) The analysis of inactivity rates across different top-level domains (TLDs) demonstrates significant variations in URL stability. TLDs with more than 50 URLs were considered to ensure accuracy and avoid skewed results due to small sample sizes.

Table 5.2 includes the results per TLD.

Domains Further analysis focused on specific domains and their URL inactivity rates. Table 5.3 shows inactivity percentages for the top domains in terms of the absolute number of inactive URLs.

5.6 Network errors

This subsection examines the distribution of HTTP and network errors identified in URL accessibility checks, providing insight into common issues impacting URL stability.

Table 5.3: URL domains by inactive URL ratio

Domain	**Inactive URLs** (count)	**Inactive URL ratio** (%)
nasa.gov	191	72.08%
cmu.edu	480	70.59%
mit.edu	173	53.23%
stanford.edu	175	51.32%
mozilla.org	250	45.54%
omg.org	442	39.32%
goo.gl	272	33.01%
ibm.com	397	31.73%
wikipedia.org	304	25.98%
eclipse.org	349	24.66%
doi.org	327	22.26%
github.io	284	21.18%
bit.ly	201	20.14%
google.com	388	15.65%
sourceforge.net	264	14.71%
microsoft.com	227	14.64%
apache.org	265	14.10%
github.com	1,624	11.17%
acm.org	436	7.30%

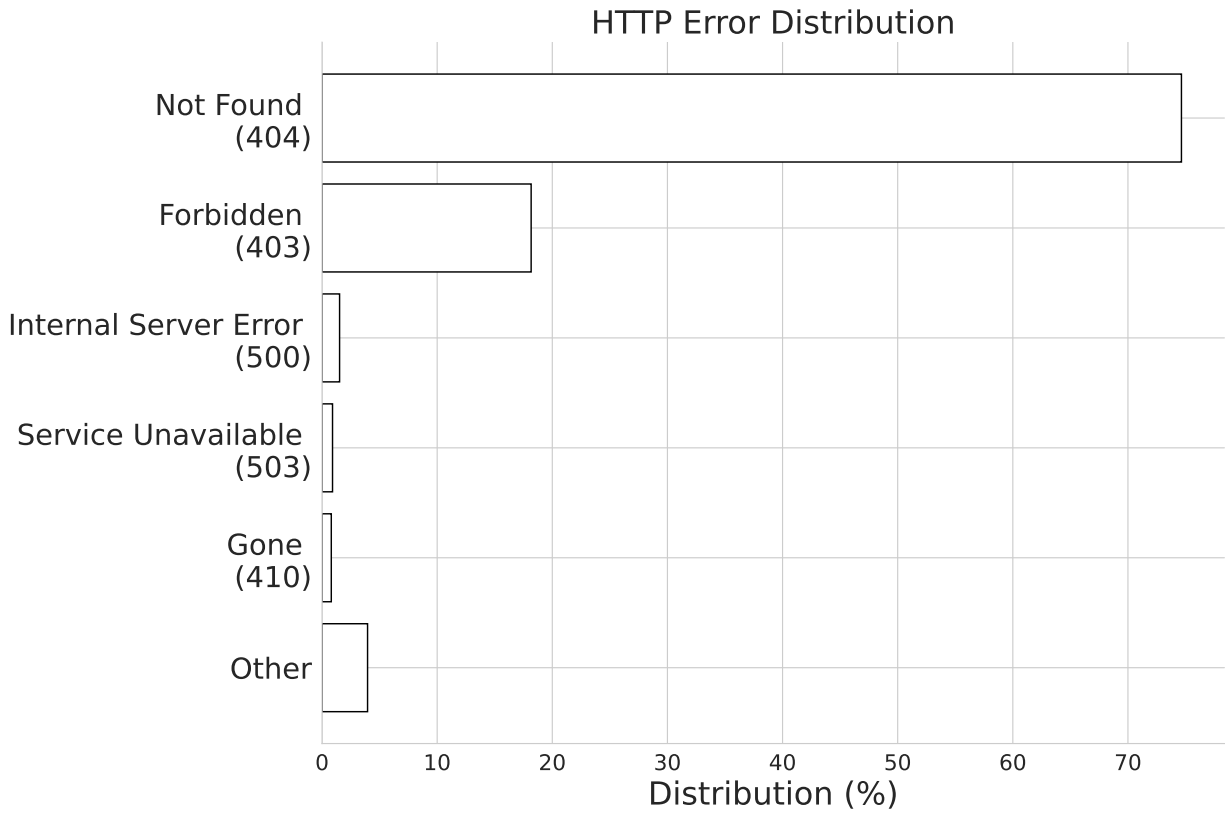

Figure 5.4: HTTP error distribution

5.6.1 HTTP Error Distribution

The analysis of HTTP errors involved categorizing the URLs based on the HTTP status codes received during accessibility checks.

Figure 5.4 illustrates this distribution, providing a visual representation of the prevalence of different HTTP errors.

5.6.2 Network Error Analysis

Network error analysis sheds light on the types of connectivity issues encountered when accessing URLs. In this study, the network errors recorded during URL accessibility checks are specifically UNIX errors. These errors provide insights into the technical challenges encountered when accessing URLs from UNIX-like systems, which were used in the automated URL verification processes in this work.

Figure 5.5 visually depicts these network errors, emphasizing their relative frequencies.

These analyses of HTTP and network errors provide valuable insights into the technical challenges affecting URL accessibility in digital scholarly communications.

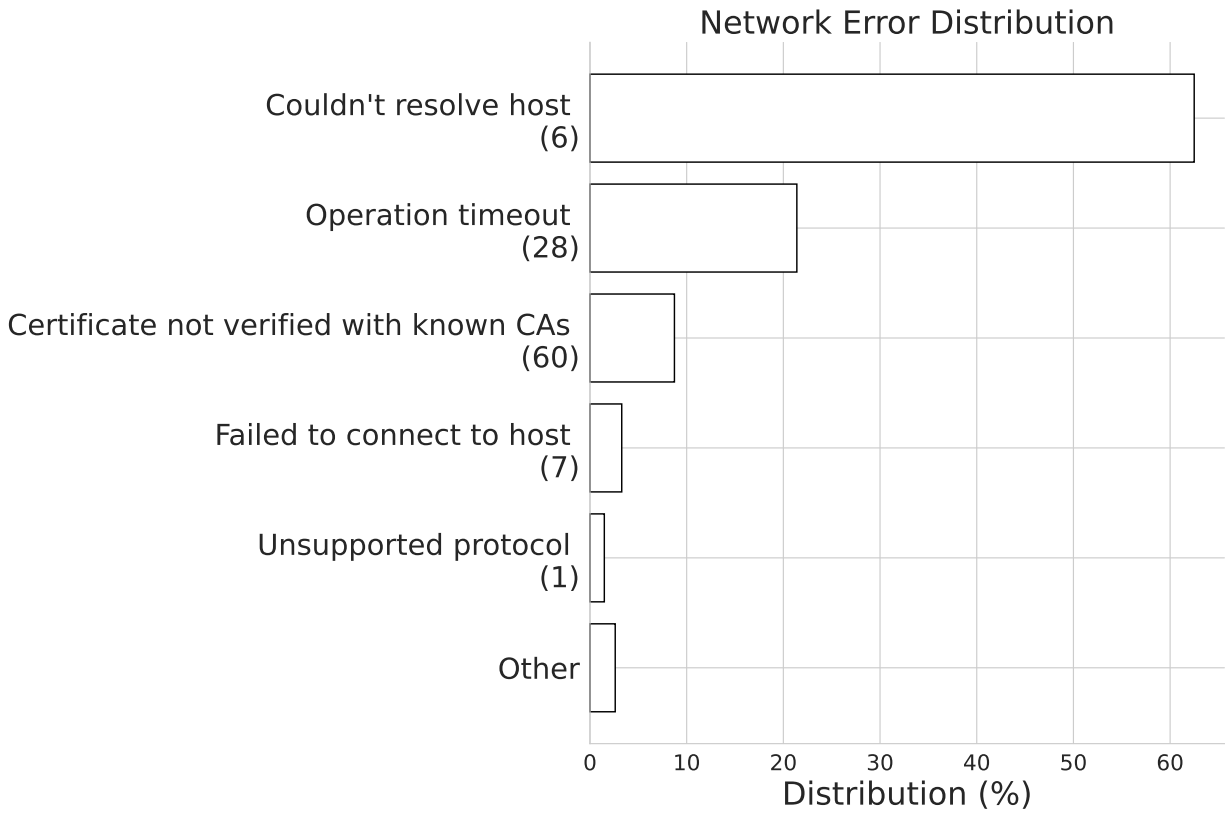

Figure 5.5: Network error distribution

Table 5.4: Linear regression coefficients for URL persistence factors

Variable	Coefficient
URL Length	-0.001
Number of Path Elements	-0.015
Journal Impact Factor	0.016
CORE Ranking	0.028

5.7 Regression analysis

The regression analysis constituted a significant component of this study, aiming to illuminate the relationships between multiple factors and the persistence of URLs in scientific publications. A linear regression model was employed to investigate these relationships, chosen for its clarity in interpreting results and its efficacy in isolating the impact of individual variables.

The regression model incorporates several key variables, including:

- **URL Length:** The total number of characters in the URL.
- **Number of Path Elements:** The count of individual elements in the URL's path.
- **Journal Impact Factor:** The impact factor of the journal where the URL was published.
- **Core Ranking:** The academic ranking of the conference associated with the URL.

Table 5.4 presents the coefficients derived from the regression model. Each coefficient reflects the influence of its corresponding variable on the likelihood of URL persistence. Negative coefficients indicate a decrease in the likelihood of URL persistence with an increase in the variable value, while positive coefficients suggest an increase in URL persistence likelihood.

This analysis provided valuable insights into how various academic and structural factors interact with URL persistence, offering a nuanced understanding of the elements contributing to digital resource longevity in scholarly communication.

Chapter 6

Discussion

6.1 Comparison with Previous Studies

Drawing upon a substantial dataset, this study contributes to the evolving narrative of URL persistence in scientific literature. The analysis contrasts with many previous studies through the use of a broader paper pool and a more inclusive criterion for active URLs. The decision to classify URLs as active in the absence of error responses aligns with the large scale of the dataset, making this approach both pragmatic and necessary.

Comparison of URL accessibility rates across years reveals significant advancements in digital resource availability. For instance, our study found a 83.77% active URL rate in 2023, a notable increase compared to the 15.97% in 1996.

Interestingly, our study's overall active URL rate of 68.78% aligns closely with certain previous findings, such as the 56.4%-81.4% range reported by Casserly and Bird [3] after reevaluation. However, it also highlights a significant increase from the much lower accessibility rates observed in earlier years, as evidenced by the 23-53% range found by Lawrence et al. [17].

A critical improvement observed in this study pertains to the half-life of URLs. Remarkably, our analysis recorded a URL half-life of 9.68 years, a substantial increase when compared to the latest half-life of 6.5 years reported by Bansal [1] in our literature review.

In summary, while this study builds upon the methodologies of prior research, it also contributes new insights, particularly in the context of recent technological advancements and changes in digital archiving practices. The evolution in URL persistence, as shown by the comparison with earlier studies, underscores the dynamic nature of digital resource management in scholarly communication.

6.2 Implications of URL Stability Trends

The analysis of URL stability trends across different years uncovers significant insights into the dynamics of digital resource availability in academic literature. Notably, the year 2020 showed a relatively high URL accessibility rate of 79.98%, suggesting an improvement in the stability of digital resources. This is in stark contrast to the earlier years, such as 1999, with only 28.88% active URLs. Such variability indicates the changing nature of web resources over the years, influenced by advancements in technology and digital archiving practices [13].

The progression from a complete absence of active URLs in the early years to higher accessibility rates in recent years mirrors the overall trend towards more stable and persistent digital references in scholarly publications. This shift can be attributed to several factors, including the widespread

adoption of more robust web technologies [33], improvements in the archiving of digital content [26], and the transition from HTTP to HTTPS protocols [30], which offer enhanced security and reliability.

A key discovery from our research is the observation of an extended half-life compared to the values documented in prior studies. The increasing stability of URLs enhances the reliability of digital references, contributing to the integrity and longevity of academic work. However, the observed variability also underscores the need for continued vigilance and adaptation in citation practices, ensuring that references remain active over time. Researchers and publishers alike must recognize the evolving nature of digital resources and embrace strategies that promote the long-term availability of web-based references.

6.3 Impact of Venue and Domain on URL Stability

Our study's examination of URL stability across different publication venues and top-level domains (TLDs) reveals clear patterns, providing insights into digital resource management in academic settings. The active URL percentages varied significantly by venue, with rates as low as 54.25% in REJ and as high as 81.37% in ICSME. This suggests that some venues may have more effective practices in place for ensuring URL accessibility, possibly due to stronger digital archiving policies or a more technology-aware audience.

In terms of TLDs, the analysis showed differences in URL stability. For example, TLDs such as '.press' and '.dev' exhibited higher activity rates, at 100% and 89.44% respectively. These differences might be attributed to varying domain management and renewal practices, which can influence the longevity of URLs associated with these domains [35].

For researchers and publishers, understanding these variations in URL stability is crucial [35]. It highlights the importance of considering the choice of domains or venues when citing digital resources, as some may offer greater URL longevity than others. Publishers and those managing digital archives are reminded of the significance of their role in maintaining digital references. Effective digital preservation strategies are essential to ensure the continued accessibility of scholarly work.

6.4 Domain Protection Mechanisms

In the course of our automated domain access attempts, we observed a notable phenomenon: a substantial proportion of domains classified as inactive were, in fact, protected by services such as Cloudflare and similar domain protection mechanisms. These services are designed to safeguard websites from a range of threats, including Distributed Denial of Service (DDoS) attacks and automated bot traffic, by acting as intermediaries between the user and the server.

The presence of these protection mechanisms introduces a significant challenge to automated URL accessibility assessments. Such services often implement verification processes, such as CAPTCHAs, that impede automated access and can result in the misclassification of domains as inactive or unavailable. This finding suggests that the increasing prevalence of domain protection mechanisms may complicate the evaluation of URL persistence, a factor that has not been extensively addressed in existing literature.

This observation points to a potential shift in the landscape of digital resource accessibility, underscoring the need for future research to consider the impact of these protection services. Accurately distinguishing between truly inactive domains and those merely shielded by security mechanisms

is essential to improve the reliability of URL stability analyses. Moreover, this insight suggests that the methodologies employed in such studies may need to be refined to account for the growing adoption of domain protection strategies.

6.5 Technical Challenges in URL Accessibility

The study's exploration into URL accessibility surfaced several critical technical challenges, providing a clearer picture of the digital landscape faced by academic literature. Predominant among these were HTTP errors, with 'Not Found (404)' and 'Forbidden (403)' emerging as the primary culprits. The 'Not Found' errors, accounting for a significant 74.65%, often reflect inactive links or moved content. On the other hand, 'Forbidden' errors, constituting 18.16%, hint at access control issues, perhaps signaling shifting permissions or privacy settings on web resources [11].

Network errors further compound these challenges, with 'Couldn't resolve host' and 'Operation timeout' standing out. These errors, which occurred in 62.49% and 21.4% of cases, respectively, underscore the delicate balance of server reliability and network stability.

Addressing these issues calls for innovative and forward-thinking strategies. For 'Not Found' errors, a potential solution could be the integration of automated link-checking tools within the publication process, offering real-time alerts to authors and editors about inactive links [6].

For 'Forbidden' errors, a more transparent and consistent policy regarding digital rights and access permissions could be beneficial. This might include clearer guidelines for authors on the use of content and the adoption of standard practices for content accessibility in academic publications.

The frequency of network errors suggests the need for a robust infrastructure. Academic platforms could explore the use of redundant hosting or content delivery networks to minimize downtime and enhance content availability [9].

6.6 Insights from Regression Analysis

The regression analysis undertaken in this study offers insights into the factors influencing URL persistence in scientific publications. Notably, the analysis included variables such as URL length, number of path elements, journal impact factor, and CORE ranking. However, contrary to initial expectations, the results did not reveal a strong correlation between these variables and URL stability.

The lack of a significant correlation suggests that URL persistence may be influenced by a complex interplay of factors not fully captured by the variables considered in this study. For instance, the absence of a strong relationship between journal impact factor or CORE ranking and URL stability implies that the prestige of the publication venue does not necessarily guarantee the longevity of its digital references. Similarly, URL length and the number of path elements did not exhibit a clear impact on URL persistence, indicating that simpler or shorter URLs are not always more stable.

These findings prompt a reflection on the multifaceted nature of URL stability in academic literature. It highlights the potential influence of other, unexamined factors that might contribute to the persistence of digital references [13]. For future research, this suggests the exploration of additional variables, such as the type of hosting platform, the frequency of content updates, or the nature of the linked content, which could provide further clarity on what contributes to the longevity of URLs in scholarly publications.

Chapter 7

Conclusion

7.1 Summary of Findings

The persistence of URLs in scientific literature is shaped by a complex interplay of factors, including technological challenges, the policies and practices of publication venues, and the inherent characteristics of web domains. This study has provided a comprehensive analysis of URL stability within software engineering scholarly works, revealing both significant progress and ongoing challenges in the field of digital preservation.

Our research has demonstrated a notable improvement in the half-life of URLs, with an observed increase to 9.68 years. This extension in URL longevity suggests that efforts to enhance the robustness of digital references are yielding positive results. However, the dynamic and often transient nature of web resources continues to pose significant challenges, as evidenced by the temporary inactivity of some URLs. These findings emphasize the need for continuous monitoring and adaptive strategies to maintain the reliability of digital references in scholarly communication.

Furthermore, our regression analysis revealed the absence of strong correlations between URL stability and factors such as journal impact factor and CORE ranking. This outcome suggests that URL persistence is likely influenced by a broader range of variables that were not fully captured in this study. The complexity of these relationships highlights the importance of further research to identify additional factors that contribute to the longevity of URLs in academic publications.

7.2 Future Work

The insights gained from this study open several avenues for future research, which are crucial for advancing our understanding of URL stability and improving digital preservation practices:

7.2.1 Investigation of Additional Variables

Future studies should explore a wider array of variables that may influence URL stability, such as the characteristics of hosting platforms, the frequency of content updates, and the specific nature of the linked content. By broadening the scope of analysis, researchers can develop a more nuanced understanding of the factors that contribute to URL persistence and identify strategies to mitigate the risks of link rot.

7.2.2 Enhancement of URL Verification Methodologies

The discovery that some domains with high inactivity rates were actually protected by services like Cloudflare points to a critical challenge in URL verification processes. Future work should focus on developing more sophisticated automated tools capable of accurately detecting and bypassing these protective mechanisms. Such advancements would improve the accuracy of URL persistence studies and reduce the likelihood of misclassification.

7.2.3 Longitudinal and Cross-Disciplinary Studies

While this study focused on software engineering, similar research should be conducted across other academic disciplines to determine whether the trends observed here are unique to this field or represent a broader phenomenon. Longitudinal studies, in particular, could provide valuable insights into how URL stability evolves over time within different academic contexts.

7.2.4 Impact of Emerging Web Technologies

As the digital landscape continues to evolve, it is essential to examine the impact of emerging web technologies, such as decentralized web protocols, on URL stability. Understanding how these technologies affect the preservation of digital resources will be critical for developing future-proof strategies that ensure the longevity of scholarly references.

7.3 Concluding Remarks

In conclusion, this study has significantly contributed to our understanding of URL stability in software engineering literature. The findings highlight both the progress made and the challenges that remain in preserving digital resources. The observed increase in URL half-life is encouraging, yet the ongoing issues of temporary URL inactivity and the lack of strong correlations in our regression analysis suggest that much work remains to be done.

As the digital environment continues to change, the academic community must remain vigilant and proactive in adapting its practices. Collaborative efforts among researchers, publishers, and digital archivists will be essential in developing and implementing effective strategies to ensure the integrity and longevity of scholarly communication. The future of digital preservation depends on our ability to anticipate and respond to the evolving challenges of URL persistence in the academic domain.

Bibliography

[1] Sonia Bansal. Decay of url references cited in desidoc journal of library & information technology. 07 2021.

[2] Sonia Bansal and Seema Parmar. Decay of urls citation: A case study of current science, 01 2020.

[3] Mary F. Casserly and James E. Bird. Web citation availability: Analysis and implications for scholarship. *College & Research Libraries*, 64(4):300–317, July 2003.

[4] croumegous. doi2pdf: A Command Line Tool to Download PDFs of Research Paper from DOI, Name or URL. `https://pypi.org/project/doi2pdf/`, 2023. Accessed: 2024-02-03.

[5] Marilyn Deegan and Simon Tanner. *Digital Preservation*. Facet Publishing, 2013.

[6] Stephan Druskat, Oliver Bertuch, Guido Juckeland, Oliver Knodel, and Tobias Schlauch. Software publications with rich metadata: state of the art, automated workflows and hermes concept, 2022.

[7] Monica Mensah Emmanuel Adjei and Eric Amponsah Amoaful. The story so far-digital preservation in institutional repositories: The case of academic libraries in ghana. *Digital Library Perspectives*, 35(2):80–96, 2019.

[8] Emily Escamilla, Martin Klein, Talya Cooper, Vicky Rampin, Michele C. Weigle, and Michael L. Nelson. *Cited But Not Archived: Analyzing the Status of Code References in Scholarly Articles*, page 194–207. Springer Nature Singapore, 2023.

[9] Jose Luis Garcia-Dorado. Bandwidth in the cloud. 2015.

[10] Carol Anne Germain. URLs: Uniform resource locators or unreliable resource locators. *College & Research Libraries*, 61(4):359–365, July 2000.

[11] Mandeep Kaur Gondara. Access control mechanisms for semantic web services-a discussion on requirements & future directions. 2011.

[12] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[13] Mat Kelly, Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson. *On the Change in Archivability of Websites Over Time*, page 35–47. Springer Berlin Heidelberg, 2013.

[14] Wallace Koehler. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2):162–180, 1999.

[15] Wallace Koehler et al. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2):9–2, 2004.

[16] Brian F. Lavoie. *The Open Archival Information System Reference Model: Introductory Guide*. Springer, 2013.

[17] Steve Lawrence, Frans Coetzee, Eric Glover, Gary Flake, David Pennock, Bob Krovetz, Finn Nielsen, Andries Kruger, and Lee Giles. Persistence of information on the web. In *Proceedings of the ninth international conference on Information and knowledge management - CIKM '00*. ACM Press, 2000.

[18] K. Leong. curl-cffi: A python library providing a cffi interface to libcurl. `https://pypi.org/project/curl-cffi/`, 2022.

[19] K. Leong. curl-cffi: A python library providing a cffi interface to libcurl. `https://pypi.org/project/curl-cffi/`, 2022.

[20] Michael Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. In Alberto H. F. Laender and Arlindo L. Oliveira, editors, *String Processing and Information Retrieval, 9th International Symposium, SPIRE 2002, Lisbon, Portugal, September 11-13, 2002, Proceedings*, volume 2476 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 2002.

[21] Fayaz Ahmad Loan and Ufaira Yaseen Shah. The decay and persistence of web references. *Digital Library Perspectives*, 36:157–166, 5 2020.

[22] Patrice Lopez. GROBID - generation of bibliographic data, 2008–2021. Accessed 2022-01-25.

[23] Julien Masanès. *Web Archiving*. Springer, 2013.

[24] Frank McCown, Sheffan Chan, Michael L. Nelson, and Johan Bollen. The availability and persistence of web references in d-lib magazine. 2005.

[25] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[26] Abderrazak Mkadmi. *Digital Archiving: Methods and Strategies*, chapter 2, pages 31–69. John Wiley & Sons, Ltd, 2021.

[27] Jeremy Myntti and Jessalyn Zoom. *Digital Preservation in Libraries: Preparing for a Sustainable Future*. ALA Editions, 2019.

[28] Beth Oehlerts and Shu Liu. Digital preservation strategies at colorado state university libraries. *Library Management*, 34(1/2):83–95, 2013.

[29] M.K. Saberi and H. Abedi. Accessibility and decay of web citations in five open access ISI journals. *Internet Research*, 22(2):234–247, March 2012.

[30] Wazen M. Shbair, Thibault Cholez, Jerome Francois, and Isabelle Chrisment. A survey of https traffic and services identification approaches, 2020.

[31] Diomidis Spinellis. The decay and failures of web references. *Communications of the ACM*, 46(1):71–77, January 2003.

[32] SQLAlchemy. Sqlalchemy: The database toolkit for python, 2024.

[33] Nicolas Tempelmeier, Elena Demidova, and Stefan Dietze. Inferring missing categorical information in noisy and sparse web markup. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, WWW '18. ACM Press, 2018.

[34] Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers, 2018.

[35] Timothy H. Vines, Rose L. Andrew, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Nolan C. Kane, Jean-Sébastien Moore, Brook T. Moyers, Sébastien Renaut, Diana J. Rennison, Thor Veen, and Sam Yeaman. Mandated data archiving greatly improves access to research data. *The FASEB Journal*, 27(4):1304–1308, January 2013.

[36] Cassie Wagner, Meseret D. Gebremichael, Mary K. Taylor, and Michael J. Soltys. Disappearing act: decay of uniform resource locators in health care management journals. *Journal of the Medical Library Association : JMLA*, 97(2):122–130, April 2009.

[37] Robin L. Wendler. *A Metadata Standard for Preservation: PREMIS*. Springer, 2013.

[38] J. D. Wren. 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, 20(5):668–672, January 2004.

[39] J. D. Wren. URL decay in MEDLINE–a 4-year follow-up study. *Bioinformatics*, 24(11):1381–1385, April 2008.

[40] Jonathan D. Wren, Kathryn R. Johnson, David M. Crockett, Lauren F. Heilig, Lisa M. Schilling, and Robert P. Dellavalle. Uniform resource locator decay in dermatology journals. *Archives of Dermatology*, 142(9), September 2006.

List of Abbreviations

URL ..... Uniform Resource Locator

HTTP ..... Hypertext Transfer Protocol

HTTPS ..... Hypertext Transfer Protocol Secure

TLD ..... Top-Level Domain

UNIX ..... Uniplexed Information and Computing Service

CAPTCHA ... Completely Automated Public Turing test to tell Computers and Humans Apart

DDoS ..... Distributed Denial of Service

DBLP ..... Digital Bibliography & Library Project

PDF ..... Portable Document Format

XML ..... Extensible Markup Language

CORE ..... Computing Research and Education

DOI ..... Digital Object Identifier