NISO TR-05-2013

IOTA Working Group Summary of Activities and Outcomes

A Technical Report of the
National Information Standards Organization

Approved: April 26, 2013

About NISO Technical Reports

NISO Technical Reports describe the work and conclusions of a working group chartered by the organization to study a particular problem or issue that is related to NISO's standardization interests. However, unlike NISO standards or recommended practices, they do not have requirements for compliance or recommend "best practices"—although their conclusions may contain some recommendations. Additionally, no consensus process of the NISO voting membership was used in their creation or approval. They are strictly informative in nature.

Published by
National Information Standards Organization (NISO)
3600 Clipper Mill Road, Suite 302
Baltimore, MD 21211
www.niso.org
301-652-2512

ISBN: 978-1-937522-17-9

Contents

Foreword

About this Technical Report

In November 2009, Adam Chandler submitted a proposal (Chandler, 2009a) to NISO's Business Information Topic Committee to investigate the feasibility of creating industry-wide, transparent and scalable metrics for evaluating and comparing the quality of OpenURL implementations across content providers. The proposed project was to build on work developed under a 2008/2009 Mellon Foundation grant, for which Chandler was the lead investigator. An existing reporting system and database were to be ported to a NISO-supported server and the new working group would collect further data, analyze the data, and develop metrics to evaluate OpenURL quality.

The proposal was approved by the Business Information Topic Committee in December 2009 and the Working Group that was formed was presented with the following problem statement:

> The OpenURL standard is a widely deployed technology to facilitate linking to resources across the library supply chain. The OpenURL-formatted URL carries the data about an item to the link resolver of the library. The resolver compares the metadata embedded within the OpenURL with what is held in the library's collection and presents the available options in a results page. For a book, there is usually a link to the library's catalog card; for an article, ideally this is a link directly to the full-text of the article. At a typical academic library thousands of OpenURL requests are initiated by patrons each week. The problem is too often these links do not work as expected, leaving patrons frustrated by a lower than desired quality of service. Periodically mention is made in the library literature to problems with OpenURL linking, but since the OpenURL standard was introduced a decade ago, no systematic method has been designed and carried out to benchmark it. This work is intended to fill the gap.

The study was to be conducted over a two-year period, concluding with a report reviewing the results and making recommendations about how or whether to continue the IOTA research. This technical report is the project's promised deliverable.

NISO Discovery to Delivery Topic Committee Members

The Discovery to Delivery Topic Committee had the following members at the time it approved this Technical Report.

Tim Babbitt
Cambridge Information Group

Pascal Calarco, Co-chair
University of Waterloo Library

Peter Murray
Lyrasis

Jeff Penka
OCLC Online Computer Library Center

Tim Shearer
University of North Carolina Chapel Hill Libraries

Lucy Harrison, Co-chair
Florida Virtual Campus

Chris Shillum
Reed Elsevier

NISO IOTA Working Group Members

This technical report was developed by the following members of the Improving OpenURL Through Analytics (IOTA) Working Group:

Adam Chandler (Chair)
Cornell University Library

Rafal Kasprowski
Rice University

Susan Marcin
Columbia University

Oliver Pesch
EBSCO Information Services

Clara Ruttenberg
University of Maryland

Elizabeth Winter
Georgia Tech Library

Jim Wismer
Thomson Reuters

Aron Wolf
Serials Solutions

Acknowledgements

The IOTA Working Group would like to acknowledge the important contributions made by Ellen Rotenberg during the first year of the initiative. We would also like to acknowledge the contribution of Phil Davis, who provided crucial advice during our experiments with statistical approaches to measuring OpenURL metadata quality.

Trademarks, Services Marks

1 Introduction

1.1 Purpose and Scope

The purpose of the IOTA (Improving OpenURL Through Analytics) project was to analyze the quality of the metadata that is passed to the link resolver from the OpenURL source and develop a standardized metric for such quality evaluation. The project focused most of its attention on a specific genre of OpenURLs—those intended to provide access to journal articles.

Essentially there are three parts to OpenURL linking: the metadata in the incoming OpenURL, matching metadata in the OpenURL to the library's knowledge base, and constructing accurate links into content provider websites.

The quality of the data in the link resolver knowledge base itself was outside the scope of IOTA. Also outside of the scope was the quality of the linkages between link resolvers and full-text content providers.

A related NISO initiative, Knowledge Base And Related Tools (KBART), is concerned with accurate knowledge base content in the link resolver. Standardizing the link syntax between the link resolver and content provider remains unaddressed by any industry effort. Chandler's 2011 *Against the Grain* article describes the relationship between these initiatives in more detail.

1.2 Terms and Definitions

The following terms, as used in this technical report, have the meanings indicated. Terms in boldface in a definition indicate the term is also defined in this section.

Term	Definition
A&I database Abstract and Index database	A content discovery product that provides descriptive information about content items but not the full text of the content. A&I databases are frequently a **source** in **OpenURL** linking.
completeness	The number of metadata **elements** provided in the **OpenURL** out of a desired or core number.
Completeness Index	A number that is attributed to a content provider (**OpenURL** r**eferrer** or **source**) to measure the **completeness** of the provider's **OpenURLs** in aggregate. It is essentially an average of the **Completeness Scores** of **OpenURLs** coming from that content provider.
Completeness Score	The measure of the **completeness** of a single **OpenURL**.

Term	Definition
DOI® Digital Object Identifier	A unique and persistent identifier for a digital content object, such as an online article, journal, chapter, book, or image, using a syntax defined in ANSI/NISO Z39.84 and ISO 26324. The DOI provides current information about the object, including where to find it on the internet. Typically registered with an authorized registration agency, e.g., CrossRef (www.crossref.org), whereby it is possible to use the DOI to look up descriptive metadata about a content item using a provided free service. Sometimes called a DOI name.
discovery platform	A website or product through which users can discover content items.
element	Descriptive metadata such as ISSN, Volume, Issue, Start Page, etc. used in an **OpenURL**.
Element Weight	A value assigned to an **element** and used in calculation of a **Completeness Score**. The value of the Element Weight represents the relative importance of the **element** to the success of the **OpenURL** in providing the user access to full text; the higher the value the more important the **element**.
Enhanced OpenURL	The set of **OpenURL elements** available after the **link resolver enhancers** have been run on the incoming **OpenURL**. Typically an Enhanced OpenURL will have more data **elements** than the original **OpenURL**.
enhancer	An automated process that enhances an **OpenURL** by using internal and external sources to supplement the **OpenURL** data **elements**. An example would be an enhancer that looks up article-level metadata from CrossRef using the **DOI** presented on an **OpenURL**.
fail	Describes the state of an **OpenURL** that does not generate any item-level links to full-text items. *See also* **success**. NOTE: Within the context of this technical report the following are not considered links to full-text items: links to a journal homepage, the table of contents for an issue, or a search page at the vendor site.
full-text target link	A link to the complete text—including all references, figures, and tables—of an article on the **target** site. If the referenced item is a journal article, the full-text article link would direct the user directly to that article at the content provider's site without requiring further navigation or searching.

Term	Definition
knowledge base	A database used with **OpenURLs** that contains information about what **targets** are available to users of the **link resolver**. Data within the knowledge base includes but may not be limited to link syntaxes and holdings with coverage details. The **link resolver** environment for a given institution can be customized to reflect that institution's collection and to only provide links to **targets** to which the library subscribes.
link resolver	Technology that controls the linking between **sources** and **targets**. The link resolver accepts and deconstructs the **OpenURL** describing a content item from a **source** and uses its **knowledge base** and associated programs to determine full text and other t**argets** appropriate for the user and create predictable links to these. The role of the link resolver is context sensitive linking to the appropriate copy of a content item.
OpenURL	As defined in ANSI/NISO Z39.88-2004, *The OpenURL Framework for Context-Sensitive Services*, a URL designed to transport metadata and thus enable linking from information resources such as **abstracting and indexing databases** (**sources**) to library services (t**argets**), such as academic journals, whether online or in printed or other formats. The linking is mediated by **link resolvers**, or link-servers, which parse the elements of an **OpenURL** and provide links to appropriate t**argets** available through a library by the use of a **knowledge base**. *Source:* Wikipedia [boldface notation added]
PubMed ID	A unique identifier assigned to a record in the PubMed databases (www.pubmed.org). A PubMed record typically describes an article or other content item. Using a PubMed ID it is possible to look up descriptive metadata of the content item through a free online service provided by the National Library of Medicine.
referrer	The identity of a website or **discovery platform** that is the **source** of the **OpenURL**. In an **OpenURL** this is represented as the SID (Source ID) or the Referrer ID. In some contexts, **referrer** is used interchangeably with **source**.
required data elements	The data **elements** that must be available to create a successful link to an **OpenURL** t**arget**. For example, if a publisher's link to a full-text article is constructed from ISSN, Volume, Issue, and Start Page values, then these fields are considered required data elements since the link will fail to access the full text if any one of them is missing.

Term	Definition
source	The website where the user discovered an item of interest and from which the **OpenURL** was initiated. The **source** is identified by the Referrer ID. In some contexts, **source** is used interchangeably with **referrer**.
success	Describes the state of an **OpenURL** that is able to generate an item-level link to one or more full-text items. The link generated is intended to populate the **link resolver** menu; however, the designation of "success" does not guarantee a user will get to the full-text item if the link is followed all the way to the full-text content provider's site. *See also* **fail**.
Success Score	A value attributed to an **OpenURL** based on its **success** in generating a link to one or more full-text items. If the **OpenURL** generates a link to a full-text item, it is given a Success Score of 1; if not, the Success Score is 0.
target	The website where the full text resides and where the user will be linked to view the item via the **link resolver**. Example targets could include content in publisher platforms, institutional catalogs or repositories, and content gateways.

2 IOTA Use Cases

2.1 Librarian Use Case

OpenURL linking is a core service for libraries around the world. For example, over 439,000 OpenURL requests were sent to Cornell's link resolver in 2012. One estimate is that across the world more than three million requests per day are sent to resolvers (MacColl, 2009). Patrons expect OpenURL links to take them to the item they are requesting: the full text online. "The most common theme related to expectations of full text: 49 percent expressed disappointment at not finding full text online for their citations" (Wakimoto, et al., 2006). A more recent study confirms that perception, finding that:

> "…end users who click on an OpenURL menu do so in order to find the full text of an article. As this and other usability studies have shown, end users focus on the links to the full text available online and overlook links to additional options such as catalog links, help text, and even information, such as the journal's peer-reviewed status that they find important." (Ponsford, et al., 2011)

While it is true that not every citation sent to a library's link resolver is for an item the library holds in its collection, it is imperative that for those the library does hold in its collection the link resolver leads the user to the item without frustration.

2.2 Content Provider Use Case

The primary target users of the IOTA quality reports are OpenURL providers interested in using the reports to monitor and improve the OpenURLs they send out to link resolvers accessed by library patrons so that when they click on the OpenURL, the links successfully go to the providers' platforms.

During the project, both the American Institute of Physics and the Hispanic American Periodicals Index (HAPI) both reached out to the IOTA Working Group and used the resources available in IOTA to make an assessment of their OpenURLs.

2.3 Link Resolver Use Case

OpenURL referring data is fragmentary and often contradictory. Different publishers and referring sources will often portray the same piece of content in radically different ways, some mutually incompatible but equally correct. In cases such as this, what data should be treated as authoritative? Cataloging principles govern bibliographic representation of electronic resources, but not at this level of granularity. In other instances, referring sources are sending metadata that is downright wrong or misleading. Following a traditional user-focused software development approach in which problems are reported as "bugs" and treated as one-time fixes can easily lead to a situation where multiple users can request changes that work contrary to one another.

As an example, the philosophy adopted by the Serials Solutions' 360 Link team to tackle this situation has been referred to as "incremental improvement." Using current resolver technology, there is no way to ensure that 100% of links produced will be successful, since successful link resolution depends on the proper functioning of systems external to the link resolver vendor. But by running large scale quality assurance testing of aggregate referring data sets, it is possible to preemptively spot potential data problems before they are reported by clients. Using a "big data" approach to this kind of testing—with test sets numbering into the thousands of queries, across multiple content provider platforms—the link resolver vendor's quality assurance team can evaluate patterns in linking best practices that match the state of affairs of the linking universe.

This philosophy dovetails excellently with the notion of an analytical approach to link resolution testing. Proactive testing like this allows for a more holistic approach to testing the health of the OpenURL ecosystem. A resource such as the IOTA repository offers to assist an approach like this by offering a readymade source for that aggregated real-world referring data that is so essential to the proper testing of the product. It means larger test sets of queries can be run to find patterns—both for each given referring source or publisher target and across multiple service providers—and to look for ways to streamline the passing of data from one publisher platform to the next. It also means that the business partner outreach team of the link resolver vendor can approach their publisher partners with empirical observations and use their findings to work together towards better product interactions across the board.

3 IOTA Project Approach

The first challenge when IOTA started in 2010 was migrating Adam Chandler's existing proof of concept software (hosted on his personal server) to a NISO-supplied server. Working with NISO's hosting company, the domains openurlquality.org and openurlquality.niso.org were registered and mapped to the NISO virtual server instance. The old website and parser code were then moved to the new server. The system requires PHP for the website, MySQL for the database, and Perl for all the backend file parsing scripts.

During this migration process, Jim Wismer joined IOTA and took over the leadership role for the backend Perl log file parsing code. Adam Chandler focused his attention on improving the functionality for the IOTA user interface. The source code for the Perl scripts is available as an SVN repository; the user interface is available as a Git repository. Currently, both can be downloaded on the IOTA assembla project site: https://www.assembla.com/spaces/oq/wiki.

During 2011, the focus of the development activity on the IOTA system was refining the reports, filtering out noise in the reports, and adding more data. By the end of that year the system contained over 20 million analyzed OpenURLs from 11 different log file providers. Transparency is a key principle guiding the IOTA effort. In that spirit, the list of log file vendors and the data sets provided by them is available on the IOTA website: openurlquality.org/logsources.php.

Work in 2012 focused almost exclusively on developing a statistically valid "Completeness Index." The Completeness Index is a number that is attributed to a content provider (OpenURL referrer or source) to measure the completeness of their OpenURLs in aggregate. It is essentially an average of the completeness scores of OpenURLs coming from that content provider. Completeness Score is the measure of the completeness of a single OpenURL. In this context, completeness refers to the number of metadata elements provided in the OpenURL out of a desired or core number. A detailed chronicle of the effort to create an industry-wide index is provided in section 5.

4 The IOTA Reporting System

4.1 Components

The IOTA system (located at openquality.niso.org) consists of the following components:

1. OpenURL log file parsing software
2. MySQL database for storing element frequency counts by provider
3. Web reports

The IOTA OpenURL parser classifies each analyzed OpenURL as an article, book, or other. As of December 2012, the IOTA repository contained over 23 million analyzed OpenURLs from 11 different link resolver sources, categorized as 41% articles, 2% books, and 57% other as determined by the algorithm in Figure 1.

```
If ( (rft_val_fmt=info:ofi/fmt:kev:mtx:journal) && (genre = article || proceeding
|| conference || preprint || null ) ) Then Type = Article
Else if ( (rft_val_fmt=info:ofi/fmt:kev:mtx:book) && (genre = book || bookitem ||
proceeding || conference || report || document || null) ) Then Type = Book
Else if ( (rft_val_fmt = null) && (genre = book || bookitem) ) Then Type = Book
Else if (rft_val_fmt = null) && (genre = article || proceeding || conference ||
preprint || null) )
Then Type = Article
Else
Then Type = Other
```

Figure 1: Genre algorithm

This algorithm is conservative to be fairly certain that the item is actually an article or book if given that OpenURL type. The reason the “other” category contains such a high percentage of OpenURLs is because a high percentage of OpenURLs today lack a correct assertion about the genre of the item. It simply is the case that many OpenURL sources do not include this element when they send out an OpenURL. There is a report in IOTA (openurlquality.org/report.php?metric=genre&bysourcereporttype=&logsource=%25&reporttype=bycore&year=all&quarter=all&openurltype=all&chooseoriginorsid=origin&min=100) that shows which vendors provide the genre element and which ones do not.

4.2 Architecture

The IOTA reporting system is a three-part architecture as described in 4.2.1 through 4.2.3.

4.2.1 Ingest

The requirement for providing a link resolver log for ingest into the IOTA system is simple. OpenURL requests should be placed in a plain text file, one OpenURL per line, as shown in Figure 2.

```
http://resolver.library.cornell.edu/net/openurl/?genre=article&issn=0145
2096&title=Diplomatic+History&volume=12&issue=3&date=19880601&atitle=THE
+FEDERALIST+PARTY+AND+THE+CONVENTION+OF+1800.&spage=237&pages=237-
260&sid=EBSCO:ahl&aulast=Rohrs
http://resolver.library.cornell.edu/net/openurl/?sid=Entrez:PubMed&id=pm
id:12889865
http://resolver.library.cornell.edu/net/openurl/?sid=google&auinit=DK&au
last=Belyaev&atitle=Domestication+of+animals&title=science+journal&volum
e=5&issue=1&date=1969&spage=47&issn=0582-2092
```

Figure 2: Sample Ingest File

The text file that contains the OpenURLs is then run through the IOTA parser, which splits on the ampersand (just like the link resolver is designed to do), then analyzes the key=value pairs for common patterns, and outputs the aggregated metrics in a tab delimited text file. The IOTA parser handles OpenURLs formatted in either 0.1 or 1.0 (Z39.88) syntax.

4.2.2 Database Structure

The delimited output file is automatically loaded into a MySQL database. Summary counts are recorded in the table called “oq” as defined in Table 1.

Table 1: IOTA oq table properties

Field	Definition
id	Auto incremented database record identifier
logsource	Provider of the log file Examples:, Cornell, EBSCO, Kansas State University A list of the log providers is available at: openurlquality.niso.org/logsources.php
year	Year of the log data source in yyyy format

Field	Definition
quarter	Quarter of the log data source There are four valid values: q1, q2, q3, q4
origin (On the public facing side this is labeled as vendor.)	OpenURLs are supposed to contain a service id, which gives the source of the OpenURL. The service id (or referrer id) contains one or more parts. These are split into two parts in the IOTA data model. In the example, `rfr_id=info:sid/www.isinet.com:WoK:WOS`, we call the first part—`www.isinet.com`—the origin. The origin is the vendor that sent out the OpenURL.
sid (On the public facing side this is labeled database.)	OpenURLs are supposed to contain a service id (sid), which gives the source of the OpenURL. The service id (or referrer id) contains one or more parts. These are split into two parts in the IOTA data model. In the example, `rfr_id=info:sid/www.isinet.com:WoK:WOS`, we call the second part—`WOK:WOS`—the sid. This second part is usually a code for the product. We recognize that calling this second element “sid” is confusing because people often use “sid” interchangeably for service id when talking about OpenURL. It is for historical reasons that we retained this term in the backend data model.
openurltype	The OpenURL type (article, book, other). Based on the presence of particular OpenURL elements, we make a guess about the format of the requested material, as described in 4.1 above.
metric	Metrics may be elements or element syntax patterns. That is, these are text patterns that are commonly found in OpenURLs. For example, the `atitle_colon` metric counts article titles elements that contain a colon.
count	The number of OpenURLs in the quarter that match a particular metric within the time frame (year, quarter).

4.2.3 Reporting

The IOTA web-based reporting system offers an easy to use interface for querying the IOTA database. There are two different reports offered: metric and source. The metric report offers the user the opportunity to compare OpenURL sources across dozens of different criteria. Some of these metrics count the presence or absence of various OpenURL elements while others look for various syntaxes within the element. These metric reports are a very illuminating way of objectively comparing the information that OpenURLs provide to libraries. Table 2 provides an example of what can be learned about vendor OpenURLs by running the metric report.

Table 2: Metric report example:
DOI frequency across select vendors, as of February 2013

Vendor	Percentage with metric
libxdoi	100%
aip.org	95%
ieee.org	85%
achs	74%
tandf	65%
embase	64%
oxfordjournals.org	63%
ingentaconnct.com	62%
cas	55%
sagepub.com	54%

The source report offers the user the opportunity to drip down into a particular OpenURL source to review the presence of elements and text patterns for a particular OpenURL provider. Figure 3 shows part of the source report for Google as the provider.

issn	**issn**		**64**	**80**
issn	issn_number	10102333	1	45
issn	issn_number_number	3333-4444	64	33
issn	issn_other	leftovers after other tests in this category	0	1
issue	**issue**		**74**	**52**
issue	issue_hasletter	summer	0	0
issue	issue_hasupplement	Suppl9	0	0
issue	issue_number	10	71	49
issue	issue_numbers_dash_numbers	11-22	0	0
issue	issue_number_dash_number	5-1	1	1
pmid	**pmid**	**info:pmid/20867593**	**4**	**9**
spage	**spage**		**79**	**55**
spage	spage_number	1067	78	53
spage	spage_number_number	416-424	0	0
spage	spage_other	leftovers after other tests in this category	0	1
spage	spage_string_w_number	e48	1	1

Figure 3: Source report: Google (as OpenURL provider)

5 Completeness Index

5.1 Description

The Completeness Index is a number that is attributed to a content provider (OpenURL referrer or source) to measure the completeness of the provider's OpenURLs in aggregate. It is essentially an average of the completeness scores of OpenURLs coming from that content provider. Completeness Score is the measure of the completeness of a single OpenURL. And by completeness we mean the number of metadata elements provided in the OpenURL out of a desired or core number. Experiments focused on OpenURLs that the parser classified as articles; "book" and "other" OpenURLs were excluded from analysis.

5.2 Method 0: Descriptive Reports (Source, Metrics)

The initial effort to judge the quality of OpenURLs across sources was with the descriptive reports that are available on the IOTA website. These reports, described in section 4.2.3, compare the frequency of data elements and text patterns across OpenURL providers. Figure 5 in Appendix A shows the relative frequency of elements. For example, date and issn appear most frequently in OpenURLs.

5.3 Method 1: Qualitative Judgment

On the surface, issues surrounding full-text linking might not seem that difficult. After all, OpenURL is only a package for transport of metadata from one web platform to another. However, due to the fact that OpenURL linking is dependent on the quality and consistency of the referring data that comes in from a wide variety of different sources, irregularities in indexing practices can often result in multiple situations that are directly contradictory to one another. Still, it is possible to develop a qualitative sense of the requirements for successful OpenURL linking. Article title, for instance, would seem to be generally unreliable, since it can be represented in a number of different ways. Does it include subtitle, for instance? Do all the other articles on the same platform also include subtitle? What about the presence or absence of non-letter characters, such as periods in abbreviations, or commas?

With these issues in mind, we can, through trial and error, make a number of useful observations that will aid in improved link resolution across the board. For instance, numbers would seem to be more reliable than words. Therefore, ideally, we would prefer to refer to a journal title by its ISSN, rather than any specific iteration of its title. To drill down to the desired level of granularity, volume, issue, and start page are much more reliable in most cases than article title and author, or even date, since they are less susceptible to the vagaries of subtitles and punctuation.

It was from this idea—that certain metadata values are more useful than others in assisting in proper link resolution—that the idea of the Completeness Index arose. If it was possible to evaluate the data in a referring URL in some way and to estimate the relative possibility of a successful link being produced by this specific set of referring data, we could "grade" a given content provider on the quality of the referring data it provides and would be able to offer concrete suggestions for improvement, if warranted. We started by choosing a number of "core" elements and giving them some weights based on our own experience.

5.4 Method 2: Correlation Method

The presence of both EBSCO and Serials Solutions representatives on the IOTA working group allowed for a unique opportunity to test link resolution against both of those knowledge bases, thus providing a wider base of data for comparison. Ideally, some common ground could be found that would allow for construction of a unified quantitative Completeness Index, featuring values that would reflect the relative likelihood of a successful article-level link being produced from the metadata values it supplied.

The large body of referring queries uploaded to the IOTA report site served as an ideal source for potential test material. These queries would be selected at random and run through each of the respective link resolvers. After success or failure of resolution was determined (based on whether the particular link resolver was able to build a full-text link to at least one instance of the relevant journal title), the pertinent pieces of metadata were extracted from the referring URL and compiled into a matrix. To begin with, a set of arbitrary values were assigned to represent the perceived importance of each metadata element, based on our "best guesses" from the initial step in qualitative analysis described in section 4.2.3. This allowed each URL to be given a Completeness Score between 0 and 1, based on the percentage of elements present, adjusted by their assigned weights.

With link resolution success or failure determined, the next step was to write a script that would parse the queries to determine which OpenURL metadata elements were present. In the context of our experience the term "success" is used to describe the state of an OpenURL that is able to generate an item-level link to one or more full-text items. The link generated is intended to populate the link resolver menu; however, the designation of "success" does not guarantee a user will get to the full-text item if the link is followed all the way to the full-text content provider's site. Since the queries in question came from a wide variety of referring sources, field name normalization was a significant problem. For instance, for the purposes of this analysis, the ISSN field and the E-ISSN field were treated as identical, since the presence of either serves the purpose of a title level object identifier. Once the sample queries were compiled, and various permutations of equivalent field names were identified and accounted for, the script then processed each referring query, looking for the determined "core" metadata values, and then scored the query based on the set of preliminary numeric rankings.

None of the primary IOTA group representatives is a qualified statistician, but the obvious choice in approaching actual quantitative analysis of OpenURL resolution data seemed to be to test for a correlation between the presence or absence of each particular commonly indexed OpenURL attribute and the successful knowledge base resolution in each of the link resolvers tested. Once the matrix had been compiled and correlated with the link resolution success rate on each of the two platforms, the relative weights given to each metadata value could be tweaked to reflect its maximum correlation coefficient presented by the matrix. In this way, the Completeness Index relative values could be derived empirically from the existing data set and compared for similarities across both of the platforms tested.

After extensive experimentation with various weighting values for each of the core metadata elements in the calculation matrix, the correlation coefficients settled in at between .4 and .5, with slight variation between the result sets achieved from EBSCO and from Serials Solutions. Table 3 illustrates the kind of results found using this method.

Table 3: Method 2 correlation results sample

Test ID	Log URL	Success?	atitle=	aulast=	date=	issn=	issue=	spage=	title=	volume=	Draft score	0.350626
1	http://ry6af4uu9w.search.seri	0	1	1	5	0	0	3	1	0	0.55	
3	http://ry6af4uu9w.search.seri	0	1	1	5	3	3	3	1	3	1	
5	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
6	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
8	http://ry6af4uu9w.search.seri	1	1	0	5	3	3	3	1	3	0.95	
10	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
11	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
12	http://ry6af4uu9w.search.seri	0	1	1	5	3	0	3	1	0	0.7	
13	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
14	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
15	http://ry6af4uu9w.search.seri	1	1	1	5	0	0	3	1	0	0.55	
16	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	
17	http://ry6af4uu9w.search.seri	1	1	1	5	3	3	3	1	3	1	

The most important metadata elements in both sets, however, were consistent. ISSN, volume, and start page were proven by the correlation numbers to be the most important in successful article resolution. Somewhat surprisingly, values that had been thought qualitatively to be important, such as date and article title, proved not to assist article resolution. DOIs and PMIDs also proved to be problematic cases, as resolution tests showed that the mere presence of such an identifier did not, as previously thought, ensure successful linking to the article referenced. In the end, it was decided that DOI and PMID values should be excluded from the Completeness Index, as their successful link resolution was dependent on external factors outside of the control of the link resolver vendor; for example, some publisher business units make journal issue table of contents—with article level DOIs included—available before the actual full text of the article is uploaded to a public server.

5.5 Method 3: Stepwise Regression

Our work in 2012 focused almost exclusively on developing a statistically valid Completeness Index. After consultation with Phil Davis, a statistician and scholarly publishing consultant, we decided that the use of correlation coefficients was not the best statistical model to approach a problem with the multivariate nature of OpenURL linking. Instead, Davis suggested a step-wise regression approach that would allow for the testing of the importance of a single variable at a time, by constructing a test set of "perfect queries" that all produced working article links on the test platform, each of which contained all the metadata values in the commonly used set. With the stepwise approach, a single metadata field was then removed from each query in the set and the queries were tested again, comparing the resolution rate to the original full set. In this way, the value that caused the largest dip in the successful link rate would be considered the most important and, down the list, based on the percentages of successful link resolution.

This method, however, showed differences in the results between the EBSCO and Serials Solutions test sets. For EBSCO's LinkSource, the two most important fields in determining successful link production were volume then start page, while for Serials Solutions 360 Link, start page then volume were the most important elements. The importance of different metadata elements to different link resolvers must be attributed to differing philosophy and code architecture between LinkSource and 360 Link. Since these are highly proprietary details, the commercial link resolver products can never really be tested under truly equivalent circumstances. At the time of our testing (an important point, because link resolver products evolve), the relative order of importance of OpenURL elements for EBSCO's LinkSource product and Serials Solutions' 360 Link product are shown in Figure 4.

```
EBSCO:
volume, spage, issn, issue, atitle, title, date, aulast

Serials Solutions:
spage, volume, issn, title, date, issue, atitle, aulast
```

Figure 4: Ranked importance of OpenURL elements for determining successful link resolution

Through the IOTA research we demonstrated that there is no statistically defensible way of deriving a single, industry-wide weight for each OpenURL element. We know which elements are most critical, but we can't weight them accurately across all the different link resolver systems. We conclude, therefore, that creating a single Completeness Index that is valid across link resolver vendors is not possible, because the relative importance of OpenURL fields varies by link resolver. Upon considering this insight, the IOTA Working Group concluded that the best course of action is to develop a NISO Recommended Practice (published as NISO RP-21-2013) for link resolver vendors to facilitate the creation of their own internal Completeness Indexes.

6 Recommendations and Next Steps

1. **Link Resolver Vendors should make use of the IOTA Recommended Practice for Link Resolver Vendors (NISO RP-21-2013).**

   We believe the method we developed to assist link resolver vendors in comparing the quality of OpenURL metadata across vendors has the potential to improve the quality of service for library patrons worldwide. Even if the improvement is incremental, the scale is so large (millions of requests per month) that the net effect is significant.

2. **Content providers should make every effort to include volume, spage, and issn in article OpenURLs. These elements are critical for success.**

   One of the discoveries we made in our empirical investigation is the critical value of particular OpenURL data elements. We would like to stress to OpenURL providers the importance of including the following key elements in OpenURL requests: volume, spage, and issn plus other standard identifiers such as ISBN, PubMed ID, and DOI. It is worth the investment to add these elements to article OpenURLs to ensure that end users can access the providers' content.

3. **Content providers and link resolver vendors and librarians should make use of the existing IOTA software and data repository to improve OpenURL metadata.**

   We will operate the IOTA system (openquality.niso.org) for six months following the publication of the NISO IOTA Recommended Practice for link resolver vendors. After six months we will review the use of the IOTA system. If it the IOTA site shows growth and community participation we will continue adding data to it.

4. **Vendors and librarians should provide log data to the IOTA repository**.

   The IOTA OpenURL repository contains over 23 million OpenURLs from 11 different log providers. Instructions for how to format log data for the project are available on the IOTA documentation site: https://www.assembla.com/spaces/oq/wiki/Log_file_specification_for_data_providers

5. **NISO should assemble a working group to investigate standardizing the link syntaxes between link resolvers and full-text content providers.**

   Even after more than 10 years of OpenURL use by vendors and libraries, link resolver vendors still struggle to find and maintain the links between link resolvers and content providers. This is a ripe, or perhaps even over-ripe, space for a library standard or best practice.

Appendix A: OpenURL Elements in the IOTA Repository

The IOTA repository contains a number of data elements as listed and defined in Table 4.

Reports available on the IOTA website, as described in section 4.2.3, compare the frequency of data elements and text patterns across OpenURL providers. Figure 5 shows the relative frequency of those elements.

Table 4: OpenURL data elements in the IOTA repository

Element	Definition
artnum*	Article number assigned by the publisher. Article numbers are often generated for publications that do not have usable pagination, in particular electronic journal articles, e.g., “unifi000000090”. A URL may be the only usable identifier for an online article, in which case the URL can be treated as an identifier for the article (e.g., `rft_id=http://www.firstmonday.org/issues/issue6_2/odlyzko/ index.html`).
atitle*	Article title.
aulast*	First author's family name. This may be more than one word. In many citations, the author's family name is recorded first and is followed by a comma, e.g., Smith, Fred James is recorded as `aulast=smith`.
authors	Author names.
btitle	Book title.
date*	Date of publication.
doi	Digital Object Identifier.

Element	Definition
eissn*	ISSN for electronic version of the journal. Although there is no distinction by format in the assignment of ISSNs, some bibliographic services now carry both the ISSN for the paper version and a separate ISSN for the electronic version. This data element is included here to allow the OpenURL to carry both ISSNs and distinguish them.
genre*	**journal**: a serial publication issued in successive parts **issue**: one instance of the serial publication **article**: a document published in a journal **conference**: a record of a conference that includes one or more conference papers and that is published as an issue of a journal or serial publication **proceeding**: a single conference presentation published in a journal or serial publication **preprint**: an individual paper or report published in paper or electronically prior to its publication in a journal or serial **unknown**: use when the genre of the document is unknown
isbn*	International Standard Book Number (ISBN). The original ISBN was a 10-digit number (with the last digit as a check digit, which may be “X”). As of January 1, 2007, all ISBNs are 13 digits, with the final value also a check digit. The number may be presented with or without hyphens, e.g., “057117678X” or “978-1-878067-73-9”.
issn*	International Standard Serial Number (ISSN). The 8-digit ISSN (with the last digit as a check digit, which may be “X”) should be presented with a hyphen, e.g., “1041-5653”.
issue*	Designation of the published issue of a journal, corresponding to the actual physical piece in most cases. While usually numeric, it could be non-numeric. Note that some publications use chronology in the place of enumeration, e.g., Spring 1998.

Element	Definition
jtitle*	Journal title. Use the most complete title available. Abbreviated titles, when known, are records in stitle. This can also be expressed as "title", for compatibility with version 0.1. Example: "journal of the american medical association"
pmid	PubMed ID. A unique identifier assigned to a record in the PubMed databases (www.pubmed.org).
spage*	First page number of a start/end (spage-epage) pair. Note that pages are not always numeric.
title	Title of journal or book.
volume *	Volume designation. Volume is usually expressed as a number but could be roman numerals or non-numeric, e.g., "124" or "VI".
*Definitions from OpenURL Registry: http://alcme.oclc.org/openurl/servlet/OAIHandler/extension?verb=GetMetadata&metadataPrefix=mtx&identifier=info:ofi/fmt:kev:mtx:journal	

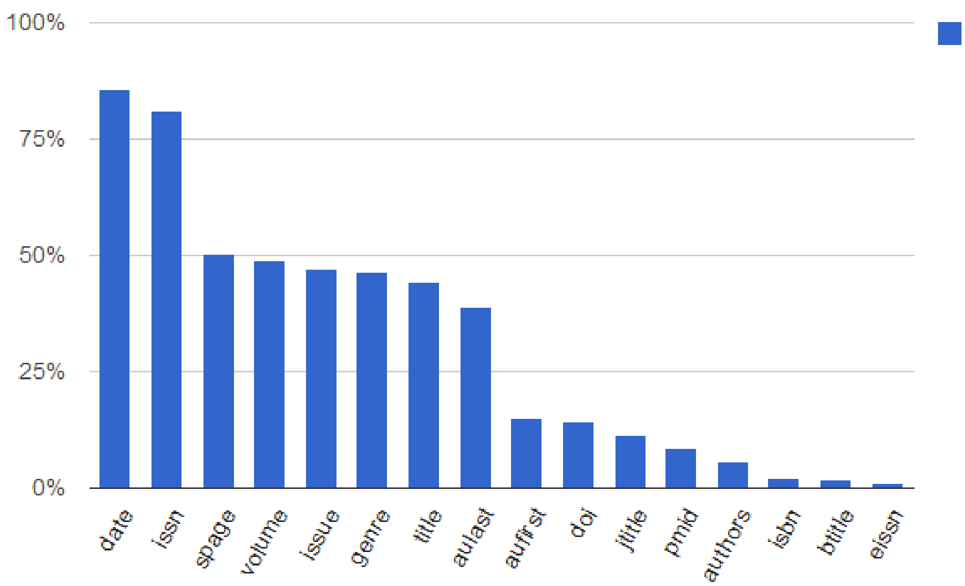

Figure 5: OpenURL elements, frequency in IOTA repository as of February 2013

Table 5: OpenURL data elements pattern metrics in IOTA repository as of February 2013

IOTA Element Pattern Metric	Example	Percent frequency in IOTA repository,
atitle_colon	a+bitter+truth:+avant-garde+art+and+the+great+war	2.76%
atitle_other		11.96%
atitle_simple	biosorption+of+precious+metals	38.83%
aulast_comma	Smith, Sue	0.08%
aulast_hyphenname	renita-schmidt	1.17%
aulast_other		6.22%
aulast_simple	Smith	31.66%
authors_comma	badcock,+johanna+c.%3bhall,+james+r	0.00%
authors_hyphenname	steven-walczak	0.01%
authors_other		5.65%
authors_simple	palmisano	0.06%
date_dateother		0.55%
date_dddd	2006	62.13%
date_dddd-dd	2007-12	2.81%
date_dddd-dd-dd	2000-08-01	8.07%
date_dddd-dddd	1986-1987	0.00%
date_dddddddd	20020801	12.25%
eissn_number	1539073x	0.48%

IOTA Element Pattern Metric	Example	Percent frequency in IOTA repository,
eissn_number_number	1365-2648	0.59%
eissn_other		0.01%
isbn_10	1557869669	0.53%
isbn_13	9780599630864	1.22%
isbn_other		0.36%
issn_number	10102333	49.27%
issn_number_number	3333-4444	30.63%
issn_other		1.14%
issue_hasletter	summer	0.25%
issue_hasupplement	Suppl9	0.19%
issue_number	10	44.34%
issue_numbers_dash_numbers	11-22	0.20%
issue_number_dash_number	5-1	1.18%
issue_other		0.81%
spage_number	1067	48.25%
spage_number_number	416-424	0.26%
spage_other		48.25%
spage_string_w_number	e48	0.94%
volume_number	16	48.45%
volume_number_number	15-16	0.06%

IOTA Element Pattern Metric	Example	Percent frequency in IOTA repository,
volume_number_number_paren	11(2-3)	0.00%
volume_other		0.23%
volume_roman	X	0.03%
volume_roman_roman	X-I	0.00%
volume_withletter	xxxii	0.09%

Bibliography

Blake, Miriam E., and Frances L. Knudson. 2002. “Metadata and Reference Linking.” *Library Collections, Acquisitions, and Technical Services,* 26 (3): 219–230. http://dx.doi.org/10.1016/S1464-9055(02)00253-1

Breeding, Marshall. 2012. *Knowledge Base and Link Resolver Study: General Findings.* Stockholm: National Library of Sweden. http://www.kb.se/dokument/Knowledgebase_linkresolver_study.pdf

Chandler, Adam. 2009a. *OpenURL Quality Metrics: A NISO New Work Item Proposal.* Submitted to the NISO Business Information Topic Committee. http://www.niso.org/apps/group_public/download.php/3021/NISOproposedworkitem_OpenURLmetrics3nov09.doc

———. 2009b. *Results of L’Année Philologique Online OpenURL Quality Investigation, Mellon Planning Grant Final Report, February 2009.* Ithaca, NY: Cornell University. http://ecommons.library.cornell.edu/handle/1813/23171

———. 2011. “NISO IOTA: Improving OpenURLs Through Analytics, in Context.” *Against the Grain,* 23 (1): 30–33.

Chandler, Adam, Glen Wiley, and Jim LeBlanc. 2011. “Towards Transparent and Scalable OpenURL Quality Metrics.” *D-Lib Magazine,* 17 (3/4). doi: 10.1045/march2011-chandler. http://dlib.org/dlib/march11/chandler/03chandler.html

Culling, James. 2007. *Link Resolvers and the Serials Supply Chain: Final Project Report for UKSG.* Oxford, Scholarly Information Strategies. http://www.uksg.org/sites/uksg.org/files/uksg_link_resolvers_final_report.pdf

Davies, Karen. 2012. “Reference Accuracy in Library and Information Science Journals.” *Aslib Proceedings,* 64 (4):373–387. http://dx.doi.org/10.1108/00012531211244734

Glasser, Sarah. 2012. “Broken Links and Failed Access.” *Library Resources & Technical Services,* 56 (1): 14–23. http://dx.doi.org/10.5860/lrts

Kasprowski, Rafal. 2011a. *NISO’s IOTA Working Group: Improving OpenURLs Through Analytics.* Presented at the UKSG Conference, April, Harrogate, United Kingdom, April 4-6, 2011. Presentation slides available at: http://www.slideshare.net/rkaspro/iota-uksg-2011

———. 2011b. NISO’s IOTA Initiative: Measuring the Quality of OpenURL Links. Presented at the NASIG Annual Conference, June 2-5, 2011, St. Louis, MO. Presentation slides available at: http://www.slideshare.net/rkaspro/iota-nasig-2011-measuring-the-quality-of-openurl-links

———. 2012a. “NISO’s IOTA Initiative: Measuring the Quality of OpenURL Links.” *The Serials Librarian,* 62 (1-4): 95–102. http://dx.doi.org/10.1080/0361526X.2012.652480

———. 2012b. *Evaluating the Quality of OpenURLs Through Analytics*. Presented at the Texas Library Association Annual Conference, April 19, 2012, Houston, TX. Presentation slides available at: http://www.slideshare.net/rkaspro/kasprowski-tla12-finalpres

Kasprowski, Rafal, and Susan Marcin. 2010. "NISO's IOTA Working Group: Creating an Index for Measuring the Quality of OpenURL." In: Beth R. Bernhardt and Leah H. Hinds, eds. *Charleston Conference Proceedings 2010: Anything Goes*, pp. 363-370. West Lafayette, IN: Purdue University Press, 2012. http://dx.doi.org/10.5703/1288284314855. Presentation slides available at: http://www.slideshare.net/rkaspro/niso-iota-openurl-quality-index-charleston-2010

Kasprowski, Rafal, Susan Marcin, and Oliver Pesch. 2012a. *NISO's IOTA Initiative: Improving OpenURLs Through Analytics*. Presented at the ALA Annual Meeting, June 24, 2012, Anaheim, CA. Presentation slides available at: http://www.slideshare.net/rkaspro/nisos-iota-openurl-quality-initiative-ala-sla-2012-13745941

Kasprowski, Rafal, Susan Marcin, and Oliver Pesch. 2012b. *NISO's IOTA Initiative: Fixing OpenURL Links Using Data Analysis*. Presented at the SLA Annual Conference, July 17, 2012, Chicago, CA. Presentation slides available at: http://www.slideshare.net/rkaspro/nisos-iota-openurl-quality-initiative-ala-sla-2012-13745941

KBART (Knowledge Base And Related Tools) [website]. A joint NISO/UKSG project. Available at: http://www.niso.org/workrooms/kbart

MacColl, John. 2009. "Herbert's Adventures In Linking." *HangingTogether [blog]*. http://hangingtogether.org/?p=616

Marcin, Susan. 2011. *NISO's IOTA: Improving OpenURLs Through Analytics*. Presented at the ALCTS-CRS Continuing Resources Standards Update Forum at the American Library Association Midwinter Conference & Exhibition, January 9, 2011, San Diego, CA.

———. 2012. *E-Books & OpenURL Linking, A Collaborative Study by the 2CUL E-Books Task Force*. Presented at the New York Technical Services Librarians Spring Program, May 10, 2012, New York. Presentation slides available at: http://www.slideshare.net/smarcin/ebooks-openurl-linking-a-collaborative-study-by-the-2cul-ebooks-task-force

NISO. 2004. *The OpenURL Framework for Context-Sensitive Services*. ANSI/NISO Z39.88-2004 (R2010). Bethesda, MD: National Information Standards Organization (NISO). http://www.niso.org/standards/z39-88-2004/

———. 2013. *Improving OpenURLs Through Analytics (IOTA): Recommendations for Link Resolver Providers*. NISO RP-21-2013. Baltimore, MD: National Information Standards Organization (NISO). http://www.niso.org/publications/rp/rp-21-2013

Ponsford, Bennett Claire, Jane Stephens, and Robin R. Sewell. 2011. "Improving OpenURL Menus: User Testing of Revisions to SFX® Menus." *Serials Review*, 37 (3): 162–170. http://dx.doi.org/10.1016/j.serrev.2011.03.016

Trainor, Cindi. 2010. *Rethinking Linking: Breathing New Life into OpenURL*. Library Faculty and Staff Papers and Presentations. Paper 1. http://encompass.eku.edu/faculty_staff/1

Trainor, Cindi, and Jason Price. 2010a. "Digging into the Data: Exposing the Causes of Resolver Failure." In: Rethinking Library Linking: Breathing New Life into OpenURL, Chapter 3. *Library Technology Reports* 46 (7): 15–26. http://alatechsource.metapress.com/content/m466322w0472g385/?p=576ed6dbbcb740a2a03ca0385d9ca2c3&pi=2

———. 2010b. "Rethinking Library Linking: Breathing New Life into OpenURL." *Library Technology Reports*, 46 (7). http://alatechsource.metapress.com/content/k38116t41000/?p=2c29f19ca8624a16a31a96d53007f680&pi=18

Van de Sompel, Herbert, and Oren Beit-Arie. 2001. "Open Linking in the Scholarly Information Environment Using the OpenURL Framework." *D-Lib Magazine*, 7 (3). doi:10.1045/march2001-vandesompel. http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html

Wakimoto, Jina Choi, David S. Walker, and Katherine S. Dabbour. 2006. "The Myths and Realities of SFX in Academic Libraries." *The Journal of Academic Librarianship*, 32 (2): 127–136. http://dx.doi.org/10.1016/j.acalib.2005.12.008

Winter, Elizabeth. 2010. *IOTA: Improving OpenURLs Through Analytics*. Presented at the NISO Webinar: It's Only as Good as the Metadata: Improving OpenURL and Knowledgebase Quality. Presentation slides available at: http://www.niso.org/apps/group_public/download.php/5146/MetadataHandouts.pdf