

Digitometric Services for Open Archives Environments

Tim Brody, Simon Kampa, Stevan Harnad, Les Carr, Steve Hitchcock

Intelligence, Agents, Multimedia Group
University of Southampton, UK
{tdb01r, srk, harnad, lac, sh94r}@ecs.soton.ac.uk

Abstract. We describe “digitometric” services and tools that add value to open-access eprint archives using the Open Archives Initiative (OAI) Protocol for Metadata Harvesting. Celestial is an OAI cache and gateway tool. Citebase Search enhances OAI-harvested metadata with linked references harvested from the full-text to provide a web service for citation navigation and research impact analysis. Digitometrics builds on data harvested using OAI to provide advanced visualisation and hypertext navigation for the research community. Together these services provide a modular, distributed architecture for building a “semantic web” for the research literature.

Introduction

In this paper we describe digitometric tools that apply and extend the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as a means of building user services for the scientific and scholarly literature.

The services described in this paper touch on a number of digital library topics: infrastructure, accessing legacy data, harvesting, online archiving, online publication, open access, scientometrics, and linking. This covers both the use of existing data through discovery and conversion, and building new data through processing and analysis.

As authors increasingly use eprint archives (built using free tools such as Southampton’s eprints.org [6]) to maximise their research impact - by maximising user access to and usage of their research through open-access - this resource is becoming an important tool for researchers. With open-access and an OAI-PMH interface any service can harvest metadata from an eprint archive and provide added-value, from simple cross-archive searching, through to advanced user interfaces and analytic tools.

The first part of this paper provides background about the OAI-PMH. We introduce the Celestial tool [4] (a cache/gateway for the OAI-PMH) and Citebase [5] (an end-user service that applies citation-analysis to existing OAI-PMH compliant eprint archives). We then analyse Citebase’s database, and summarise the findings of a user survey conducted by the Open Citation Project [7]. Finally we introduce some of the new directions arising out of this work - creating a knowledge environment built on the OAI-PMH.

Open Archives Initiative

The Open Archives Initiative [13] Protocol for Metadata Harvesting (OAI-PMH) is designed to address the need to expose metadata - titles, authors, abstracts etc. - from research literature archives in a structured form. An XML protocol built on the HTTP standard, OAI-PMH is in effect a CGI interface to databases. Based on 6 commands (or “verbs” in OAI parlance) OAI-PMH allows metadata to be incrementally harvested by *service providers* (the HTTP client) from *data providers* (the HTTP server).

There are 62 OAI-registered publicly accessible data providers (plus another 98 unregistered ones), exposing around two million records covering research literature (e.g. arXiv.org), music manuscripts (e.g. Library of Congress), theses, and others. Some service providers have been developed or adapted to make use of OAI-PMH, that allow users to search both commercial abstract databases and the freely available abstracts from public data providers (e.g. Scirus [25]). In the USA OAI-PMH is being used to build a large-scale distributed library system, NSDL [14].

The OAI-PMH allows the transfer of *metadata records* encoded in XML. To be OAI-compliant a data provider must expose their records in Dublin Core, but they can expose their data in any format that can be

encoded in XML. The metadata records that describe a single entity form an *item*, identified by a unique *identifier*.

The OAI-PMH is being used to transfer sizeable amounts of data - in the case of <http://arXiv.org/> some 230,000 metadata records (Figure 1 shows the increase in records for all OAI archives cached by ‘Celestial’ – see next section). As the number of OAI-PMH sites increases, and the size of the data provider databases grows (Figure 2), there is a growing need to build scalable infrastructures to support the transfer of data from data providers to service providers (a many to many relation). Caching is a useful method to distribute the load within such distributed systems using tools like Celestial.

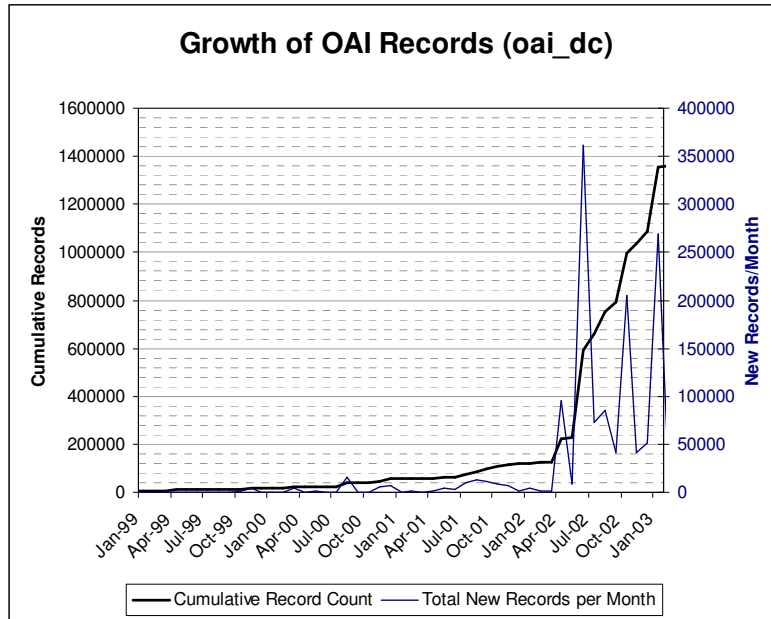


Fig. 1. Celestial attempts to harvest records from 161 OAI archives. Each OAI record harvested contains a datestamp (when the OAI record was created or last updated). A histogram of these datestamps plots the growth of OAI records over time. The blue line is the number of new OAI records per month (according to the datestamp) and the black line the cumulative number of records. The peaks in new records shows when large, new archives come online and expose a large back-catalog of pre-existing records. (“Record” usually – but not always – means a full-text paper.)

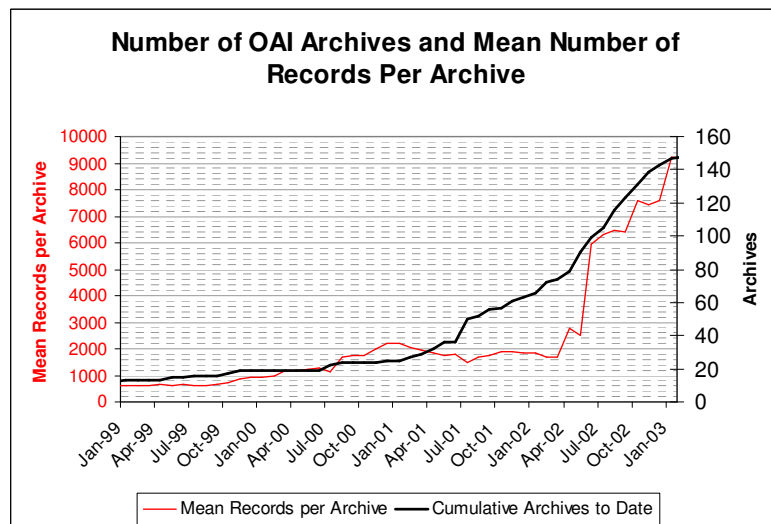


Fig. 2. A rough estimate of when OAI archives have come online can be calculated by taking the earliest timestamp from that archive. The black line shows the cumulative number of available OAI archives, and the red line shows the mean number of records in those archives. Note that both the number of archives and the number of records in those archives is increasing.

Celestial

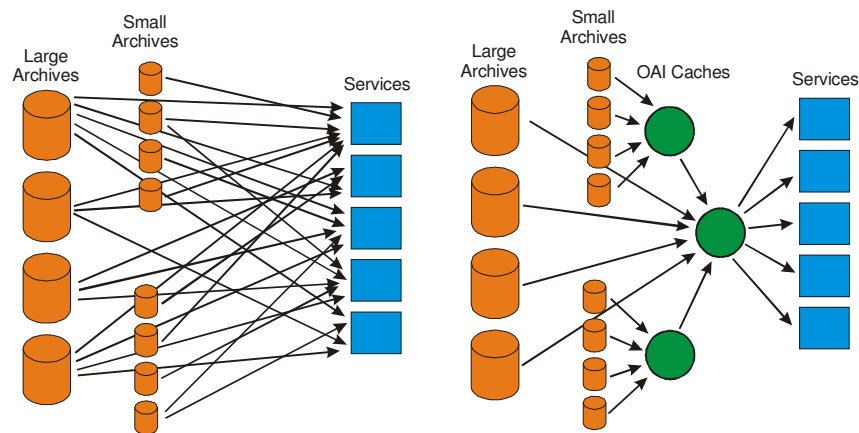


Fig. 3. Non-Aggregated vs. Aggregated OAI harvesting.

Celestial is software that supports the caching of metadata from OAI archives, gateways between legacy (1.0 and 1.1) and current (2.0) OAI implementations, and attempts to correct incorrectly implemented OAI archives.

In a distributed environment caching moves processing and network load away from the source and closer to the target (Figure 3). As OAI archives are often small and low-performance, reducing the load on them can be important – especially where the OAI-PMH interface may be seen to interfere with other services. To support the caching of OAI responses Celestial acts as an *OAI cache/proxy*. Working at the application-level it harvests records from data providers using the OAI-PMH, and re-exposes them to service-providers through its own OAI-PMH interface. Celestial is able to make a complete copy of an OAI archive, including all the metadata records, and set memberships associated with an item. Should the data provider become unavailable, Celestial is able to act as a surrogate.

By using the incremental, timestamp-based harvesting ability of OAI-PMH, Celestial only harvests those records that are new or have changed from a data provider. By comparison an HTTP cache would have to query all records to determine whether they had altered from a prior harvest.

Celestial is designed to provide as high performance as possible. It achieves this by trading storage space for performance. A significant overhead with any XML-based application is generating the XML tag structures. To avoid this Celestial stores the OAI header and metadata as XML. When generating a response Celestial prints the raw data, and only needs to generate XML tags for the OAI protocol components (e.g. the request header, and flow-control tokens).

OAI-PMH flow-control is handled using stateless cursors. Celestial assigns each record a timestamp and unique identifier. These two values are joined to form an index into the record list. As a harvester retrieves records Celestial moves a cursor along this index, and at the end of a partial list Celestial provides the harvester with the current cursor (the timestamp plus unique identifier), and an encoding of the original request (which might include a set or timestamp filter) in the OAI-PMH resumption token. Given a resumption token Celestial can jump straight to the end of the previous partial list by using the index key.

If new records are added to Celestial during a harvest they will be returned at the end of the harvest, as the new record's timestamp will be greater than any previous records. This makes the resumption tokens generated by Celestial stateless, as no changes can occur that would make the result set inconsistent.

OAI archives that have not upgraded to 2.0 have been removed from the official OAI-compliant list (and hence unlikely to be included in new OAI services). As Celestial provides an OAI 2.0 (the current version)

interface to harvesters, but can itself harvest from version 1.0, 1.1 or 2.0, it acts as an *OAI gateway* between non-upgraded data providers and upgraded service providers. In OAI 2.0 each record has the set membership of that record. To provide the set hierarchy to OAI 2.0 harvesters Celestial inverts the set membership exported by an OAI 1.x archive. For OAI 1.x this set membership is found by exhaustively querying each set, building up the set membership for each item.

Often data providers will export records from sources that are not Unicode-based. If a data provider does not convert and check these records before exporting them, bad characters can appear in the data provider's OAI-PMH export, preventing XML parsing. Celestial makes a best-effort to correct these errors by replacing the location of bad characters (as reported by the XML parser) with a valid character, "?". The process of XML parsing, correcting characters, and re-parsing can be repeated until either the OAI-PMH response can be parsed or the act of replacing encroaches on the XML tags and makes the response unrecoverable.

As well as attempting to fix OAI-PMH responses in real-time, Celestial records errors that occur during harvesting. An archive administrator can use these harvest logs to correct mistakes in their implementation, or underlying data records. As the OAI-compliance tests do not make a full harvest of archives, this can often highlight problems (e.g. with flow-control) that the OAI registration process does not.

Celestial implements the OAI provenance schema. This records the path that records have taken through OAI proxies, caches and aggregators, by storing with the metadata record the location from which the record was harvested, when it was harvested, and whether any alterations have been made. Provenance data can be used by service providers to "de-dup" the same record, if the service harvests from multiple sources.

A promising possibility for Celestial is as a tool for exposing any data source via an OAI-PMH interface. Out of the box, Celestial only supports getting data via OAI. It is relatively easy, however, to create a system that would insert records directly into Celestial's back-end database, which can then be served through the OAI-PMH interface.

While Celestial is a distinct, freely-downloadable software package, at Southampton University [3] a mirror of Celestial hosts a copy of the metadata from 161 different OAI archives (OAI-registered archives (including the OAI-registered eprints.org archives), plus any unregistered eprints.org installations found, and active archives registered with the Repository Explorer [9]).

The Celestial mirror is used within Southampton by Citebase Search. As a developing service Citebase often needs to completely re-harvest its metadata, and using a local mirror avoids repeatedly making very large requests to source archives.

Citebase Search

Citebase, more fully described by Hitchcock et al. [1], allows users to find research papers stored in open access, OAI-compliant archives - currently arXiv (<http://arxiv.org/>), CogPrints (<http://cogprints.soton.ac.uk/>) and BioMed Central (<http://www.biomedcentral.com/>). Citebase harvests OAI metadata records for papers in these archives, as well as extracting the references from each paper. The association between document records and references is the basis for a classical citation database. Citebase is best viewed as a kind of "Google for the refereed literature", because it ranks search results based on the number of references to papers or authors (although it is not - currently - using a hub-authority graph algorithm to rank). Citebase contains 230,000 full-text eprint records, and 6 million references (of which 1 million are linked to the full-text).

Citebase was developed as part of the JISC/NSF Open Citation Project, which ended December 2002. As part of the project report a user survey [23] was conducted on Citebase. This was used both to evaluate the outcomes of the project, and to help guide the future direction of Citebase as an ongoing service. The report found that "Citebase can be used simply and reliably for resource discovery. It was shown tasks can be accomplished efficiently with Citebase regardless of the background of the user."

Primarily a user-service, Citebase provides a Web site that allows users to perform a meta-search (title, author etc.), navigate the literature using linked citations and citation analysis, and to retrieve linked full-texts in Adobe PDF format. Citebase also provides a machine interface to the citation data it collects through its own OAI-PMH interface using the Academic Metadata Format (AMF) [10], a new XML format for scholarly literature. As part of the development of Citebase we have looked at the relationship between

citation impact (“how many times has this article been cited”) and web impact (“how many times has this article been read”).

Citation-navigation provides Web-links over the existing author-generated references. First, wherever possible, Citebase links each reference cited by a given article to the full-text of the article that it cites (if it is in the database). This fan-in (“citations-from”) and fan-out (“citations-to”) then provides the user with links to all articles (in the database) that have cited a given article, as well as to all articles that have been co-cited alongside (hence are related to) the given article. This allows the user to navigate back in time (articles referred-to), forward in time (cited-by), and sideways (co-cited alongside).

Citebase provides information about both the citation impact and usage impact of research articles (and authors), generated from the open-access pre-print and post-print literature that Citebase covers. The citation impact of an article is the number of citations to that article. The usage impact is an estimate of the number of downloads of that article (so far available for one arXiv.org mirror only).

Citebase’s Web Interface

The front-end of Citebase is a meta-search engine. This allows the user to search for articles by author, keywords in the title or abstract, publication (e.g. journal), and date of publication. After generating a search, Citebase allows the results to be ranked by 6 criteria: citations (to the article or authors), Web hits (to the article or authors), date of creation, and last update. The by-author ranking is calculated as the mean number of citations or hits to an author (e.g. total citations divided by total papers to author “Hawking, S”). A per-article author-impact is then calculated by taking the mean author-impact of all the named authors. Citebase currently uses only the family name and the first initial to identify authors; as these services develop it is hoped that algorithms (to be developed in collaboration with the Institute for Scientific Information, ISI) for recognizing and distinguishing authors with the same or similar names will improve this metric.

From the meta-search users can either choose to view an abstract page, or jump directly to a cached full-text PDF (if available) for each matching article.

The abstract page displays a full meta-record (title, authors, abstract, rights etc.), the articles cited by the current article, articles that have cited the current article, and articles co-cited alongside the current article. In addition to listing the citing articles, Citebase provides a summary graph that shows over time when the citing articles have appeared, and when the current article has been downloaded (e.g. see Figure 4). This provides a visual link between the citation and web impacts.

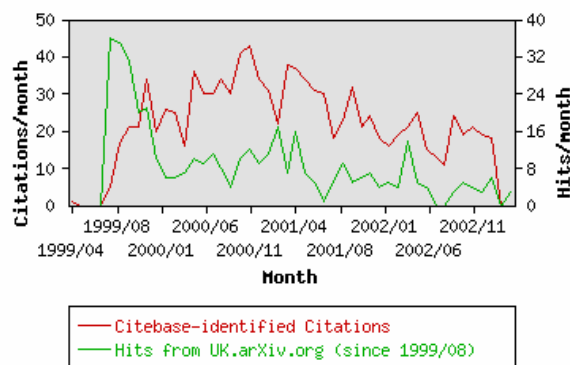


Fig. 4. The Citebase abstract page displays a histogram of references and web hits to the current article over time. The references are counted as occurring on the day that the citing article was deposited in the archive. This particular article shows the burst of downloads that articles receive soon after they are deposited, which then drops to either nothing (if the paper achieves little impact), or continues at a lower level. The citation impact of an article peaks after a delay of between 3 months to a year (depending on the speed of the publication cycle) before slowly dropping as the paper gets older and less relevant.

When viewing a cached full-text PDF Citebase overlays reference links within the document, so a user can jump from viewing a full-text to the abstract page of a cited article.

Like the archives it harvests from, Citebase provides an OAI-PMH interface to the data that it contains. Along with re-exposing the Dublin Core metadata (title, author, abstract), Citebase provides records in the Academic Metadata Format (AMF) [10]. AMF encapsulates the relationships within the scholarly research: between authors, articles, organisations, and publications. Other services can harvest this enhanced metadata from Citebase to provide a reference-linked environment, or perform further analysis (or they can be harvested by the source archives to enhance their own data).

Analysing Citebase's Database

Since the creation of Garfield's science citation index [21,22] researchers have analysed the "citation web" to look for patterns in the growth and direction of scholarly research – most often to determine how the ISI Journal Impact Factor is related to a particular subset of the research literature. With the research literature moving online a new metric for impact can be measured: web hits – an indicator of how much an article has been accessed.

Web hit data can be subject to inaccuracies and noise. For example, if an arXiv paper is referenced by Slashdot [12] it will receive many hits from casual users (the "Slashdot effect"), which probably do not reflect true impact on other researchers. Citebase filters Web logs by removing known Web crawlers (e.g. Googlebot), then only counting one hit from one location per day. This is probably an over-correction, for although it removes most "unwanted" hits, it also excludes valid hits from users who may be sharing a single machine, or Web proxy.

Given the success of the arXiv.org online archive, it is not surprising that citation-impact and usage-impact (measured from the UK arXiv.org mirror) are related. The effect is probably bi-directional: the more something is cited, the more it is read as researchers follow citations, and the more something is read, the more likely it is to be cited as researchers cite what they have read. (We are teasing apart these two effects, as they probably have different time-constants, citation-to-reading being a short-latency effect, occurring soon after the time the citing article appears, and reading-to-citation being a longer-latency effect, requiring the reader to first write a paper of his own.) As researchers use the Web more, and the amount of literature available online increases, these trends are likely to grow (Table 1). It is the few, very high-impact articles, for which the correlation between citation and web impact is highest (Table 1). In order to analyse the correlation between citation and web impact a Web tool "Correlation Generator" [24] was created that allowed various parameters to be altered, and a correlation to be generated on-the-fly.

All	$r=.27, n=219328$
Q1 (lo)	$r=.26, n=54832$
Q2	$r=.18, n=54832$
Q3	$r=.28, n=54832$
Q4 (hi)	$r=.34, n=54832$
hep	$r=.33, n=74020$
Q1 (lo)	$r=.23, n=18505$
Q2	$r=.23, n=18505$
Q3	$r=.30, n=18505$
Q4 (hi)	$r=.50, n=18505$

Table 1. High Energy Physics (HEP) is the longest established and most comprehensive of the subject areas within arXiv. It is the best example of what a completely open-access corpus would look like. While the other subject areas covered by arXiv are incomplete they provide a comparison for partial-coverage. The articles contained in the entire archive (“All”) and HEP were separated into quartiles (of size n) according to their citation impact (Q4 is highest impact quartile, Q1 lowest). The correlation r was then found between the citation impact and usage impact of the articles contained within each quartile. HEP shows the highest correlation, with .5 for the highest citation impact quartile.

The correlation between citation and usage impact is highest at the high end of the citation impact spectrum. Low impact articles disappear into obscurity, little read and little cited, while a few high impact articles continue to be read and cited for longer periods of time. But regardless of whether a article is destined for obscurity or fame, it still gets a burst of Web impact soon after it is first released on the web (probably coming from users who are following that topic either through the arXiv’s automatic daily/weekly new-paper alerting service or through regular active browsing of new contents).

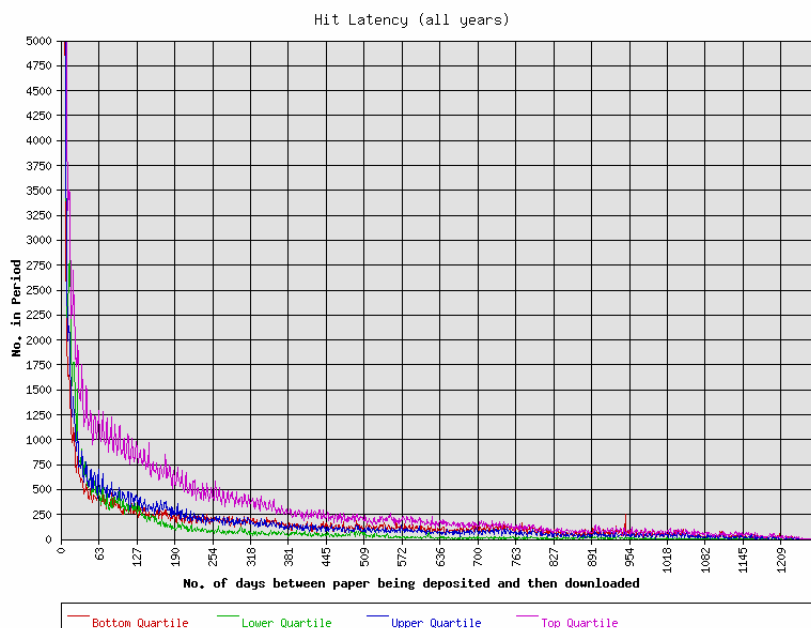


Fig. 5. The bi-directional influence of citation and usage impact is a cycle: an article is written, archived, “advertised” by an alerting service, read, and perhaps eventually (after several months) cited by a new article, a new “advertisement” that loops back to make it read more again, etc. An important article, destined for higher citation impact, will already have been read more; it will also continue to be read more over time than a low citation impact article (as users loop back to it from citations appearing in newer articles). Dividing the arXiv.org articles into citation-impact quartiles, a histogram of hit data for each quartile can be plotted against the time delay between the article being deposited and later downloaded (the “hit latency”). While all 4 quartiles show a large number of downloads within the first few days of being available, higher impact (the upper and top quartiles) show double the hit rate over a longer period of time.

Digitometric Services for OAI

Scholarly research consists of systematic investigation, gathering the empirical, theoretical, and scholarly information relevant to a particular research field. These disparate activities include detective work - becoming proficient in a field and understanding its present and past literature’s landscape. This is a continuous task that requires a scholar’s constant and full attention.

The Digitometric Framework provides advanced services over research metadata collected from various OAI-compliant archives. The Framework exposes an open and extensible interface where services are attached to enable researchers to better investigate and understand their research field. Basic visualisations

(e.g. bar charts of papers/author) to advanced visualisations (e.g. co-citation maps), knowledge services (e.g. identifying the most prominent researchers), and hypertext linking of different research artifacts have been implemented and added as services.

The experimental Digitometric software consists of a back-end database to store metadata harvested from OAI archives, software to interact with the database and provide the base functionality of the framework, and several interchangeable and extensible modules representing each of the analysis and visualization services (Figure 6). Harvested metadata are analysed, enhanced (e.g. adding data for visualisations), and converted into a native format for use in the framework. The new metadata are also exposed for other services to use and build on. The framework therefore represents an open and extensible solution to managing and publishing scholarly metadata.

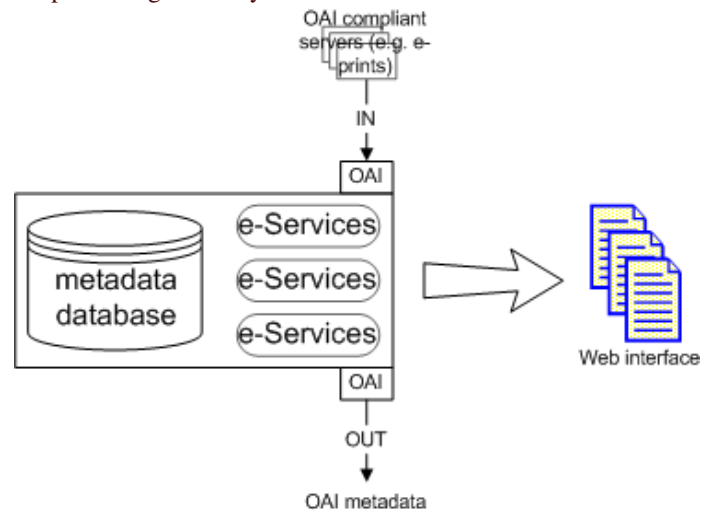


Fig. 6. Architecture of the Digitometric Framework. “e-Services” are, for example, co-citation visualization tools.

Initially, the database was populated with the entire collection of papers in the arXiv.org archive. This provided a large database to create and explore detailed visualisations of the research landscape.

Visualising Research

The Digitometric services can be used to dynamically visualise different aspects of research metadata. For example, users can retrieve basic graphs illustrating the highest publishing authors or the citation network of a particular publication (Figure 7).

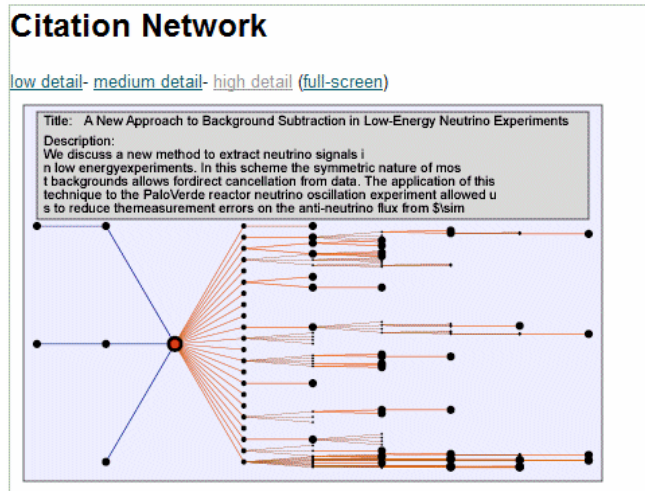


Fig. 7. A dynamically generated citation network. The current article is shown in the middle (a single red dot), while articles that have cited the current article are linked by blue lines (and articles that have cited those), and articles that the current article has cited are linked by orange lines (and articles cited by those article).

More interestingly, co-citation maps can be dynamically generated and displayed. Two papers are co-cited when a third paper cites them both. Co-citation analysis therefore relates bibliographic data based on co-citation strengths (i.e. the number of times two papers are cited together). The premise is that similar papers, for example papers that discuss or raise the same issue or methodology, are frequently co-cited. When co-citation values are used as proximity measures in visualisations where papers that are frequently co-cited are plotted near each other, the resulting graph enables "research fronts" (speciality fields) to be identified. Research fronts will usually emerge around a few seminal (or core) papers that are heavily co-cited [20]. These graphs can be evaluated, for example, to determine how a particular idea has influenced a field. In this case, several clusters may arise around the corresponding publication, each containing papers that share a common interpretation or understanding of the idea. Furthermore, researchers could analyse the clusters over different time periods to understand how research "hot-spots" have changed and been absorbed into the corpus.

Co-citation analysis has been criticised [16,19] for over-simplifying of the citation link, for technical problems (e.g. inaccurate citations), and for focusing only on citations when other factors (e.g. social/political motivation behind the citation) should be taken into consideration. However, Garfield [17] nevertheless notes that co-citation provides a useful and predictive perspective on scholarly material when used cautiously and wisely. He shows convincingly how he has uncovered important historical links between research fronts using co-citation analysis that scholars had previously overlooked [18].

Users of the Digitometric services can retrieve a co-citation map at any time and even use a particular publication as the launch point for navigating the co-citation space. Different variables are set to define the accuracy, size and co-citation threshold to be used, each having a significant bearing on the time required to construct the maps.

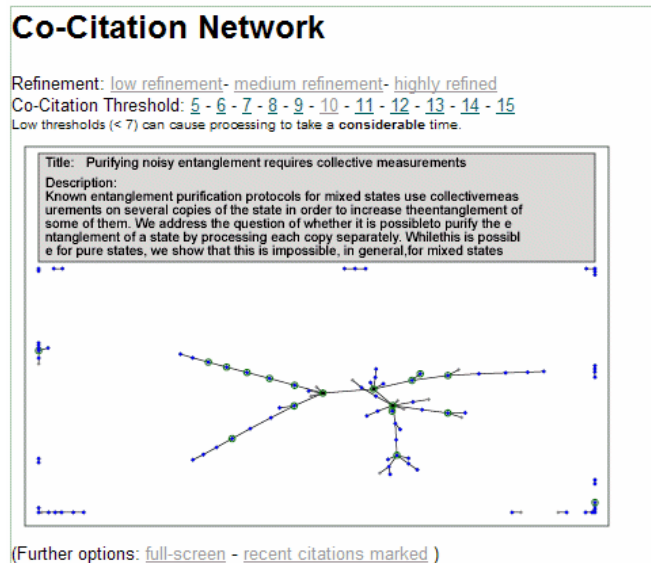


Fig. 8. A co-citation map embedded within the Digitometric user interface. The nodes on the map represent individual publications. By hovering with the mouse pointer over a node, the user can generate details (title, author, abstract) in the information box. The arcs between the nodes represent a co-citation relationship. A cluster of related publications (perhaps around a project, or theorem) are evident in the centre of the map. Four distinct paths emanate out of this indicating the possibility of specialty fields arising out of the main cluster.

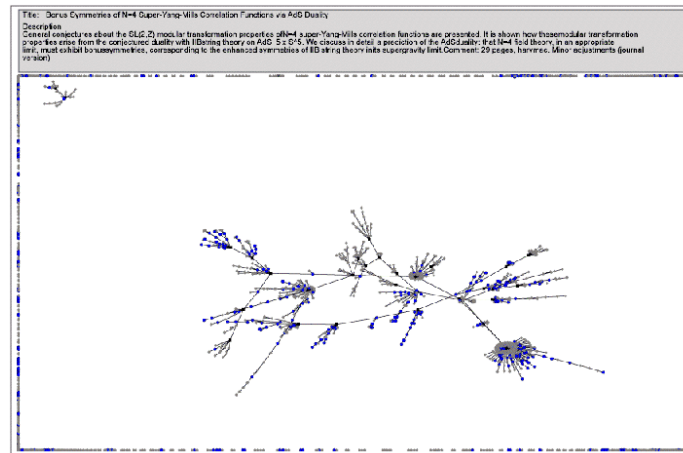


Fig. 9. A full-sized co-citation map with a lower co-citation threshold resulting in more nodes being included. Several clusters (research fronts) are evident, in particular the large cluster towards the bottom right of the map. Researchers may get a better understanding of their research landscape by exploring these clusters and the relationships between them. Different colours are also used to indicate which nodes have been recently highly cited, paving the way for up-and-coming (or dying) research fronts to be identified. There are also several occurrences of 5 or 6 nodes emanating sequentially out of a single node, indicating a sequence of papers being published that address a common problem or theme.

Knowledge Questions

Digitometrics can also analyse the metadata and infer new facts. Such capabilities are central to the Semantic Web initiative. For example, based on the citation patterns between papers, the most significant papers and prominent researchers can be detected. The contributions to a particular line of research or researcher can be mapped by analysing and displaying co-authorship patterns (Figure 10). With further

high quality metadata available, researchers can raise increasingly subtle questions about the direction and time-course of developments in their research field, such as how perspectives have changed and how a particular methodology has affected a research area.



Fig. 10. The collaborators for a researcher “Tollestrup, A” are calculated and presented by analysing co-authorship patterns.

Hypertext Navigation

Digitometrics enables complete hypertext linking between different research artifacts (e.g. researcher, publication, project, organisation, publication medium, text). When a service is offered for a particular artifact (e.g. collaboration lists among researchers, co-citation maps for literature) these are available to authors, researchers, or evaluators (Figure 11).

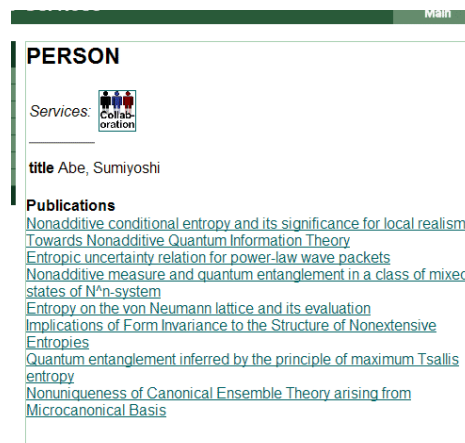


Fig. 11. Typical information screen for a researcher. This illustrates the metadata gathered for a particular researcher, including a list of all their known publications, plus the collaboration measure is offered as a service (Figure 10).

Conclusion

As the coverage of the open-access archives increases so will the technology to process and parse the online corpus, to provide hypertext-navigation (between articles, authors, institutions, and collections), and to provide real-time analysis (“digitometrics”) of the research world, for example to improve the accuracy and breadth of research assessment.

The potential benefit for research literature – and research itself - through open-access, distributed archives is only beginning to appear. Authors maximise their research impact by maximising the visibility of their work, and at the same time provide a huge resource to build services on. Celestial, Citebase, and

Digitometric Services demonstrate some of this potential – even with the limited amount of literature available now.

References

1. Hitchcock, S. et al: "Open Citation Linking: The Way Forward". D-Lib Magazine, Vol. 8, No. 10, October 2002 (available at <http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html>)
2. Liu, X., Brody, T. et al: "A Scalable Architecture for Harvest-Based Digital Libraries" D-Lib Magazine, Vol 8, No. 11, November 2002 (available at <http://www.dlib.org/dlib/november02/liu/11liu.html>)
3. University of Southampton Celestial Mirror <http://celestial.eprints.org/cgi-bin/status>
4. Celestial <http://celestial.eprints.org/>
5. Citebase Search <http://citebase.eprints.org/>
6. EPrints.org <http://software.eprints.org/>
7. Open Citation Project <http://opcit.eprints.org/>
8. Open Archives Initiative <http://www.openarchives.org/>
9. Repository Explorer http://www.purl.org/NET/oai_explorer
10. Academic Metadata Format <http://amf.openlib.org/doc/ebisu.html>
11. Dublin Core <http://dublincore.org/>
12. Slashdot <http://www.slashdot.org/>
13. van de Sompel, H., Lagoze, C.: "Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative" ECDL 2002, in LNCS 2458, pp. 144-157
14. Lagoze, C., Arms, W., Gan, S., Hillmann, D., Ingram, C., Krafft, D., Marisa, R., Phipps, J., Saylor, J., Terrizzi, C., Hoehn, W., Millman, D., Allan, J., Guzman-Lara, S., Kalt, T. (2002) "Core Services in the Architecture of the National Digital Library for Science Education (NSDL)" Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries 201-209 <http://arxiv.org/abs/cs.DL/0201025>
15. AKT reference <http://www.hyphen.info/>
16. Edge, D.: "Why I am not a co-citationist?", Society for Social Studies of Science Newsletter, 2, pp13-19, 1977.
17. Garfield, E.: "Scientists should understand the limitations as well as the virtues of citation analysis", The Scientist, 7 (14), p12, 1993.
18. Garfield, E., Sher, I., and Torpie, R.: "The use of citation data in writing the history of science, Institute for Scientific Information", Philadelphia, 1964. (available from: <http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf>)
19. MacRoberts, M. and MacRoberts, B.: "Problems of citation analysis: a critical review", Journal of the American Society for Information Science, 40 (5), pp342-349, 1989.
20. Small, H.: "Co-Citation in the Scientific Literature: A New Measure of the Relationships Between Two Documents", Journal of the American Society for Information Science, 24, pp265-269, 1973.
21. Harnad, S., Carr, L.: "Integrating, navigating, and analysing open Eprint archives through open citation linking (the OpCit project)" Current Science, Vol 59, No. 5, September 2000 (available from <http://tejas.serc.iisc.ernet.in/~currsci/sep102000/629.pdf>)
22. Garfield, E.: "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas" Science, Vol. 122, No. 3159, pp108-111, 1955 (available from [http://www.garfield.library.upenn.edu/papers/science_v122\(3159\)p108y1955.html](http://www.garfield.library.upenn.edu/papers/science_v122(3159)p108y1955.html))
23. Hitchcock, S., Wookou, A. et al: "Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service" 2002 (available from <http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report.html>)
24. Citebase Correlation Generator <http://citebase.eprints.org/analysis/correlation.php>
25. "Scirus, for scientific information only" <http://www.scirus.com/about/>