Access to Citation Data: Cost-benefit and Risk Review and Forward Look

Dr Max Hammond
Professor Charles Oppenheim
Dr Geoff Curtis
Executive summary

Context

1 Citation data provides a basis for business intelligence for academic research. At its simplest, (De Bellis, 2009) [1] a citation datum is an assertion that, say, two published journal articles, A and B, have a specific relationship. For example, the relationship might be “A is built on B” or “A refutes B”. Large-scale aggregations of citation data can be exploited to show relationships between researchers such as influence and collaboration and provides measures of, for example, research and researcher impact.

2 There is increasing demand for access to citation data in order to inform research, its planning, strategic direction and management. It is therefore timely to examine how citation data may develop and be used and Jisc [2] commissioned this review of future directions for citation data.

Current position

3 Large-scale aggregations of citation data [3] derived from peer-reviewed academic research articles are increasingly exploited to inform the conduct of research, performance management and business intelligence. Given the move to open access (OA) publication, it is timely to examine how citation data may develop and be used, which is why Jisc has commissioned this review of future directions for citation data.

4 Key usage scenarios for citation data require comprehensive data. This is data with the right coverage of works, the right time span of current and historical citations and the right quality of journals/books. The creation, dissemination and exploitation of citation data are global. It’s clear, too, that the majority of the effort invested, and the value generated, in the citation data lifecycle relates to collating and indexing work on an international scale.

5 There are two well-established paid-for high quality citation data services, Elsevier’s Scopus and Thomson Reuters’ Web of Knowledge, and two relatively new free automated services, Google Scholar and Microsoft Academic Search, that undertake less quality-control, and are perceived by users as providing lower quality results. There are also global community initiatives eg CrossRef driven by publishers, and ORCID driven by institutions and publishers, which are successful in improving the consistency and accuracy of citation data and in reducing costs to citation data service providers.

Assessment

6 Overall, the provision of indexed citation data in the UK works well [4]. Despite some significant niggles detailed below, there is generally a high level of satisfaction within the community for the current

---

1 Citation data and citation metrics are introduced at Section 2; the citation data lifecycle is considered more fully at Annex A. Current citation data services and other relevant initiatives and services are described at Section 4.

2 The work was commissioned by the Joint Information Systems Committee (JISC), which has now become Jisc. See www.jisc.ac.uk [accessed 9 June 2012].

3 Citation data and citation metrics are introduced at Section 2; the citation data lifecycle is considered more fully at Annex A. Current citation data services and other relevant initiatives and services are described at Section 4.

4 The market analysis is set out at sub-section 4.5. The overall findings are set out at sub-sections 6.2 and 6.3.3.
approach, which is considered to be robust, trusted, well-understood and meeting current usage scenarios. There were few complaints regarding the cost of commercial citation data services. At present, the free services are not perceived to provide sufficient quality, or to operate in a manner that is sufficiently transparent to meet research evaluation and business intelligence usage scenarios.

Publishers interviewed tended to value discoverability more highly than the need to enhance revenue streams by selling bibliographic and citation data. There is a possibility that publishers may agree to make their citation data openly available – indeed, some already do.

Widespread access to the outputs of research through OA publishing will undoubtedly enable broader re-use of this data for citation analysis. With time, the importance of older, non-OA publications will reduce. There may be some new entrants to the citation data service market. Initially, these are likely to offer subject-specific abstract and indexing (A&I) rather than comprehensive citation data services.

Potential for open citation data

Citation data is valuable, and it is logical to assume that there is more benefit to the society that paid for the research when more users have access to the citation data. Although this argument is valid, it seems inappropriate to attribute the same level of benefits to citation data as to those that accrue from openness of the data itself. Moreover, the value of citations from a single paper is very limited. The value is mainly created by the organisations that collate and index the data across many countries, publishers and disciplines. This value is only accessible after this processing, rather than from the article citation data itself.

Even if OA publishing becomes predominant, there are significant limitations to how the open model can work for citation data. The study has analysed two different alternative end-to-end business models [5] that reflect different levels of openness for citation data. While these options are viable, the study team does not believe that they offer a likely or preferable future environment. This does not mean that change should not happen, but rather that it is likely to be gradual rather than disruptive. The key inhibitors of change are:

- The level of difficulty in extracting and interpreting citation data from original publications
- The high cost of beginning the process of developing suitable scale systems for citation data
- The challenges in accessing, collating and indexing historical citation data
- The market for the supply of indexed citation data to end users in the UK is functioning well
- General, although not total, satisfaction with the current systems
- The global nature of citations and the fact that the UK represents just a small part of the citation landscape

Based on these inhibitors, and the overall positive assessment of the current system alongside the relatively weak openness argument, the study team conclude that an imposed radical intervention in the indexed citation data market would be difficult to justify. In the team’s opinion it would be better to encourage the opening of citation data through the opening of publications themselves, and make limited interventions that support this aim while enabling the current system to work more effectively.

Section 5 examines possible future business models with Annex B providing the detailed analysis.
What could improve the system?

12 Jisc should continue with its emphasis on open publication and open access. It should actively engage with the stakeholders and actively monitor the citation data situation and change its approach in response.

13 Jisc should use its best endeavours to encourage stakeholders such as publishers and citation data service providers, to move toward a standardised syntax for citations, or develop an approach to embedding citation metadata in authors’ drafts. This will help improve the effectiveness of citation data services by reducing the complexities in processing publications.

14 Jisc should convert its support for ORCID into active engagement with, and commitment to, the further development of the service.

15 Jisc should commission research to develop best practice guidance for use of simple citation metrics for evaluation of institutions, departments and individuals. This would reduce the concerns of many stakeholders regarding inappropriate decisions based on citation data.

16 Jisc should work with representative publisher bodies to develop and promulgate the business case for moving to an essentially digital but lightweight citation data workflow, and help their members implement this. The intention is to make the citation data workflow more efficient and therefore more economic.
Introduction

1.1 Aim, scope and focus of the study

Aim

1.1.1 The overarching aim of the report is to explore and suggest practical directions and actions to move toward more cost-effective creation, dissemination and exploitation of citation data in the context of current and potential future usage scenarios. A further aim is to propose the roles that Jisc and others might play in this system in future.

Scope and focus

1.1.2 The general scope of this work is the creation and exploitation of citation data derived from peer-reviewed academic research articles. However, the citation of datasets is specifically excluded. While bibliographic metadata associated with the referencing and referenced outputs is clearly relevant, it is not the focus of the work. Approaches to exploitation of citation data are also not the focus of the review, except in as much that this might increase or change the demand for different types of citation data.

1.2 Study approach

1.2.1 The study approach has been informed by the JISC invitation to tender (JISC Executive, 2012) and was conducted in three phases. In the first phase, desk research and an intensive series of interviews were undertaken with key stakeholders including publishers, citation data providers, citation data users, research funders to understand the strategic drivers. The information gathered was used to identify possible outline usage scenarios and develop a business model framework together with an initial view of the pros and cons of each scenario.

1.2.2 In the second phase, an agreed set of usage scenarios and business models were developed and refined in consultation with users of citation data, publishers and citation data providers. In addition, a DevCSI developers' workshop or Hack Day was held on 27 September 2012 [6] to explore usage of citation data through short trials or pilots using real citation data. This brought together a group of domain experts, users and developers to explore ideas related to potential real world uses of citation data and to prototype potential solutions. The group investigated aspects of citation data and its use, including the properties of sparse networks of data, and considering new ways to visualise citation data. The event provided some interesting perspectives on the use of citation data, and in particular supported the design of the 'open' processes.

1.2.3 In the final phase, the results of the second phase have been used to develop options for a viable practical direction and set of actions for taking the use of citation data forward. It is planned to seek feedback and agreement with stakeholders at a final stakeholder meeting.

---

6 The outcomes of the event are described in detail at [http://devcsi.ukoln.ac.uk/2012/11/05/citation-data-hack-event-report/](http://devcsi.ukoln.ac.uk/2012/11/05/citation-data-hack-event-report/).
2 Background - citation data and citation metrics

2.1 Introduction

Overview

2.1.1 The idea of collecting and using citation information to aid scholarly endeavour is not new. The very earliest citation index dates back to mediaeval times and relates to theological questions. The well-established Shephard’s Citations US legal citation service dates back to 1873. Nonetheless, in most people’s eyes citation analysis is indelibly associated with Eugene Garfield and his ideas for a citation index for science and technology, first expounded in a famous article in Science in 1955.

2.1.2 In the late 1950s, supported by funding from the US National Institutes of Health, Garfield proved the concept of citation indexing could be used to identify associations of scientific and technical ideas. Having proved the concept, and without any further external funding, in 1963 Garfield launched the first edition of Science Citation Index covering the literature of 1961. This formed the basis of citation indexing services that exist today. What is not so well known is that Garfield’s first experiment with a commercial citation service was not in fact Science Citation Index, but with a patent citation index in 1957. However, the US Patent and Trademark Office, whose support and co-operation would have been crucial to keep the service going, was not interested in the idea, so Garfield switched his attention to scholarly journal articles instead.

Changes over time

2.1.3 Initially, citation indexes, which were like the abstracting and indexing services of the time, only available in print form, were used as a novel information retrieval service. If someone was interested in the subject of the structure of DNA, rather than looking up the term ‘DNA’ in the printed index of an abstracting and indexing service, they would look at who had cited a key initial article on the topic, such as the Watson & Crick article in Nature, and would then follow up on the details of those articles that had cited Watson & Crick. It was widely seen as an effective alternative to subject index searching in science-based subjects.

2.1.4 Over the years, Garfield’s Institute for Scientific Information (ISI) both expanded the scope of its services, and made computer files available for searching, initially on magnetic tape, later via online host services, and nowadays using the internet.
2.2 What is a citation?

Introduction

2.2.1 A citation, in essence, is a relationship between two documents – and in the context of this report, at least one of those documents is likely to be a piece of academic work. The citation has a direction – one piece of work cites the other. As a minimum, a citation can be defined by two pieces of information – a reference for the cited work, and a reference for the citing work. There are significant challenges in defining these references, and these are discussed at paragraphs 2.2.5 - 2.2.11.

2.2.2 One key issue with the practical use of citation data is that the majority of usage scenarios require ‘reverse lookup’. It is simple to discover which works are cited by any given work, but discovering all the works that cite another item is much harder. The information about ‘incoming’ citations is stored within or alongside each citing document, and these are likely to be distributed among different locations.

2.2.3 When discussing citation data, several meanings are often used interchangeably. This can lead to confusion, and it is necessary to be explicit about what is meant by ‘citation data’ in any given case. This could include:

- A single citation
- All the citations from or to a work
- A corpus of citations regarding different works (the boundaries may be a publisher, a subject, everything available, or some other factor)
- The number of citations that a work has received
- Derivative metrics such as h-index or Impact factor
- Other information about the works involved (see discussion on bibliographic data at paragraphs 2.2.12 - 2.2.13)

2.2.4 Further, citations are typically considered within the context of academic publications, and primarily journal articles. However, a range of other types of work could be cited, or could in turn cite an academic publication. There are a number of issues in the interpretation of broader citations – they clearly represent some impact of the work, but this is not captured in the existing citation systems. What value does an academic paper being mentioned in a blog have, compared to being mentioned in another academic paper? This report focuses on ‘traditional’ citations – including journal articles, books, conference proceedings and other well-accepted academic outputs.

Identifiers

Works

2.2.5 In order for a citation to be meaningful, it must as a minimum identify the cited work, (which may be implicit, as the citation may be contained within the work itself). People have traditionally constructed these references from bibliographic information for the cited work including author(s), article title, journal title, journal issue and volume number, book title, page numbers, cited page(s) and many other possibilities. The exact format used varies strongly between different publications. This is particularly challenging given the range of works that could be referenced. Although the bulk of citations in
 academic literature are between peer-reviewed journal articles, they could also include many other types of information [7].

2.2.6 In addition, as authors typically create these bibliographic references manually, there is a considerable opportunity for errors to be introduced during the drafting process. No reliable figures for the proportion of errors have ever been identified, but it is plausible to assume the proportion of errors could be as high as 10% of all references in any given output. There is also the problem of a lazy author who does not check the original item they are citing, and simply reproduces someone else’s citation to a source document, thereby repeating any error made by the earlier author.

2.2.7 Some works will have been issued an unambiguous identifier. The most common scheme for this is the Digital Object Identifier (DOI) (DOI, 2012a). This system allows organisations to issue unique identifiers for objects, and not just digital objects – DOI is a digital identifier of an object, rather than just an identifier of a digital object (DOI, 2012b). Typically, this happens when publishers issue DOIs for their own publications. A DOI has a format such as ‘10.1021/ja02105a700’. Whereas the vast majority of articles published now are issued with DOIs, the picture is considerably more variable for older works.

2.2.8 Related to the DOI initiative is CrossRef (CrossRef, 2012a), which manages the DOI registration process for publishers, and acts as a ‘resolver’ – given a DOI, CrossRef can provide a URL and some other metadata for the object to which that DOI refers.

Author

2.2.9 In the context of citation data and more broadly (UKOLN, 2012), it is essential to identify the authors of works uniquely. This is necessary for processing data to generate statistics for performance measurement and business intelligence uses. Automatically assigning publications to authors is challenging – for example, their name is most likely not unique, their institutional affiliations may change throughout their career, and they may collaborate with other institutions.

2.2.10 Both Thomson Reuters and Elsevier have established identification mechanisms for authors to claim their own work. These are known as ResearcherID (Thomson Reuters, 2012c) and Scopus Author Identifier (SciVerse Scopus, 2012a) respectively. Both have had only modest uptake by authors.

2.2.11 A recently launched scheme called ORCID (ORCID, 2012) has the support of Thomson Reuters, Elsevier, CrossRef, and a range of publishers and major research institutions. ORCID was recently selected by Jisc as the preferred solution for the UK (JISC, 2012a). ResearcherID and Author Identifier are increasingly being linked to ORCID.

Relationship to bibliographic data

2.2.12 Although, at a minimum, citation data contains the link between two identifiers, in practice it is usually associated with broader information on the objects at the ends of the link. Basic bibliographic information is included within a traditional citation (frequently including authors, journal/book title, issue, volume, article title, page numbers, year), and this level of information is included within all the major citation databases.

---

for example: a book, a book section, a range of pages from a book, a website, a blog posting, a database, conference proceedings or a single contribution in a set of conference proceedings, software, legislation, a single paragraph from a law, a court case, a television or radio programme, a patent specification, a technical report, an item in a museum collection, a scientific specimen, and a wide range of other things including personal communications and unpublished works.
From one perspective, citation data and bibliographic data could be entirely separate. Citations could be processed based on their identifiers, and without any understanding of the objects to which they refer. From another perspective, citation data is one type of bibliographic data relating to the objects involved, and has no meaning for people who don’t understand the objects. Although this may seem to be a subtle distinction, it has implications for the design of systems that distribute and process citation data, as it is clearly necessary to match the citations up with the objects they relate to before the citations can be used meaningfully (Annex A).

**Data completeness**

When considering corpuses of citation data – all the citations relating to a paper, an author, or other criteria – it is important that the corpus should be ‘complete’. This raises questions about what completeness means, however. Clearly, all relevant citations should be included – but what is relevant depends on how the data will be used.

There are two distinct aspects to this:

- **Scope**: the sources from which citations are included in the corpus
- **Accuracy**: whether the citations within scope are correct and assigned correctly to authors, works, etc.

The traditional scope for citation data corpuses is that they cover citations within and between journals and conference proceedings. Scopus also includes trade publications and books. There is a strong argument that while these scopes are appropriate for natural sciences, they are insufficient to cover the range of academic output within other domains [8]. Google Scholar and Microsoft Academic Search include a wider range of outputs, but whether this is beneficial depends on the final use of the data. For example, it is unclear how these ‘non-traditional’ citations could or should affect measures of research impact.

**The uses of citation data**

Right from the launch of *Science Citation Index*, Garfield had personally investigated the use of the citation data collected for uses other than simply subject searches, and published hundreds of short articles, and many longer ones, describing the results of his research. In 1979, he wrote a seminal book, *Citation Indexing – its theory and application in Science, Technology and Humanities*. Only one chapter of the book is devoted to using citation indexes for searching, whilst five are devoted to other uses, i.e., as a science management tool, for research into the history of science, for mapping the structure of science, for assessing the impact of journals and for assessing the impact of individuals. At the time the book was written, the extension of ISI citation database scope to the Social Sciences and the Arts and Humanities was relatively recent, so the emphasis on science is not surprising.

In the 1980s and beyond, many researchers undertook citation analysis studies using their access to the ISI databases, now known as Web of Knowledge and acquired by Thomson (now Thomson Reuters). In recent years, such researchers have acquired further tools for citation analyses, including Reed-Elsevier’s Scopus service, Thomson Reuters’ Web of Knowledge and Web of Science, Google Scholar, Microsoft Academic Search and CiteSeer*. Between them these services offer a range of prices and quality in terms

---

8 Some of the issues are raised by (Harzing, 2007).
of their accuracy, scope of sources and transparency. While each of these services has its pros and cons, their widespread availability has certainly given an impetus to citation analysis studies.

Citation analysis

2.3.3 As already noted by (Garfield, 1979), one important use of citation analysis is the evaluation of the impact and by implication, the importance of research. This could be at the level of an individual scholar’s research, a group of scholars’ research, say in a university department, an entire university’s research, or an entire country’s research. The impact might then be used to decide on funding priorities [9], on appointments and promotions [10] or on deciding which journal to submit a piece of research to for publication.

2.3.4 It is now common practice for universities wishing to give tenure to, promote, or head hunt an academic to consider the individual’s citation statistics as well as their publication and research income profile. Indeed, many years ago there was an employment case in the US (Fisher vs. Vassar College (Leap, 1995)) [11] where the candidate with the higher citation count won a claim of sexual discrimination against a university that had given an appointment to another candidate with a significantly lower citation count, so there is legal precedent for saying that citation counts are a reliable method of assessing a person’s quality.

h-index

2.3.5 This sexual discrimination case occurred many years before the development of the h-index, a measure of an individual’s citation impact, which has become extremely popular method for evaluating scholars. An individual has an h index of n if they have at least n publications, each of which has been cited at least n times. The h-index has the great advantage that it is very simple to calculate, either manually or by using software provided by the citation database producers and others. An individual’s h-index has to be a whole number. The lowest possible h-index is 0, for someone who has not received any citations to their papers at all. The highest h-index varies from subject to subject, but can reach the low hundreds in some science areas.

2.3.6 One disadvantage of the h-index is that it favours those who have been publishing and receiving citations over many years rather than early career researchers. On the other hand, it does avoid the distortion of, say, one individual who has received thousands of citations to one paper, but none to any other paper. On a simple citation count basis, they have done extremely well, but their h-index score is just 1. Thus the h-index favours the person who produces a steady stream of reasonably well-cited papers over someone who has one or two extremely highly cited papers but little else. In recent years, many alternatives to the h-index have been proposed, but none offer the simplicity of approach of the original h-index (De Bellis, 2009).

2.3.7 The h-index is generally regarded as a robust way of comparing the performance of scholars in the same subject area who have been active for similar periods of time. The h-index should be used only with caution or not at all when either or both of these two criteria do not apply.

---

9 Those with an excellent research impact are likely to get more money, or possibly the reverse if one wants to help identify and then improve weak research.

10 Perhaps giving preference to the person with the higher citation count.

11 While the ruling was overturned on appeal, the use of citation data as a means of characterising academic performance was not (United States Court of Appeals, Second Circuit, 1997).
**Journal impact factor**

2.3.8 Another measure that has received a lot of attention is the journal impact factor. The impact factor of a journal is the average number of citations given in articles in that journal published in one year over a certain period. This is typically the two years following the year of publication of the journal articles, but many other time periods have been used. Thus, if a particular journal published 50 articles in 2009, and those articles in total received 200 citations in the period 2010-2011, then that journal’s impact factor is 200/50 = 4.0. Again, the implication is that the higher the impact factor, the more prestigious or important the journal is. Commercial citation services offer impact factor calculators, which are very easy to use, as well as publishing annual listings of the impact factor of the journals they are aware of.

2.3.9 The same caveats apply to impact factor as to the h-index. It is a useful tool to compare two journals in the same subject area as long as the method of calculation is consistent. It is not appropriate at all for comparing journals in different subject areas, because the citation patterns in different subject areas vary so much. However, the biggest concern about impact factor is the way it is used by editors, authors and employers. Journal publishers are always quick to trumpet their impact factor if it has gone up in the rankings for its subject area.

2.3.10 There are persistent rumours that some editors induce authors to make extra citations to their own journal if the article is to be accepted, for the sole motive of improving their journal’s impact factor. There are also great concerns that universities are trying to pressurise their academic staff to only submit articles to high impact factor journals in their field, in the mistaken belief that because the impact factor is high, then the article when published will automatically receive a lot of citations. All too often, the high impact factor journal is quite inappropriate for the particular subject the academic has been researching. In addition, such pressure just adds to editors’ workloads, as increased numbers of submissions cannot be accommodated easily by the journal in question.

2.3.11 While some question the value of the impact factor, it is regularly used and is regarded as a useful indicator by authors, publishers, research managers and funders.

**Interpretation of impact**

2.3.12 The use of citation analysis to evaluate the impact, and by implication the quality of research, has always been controversial, and that controversy has become more acute as such analysis has increased. The controversy revolves around a number of factors:

1) Scholars in the arts, social sciences, pure sciences and technologies employ citations differently.

2) Most citation indexes rely heavily on journal articles as the primary input, but in some subject areas monographs, reports, patents, etc., are far more important sources than journals.

3) Even in an individual article, the reason why the author has cited each reference may well differ, eg, for historical background, as the inspiration behind this research, for comparing the results found with earlier ones, in order to criticise the earlier paper, to appease the journal editor or referees, promote one’s own previous research, and so on.

4) Some key papers no longer get cited because they are so well-established that it is assumed the reader will know about them. The classic example of this is allegedly Einstein’s paper where he introduced the formula $E = mc^2$. There is no question that many authors no longer bother to cite the original article when using the equation.
5) Most citation indices reduce quality to a single, crude number, when many complex factors are involved in assessing whether a piece of research is a real advance. More sophisticated measures, mostly still using citations as the raw material, have been proposed, but are not used nearly as frequently.

6) Some key papers only become recognised as important many years after they were first written, whereas evaluations tend to focus on recent citations.

7) The focus on impact factors for evaluating journals is forcing authors to submit articles to high impact factor journals even if they are not the most appropriate for the particular research output.

8) There are problems in identifying the correct citations to an individual with a particularly common name, such as John Smith. While all of the citation services apply various degrees of algorithmic processing to try and minimise these errors, the results are still imperfect.

9) Google Scholar in particular contains many obvious errors, but all the services do make mistakes. Some of these are keyboard or transcription errors, and some are mistakes in citations made by an author that are faithfully recorded by the citation indexing service.

10) The different and sometimes unrevealed coverage of the available services, alongside problems with author disambiguation, etc mean that different services will give different citation analyses and calculations of impact factor and h-index.

11) With the increasing use of repositories and other websites for publishing scholarly materials, the importance of journal articles is declining; there is no common approach to understanding the relevance of the non-journal citations, indexing them, and valuing them.

**Journal usage as a measure of impact**

2.3.14 Although the relatively few studies that have been carried out clearly show that citation counts are strongly correlated with subjective measures of eminence and importance in all subject areas, scepticism about citation analysis remains high.

While out of scope of this study, journal usage can provide useful business intelligence alongside, for example, the journal impact factor. Usage can be measured in a number of ways:

- the number of downloads for electronic materials
- the number of borrowings for physical journals
- the time spent reading an electronic item
- the number of document supply requests eg from the British Library Document Supply Centre, etc. However, these can all be fraught with difficulties. For example, a document might have been read by serendipity and then discarded, or the reader might have been distracted while reading the item.
2.3.15 As examples, publishers measure the usage of their own journals and use these as well as impact factor to help with business decisions. Librarians [12] measure the usage of journals etc, within their institutions and use this in conjunction with the impact factor to support decisions on which journals to subscribe to.

2.3.16 Even where statistics on journal usage are reliable, they do not measure the impact of the item, but rather the interest that item has generated in the mind of the reader. Therefore even though the few studies that have been carried out show that there is a correlation between citation statistics and journal usage statistics, it is likely that the two sets of figures represent two different aspects of the same output.

2.4 Citation data services

2.4.1 The two major commercial suppliers of citation data [13], Scopus and Web of Knowledge, were quick to recognise the commercial opportunities for exploiting the citation and other data they held for evaluation and assessment purposes. Both offer value-added services, some charged for, some bundled in with the basic subscription, for analysing citation data in a variety of ways.

2.4.2 While some suppliers, including Google Scholar and Microsoft Academic Search, offer free services, they do not provide similar value-added services. However, (Harzing, 2007) has developed her Publish or Perish software to make it easier to analyse citations using Google Scholar.

2.4.3 These various services have enhanced interest of both employers and funders in exploiting citation data. These developments have coincided with the development of research assessment exercises in a number of countries. Such exercises, typically undertaken by the Ministry of Education or the relevant major funding agency for higher education, help the Ministry or agency to reward universities offering excellent research financially. In effect, this penalises those that do not undertake excellent research.

2.4.4 In Australia, the Excellence in Research in Australia initiative is explicitly heavily dependent on citation data. In contrast, the UK’s Research Excellence Framework (formerly the Research Assessment Exercise) has in the past avoided citation data as an explicit criterion, though in the forthcoming 2014 Research Excellence Framework, citation data will be a component of the evaluation in some science subject areas.

---

12 To help libraries with decision-making on what e journals, e-books and databases to subscribe to and which not to, the COUNTER Project (Counter, 2012) has developed a series of widely-used and supported protocols on how statistics should be collected and presented to libraries for easy analysis. This is an important initiative and one that is widely used by libraries to assess what the most popular resources are, but the measure is primarily used as a purchase decision tool rather than as a measure of the ultimate impact of an item.

13 These services are described in detail at Section 4.
3 Citation data usage scenarios

3.1 Introduction

3.1.1 To understand and address the working of the overall citation data system, it is first necessary to understand the ways that citation data are used, and any changes anticipated in the future. The usage scenarios considered are listed at Table 3-1 together with priorities for the citation data required to support the scenario. These are discussed further.

<table>
<thead>
<tr>
<th>Usage scenario</th>
<th>Priorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conducting research</td>
<td>‘Comprehensive’ ie cover the relevant and important journals</td>
</tr>
<tr>
<td></td>
<td>Retrospective ie go far enough back in time to ensure all relevant articles may be found</td>
</tr>
<tr>
<td></td>
<td>Current ie be up to date</td>
</tr>
<tr>
<td>Performance management</td>
<td>Transparent ie have a clear and published definition of the data sources and processes used</td>
</tr>
<tr>
<td></td>
<td>Trustworthy ie operate in a manner such that stakeholders trust the result</td>
</tr>
<tr>
<td></td>
<td>Reproducible ie provide similar results for other citation data sets with any differences being explainable</td>
</tr>
<tr>
<td>Business intelligence</td>
<td>Contextualised ie be connected with bibliographic and other data</td>
</tr>
<tr>
<td></td>
<td>Current</td>
</tr>
<tr>
<td></td>
<td>Usable with other systems</td>
</tr>
</tbody>
</table>

Table 3-1: Summary of usage scenarios for citation data

3.2 Conducting research

The requirement

3.2.1 The original and most obvious use of citation data is in the conduct of research itself. Citations represent direct and explicit connections between academic publications and can be used to find useful articles, other workers in the field or identify potential collaborators.

3.2.2 In this usage scenario, the specific citations are of primary importance, rather than any metrics derived from them. Citations are typically used in a very personal way – a researcher will follow a citation trail to understand the domain. This understanding may be retrospective or current.

- The researcher may retrospectively follow historical citations in order to understand how a field has developed. This is typically part of a much broader literature search– the requirement is to understand the field, and citations are one way of doing so.
- They may follow current citations in order to understand the current environment – who is working in a field of interest to them, and what they are working on. They may track who cites their own work, and who cites their ‘competitors’. Tracking current citations helps researchers understand who they have influenced, and how. Some subjects have a tradition of conducting discussions through follow-up articles and letters to editors. In other cases the follow up may come months or even years later, when new work is published that confirms, refutes, extends or constrains the
original piece. In some published papers, an author will cite a forthcoming publication and thereby can provide clues as to where that particular researcher’s interests are developing or changing.

3.2.3 For the conduct of research, the existence of a citation is only a pointer—a way to find the citing or cited material.

3.2.4 For this usage scenario, the citation data needs to be comprehensive, retrospective, and current.

The current system

3.2.5 Academics typically identify citations through a commercial citation data service to which their institution subscribes, and use these services interactively through their websites. Some of these services can be used to notify the user when a particular article is cited. Identifying and following citations forms part of a much broader review including bibliographic data and the referenced publications themselves.

Trends

3.2.6 The use of citations in this way is almost as old as research itself, and changes to uses of citations for these purposes are relatively minor. They represent changes in process and use of technology and software, rather than any fundamental shifts. The easy and timely availability of citation information permits researchers to follow domains with greater ease than before, and to discover citations to their own and others’ work more rapidly than was previously possible. This is analogous to the overall move to electronic publishing – the process is essentially the same, but has become faster and more accessible.

3.3 Performance management

3.3.1 This usage scenario requires the use of citation data to measure and compare research. This can be conducted at a number of scales, from individual researchers to entire grant portfolios. There is a fine line between performance management and business intelligence – see below (sub-section 3.4).

<table>
<thead>
<tr>
<th>Whose performance is measured</th>
<th>Who is measuring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual researchers</td>
<td>Department, institution, funding body</td>
</tr>
<tr>
<td>Research projects</td>
<td>Project leader, funding body</td>
</tr>
<tr>
<td>Departments</td>
<td>Institution, funding body</td>
</tr>
<tr>
<td>Publications</td>
<td>Publishers</td>
</tr>
<tr>
<td>Research institutions</td>
<td>Funding body</td>
</tr>
<tr>
<td>Grant portfolios</td>
<td>Grant-awarding organisation</td>
</tr>
</tbody>
</table>

Table 3.2: Performance measurement

3.3.2 For use in performance management, citation data is typically used at an aggregate level – little attention is paid to the individual citations, but rather it is the number of citations that matters.

3.3.3 Performance management, including using citation data, has significant influence. For example, decisions made by funding bodies affect the allocation of billions of pounds of funding for research each year, and decisions made by university departments can significantly affect the career progression of individual researchers. As such, for reasons of both public accountability and individual fairness, it is
fundamentally important that the process used, and the data included are transparent, trustworthy and reproducible and to a lesser extent comprehensive.

The current system

3.3.4 Citation data is treated as one signal within the complex set of measures and indicators that give some idea of the quality of research or researchers. At an informal level, individuals such as researchers, project leaders, and potentially other institutional staff may use the same tools they use to investigate citations for research (see above) to generate some insight into their own performance.

3.3.5 On the other hand, formal systems require a corresponding increase in the formality with which citation data is obtained. It is difficult to generalise the systems, but the requirements for data to be transparent and reproducible typically lead people to use one of the major citation data providers. These services are familiar to stakeholders, are mostly transparent about their approaches and data sources, and are willing to enter into contractual agreements for the supply of data. Robust systems will typically define in detail the analysis approach, and may constrain the citations or publications that can be included, for example, by excluding review papers, which are typically very highly cited and typically include a lot of citations themselves. This service will typically perform analyses according to a fixed and predetermined procedure and schedule.

3.3.6 As mentioned previously, citation analysis now forms a part of the research assessment process in the UK (REF) and Australia (ERA). Whereas citation analysis plays a limited role in the REF, it has taken an increasingly important role in the ERA.

3.3.7 One specific use of citation data for performance management comes during recruitment of researchers. Citation data, and primarily the h-index, often forms part of that assessment. The way that this information is used depends on the appointment panel, but it is unlikely to be used as an absolute criterion, but rather as a single piece of information that must then be analysed and placed in context. (cf paragraphs 2.3.5 - 2.3.7).

3.3.8 Publishers use citation data alongside other bibliometric measures as an important indicator of the success of their publications as we have explained, the impact factor is still seen by many as a key indicator of the quality of a journal.

Trends

3.3.9 The use of citation data for performance measurement is not new, but has been steadily growing in importance, in particular over the last decade, at all levels from individuals to countries. The options and limitations for the use of citation data in performance measurement are now relatively well understood by specialists, although there are frequent misunderstandings and misconceptions amongst researchers and research managers.

3.3.10 Some consider that citation metrics do not describe the full range of research outputs or impacts and may view them with particular scepticism, particularly by scholars in the arts and humanities. There is a growing movement [14] to develop additional metrics that provide broader or alternative perspectives.

---

14 For example, the altmetrics manifesto (Altmetrics, 2012) – although this is grander in scope, and considers metrics for peer review rather than simply research impact.
3.3.11 It seems likely that we are nearing the limits of straightforward citation analysis. Given that maintaining trust in the systems is critically important, analysis can only use relatively simple metrics like citation counts, h-index, etc, which can be clearly explained and understood. The use of more-complex measures, which typically have more indirect, strategic implications, is the domain of business intelligence.

3.4 Business intelligence

**The requirement**

3.4.1 Business intelligence is the use of data to support better decision making in business. This typically means using a large volume of data from across the business to measure and analyse operational and strategic performance and provide actionable information.

3.4.2 There is a fine line between business intelligence and performance management, and in reality they represent different ends of a spectrum of analysis, with performance management focusing on simple metrics, and business intelligence focusing on more-complex analysis.

3.4.3 The nature of business intelligence analysis is that it is most likely to be conducted by research institutions and publishers. Funding bodies are typically driven by strategic and policy factors, rather than a requirement to maximise output in its own right, and individual researchers are unlikely to have the expertise or time to interrogate business intelligence systems.

3.4.4 Some institutions use such intelligence to identify the optimum journal for a researcher to publish in order to get the optimum impact both for them and for the institution.

3.4.5 The nature of business intelligence is that citation data on its own is of little use; it is one piece of data that can be incorporated into a larger range of data. Other sources will include bibliographic data, text-mining of academic outputs, funding and grants information, competitive analysis, league tables and their underlying models, and any other relevant sources of information.

3.4.6 The specific questions that business intelligence tools could be used to answer varies depending on the institution, and the usage of business intelligence systems within higher education institutions is currently limited. However, citation (and other bibliographic) data provide a convenient and self-contained perspective on research, and services exist to provide business intelligence views on this data without broader integration into institutional systems.

3.4.7 Publishers use citation data alongside other bibliographic and bibliometric data as key tools to understand the impact and markets for their publications. Some go further, analysing citation networks to support the process of selecting editors or reviewers for their publications.

3.4.8 The data used for business intelligence must be **contextualised** with bibliographic and other data, current, and usable within other systems.

**The current system**

3.4.9 Current systems vary widely between organisations, depending on organisational strategy and resources. Some organisations obtain basic citation data that they then process within their own
systems. This may include putting directly into the local Current Research Information System (CRIS) for a research institution, or a production system for a publisher. Other organisations subscribe to analytics services provided by the major citation data suppliers.

**Trends**

3.4.10 The major suppliers of citation data are investing strongly in business intelligence products, and a range of higher education institutions are using these tools to develop and understand their own, and competitors’ research profiles. This trend seems likely to continue, and may well expand to utilise metrics other than citation data. The extent to which this happens will be driven by the policies of government and research funders as well as by the attitudes of research leaders in higher education institutions.
4 Current citation data services

4.1 Introduction

4.1.1 A number of tools and services have been developed to support the use of citation data for a range of purposes. While these services all provide access to citation data and tools to support business intelligence, it is important to understand that their primary purpose is to provide search, abstracting and indexing services to allow academics to find articles and potential collaborators that are relevant to their research.

4.1.2 Citation data services differ significantly from one another. The differences may be related to price (subscription based or free services), or non-price factors such as the content covered, the tools available for analysis, the quality of the result, the limitations on use, or the type of infrastructure for aggregating citations.

4.1.3 What follows is an overview of current paid-for and free services that provide citation data at sub-section 4.2. CrossRef has an increasingly important role in the delivery of citation data services and is described at sub-section 4.3. While Mendeley does not have a current citation data offering, it is considered to have potential to develop in this way and is detailed at sub-section 4.4. We also discuss the important issue is whether the citation data market works well and is addressed at sub-section 4.5. Finally, sub-section 4.6 provides an application of the citation data lifecycle model (Annex A) by overlaying current providers of existing service elements onto the model.

4.1.4 The overviews and analysis are based on openly available information, that is, marketing materials supporting a particular service and academic comparisons of the services and, where possible, interviews with the services themselves. The services are continually evolving in scope and capability as part of the competition between them, so what is stated below is likely to be soon out of date.

4.2 Current services

Paid-for services

4.2.1 The key paid-for services are Elsevier’s SciVerse Scopus and Thomson Reuters Web of Knowledge. These are summarised at Table 4-1. Both claim to be the world’s largest citation data service and both argue that their coverage and selection process mean that results come from the highest quality journals. The price of these services was not raised as an issue by any of the stakeholders interviewed.

Free services

4.2.2 There are a number of free citation data services, including Google Scholar, Microsoft Academic Search, CiteSeerX and the JISC Open Citations Corpus.
# Access to Citation Data: Cost-benefit and Risk Review and Forward Look

## Paid-for services

<table>
<thead>
<tr>
<th>Provider</th>
<th>Elsevier</th>
<th>Thomson Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overview</strong></td>
<td>See (SciVerse Scopus, 2012b)</td>
<td>See (Thomson Reuters, 2012a)</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Europe</td>
<td>United States</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Subscription service</td>
<td>Subscription service</td>
</tr>
<tr>
<td><strong>Launched</strong></td>
<td>2004</td>
<td>Web of Knowledge: 2002, Web of Science: 2004, (as ISI from 1960s)</td>
</tr>
<tr>
<td><strong>What it is</strong></td>
<td>Citation index and database of peer-reviewed literature in English and 30 other languages</td>
<td>Academic citation index and database of peer-reviewed literature in English and 45 other languages</td>
</tr>
<tr>
<td><strong>Scope</strong></td>
<td>Scientific, technical, medical and social sciences, arts and humanities</td>
<td>The sciences, social sciences, arts and humanities</td>
</tr>
<tr>
<td></td>
<td>SciVerse ScienceDirect provides access to full-text science articles</td>
<td>It offers access to journal articles in the sciences, social sciences, arts and humanities</td>
</tr>
<tr>
<td><strong>Database includes (as at December 2012)</strong></td>
<td>– 20,500 peer-reviewed journals from 5,000 publishers worldwide</td>
<td>– 23,000 academic and scientific journals (including Web of Science journal listings)</td>
</tr>
<tr>
<td></td>
<td>– 1,800 open access journals</td>
<td>– 23 million patents</td>
</tr>
<tr>
<td></td>
<td>– 360 book series</td>
<td>– 110,000 conference proceedings</td>
</tr>
<tr>
<td></td>
<td>– 400 trade publications</td>
<td>– 9,000 websites</td>
</tr>
<tr>
<td></td>
<td>– 5.3 million conference papers</td>
<td>– 148,000 conference proceedings from 1990</td>
</tr>
<tr>
<td></td>
<td>– 49 million records with 26 million records after 1996 and 22 million records with references between 1823 and 1995</td>
<td>– Coverage from 1990</td>
</tr>
<tr>
<td></td>
<td>– Over 87 million source items</td>
<td>– Over 700 million cited references over 256 scientific disciplines</td>
</tr>
<tr>
<td><strong>Tools/features</strong></td>
<td>– Linking to full-text articles and other library resources</td>
<td>– Linking to full text articles</td>
</tr>
<tr>
<td></td>
<td>– Author profiler/ORCID</td>
<td>– Cited reference searching</td>
</tr>
<tr>
<td></td>
<td>– Identifier including h-index</td>
<td>– Author identification tools</td>
</tr>
<tr>
<td></td>
<td>– Citation Tracker</td>
<td>– Analysis tools</td>
</tr>
<tr>
<td></td>
<td>– Affiliation Identifier</td>
<td>– Key indicators: impact factor and h-index</td>
</tr>
<tr>
<td></td>
<td>– Journal Analyser</td>
<td>– Journal evaluation</td>
</tr>
<tr>
<td></td>
<td>– Alerts and RSS feeds</td>
<td>– Performance evaluation</td>
</tr>
<tr>
<td></td>
<td>– Interoperability with SciVerse ScienceDirect, Reaxys and ProQuest’s CSA Illumina</td>
<td>– Visualisation tools using Citation mapping</td>
</tr>
<tr>
<td></td>
<td>– Data export via bibliographic managers such as RefWorks, EndNote and BibTeX</td>
<td>– Bibliography management alerts and RSS feeds</td>
</tr>
<tr>
<td><strong>Selection process</strong></td>
<td>Content is reviewed by independent Scopus Content Selection and advisory board to avoid potential conflict of interest between Elsevier as an international publisher of scientific journals, and Scopus</td>
<td>Editors evaluate journals for coverage in Web of Science and cover the journal's basic publishing standards, its editorial content, the international diversity of its authorship, and associated citation data</td>
</tr>
<tr>
<td></td>
<td>Main selection categories: journal policy, content, journal standing, regularity and online availability</td>
<td>Commitment to “comprehensive” coverage but this does not necessarily mean all-inclusive (Garfield, 1990), (Thomson Reuters, 2012b)</td>
</tr>
<tr>
<td><strong>Business intelligence</strong></td>
<td>SciVal provides business intelligence capabilities for research institutions and funders</td>
<td>A range of evaluation tools, including InCites, Essential Science Indicators and Journal Citation Reports</td>
</tr>
<tr>
<td><strong>Usage</strong></td>
<td>Not revealed</td>
<td>20 million researchers from over 4,700 institutions in 98 countries use Web of Knowledge</td>
</tr>
<tr>
<td><strong>Potential issues</strong></td>
<td>Scopus contains fewer pre-1996 records than Web of Science</td>
<td>In a comparison of citation data services, (Falagas, et al., 2007) observe that for citation analysis, Scopus offers about 20% more coverage than Web of Science. The converse view is that Web of Science has a smaller number of higher quality journals</td>
</tr>
</tbody>
</table>

Table 4.1: Current paid-for data citation services
### Current Citation Data Services: Google Scholar and Microsoft Academic Search (Part 1 of 2)

<table>
<thead>
<tr>
<th>Provider</th>
<th>Google</th>
<th>Microsoft Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>See (Google, 2012)</td>
<td>See (Microsoft, 2012)</td>
</tr>
<tr>
<td>Location</td>
<td>Global</td>
<td>Global</td>
</tr>
<tr>
<td>Type</td>
<td>Free service</td>
<td>Free service</td>
</tr>
<tr>
<td>Launched</td>
<td>Released in beta in November 2004</td>
<td>Released in beta in 2009 replacing earlier service</td>
</tr>
<tr>
<td>What it is</td>
<td>Primary mission is to enable users, whoever and wherever they are, to find relevant research through the web. It is a freely accessible web search engine that indexes the full text of scholarly literature [15] using an entirely algorithmic approach [16]. This stated mission seems likely to continue</td>
<td>Freely available search service to allow scholars, scientists, students, and practitioners to find quickly and easily academic content, researchers, institutions, and activities</td>
</tr>
<tr>
<td></td>
<td>Allows users to search for digital or physical copies of articles, whether online or in libraries in English or any other language used by Google</td>
<td>Indexes academic papers but also surfaces key relationships between and among subjects, content, and authors in a manner that highlights critical links that help define research</td>
</tr>
<tr>
<td></td>
<td>Uses web crawler to access available publishers’ websites etc. Quid pro quo is traffic generated for publisher’s website</td>
<td>Content updated weekly</td>
</tr>
<tr>
<td>Scope</td>
<td>Coverage includes biology, life sciences and environmental sciences, business, administration, finance, economics, chemistry and materials science, engineering, pharmacology, veterinary science, social sciences, arts and humanities</td>
<td>Coverage includes 15 different disciplines and more than 200 sub-domains over science, technology, and medicine, the social sciences, and the humanities</td>
</tr>
<tr>
<td>Database includes</td>
<td>No detailed data is provided on coverage. It indexes “full-text journal articles, technical reports, preprints, theses, books, abstracts and other documents, including selected Web pages that are deemed to be scholarly”</td>
<td>As of December 2012, coverage included:</td>
</tr>
<tr>
<td></td>
<td>Google scholar’s coverage of US legal cases is extensive</td>
<td>– 48,774,208 publications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– 21,931,011 authors</td>
</tr>
<tr>
<td>Tools/features</td>
<td>Tools provided claimed to be as easy to use as a Google web search and includes an option which automatically narrows search results for a specific journal or article (Advanced search), gives a link to the journal article (Group of), gives cited-by links (Cited-by), and related articles (Related articles)</td>
<td>Besides search functionality, it provides information on all entities covered (Profile page) and the connection between scholars (Co-author path)</td>
</tr>
<tr>
<td></td>
<td>Instead of using one factor to rank results, Google Scholar ranks results with a combined ranking algorithm weighing the full text of each article, the author, the publication impact and how often the piece has been cited. There is a high weighting on citation counts and document titles</td>
<td>Visualisation tools cover co-authors (Co-author graph), researcher collaboration, cited and cited-by links (Citation graph), mentors and advisees (Genealogy graph) and maps of scholarly output within organisations (Academic map). Other tools analyse research trends by domain (Domain Trend), compare organisations (Organisation Comparison) and provide details such as event dates, locations, and calls for papers (Call for papers)</td>
</tr>
<tr>
<td>Selection process</td>
<td>Algorithmic</td>
<td>Works with publishers and other content providers to increase data coverage. Content providers include Elsevier and CrossRef on behalf of its members</td>
</tr>
</tbody>
</table>

---

15 Scholarly literature includes peer-reviewed articles together with books and other non-peer-reviewed materials.  
16 A particular issue concerns normalisation of citation data arising from different practices adopted by disciplines in the paper age (eg use of title or not, lack of standardised abbreviations, etc). While Google would be pleased to see the opening of citation data, there is still the normalisation problem. Google thus sees standardisation of citations as useful.
| Potential issues | Limited access by users to subscription publications (Falagas, et al., 2007). Lack of transparency on selection policy and coverage. Coverage varies by discipline compared to other databases. Metadata is sometimes inadequate and less often updated than for other services (Jacsó, 2012). No API (application programming interface software) exists to allow users to access Google Scholar data from other systems but Publish or Perish (Harzing, 2007) provides suitable analytics | Limited access by users to subscription publications (Falagas, et al., 2007) Coverage is not as comprehensive as the subscription services but is currently increasing rapidly (Jacsó, 2011) |

*Table 4.2: Current free data citation services: Google Scholar and Microsoft Academic Search (part 2 of 2)*
## Jisc Open Citations Data Corpus

<table>
<thead>
<tr>
<th>Free Services</th>
<th>Jisc Open Citations Data Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Provider</strong></td>
<td>The College of Information Sciences and Technology</td>
</tr>
<tr>
<td><strong>Overview</strong></td>
<td>See (CiteSeerX, 2012)</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Pennsylvania State University</td>
</tr>
<tr>
<td></td>
<td><a href="http://citeseerx.ist.psu.edu/index">http://citeseerx.ist.psu.edu/index</a></td>
</tr>
<tr>
<td><strong>Subscription type</strong></td>
<td>Free service</td>
</tr>
<tr>
<td><strong>Launched</strong></td>
<td>Released 2008, based on CiteSeer originally launched in 1997</td>
</tr>
<tr>
<td><strong>What it is</strong></td>
<td>A not-for-profit scientific literature digital library and search engine</td>
</tr>
<tr>
<td></td>
<td>Aims to improve the dissemination of scientific literature and provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency and timeliness in the access of scientific and scholarly knowledge</td>
</tr>
<tr>
<td></td>
<td>It harvests information using a focused web crawler (citeseerxbot)</td>
</tr>
<tr>
<td></td>
<td>It is often considered to be the first automated citation indexing system. It is also a predecessor of academic tools such as Google Scholar and Microsoft Academic Search; a past sponsor of CiteSeerX was Microsoft Research</td>
</tr>
<tr>
<td><strong>Scope</strong></td>
<td>Focused primarily on computer and information science</td>
</tr>
<tr>
<td><strong>The database includes</strong></td>
<td>Harvested from reference lists of all open access articles in PubMed Central</td>
</tr>
<tr>
<td><strong>Tools/ features</strong></td>
<td>Besides autonomous citation indexing and automatic metadata extraction, it provides citation statistics for all cited articles, links to references, author disambiguation, the citation context, automatic notification of new citations to given papers, and new papers matching a user profile as well as location of related documents and full text indexing</td>
</tr>
<tr>
<td></td>
<td>CiteseerX makes available resources such as algorithms, data, metadata, services, techniques, and software that can be used to promote other digital libraries</td>
</tr>
<tr>
<td><strong>Selection process</strong></td>
<td>CiteSeerX automatically harvests research papers from the public Web but also accepts submissions through a document submission system</td>
</tr>
<tr>
<td><strong>Potential issues</strong></td>
<td>Any material within PubMed Central that meets selection criteria</td>
</tr>
<tr>
<td></td>
<td>Current coverage is limited to biomedical sciences</td>
</tr>
<tr>
<td></td>
<td>Long-term sustainability depends on Jisc funding continuing</td>
</tr>
</tbody>
</table>

*Table 4.3: Current free data citation services: CiteSeerX and Jisc Open Data Citations Service*
4.3 CrossRef

4.3.1 Launched in early 2000, CrossRef (CrossRef, 2012b) is an independent, not-for-profit publishers’ collaboration that provides an infrastructure for linking citations across publishers. It does this through DOIs and is the only full-scale implementation to date. As of 5 November 2012, there were 56,887,098 registered CrossRef DOI links with 28,176 journals covered.

4.3.2 Participation in CrossRef comes through a variety of organisations. Any publisher of primary research material in digital format can register content by joining its membership. Libraries also use CrossRef to enrich online catalogues and databases with links to their own full-text holdings where appropriate. Furthermore, intermediaries, including secondary publishers and journal hosting services, use CrossRef as affiliates, enhancing their own products and content offerings with DOI-based citation links.

4.3.3 CrossRef works by assigning a DOI prefix to publishers of electronic scholarly content. For each content item publishers want to register, the system creates a unique DOI. CrossRef then registers each article’s DOI and URL in a central DOI directory. This allows for permanent inbound links to the content. Publishers also submit the citations contained in each deposited article to the Reference Resolver. This allows publishers to add outbound links to citations that are linked to content already registered in CrossRef. Publishers also submit necessary metadata for the scholarly content.

4.3.4 Pricing is based on multiple tiers with publishers, affiliates libraries and researchers paying differing amounts for the service. Publishers (CrossRef, 2012c) pay an annual fee based upon their publishing revenue (between $275 for a turnover below $1m and $50k for a turnover in excess of $500m) plus a transaction fee for each DOI deposit which is currently $0.20 per current item and $0.02 per back file item. Affiliates (CrossRef, 2012d) pay an annual administrative fee ($500 for a turnover below $1m and $10k for a turnover in excess of $10m) as well as a DOI retrieval fee. However, libraries and researchers can access the system for no charge.

4.3.5 Two key CrossRef services that support citation data services are CrossRef Metadata services and Cited-by Linking, which are described further below.

CrossRef metadata services

4.3.6 CrossRef metadata services offer a paid-for suite of tools for authorised partners to collect metadata from different publishers to streamline their own crawling, indexing and linking services. It includes a variety of metadata dissemination tools and provides standard terms and conditions for partners using the cross-publisher metadata. The service also removes the need for bilateral agreements covering use of metadata between publishers and partner service providers.

4.3.7 There are two forms of the CrossRef metadata service: basic and enhanced. The basic service is aimed at any non-publisher organisation linking to publisher full text through CrossRef DOIs and wanting to host the CrossRef data locally (CrossRef, 2012f).

4.3.8 Enhanced membership allows affiliates to access over 50 million metadata records registered in CrossRef via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface. This interface serves as the central point for the distribution of metadata from participating publishers, using a widely adopted technology targeted at consumers of large quantities of metadata. Access to the CrossRef metadata repository is controlled by Internet Protocol authentication and can be tailored to provide specific content from selected publishers to each authorised recipient (CrossRef, 2012g).
Cited-by Linking

4.3.9 The CrossRef Cited-by Linking (formerly Forward Linking) (CrossRef, 2012e) service is built on top of the DOI infrastructure and allows direct primary publisher-to-publisher linking. It allows a CrossRef member to find out how a publication is being cited. This information may only be displayed alongside primary articles that they publish. While CrossRef Cited-by Linking is a free service, members that use the service have to participate in the process by submitting metadata and references for the items they publish. Once a member is contributing, they can draw down data for their content. Publications include journals, articles, monographs, reference works, etc.

4.3.10 The usefulness of Cited-by Linking depends on the uptake of membership and although CrossRef staff actively promote Cited-by Linking to its membership through webinars and email reminders, participation in the service has been limited. This in turn means that the use of the ‘Cited By’ information varies across disciplines and publishers.

4.4 Mendeley

4.4.1 While not offering a current citation data service, Mendeley (Mendeley, 2013a) has the potential to take part in this area.

4.4.2 Mendeley is a reference manager and academic social network that helps users store and organise research materials, discover relevant research and collaborate with others online. The basic personal service is free and provides 1 GB of personal library space. Mendeley helps users by proving algorithms to extract bibliographic data from items under management and to allow annotation of the materials.

4.4.3 Mendeley makes available suitably anonymised bibliographic and metadata regarding the research materials in its database through an API. There is a thriving third-party app development market (Butcher, 2012).

4.4.4 Paid-for versions have a monthly subscription for which Mendeley provides further library space and collaboration tools for closed groups, depending on the plan selected. The Mendeley Institutional edition is an analytics tool built on top of Mendeley that facilitates collaboration and helps librarians, research directors and others understand the research activity and scholarship output of their community.

4.4.5 Mendeley holds some 340 million items of which around 75 million are unique with a significant majority being peer-reviewed. It has about 2 million unique logins with 20% of these being in the US and 10% in the UK.

4.4.6 Research demonstrates that analysis of the bibliographic, metadata and usage data for the research materials held in Mendeley provides information about the importance of an item (Bar-Ilan, 2012), (Li & Thelwall, 2012), (Li, et al., 2012), (Van Noorden, 2012). However, this is not the same as traditional citation counts. Thus, Mendeley has potential for offering an alternative metric of impact/importance of an article. However, until Mendeley (or a third-party) develops a clear set of tools for exploiting Mendeley data, it remains simply a potentially interesting alternative.

4.4.7 We also understand that Mendeley has considered exploiting the citation data within its corpus of research materials and has developed algorithms to help with this. It is understood that currently there are no plans for any citation data-based service.
4.4.8 During this project, Mendeley has been acquired by Elsevier (Mendeley, 2013b). It will be interesting to see how Mendeley will be integrated with Scopus.

4.5 Does the citation data market function well?

4.5.1 Although the system works well, it is instructive to consider whether it forms a well-functioning market. Well-functioning markets are essentially those that minimise price and maximise competition including innovation. A full evaluation of the market for citation data and detailed econometric analyses are not feasible within this project, but this sub-section will explore some of the key aspects. The framework adopted is the Competition Commission’s draft Guidelines for Market Investigations (Competition Commission, 2012). The market considered is the supply of indexed citation data to end users in the UK. Each of the following sub-sections briefly considers one of the “theories of harm”, and whether they may be active in this marketplace.

4.5.2 A key challenge in conducting this analysis is separating the UK market for citation data from the global market for A&I services more generally. This analysis focuses on the former, but must be considered within the larger picture.

**Weak rivalry**

4.5.3 There do not appear to be supply-side or capacity constraints; the products available are differentiated but still broadly substitutable and competitive.

4.5.4 The most significant potential factor is the small number of providers of indexed citation data – essentially four major providers (Scopus, Web of Knowledge, Microsoft Academic Search and Google Scholar). Scopus and Web of Knowledge both have high market share, and the majority of institutions that subscribe to either actually subscribe to both. This leads to a potential for excess concentration of market power in these providers.

4.5.5 It is challenging to consider the degree of competition between these providers. Both charged-for services have a high market share, and anecdotal evidence suggests that Google Scholar is widely used by end-users, while Microsoft Academic Search is growing in use. Both of the charged-for services are licensed through Jisc Collections, which potentially limits competition on price, and should ensure pricing and service transparency.

4.5.6 We can investigate the market power of these suppliers by looking for indicators that each is exploiting their high market share. If this were the case, we might expect to see a lack of innovation, predatory pricing, or unreasonable negotiations with suppliers. But stakeholder views suggest that there are no significant concerns about any of these aspects.

4.5.7 There is one exception. Several publishers raised concerns about the approach Google took to negotiating access to the publishers’ content. This is not an aspect of the citation data market, but one of web search. Google appears to have been able to exploit its dominance of the web search market to gain access to information that is necessary for Google Scholar, and so allow it to function under favourable conditions. Innovation appears high in the market, with all providers working to expand coverage, develop new tools and metrics, and better meet the needs of their customers.
Restrictions on entry and expansion

4.5.8 The cost of entry to the market for providing indexed citation data is moderate to high. These costs primarily reflect the effort required for negotiating with publishers for their content, and the technological infrastructure required for processing data on the scale necessary. It is not possible to create a niche database of citation data – meaning that new entrants have to compete on a large scale from the outset. This investment will be to a large degree ‘sunk’ – an investment that could not be recovered even if the original investor chooses to leave the market.

4.5.9 There are strategic advantages to the incumbents in the indexed citation data market. There is already sunk investment in R&D, and they can also benefit from existing reputation, relationships and technologies. There is no evidence for more aggressive exploitation of the incumbency, for example long-term contracts with customers or interference with cost structures of competitors.

4.5.10 The costs for customers switching between providers are low, but there may be some indirect effects caused by differing coverage between providers for example, calculation of a metric may yield a different answer if different publications are included within the scope of the search.

4.5.11 Although there are real barriers to entry into this market, three of the four major players have entered in the last decade. This demonstrates that entry is feasible for those organisations that can operate at the scale required – this scale is an intrinsic factor of the product, rather than the market that exists to provide it.

Coordinated conduct

4.5.12 The market for indexed citation data is concentrated and transparent. There are relatively few products available, and their capabilities and price points are well known by competitors. This situation can lead to tacit coordination between providers that can harm competition.

4.5.13 On the other hand, several aspects tend to inhibit coordination in this market. The economic situation for these services is unstable – neither Google Scholar nor Microsoft Academic Search charge fees to their users. This represents a radically different business model, which challenges and may disrupt the charged services. Furthermore, the products available are differentiated and evolving in their coverage and capabilities. Although customers for citation data may be easily segmented in a number of ways, all the providers appear to be vigorously competing in all segments.

4.5.14 In summary, despite some features that may promote coordinated conduct, others inhibit it. There appears to be no evidence of coordination in practice.

Vertical relationships

4.5.15 Up to the point of creating Indexed Citation Data, there is a limited degree of integration between the providers of indexed citation data, in that Elsevier is a major publisher, as well as the provider of Scopus. However, the nature of the market for indexed citation data is that it is necessary to work with all major publishers. Consequently, internal integration may bring some benefits, but these are probably not significant. It is impossible to determine whether Elsevier would or does restrict access to the data from their publications in order to support the Scopus service. Elsevier data is available in a wide range of A&I services, including the other major providers of indexed citation data, but we are not privy to the conditions attached to these supplies of data.
4.5.16 It is of greater potential concern that the suppliers of indexed citation data can control what their users can do with the service they receive. All the suppliers’ offerings have integrated collation and analysis, so they intend customers to use the tools that they provide to conduct analysis of the data, although all the suppliers also enable their clients to integrate the raw data directly into their own systems. There are restrictions on what customers can do with any tools they generate to process indexed citation data. For Web of Knowledge and Scopus, these tools could be in competition with the analysis tools provided by the suppliers themselves. Elsevier permit the creation of chargeable ‘apps’ that use their data, whereas Thomson does not; use of their data is limited to internal systems. This inhibits competition in the potential secondary market which would involve the resale of enhanced indexed citation data from these providers, but not in tools to analyse the data.

4.5.17 No stakeholders raised concerns about the providers attempting to form arrangements that would reduce competition, such as exclusive supply or purchasing, tying and bundling, pricing arrangements and so on). This probably reflects that the requirements of users are being satisfied since the current uses of citation data are internal, and that users don’t feel the absence of a secondary market. Nonetheless, this so-called ‘foreclosure effect’ potentially limits innovation in the analysis of indexed citation data. Detailed analysis of whether this is significant is outside the scope of this report.

Weak customer response

4.5.18 There is good information available regarding the different indexed citation data products on the market, and switching costs are low.

4.5.19 The factor most likely to weaken customer response arises when a research evaluator such as a national body or recruitment panel decides to base their evaluations on a single source of citation data. Research organisations that are evaluated using this data may feel that it is necessary for them to have access to the same data. This risks distorting the market for indexed citation data. The evaluator is a customer within the market, and so long as they conduct their procurement in a transparent and fair manner, this distortion is probably unavoidable.

4.5.20 The study team has heard some anecdotal evidence of institutions that prefer to use a different supplier of indexed citation data from their evaluators, as they prefer the analytical tools provided by their own supplier. Those institutions are using different data to make decisions that will affect evaluation – suggesting that they have judged the data sources to be interchangeable, and the choice of supplier is made on other factors to the research evaluators’ factors.

Market assessment

4.5.21 Based on the discussion above, the study team concludes that the market for the supply of indexed citation data to end users in the UK is functioning well.

4.6 Application of the citation data lifecycle model

4.6.1 As an illustration of the citation data lifecycle model at Annex A, existing service providers are overlaid onto this model as shown in Figure 4-1.
Access to Citation Data: Cost-benefit and Risk Review and Forward Look

Figure 4.1: Mapping of existing services to model [17]

For meaning of the colours, see Table A.1.
5 Future models

5.1 Introduction

5.1.1 The study has examined the strengths and weaknesses of the existing citation data ecosystem, and investigated how the ecosystem may evolve in future by identifying and assessing alternative approaches. There are already a number of different processes by which citations are passed from their authors to users, and the range of possible processes is likely to grow. The details of the analysis are provided at Annex B and summarised in this section.

5.1.2 The approach is based on the citation data lifecycle model set out at Annex A and comprises two stages:

- Firstly, identifying and assessing alternative approaches to each process element set out at Table A-1 (sub-section 5.2);
- Secondly, identifying and assessing the future business model by investigating alternative end-to-end options (sub-section 5.3);
- In practice, it is likely that for the foreseeable future, these routes will all operate in parallel.

5.1.3 This analysis of citation data covers recognised academic papers or books, which may or may not be published through the typical process. Broader issues of research evaluation, publishing in popular rather than academic, or trade literature etc, are out of scope.

5.2 Alternative approaches to process elements

5.2.1 Table 5-1 summarises our investigation of the process element options. We only address the current situation for ‘analyse and exploit’ process elements, as this is out of scope for this study.

5.2.2 A detailed analysis of the options is provided at Annexes B.2-B.5. The study team’s opinion is that with the exception of the federated search option for collate and index, all of the options are potentially viable. Moreover, it is considered likely that they are not mutually exclusive, so that different instances of these options will exist simultaneously, as is already the case for the various publication options.
5.3 Analysis of future business model

5.3.1 This sub-section examines ‘end-to-end’ models of how citation data can be created, processed, and made available to users. These options represent three realistic and interesting cases, which are summarised at Table 5-2 in terms of the process elements from Table 5-1 [18]:

- The current model
- An open model, where article metadata, citation data and abstracts are freely available from each publisher
- A central, open distribution model, where article bibliographic metadata and citation data are freely available from an open centralised database. One possibility is that CrossRef might undertake this role, building on the data it already holds. This option is intermediate between the current and open models

<table>
<thead>
<tr>
<th>Process element</th>
<th>Current model</th>
<th>Open model</th>
<th>Central, open distribute model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
<td>Existing processes</td>
<td>Existing processes</td>
<td>Existing processes</td>
</tr>
<tr>
<td>Publish</td>
<td>Traditional publication</td>
<td>Non-traditional publications</td>
<td>Traditional publication</td>
</tr>
<tr>
<td>Distribute</td>
<td>Traditional distribution (including Gold OA)</td>
<td>Open metadata and citations</td>
<td>Traditional distribution (including Gold OA)</td>
</tr>
<tr>
<td>Collate &amp; Index</td>
<td>Existing systems</td>
<td>A range of providers may offer centralised or distributed options</td>
<td>Open centralised database</td>
</tr>
<tr>
<td>Analyse &amp; Exploit</td>
<td>See Usage scenarios (Section 3)</td>
<td>See Usage scenarios (Section 3)</td>
<td>See Usage scenarios (Section 3)</td>
</tr>
</tbody>
</table>

Table 5-2: Comparison of end-to-end options considered

18 Not all process element options have been included (e.g. create: structured citation capture during authorship is considered to provide only a variation of these three options and the federated search option is technically infeasible).

19 Alternatives are out of scope. The analysis and exploitation of citation data depend strongly on the requirements of the specific user. The usage scenarios considered are discussed in Section 3.
Overall assessment

5.3.2 These end-to-end options are described and analysed in detail at Annex B.6. The overall assessment is that these options are all viable. In practice, the situation is likely to be complex with these options existing in parallel and interacting. The options described are not deterministic. The study team does not believe that any of these options is a likely or preferable future environment.
6 Conclusions and way forward

6.1 Context

6.1.1 Over the last decade or so, the principle of open access (OA) has become broadly accepted in a range of contexts including public sector data (HM Government, 2012) and the outputs of publicly funded research – both research data itself (RCUK, 2012a), (The Royal Society, 2012) and the publications created from the research (RCUK, 2012b). Two major studies (Houghton, et al., 2009) and (Finch, 2012) have demonstrated the economic benefits of OA, and driven UK government policy (BIS, 2012), respectively. Similar policy shifts are occurring within the European Commission Framework Programme (European Commission, 2012) and in most developed nations.

6.1.2 The funding changes in higher education in England and the continuing downward pressure on university finances are driving the trend for increased use of citation data for performance management and business intelligence, for example by improving their standing using citation data to provide business intelligence, identifying appropriate journals for a department's academics to publish in, looking at comparative impact of different departments and institutions.

6.1.3 These factors suggest that it is a good time to take a strategic look at how the creation, dissemination and exploitation of citation data might be enhanced and what Jisc and the UK academic community might be able to do to help this process.

6.2 Overview of the current situation

How citation data is used

6.2.1 The key usage scenarios for citation data require comprehensive data. This means:
  – With the right coverage of works
  – With the right time span – both current and historical citations

6.2.2 In some cases, the ‘right’ coverage is very limited. When conducting research in some fields for example all citations may be generated within a small number of publications. In other cases, citations may be created in an extremely wide range of publications.

6.2.3 The ‘right’ coverage is also influenced by the quality of publications. The major A&I service providers are selective, and include a process to assess the quality of journals and more recently, books before including them.

How it works

6.2.4 Citation data flows through a range of processes from creation to use. The key process involve publishers transferring data on their content to A&I providers, in bulk. The A&I providers extract data, including citation data from the publications, and then index this data to make it useful.
6.2.5 Primarily due to the historical development of academic publishing in different domains, there are extensive variations in approaches to citation, and the ways that citations are described and documented. Most of the effort invested in the lifecycle, and the value generated from it, lies in the collating and indexing process, which involves publications on a global scale.

6.2.6 Research evaluation based on citation data is bound to the selection of journals that the citation data provider used to calculate performance. There are much broader questions around the concept of research outputs and the evaluation of research involved here. Although these are out of scope for this study, it is necessary to recognise the relationship and indeed, arguably, the tension, between the selection made and the quality of research evaluation.

**Current services**

6.2.7 There are two well-established paid-for citation data services (Elsevier’s Scopus and Thomson Reuters’ Web of Knowledge) and two relatively new free services (Google Scholar and Microsoft Academic Search). Google Scholar uses an entirely algorithmic approach and Microsoft Academic Search makes more extensive use of IT than the paid-for services to reduce the costs of acquisition of citation data. However, the resulting quality in terms of accuracy, scope of sources and transparency is perceived by some to be inadequate for many usage scenarios.

6.2.8 There are global community initiatives eg CrossRef driven by publishers, and ORCID driven by institutions and publishers), which are successful in improving the consistency and accuracy of citation data and in reducing costs to citation data service providers. CrossRef provides a single database for bibliographic information and some citation information, and ORCID aids disambiguation of author identity.

6.3 Current approach

**Assessment**

6.3.1 The market for the provision of indexed citation data in the UK works well (sub-section 4.5). A range of providers compete in different parts of the market for citation data and A&I services more broadly, ranging from expensive high-quality data that is the product of significant manual and computerised intervention to free-of-charge, automatically processed data. All the usage scenarios identified in Section 3 are satisfied by the existing systems. Moreover, the feedback that the study team received from stakeholders indicates, despite some niggles, a high level of satisfaction with the current market offerings of citation data.

6.3.2 Moreover, there were few complaints [20] from stakeholders regarding the cost of accessing the commercial services. There appears to be a general view that these service providers add value to the research outputs by indexing and analysing them, and that the price of these services is accepted as appropriate for the quality of their offerings.

6.3.3 Publishers see some value in the citation data in their publications, but must balance any revenue from the sale of their data against the benefits of enabling discovery of their content. The publishers that

---

20 The cost of one of the services for a major higher education library was around 1% of the journals’ budget. The exceptions were regarding small and medium enterprises’ access to paid-for citation data services.
spoke to the study team tended to value discoverability more highly than the need to enhance revenue streams by selling citation data. This leads to the possibility that publishers may agree to make their citation data openly available – indeed, some already do.

The potential for open citation data

6.3.4 Citation data is valuable, and it is logical to assume that the same arguments apply to opening citation data as to opening primary research data and publications. Namely, the more users that have access to it, the more benefit is generated within the society that has also paid for the research. This applies especially in the context of the ‘conducting research’ usage scenario (sub-section 3.2).

6.3.5 This study considers that this argument is valid, but is much weaker than that for access to the original research outputs: it seems inappropriate to attribute the same level of benefits to its openness as to those that accrue from openness of the data itself. Moreover, indexed citation data—the kind that is applicable to realistic usage scenarios—is not created by the authors or the publishers of research outputs, nor is it paid for directly by the research funders. The added value of citation data is mainly created by the organisations that collate and index the data across many countries, publishers and disciplines. This value is only accessible after this processing, rather than from the article citation data itself.

Challenges of OA for citation data

6.3.6 A key consideration for the future of the citation data ecosystem is the impact that OA publishing will have. Will widespread access to the outputs of research enable a broader range of re-use of this data for citation analysis? Undoubtedly – but in practice there are significant limitations to how the open model can work for citation data, the key examples being:

- **The scale required**: for the research usage scenario, and for many important metrics (eg h-index) it is necessary to have access to essentially all publications including those published in the past which are not currently, and are unlikely to become, open. This is different from the position for research outputs, where the value of a single item is higher. Essentially, the network effect for citation data is much stronger and in many cases, it is only useful when you have most of it.

- **The cost and complexity of collation and indexing**: in order to meet the requirement for scale, the collation and indexing activities are intensive processes that demand high inputs of effort and technical infrastructure

6.3.7 Over time, increasing compute and storage capabilities should reduce the cost and complexity, but only to a point. Perhaps more importantly, the passage of time will reduce the importance of older, non-OA publications – but this process is likely to require several decades from the point when most publications are open, that is, the duration of a career.
6.4 Future systems

Background

6.4.1 There are significant challenges in influencing any significant changes to the systems that create, process, and analyse citation data. The key inhibitors of change are:

- The level of difficulty in extracting and interpreting citation data from original publications
- The high cost of entry in developing systems of sufficient scale to process sufficient data to be useful to stakeholders
- The challenges in accessing, collating and indexing citation data retrospectively
- The market for the supply of indexed citation data to end users in the UK is functioning well (subsection 4.5)
- General although not total satisfaction with the current systems
- The global nature of citations and the fact that the UK represents just a small part of the citation landscape.

6.4.2 As discussed above, the move to open access publishing is likely to alter the way that citation data systems work. As more data are available openly, it is likely that some new entrants will enter the system. However, these are more likely, at least initially, to be subject-specific A&I providers, rather than those offering the kinds of general and extensive coverage necessary to satisfy the majority of usage scenarios for citation data.

The argument against radical intervention

6.4.3 The study team has examined future business models (sub-section 5.3) that have different levels of openness for citation data:

- The current model
- An open model, where article metadata, citation data and abstracts are freely available from each publisher
- A central, open distribution model, where article bibliographic metadata and citation data are freely available from an open centralised database; one possibility is that CrossRef might undertake this role, building on the data already held. This option is intermediate between the current and open models

6.4.4 The usage scenarios for citation data strongly inhibit the development of an open model for citation data. This is due to the sheer number of data sources that would need to be included from several thousand publishers with tens of thousands of publications, going back many years. Even if every publisher provided citation data openly, it would still be necessary in practice to create a centralised store of this data to enable effective processing, with all the costs that would be associated with such an undertaking. A central, open distribution model would ease the collection challenge, but would still require extensive investment to process the data.
Based on the overall positive assessment of the current system (paragraph 6.3.1) and the relatively weak openness argument, we conclude that an imposed radical intervention in the indexed citation data market would be difficult to justify. In the study team's opinion it would be better for Jisc to continue with its emphasis on open publication and open access and make limited interventions to help the current system to work more effectively.

**What could improve the system?**

In order to improve the system, the study team suggests Jisc adopts policy and a small number of interventions concerned with improving data flows and users' understanding. None of these interventions are simple, and there is currently no single entity or grouping that could take responsibility for implementing them.

The possible interventions considered are:

- Supporting moves toward a uniform citation syntax
- Encouraging widespread adoption of ORCID
- Improving data users’ understanding of the strengths and weaknesses of citation metrics
- Improving citation data extraction, collation and indexing workflows

**Uniform citation syntax**

If journals adopted uniform syntax for citations, the complexities in processing publications would be significantly reduced. Many thousands of different styles are used, and parsing them is not easy. It may prove infeasible to agree a single syntax for all publications, but any efforts to reduce the number of styles in use would be beneficial. Publishers or publishers' associations, or learned societies, could most usefully lead any moves to achieve this.

No doubt at some point in the future, information about the context of the citation, whether critical, supportive or historical, will become available to those evaluating citation data. It would therefore be useful if the developers of word processing and bibliographic management applications provided an agreed standardised method, acceptable to publishers, for enriching the metadata associated with citations. This might mean that when an author inserts a citation into a draft publication, the context could be indicated in a quick and simple way. Lists of citation contexts have been published, and it would not be difficult to reach a consensus on a list that covers all possibilities, while not being too lengthy or complex to understand or implement.

**Widespread adoption of ORCID**

ORCID represents a significant opportunity for more effective processing of current publications, and some historical publications. One of the key challenges is to identify authors uniquely while indexing publications. ORCID will provide high quality data to support this assignment, if it is adopted widely, and if users choose to share their ORCID data. In the long term, this will reduce the importance of the work that the A&I providers undertake on citation data. It is pleasing that Jisc has expressed its support for ORCID, and the study team recommends that Jisc convert this support into active engagement with, and commitment to, the further development of the service.
Improved understanding of strengths and weaknesses of citation metrics for citation data users

6.4.11 Many stakeholders expressed concerns that users of citation data, in particular at the institutional level, are making inappropriate choices based on citation data. There is no question that much of the scepticism about the use of citation data by management, stems from the reliance by senior managers in higher education institutions on simple metrics such as the impact factor or h-index. There is particular concern about the way that journal impact factors are often used as a criterion in where to publish material. Whereas the major citation data providers publish a range of material regarding the use of such data, it is clear that some organisations are applying metrics based on the information that is easily available, rather than that which truly reflect the quality of research, publications, or even institutions.

6.4.12 The study team recommends that Jisc commissions research to demonstrate the risks involved in over-reliance on impact factors. It also recommends that Jisc uses this research to issue guidance explaining best practice for using the Impact factor in conjunction with other information for policy-making regarding where citation information should be published.

Improved citation data extraction, collation and indexing workflows

6.4.13 There is evidence that the citation data workflow could be more efficient and therefore more economic in the way that the data is extracted, collated and indexed. For example, there is still a significant element of paper being manually processed by the major indexed providers for citation extraction. This cannot be the most efficient, effective and economic way of doing this. Representative publisher bodies should develop and promulgate the business case for moving to an essentially all digital but lightweight citation data workflow, and help their members implement this.

6.5 Next steps

Uniform citation syntax

6.5.1 Jisc, publishers and citation data service providers should use their best endeavours to encourage stakeholders such as publishers and citation data service providers to develop, agree and adopt a standardised syntax for citations, or approach to embedding citation metadata in authors’ drafts. Possible strands of activity for Jisc are:

- Based on previously published lists, investigate possible approaches to uniform citation syntax, in conjunction with stakeholders
- Develop a Jisc policy for encouraging the move to a uniform citation syntax
- Encourage and support stakeholders to establish a body to take responsibility for and implement the move to the uniform citation syntax and citation embedding approach

6.5.2 It is estimated that the overall goal might take 3 – 5 years to be achieved. The suggested timescale is for Jisc to have completed the investigation of possible approaches and published its policy by end December 2014.
Widespread adoption of ORCID

6.5.3 Jisc should convert its support for ORCID into active engagement with, and commitment to, the further development of the service. This requires Jisc to develop and implement a plan:

- Identifying how best to do this within the context of its own programme and remit, including the core message to be communicated
- Engaging with other UK bodies (e.g., UUK, RCUK, RIN, RLUK, DCC etc) and with ORCID
- Developing a communications plan for its own staff and academics to inform them of Jisc’s policy

6.5.4 The suggested timescale is for the plan to be published by June 2013 and to be completed by the end of March 2014.

Improved understanding of strengths and weaknesses of citation metrics for citation data users

6.5.5 Jisc should investigate the risks of over-reliance on impact factors for evaluation of institutions, departments and individuals. The output should be in the form of best practice guidance for a range of stakeholders, including senior institutional managers.

6.5.6 The suggested timescale is for the research to be commissioned and completed by the end of December 2013.

Continue to engage with stakeholders

6.5.7 Jisc should actively engage with stakeholders in this area to ensure there is mutual understanding of Jisc’s and the stakeholders’ interests and concerns and to ensure that the organisation keeps abreast of the stakeholders’ strategies. The engagement needs to be continual and carried out at least annually. It should include discussion of how Jisc might help the process for example regarding the possibility that publishers may agree to make their citation data openly available. There is no timescale suggested for the completion of this activity.

Monitor the current situation

6.5.8 Jisc should actively monitor how the citation data infrastructure develops [21] and develop its overall policy to citation data, appropriately.

6.5.9 Such monitoring needs to be continual but with an assessment of trends and directions and an indication of policy implications being produced annually. There is no timescale suggested for completion of this activity.

---

21 For example, the takeover of Mendeley by Elsevier may lead to significant changes.
References

Available at: http://altmetrics.org/manifesto/
[Accessed 26 November 2012].

Bar-Ilan, J., 2012. JASIST@mendeley. [Online]
Available at: http://altmetrics.org/altmetrics12/bar-ilan/


Butcher, M., 2012. Mendeley's open API approach is on course to disrupt academic publishing. [Online]
Available at: http://techcrunch.com/2012/08/22/mendeleys-open-api-approach-is-on-course-to-disrupt-academic-publishing/

CiteSeerX, 2012. About Cite SeerX. [Online]
Available at: http://csxstatic.ist.psu.edu/about
[Accessed 20 November 2012].

Available at: http://www.competitioncommission.org.uk/assets/competitioncommission/docs/2012/consultations/market_guidlines_main_text.pdf
[Accessed 17 January 2013].

Available at: http://www.projectcounter.org/
[Accessed 14 December 2012].

CrossRef, 2012a. crossref.org. [Online]
Available at: http://www.crossref.org
[Accessed 12 December 2012].

Available at: http://www.crossref.org/01company/16fastfacts.html
[Accessed 8 November 2012].

CrossRef, 2012c. Publisher fees. [Online]
Available at: http://www.crossref.org/02publishers/20pub_fees.html
[Accessed 14 December 2012].

CrossRef, 2012d. Affiliate fees. [Online]
Available at: http://www.crossref.org/04intermediaries/34affiliate_fees.html
[Accessed 14 December 2012].

CrossRef, 2012e. Cited-by linking. [Online]
Available at: http://www.crossref.org/citedby/index.html
[Accessed 14 December 2012].

CrossRef, 2012f. Information for affiliates. [Online]
Available at: http://www.crossref.org/04intermediaries/index.html#basic_affiliate
[Accessed 4 December 2012].


Available at: http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier
[Accessed 22 November 2012].

Available at: http://www.info.sciverse.com/scopus
[Accessed 19th November 2012].

Available at: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
[Accessed 17 January 2013].

Thomson Reuters, 2012a. What it is. [Online]
Available at: http://wokinfo.com/about/whatitis/
[Accessed 19th November 2012].

Available at: http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process/
[Accessed 19th November 2012].

Thomson Reuters, 2012c. What is ResearcherID?. [Online]
Available at: http://www.researcherid.com/
[Accessed 19 November 2012].

Available at: http://technicalfoundations.ukoln.ac.uk/node/92
[Accessed 12 December 2012].

Van Noorden, R., 2012. What were the top papers of 2012 on social media? [Online]
Available at: http://blogs.nature.com/news/2012/12/what-were-the-top-papers-of-2012-on-social-media.html

Available at: http://en.wikipedia.org/wiki/CiteSeer
[Accessed 14 December 2012].
A Citation data lifecycle model

A.1 Introduction

A.1.1 The ecosystem for citation data is large and complex, so understanding and describing it is difficult. The approach used in this here is to build on the basic concepts of citation data (Section 2) to set out a simple lifecycle model for citation data services. The model comprises three views:

- Lifecycle process view (Annex A.2)
- Conceptual information view (Annex A.3)
- Added-value view (Annex A.4)

A.1.2 This model identifies the key features of these three areas. The actual business process may differ considerably in detail to take account of, for example, the approach to citations in particular subject domains or whether an article is available only as a printed item or in electronic form. Moreover, a consequence of the increasing use of electronic media and IT is that the detailed business process can change. For example, publishers may tag lists of citations using XML, rather than this work being done as part of the index process.

A.1.3 Sub-section 4.6 uses the citation data lifecycle model to describe and characterise the existing data citation services.

A.1.4 The Figures and Tables in the model use the same colour scheme to identify elements of the lifecycle process, information and value added views.

A.2 Lifecycle process view

A.2.1 The citation data lifecycle (Figure A-1) from creation of the original article with included citations to exploitation of citation data is described in terms of the process elements set out at Table A-1. Not all elements need be present, eg citation data might be collated and indexed but never analysed or exploited.

A.2.2 For each process step, the following factors need to be considered:

- The specific actors that undertake the process element eg the publish process might be undertaken by a publisher or the author or the author’s institution
- Each process element requires appropriate quality assurance and so quality assurance activities do not form a separate process element
- Potentially, any and each process element could add value to the data and thus provide a means of generating revenue. The financial aspects of the business model thus overlay the overall process model (Annex A.4)
Access to Citation Data: Cost-benefit and Risk Review and Forward Look

Figure A-1: Lifecycle view

<table>
<thead>
<tr>
<th>Process element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
<td>The work done by the ‘citing party’ – typically the author of a publication – to identify and reference the cited work</td>
</tr>
<tr>
<td>Publish</td>
<td>The work done to convert a document into a publication, which may include peer review. A publisher typically conducts this, but some documents may be self-published</td>
</tr>
<tr>
<td>Distribute</td>
<td>The distribution of publications to those who wish to use them. Most commonly, this is through the ‘normal’ publication channels – journals, reports, etc. Other channels such as institutional and subject repositories, and authors’ own websites are also available. This distribution may be physical or electronic, and the electronic form can consist of a range of different formats, most commonly PDF and XML. Publications (or just their citation data) may also be distributed through specific channels for collation, such as CrossRef, Microsoft Academic Search, Google Scholar, Scopus and Web of Knowledge.</td>
</tr>
<tr>
<td>Collate</td>
<td>The work done to bring together the citation data from the various publications within the scope</td>
</tr>
<tr>
<td>Index</td>
<td>The work done to process the raw citation data from publications, to produce useful information. This may include identifying citations within publications, extracting, interpreting and normalising their formats, and creating a data store than can then be used to process this data</td>
</tr>
<tr>
<td>Analyse</td>
<td>The work done to draw meaning from the normalised citation data. This may include identifying individual citations of a document, a publication, an author, an institution or any other criterion; generating statistics (eg h-index, g-index, impact factor), or much more complex analyses</td>
</tr>
<tr>
<td>Exploit</td>
<td>The work done to take the analysed information, and make business-relevant decisions</td>
</tr>
</tbody>
</table>

Table A-1: Process elements

For reference, Table A-2 describes the actors that participate in the citation data lifecycle model.
A.3 Conceptual information view

A.3.1 As a minimum, a citation can be defined by two pieces of information – a reference for the cited work, and a reference for the citing work. This definition is clear, but it has been the strong view of interviewees during this project that without further information about the objects involved, this data is of extremely limited use. Usable citation data might include bibliographic data, article abstracts, complete article content, etc. Typically, this data might be created by the author, or be derived from the published article in paper or electronic form by human or machine reading/scanning/processing, including use of text mining and text analytics. Although the bibliographic metadata is not the focus for this project, there are fundamental questions about the usefulness of citation data without it.

A.3.2 While, for example, the Open Citation Project (JISC, 2010) essentially only stores data on the citations themselves, most other citation data services are in fact Abstracting and Indexing (A&I) services [22] that index the text of articles comprehensively in most cases [23]. These latter services see citation data as merely one aspect of identifying and analysing research publications. To overcome these difficulties of what is included within citation data, the model is designed to be essentially independent of the details of the citation data.

A.3.3 In parallel to the citation data lifecycle model, Figure A-2 provides a conceptual information model, which illustrates the involvement of citation data at each stage of the lifecycle process view. It is important to note that the refinement and processing of data adds value and is the basis for the revenue stream view at Annex A.4.

---

22 Providers of A&I service includes citation data service providers such as Scopus, Web of Knowledge, Google Scholar and Microsoft Academic Search.

23 Section 4 includes an overview of the key citation data services: Microsoft Academic Search, Google Scholar, Elsevier’s SciVerse Scopus and SciVal, and Thomson Reuter’s Web of Knowledge.
Information element | Description |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Article data</td>
<td>Article data comprises the basic content together with direct cross references to other articles and implicit citation data. Implicit citation data includes, for example, the context of the citation whether it refers to previous work that the current article is built on or a piece of work that the article refutes. The article data will depend on many factors, including publication venue for example whether it is released by a publisher or self-published and the publisher’s own citation and metadata requirements</td>
</tr>
<tr>
<td>Publisher data</td>
<td>Publisher data comprises all types of data associated with publishing and distributing a submitted article. This includes bibliographic data, the article’s content and citation data for that article</td>
</tr>
<tr>
<td>Collated citation data</td>
<td>As a minimum, collated citation data comprises the results of collating citation data within the specified scope. It might also include related bibliographic data</td>
</tr>
<tr>
<td>Indexed citation data</td>
<td>Indexed citation data builds on collated data including, as a minimum, cited and citing articles in a normalised format and an index of citation data. It might also include links to related bibliographic data, and may have been indexed by such data, eg by identifying authors and institutions. The indexed citation data is the form of data that can be analysed to answer business-relevant queries</td>
</tr>
</tbody>
</table>

**A.4 Added-value view**

**A.4.1** This view describes where value is added as part of the citation data lifecycle. Adding value provides opportunities to generate revenue streams.

**Added-value**

**A.4.2** Table A-4 analyses where value might be added in the citation data lifecycle. This looks at actors adding value by undertaking specific activities and identifies to whom the added-value accrues benefit. The added value might be a cashable benefit (eg allow generation of a revenue stream) or non-cashable (eg...
improve an actor’s brand). As well as the added-value activities, general business activities such as developing service improvements and carrying out marketing, etc also help to create added-value. For simplicity, these are not shown.

<table>
<thead>
<tr>
<th>Process element</th>
<th>ID</th>
<th>Actor</th>
<th>Value added by</th>
<th>Value accrues to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
<td>V1</td>
<td>Author</td>
<td>Creating citations in an article, book, etc</td>
<td>Author</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>Author</td>
<td>Checking validity of citation</td>
<td></td>
</tr>
<tr>
<td>Publish</td>
<td>V3</td>
<td>Publisher</td>
<td>Processing draft articles and organising peer review, including checking validity and format of input citations, etc</td>
<td>Publisher</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>Peer reviewer</td>
<td>Undertaking peer review to achieve article and journal quality, including checking validity of input citations</td>
<td>Peer reviewer</td>
</tr>
<tr>
<td>Distribute</td>
<td>V5</td>
<td>Publisher</td>
<td>Distributing published materials, including citation data</td>
<td>Publisher</td>
</tr>
<tr>
<td></td>
<td>V6</td>
<td>Intermediary</td>
<td>Common services for distribution of citations (eg CrossRef, ORCID)</td>
<td>Intermediary</td>
</tr>
<tr>
<td>Collate</td>
<td>V7</td>
<td>Citation data service provider</td>
<td>Selecting journals, books etc for use in collation</td>
<td>Author, Publisher</td>
</tr>
<tr>
<td></td>
<td>V8</td>
<td>Aggregator</td>
<td>Collating raw citation data for a group of publishers, including digitisation, re-keying, etc</td>
<td>Publisher</td>
</tr>
<tr>
<td>Index</td>
<td>V9</td>
<td>Citation data service provider</td>
<td>Indexing of collated citation data, Including assigning works to authors, institutions, etc</td>
<td>Citation data service provider</td>
</tr>
<tr>
<td>Analyse</td>
<td>V10</td>
<td>Citation data service provider</td>
<td>Providing citation data for use by others (eg other A&amp;I services)</td>
<td>Tool provider</td>
</tr>
<tr>
<td></td>
<td>V11</td>
<td>Tool provider</td>
<td>Creating and enhancing tools to provide metrics</td>
<td></td>
</tr>
</tbody>
</table>

Table A-4: Where added-value is created in the citation data lifecycle

(✓ - cross-impact, ➔ - enables later stages in lifecycle, ● - end benefit)

The majority of the effort invested, and the value generated, in the citation data lifecycle relates to collating and indexing citations from publications on an international scale.
Revenue streams

A.4.4 Potential revenue streams are identified in Table A-5, which cross-references the added-value activities described in Table A-4. The revenue streams are focused on those related to citation data. The potential revenue streams and beneficiaries across the citation data lifecycle are illustrated in Figure A-3. Not all of these are currently used by every beneficiary: for example, not every publisher charges for their raw citation data, arguing that they receive a benefit in kind by increased website traffic. Each revenue stream corresponds to activities that add value.

A.4.5 Possible beneficiaries have been identified by function eg author, publisher, institution, peer reviewer, aggregator; intermediary eg CrossRef or ORCID; citation data service provider; business intelligence provider; funder; research council and researchers. It is important to note that some of these may be different facets of the same organisation. For example, a major publisher may have the roles of publisher, aggregator, citation data service provider and business intelligence provider both internally and externally. Similarly, some of the functions may be carried out in-house by an institution eg business intelligence provider.
### Table A-5: Possible revenue streams

<table>
<thead>
<tr>
<th>Process element</th>
<th>Revenue stream</th>
<th>Description</th>
<th>Added-value stream/comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
<td>S1</td>
<td>Author: Revenue stream from fees received (eg incentive fees for creating/validating citations, etc)</td>
<td>V1, V2/ Not done in current peer review model</td>
</tr>
<tr>
<td>Publish</td>
<td>S2</td>
<td>Publisher: Fees received for publishing an article (eg from articles published under Gold Open Access)</td>
<td>V3, V7/ Premium from quality assured citations, etc</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>Peer reviewer: Incentive fees received for checking validity of citations as part of carrying out peer review</td>
<td>V4/ Not done in current peer review model</td>
</tr>
<tr>
<td>Distribute</td>
<td>S4</td>
<td>Publisher: Additional fees received for access to journals, payments for books, etc either on a one-off or subscription basis arising from availability of citation data</td>
<td>V5/ Significant fees can be generated by publisher website visits driven by academic searches and citation data</td>
</tr>
<tr>
<td></td>
<td>S5</td>
<td>Intermediary: Fees for use of common citation services eg CrossRef Cited By</td>
<td>V6/ Membership fees and usage fees</td>
</tr>
<tr>
<td>Collate</td>
<td>S6</td>
<td>Aggregator: Fees received from citation data service provider for collated citation data from a group of publishers</td>
<td>V8/</td>
</tr>
<tr>
<td>Index</td>
<td>S7</td>
<td>Citation data service provider: Fees received for access to citation data either on a one-off or subscription basis</td>
<td>V7, V9, V10/</td>
</tr>
<tr>
<td>Analyse</td>
<td>S8</td>
<td>Tool provider: Fees received from institutions, etc to use provided metrics eg h-index</td>
<td>V11/ Tool providers include citation data service providers</td>
</tr>
<tr>
<td>Exploit</td>
<td>S9</td>
<td>Citation data service provider: fees obtained from institutions for specialist advice, value-added services and reports, and support in decision making</td>
<td>Realisation of added-value</td>
</tr>
<tr>
<td></td>
<td>S10</td>
<td>Business intelligence provider: fees obtained from institutions for specialist advice and support in decision making</td>
<td>Realisation of added-value</td>
</tr>
<tr>
<td>Systemic</td>
<td></td>
<td>Author: increased grant income, salary, or promotion due to recognition of work</td>
<td>These revenue streams are indirect, and are enabled by the system functioning as a whole – they are so-called emergent properties</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Publisher: Income received from increased web-site traffic for providing collated citation data for that publisher</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Publisher: income from increased subscriptions or article processing charges due to increased prestige of journal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>User, which may include publishers and authors: increased ability to fulfil the usage scenarios relevant to them</td>
<td></td>
</tr>
</tbody>
</table>
Access to Citation Data: Cost-benefit and Risk Review and Forward Look

Figure A-3: Revenue streams view
B  Options and assessment for future data citation models

B.1 Introduction

B.1.1 This annex assesses the strengths and weaknesses of the existing citation data ecosystem, and investigates ways it may evolve in the future.

Approach

B.1.2 The ecosystem is large and complex, so the approach adopted is to investigate options for each of the process elements within the citation data lifecycle model (Table A.1). Possible end-to-end models are then constructed from the process element options and analysed (Annex B.6).

B.1.3 It is important to understand that the current ecosystem is heterogeneous with process elements being carried out differently by different parts of the community. The processes change as competitors seek commercial advantage and the balance moves from paper to electronic media. It is very likely for the foreseeable future that all the end-to-end models described below would operate in parallel with all the current systems. For simplicity, the analysis is carried out as if only one option is present.

Process element options

B.1.4 Table B.1 summarises the process element options investigated. Options for collate and index are addressed together here, as in practice it is not sensible to collate data without indexing it. Only the current situation is addressed for analyse and exploit process elements, as this is out of scope for this study.

B.2 Create

B.2.1 The create process element covers the work done by the ‘citing party’ – typically the author of a publication – to identify and reference the cited work.

Existing processes

B.2.2 Specific practices of how documents are drafted vary strongly between individuals, research groups, subject disciplines, institutions, and over time. This section describes what is considered as probably the most common approach.

B.2.3 Typically, a paper will be drafted using a word processor – Microsoft Word is common, others may use online services such as Google Docs, and others - particularly in engineering, computer science, and physical sciences - may use one of a range of LaTeX editors. Citations are entered into the document using the format and syntax required by the intended publisher – these are human-readable formats.
Options and assessment for future data citation models

<table>
<thead>
<tr>
<th>Process element</th>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Create</strong></td>
<td><strong>Existing processes</strong></td>
<td>The ‘citing party’ – typically the article author - identifies and references a cited work. Specific practices vary strongly between individuals, research groups, subject disciplines, institutions, and over time. Typically, a paper is drafted using a word processor (e.g. Microsoft Word, Google Docs, LaTeX editor particularly in engineering, computer and physical sciences) Citations are included in the document using the format and syntax of the intended publisher. Some authors use tools such as Endnote, Zotero, or Mendeley to manage citations. The citation data in the document is a human-readable bibliographic reference in text form, rather than as a structured unique identifier.</td>
</tr>
<tr>
<td><strong>Structured citation capture during authorship</strong></td>
<td></td>
<td>Tools could be used during the drafting process to capture structured citation data in the document, which could be passed on in a well-indexed format. This could be standalone or integrated with existing citation-management tools.</td>
</tr>
<tr>
<td><strong>Publish</strong></td>
<td><strong>Traditional publication</strong></td>
<td>This is the work done to convert a document into a publication, which may include peer review. This represents the majority of academic publishing in most domains at present. The draft prepared by the author is supplied to publisher, typically in Word/PostScript/RTF/LaTeX format. The draft is reviewed by the journal editor and reviewers and revised, as necessary. It is then copy edited and prepared for publication (typescripting, checking image formats, converting to XML and tagging, etc). A final version of the document is checked by the author, and then published.</td>
</tr>
<tr>
<td><strong>Non-traditional publications</strong></td>
<td></td>
<td>Alternative publishing models have been discussed for many years (D-Lib magazine, 2004) but are still evolving and currently represent only a small minority of academic publications. These may change the peer review concept with the related issue of how to value such citations. Perhaps the most widespread new approach in many subject domains is the establishment of repositories to share papers for that subject. These repositories usually contain a mix of pre-prints, drafts, ‘traditionally’ published work, and other types of information. In most cases, these are seen as secondary to the traditional publishing route, but for some they are an important primary route.</td>
</tr>
<tr>
<td><strong>Distribute</strong></td>
<td><strong>Traditional distribution</strong></td>
<td>This is the distribution of publications, or the citation data they contain to those who wish to use them. The publishers host PDF and HTML versions of the articles on their own platform which may be outsourced, and will distribute paper copies to those subscribers who want them. Publishers distribute article data to other organisations according to established relationships. Typical recipients include A&amp;I services (e.g Scopus and Web of Knowledge), Microsoft Academic Search and CrossRef.</td>
</tr>
<tr>
<td><strong>Gold Open Access (OA) (pay to publish)</strong></td>
<td></td>
<td>For fully-gold OA, the entire publication is freely available online. Any user may ‘pull’ the material to process as they wish. In practice, publishers of fully-gold publications typically also participate in the traditional distribution mechanism, by ‘pushing’ content to established partners. Hybrid publications may operate in a range of ways, but typically OA papers are treated the same as closed papers, except that they are available to all users of the publisher’s platform. Article content and metadata is transferred through the traditional distribution mechanism, to the same users. As far as the study team is aware, there are no hybrid journals that allow bulk downloads of just the open access portion of their content.</td>
</tr>
<tr>
<td><strong>Open metadata and citations</strong></td>
<td></td>
<td>This option considers the situation where the content is (or may be) closed, but the article metadata, reference list and abstract are openly and freely available for users to ‘pull’ the citations and metadata. Interactive querying through an API is unlikely to be feasible for realistic usage scenarios (see Federated Search below). A system could be constructed whereby citation data alone would be made open by the publishers, leaving users to obtain other metadata through another channel. Stakeholders interviewed considered this approach to be illogical and convoluted because of the need to aggregate both data sources, with attendant risks to data quality. This is not considered further here.</td>
</tr>
<tr>
<td><strong>Self-publication/self-archiving</strong></td>
<td></td>
<td>An alternative to Gold OA (pay-to-publish) is so-called Green OA – where an author can lodge papers in an appropriate repository (including subject and institutional repositories, but potentially simply on their own website). There is overlap with Non-traditional publications, in that many of those publications are in practice operated as repositories.</td>
</tr>
</tbody>
</table>

Table B-1: Summary of current and alternative process models for citation data (part 1 of 2)
Access to Citation Data: Cost-benefit and Risk Review and Forward Look

<table>
<thead>
<tr>
<th>Process element</th>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collate &amp; Index</td>
<td></td>
<td>This includes the work done to bring together the citation data from the various publications within the desired coverage, and the work done to process the raw citation data from publications to produce useful information. This may include identifying citations within publications, extracting, interpreting and normalising their formats, and creating a data store that can then be used to process this data.</td>
</tr>
<tr>
<td>Existing systems</td>
<td>A range of systems already exists for collation and indexing of citation data (Section 4). The key points to note are:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– All existing systems have established relationships with publishers, although some also take data from ‘open’ websites</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– All existing systems rely strongly on bibliographic data and do not treat citation data as separate from the article’s bibliographic data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– All existing systems bar the JISC Citations Data Open Data Corpus are designed to index and enable discovery of academic literature. They treat citation data as one piece of information about the literature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– The services which are provided on a charged basis (Scopus and Web of Knowledge) have selective coverage. Both take data from a range of distribution channels including around 35% in paper format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– Some publishers raise revenue by charging Elsevier for access to their data</td>
<td></td>
</tr>
</tbody>
</table>

The bulk of the effort undertaken by the existing services is in parsing and indexing citations that are provided in many thousands of different formats, and with varying quality. Use of ORCID should ease the challenges of accurately assigning works to authors but it will not help in assigning citations to works. Further, ORCID is user-centric, and is not designed to be retrospective: only active researchers are likely to register with ORCID.

Open centralised database

This option considers whether a central, shared data store such as CrossRef could be made open. The infrastructure already exists for publishers to make metadata and citation data available without charge – CrossRef already holds essentially all publication metadata, and a large proportion of current and recent citation data.

Federated search

It is possible to envisage a system that does not depend on a central coordinating body to collate citation data. Rather, “collation” would be conducted in real time by searching APIs provided by data holders – traditional and non-traditional publishers. This is known as a federated search. This option would make most sense if distribution is through an open metadata and citation data distribution channel, but in principle it would be possible to include open and closed sources within a federated search.

Analyse & Exploit

The analysis and exploitation of citation data depend strongly on the requirements of the specific user. The usage scenarios considered are discussed in Section 3.

Table B-1: Summary of current and alternative process models for citation data (part 2 of 2)

B.2.4 Some authors use tools such as Endnote, Zotero, or Mendeley to manage citations. These tools vary in scope, but can typically:

- Import information for a reference from a range of online databases, or allow the user to input information directly
- Store the metadata describing the reference in a database
- Insert a formatted reference into a word processor document
- Manage the creation of footnotes, endnotes, bibliographies, I

B.2.5 Regardless of whether these tools are used to create citations, the data describing these references is incorporated in the document as a bibliographic reference intended to be read by humans, rather than as a structured unique identifier.
Structured citation capture during authorship

B.2.6 It would be possible to use tools during the drafting process to capture structured information regarding the citations in the document. This could also be used to pass on this information in a well-indexed format. This could be standalone, or could be integrated with existing citation-management tools.

B.2.7 The exact capabilities of these tools would need to be developed, but could include:

- Using high-quality databases to allow authors to check bibliographic data (e.g. using CrossRef) when it is created, or in a separate quality assurance process. Existing tools can import data from a range of sources, but the study team is unaware of any tools that check existing data, in order to support accuracy. This is of course dependent on the user of the tool having access to the service that is providing the bibliographic data.

- Collecting extended information on the citations entered. As discussed above, traditional citation styles do not include any description of what the reference means. That information is included – implicitly or explicitly – in the content itself. The most practical point to collect this information from the author is during the writing process.

B.2.8 Authors of a work are the only actors in the system who actually know what work they intended to cite, and why they are citing it.

B.2.9 Table B-2 sets out the assessment for this option.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>If authors describe their citations in an unambiguous and easy-to-process format, it will aid processing later in the system. However, this will probably need very high levels of adoption to drive any real change at the system level.</td>
</tr>
<tr>
<td>Benefits</td>
<td>Unambiguous information requires less processing later in system. This will reduce the cost of obtaining citation data from article data. Higher accuracy in citations will increase overall trust in the system.</td>
</tr>
<tr>
<td>Complexity</td>
<td>Increased. Publishers, citation data providers and tool vendors will need to agree common formats and interchange standards for the description of citations.</td>
</tr>
<tr>
<td>Data</td>
<td>This option increases the amount of data available later in the system.</td>
</tr>
<tr>
<td>Disruption</td>
<td>Significant changes will be required to the authoring and publication workflows, as there is no simple way to embed extended citation information in the range of word processor documents that are in common use.</td>
</tr>
<tr>
<td>Financial</td>
<td>There would be a cost in developing and maintaining the tools to collect structured citation data during authorship.</td>
</tr>
<tr>
<td>Fragility</td>
<td>Tools which could increase citation data quality by checking against other data sources depend on the continuing availability of these data sources.</td>
</tr>
<tr>
<td>Technology</td>
<td>Tools for managing citation data already exist (e.g. EndNote, Mendeley, etc), but would require adaption to include database search as a quality assurance (QA) activity. To share extended and/or unambiguous citation data requires an agreed format for representing and transferring this information and interchange standards, which must be adopted by all manufacturers of tools, and all publisher involved in the system. This is likely to prove extremely challenging.</td>
</tr>
<tr>
<td>Trust</td>
<td>Potentially increased trust, through increased quality of the data entering the system.</td>
</tr>
</tbody>
</table>

Table B-2: Create: structured citation capture during authorship option assessment
B.3 Publish

B.3.1 This is the work done to convert a document into a publication, which may include peer review.

B.3.2 This sub-section does not consider the broader issues of research evaluation, publishing in popular rather than academic literature, trade literature etc. There are much broader questions around the concept of research outputs in play – but for the purposes of this analysis of citation data, outputs are only considered that could be recognised as academic papers or books, but may not be published through the typical process.

**Existing process – ‘traditional’ publication**

B.3.3 A range of workflows is used by different publishers. As a baseline the study considers the ‘traditional’ publication workflow, which represents the majority of academic publishing in most domains at present. This process is essentially the same for peer-reviewed open access journals (see Annex B.4).

B.3.4 The draft prepared by the author is supplied to publisher, typically in Word/PostScript/RTF/LaTeX format. Many publishers receive these files directly into a publication management system, but smaller publishers may still handle them manually. The author will provide some document-level metadata such as title, subtitle, authors, author affiliation, etc.

B.3.5 The journal editor and external peer reviewers will conduct a peer review process. The details of this process vary, but it is likely to include revisions to the work. The review process may identify new citations, and it is possible that the reviewers will identify errors in citations at this stage.

B.3.6 After the work has been reviewed and (if required) revised, it will be copy edited and prepared for publication (typesetting, checking image formats, converting to XML and tagging, etc). Whereas copy editing may be conducted in house or outsourced, most UK publishers outsource typesetting to one of a small number of service providers based overseas. Depending on the publication, references may be checked for accuracy during copy editing, and will probably be checked for style during typesetting. A final version of the document is often checked by the author, and is then published.

B.3.7 Table B-3 sets out the assessment for this option.

**Non-traditional publications**

B.3.8 A range of alternative publishing models are currently evolving, many of which change the way the peer review is conducted. This raises fundamental issues for citations to and from these publications – should citations from non-traditional publications be considered as equal or similar to those in traditional publications?

B.3.9 Non-traditional publishing pathways have been discussed for many years (D-Lib magazine, 2004) and represent ecosystems of varying complexity. Despite vocal complaints and dissatisfaction with the traditional publishing process, these new approaches currently represent only a small minority of all academic publication.

B.3.10 Perhaps the most widespread new approach is that in many subject domains, repositories have been established to share papers for that subject. These repositories usually contain a mix of pre-prints, drafts,
‘traditionally’ published work, and other types of information. In most cases, these are seen as secondary to the traditional publishing route, but for some they are an important primary publication route – and it is this direct-to-repository publication that is considered here. The best known—and probably most important to its communities—example is arXiv, which includes papers from a range of scientific fields, with an emphasis on physics. Whereas most papers in arXiv are published elsewhere, some – including a number of very important works – are only published to in arXiv.

B.3.11 Typically, these non-traditional publication routes involve much less input by the publisher to the work than the traditional processes – there is unlikely to be any copy editing or typesetting beyond that conducted by the author. In many cases, however, it is possible and even encouraged to update

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Publications within these non-traditional publication channels are usually open access, although not necessarily with a permissive licence. Anyone interested in the content of these publications can obtain it, and could parse the contents to extract citations. The content held in these publications is typically in a presentation format such as PDF or LaTeX, and does not have references marked up (in XML, for example).</td>
</tr>
<tr>
<td>Benefits</td>
<td>None directly, although these publishers are open access, which enables open routes for the distribution process.</td>
</tr>
<tr>
<td>Complexity</td>
<td>Using these publication routes raises a range of issues in terms of access to citation data. Firstly, these repositories overlap with traditional publications, which will require de-duplication of content and references to the content. Secondly, there may be different versions of papers available in the archive and these may also differ to any final published version. How can citations to different versions be considered? Thirdly, although the content can be parsed for citation information, this complex task is typically pushed ‘downstream’ to the indexing stage, where it has lost the context of its original publication.</td>
</tr>
<tr>
<td>Data</td>
<td>Repositories such as arXiv typically do not process publications – so third parties must analyse references embedded within the full text of the documents. Utilising these repositories as sources of citation data causes challenges in data cleaning and interpretation (see complexity above), and raises new issues of completeness: is it possible to gather sufficient information from open publications to create useful citation databases? It is unlikely that any system gathering data from only these publishers would collect enough data to be useful for realistic citation data usage scenarios. As criteria for acceptance in these publication routes vary so strongly, it is difficult to compare the ‘types’ of content they publish when considering what value to place on citations made by this content.</td>
</tr>
<tr>
<td>Disruption</td>
<td>In some ways, collecting content from this kind of publication represents relatively minor disruption to the process currently used by the major aggregators and indexers – full-text content is collected, collated and analysed. However, the questions of provenance and de-duplication likely mean that this kind of data must be treated separately – for example, Scopus indexes arXiv content, but within a separate “web” section rather than the main database.</td>
</tr>
<tr>
<td>Financial</td>
<td>Publications such as arXiv typically have significant running costs. Although access to the contents is open, the repository itself must be funded. Some repositories are operated by or on behalf of funding bodies, institutions operate others, and others are forming new business models. ArXiv is operated by Cornell Library, and is now funded by contributions from the institutions that make the most submissions.</td>
</tr>
<tr>
<td>Fragility</td>
<td>Many subject repositories are run on a goodwill basis, and their medium to long term survival and status is not assured.</td>
</tr>
<tr>
<td>Technology</td>
<td>No new technology would be required in the publication stage, unless structured citation data is captured during authorship.</td>
</tr>
<tr>
<td>Trust</td>
<td>Relying on or including non-traditional publications in citation data will probably weaken trust in the quality of metrics derived from that data in the current publishing environment. If non-traditional publications become established widely, and in many subject domains, this issue will probably be of less concern – depending on broader changes in the concepts of what constitute research outputs.</td>
</tr>
</tbody>
</table>

Table B-3: Publish: non-traditional option assessment
B.4 Distribute

B.4.1 The distribute process element covers the distribution of publications, or the citation data within, to those who wish to use them.

**Traditional distribution**

B.4.2 Once articles have been completed, they are published. Many publishers make a version available online rapidly, before including the paper in an issue of the final publication. Articles and publications are typically distributed in several formats, including PDF, HTML, XML and paper copies.

B.4.3 The publishers will host PDF and HTML versions of the articles on their own platform, which may be outsourced, and will distribute paper copies to those subscribers who want them. Publishers distribute article data to other organisations according to established relationships. Typical recipients include abstracting and indexing services including Scopus and Web of Knowledge, Microsoft Academic Search, Google Scholar and CrossRef.

B.4.4 Exact details of the distribution channels vary, but they usually operate in a ‘push’ manner as the publisher sends data to the recipients.

B.4.5 Table B-4 sets out the assessment for this option.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>This is the channel by which the vast majority of citation data becomes available for use – publishers deliver structured sets of data, including citation data, to their partners, in known formats and normally according to a known schedule.</td>
</tr>
<tr>
<td>Benefits</td>
<td>The relationships are well defined – so it is possible for publishers to work directly with their partners to resolve issues in the transfer or processing of data. Publishers hold comprehensive information for their publications – by definition. This distribution channel ensures that downstream users of the data can be confident in the coverage of the data, the provenance of the data, and the degree of completeness of the data.</td>
</tr>
<tr>
<td>Complexity</td>
<td>This distribution channel is relatively straightforward – there are a series of bilateral formal or informal agreements between publishers and users of their data.</td>
</tr>
<tr>
<td>Data</td>
<td>This distribution channel can deliver comprehensive structured full text information for that publisher’s content. Smaller publishers may find it impossible or impractical to provide some formats (for example, XML) – but this distribution channel is robust enough to respond to this by allowing PDF or paper content.</td>
</tr>
<tr>
<td>Disruption</td>
<td>N/A</td>
</tr>
<tr>
<td>Financial</td>
<td>There are financial transactions involved in this distribution chain – some publishers charge some recipients for their content. From the publishers’ perspective, it is a balance between the value of the content to the users, and the value to the publisher of exposing their content to these users.</td>
</tr>
<tr>
<td>Fragility</td>
<td>The current system is relatively robust – there are many publishers, and a range of A&amp;I/citation data services available. The balance of interests described above makes it unlikely that a major publisher would withdraw services from major A&amp;I services, and vice versa.</td>
</tr>
<tr>
<td>Technology</td>
<td>N/A</td>
</tr>
<tr>
<td>Trust</td>
<td>This system helps to engender trust in the process – contractual and defined relationships between the publishers and the users of the data, and structured, comprehensive data transfers help users to trust the quality of data.</td>
</tr>
</tbody>
</table>

*Table B-4: Distribute: traditional publication option assessment*
Gold OA

B.4.6 Gold OA is the term given to publications for which a fee (an Article Processing Charge or APC) is charged for the publication in exchange for the outputs becoming open access – in effect, a change from the dominant pay-to-read system based on journal subscriptions, to a pay-to-publish system.

B.4.7 Publications may be fully-gold, whereby all of their content is open after publication, or may adopt a hybrid model where authors have the choice to pay the APC for gold publication, or not.

B.4.8 The situation for fully-gold publications is straightforward – the entire output from the publication is available online. Any user may take the material to process as they wish, where the publisher makes the content available – a ‘pull’ relationship. In practice, publishers of fully-gold publications typically also participate in the traditional distribution mechanism, by ‘pushing’ content to established partners (eg A&I services).

B.4.9 Hybrid publications may operate in a range of ways, but typically OA papers are treated the same as closed papers, except that they are available to all users of the publisher’s platform. Article content and metadata is transferred through the traditional distribution mechanism, to the same users. As far as the study team is aware, there are no hybrid journals that allow bulk downloads of just the open access portion of their content. In some cases, hybrid journals will accept papers with specific requirements based on the research funder’s needs. For example, whereas most Elsevier journals will accept APCs to publish any article openly, some will only accept them in order to comply with specific funder requirements (Elsevier, 2013). In some cases, publishers have agreements with the funding councils to automate elements – for example by transferring Wellcome Trust-funded outputs to PubMed Central.

B.4.10 Table B-5 sets out the assessment for this option.
Open metadata and citation data

B.4.11 It would be possible for publishers to release the metadata and citation data for their publications openly, irrespective of the licence conditions in the content itself. Open access to the full text of articles is a broader question, and is discussed above. This model considers the situation where the content is (or may be) closed, but the article metadata, reference list and abstract are available openly.

B.4.12 This is a ‘pull’ model, where the publishers make data available, and users can harvest the citations and metadata. It would be possible to provide interactive querying through an Application Programming Interface (API), but in practice this would be unhelpful for realistic usage scenarios (see paragraph B.5.9).

B.4.13 In principle, it would be possible to distribute citation data without metadata (e.g., as sequences of DOIs), but in practice citation data is treated as one aspect of article metadata, and the references within articles are expressed as bibliographic data (see also paragraphs 2.2.12 - 2.2.13 and Annex A.3). All of the usage scenarios identified for citation data depend on bibliographic data for the articles involved, and most require extended data including at least the article abstract, and preferably the full text.

B.4.14 It would be possible to construct a system whereby citation data could be made open by the publishers, leaving users to obtain other metadata through another channel. Stakeholders interviewed could see no benefit to this kind of approach, and considered it to be illogical and convoluted – it would be necessary
to aggregate both data sources before processing, with attendant risks to data quality. For these reasons, this approach is not considered further here.

Table B-6 sets out the assessment for this option.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Access to citation data would be improved, as a range of alternative providers could consume article citation and metadata to create new services.</td>
</tr>
<tr>
<td>Benefits</td>
<td>The cost of access to raw citation data – at least for current publications – would be reduced. This could enable new entrants to the market to create services based on this data. There would be broader benefits to the discovery of academic publications. Increased discoverability is a benefit for the publishers as well as users of the data.</td>
</tr>
<tr>
<td>Complexity</td>
<td>Participating publishers would make their own metadata available. Aggregation would require knowledge of which publishers to harvest data from. Each data user would need to draw data from all the publishers of interest to them; for citation data this usually means all publishers. This comprises many thousands of organisations, and tens of thousands of publications. Agreeing and specifying the technologies required for this open data will take time, and will lead to significant inertia in changing the structures in future.</td>
</tr>
<tr>
<td>Data</td>
<td>Major publishers would be able to produce data in a usable format with minimal extra effort – their production workflows already manage multiple data formats and export paths. Smaller publishers may not be able to provide data in a structured format. Issues may arise for data regarding historical articles, where different publishers have digitised different quantities of their back files, and may have different metadata available for these historical publications.</td>
</tr>
<tr>
<td>Disruption</td>
<td>The current A&amp;I providers would probably see challenges to their business models, and may find that they come under pressure from free or lower-cost alternatives. Both Scopus and Web of Knowledge focus on the quality of their data, and both ingest and maintain data that is not currently available in electronic form from the publisher – so it is unlikely that this distribution channel would immediately disrupt the existing systems.</td>
</tr>
<tr>
<td>Financial</td>
<td>Revenue streams from A&amp;I services to publishers would by definition be eliminated. Enabling lower-cost alternatives to the A&amp;I providers may cause some pricing pressure, but would not directly challenge their services, which exploit full text for discovery and analytics. It seems likely that both Scopus and Web of Knowledge would further shift their focus toward analytics and business intelligence services. Conducting analytics using datasets of the size of either of these services is compute- and storage-intensive, and the barriers to entry – even with free, well-structured metadata – are significant.</td>
</tr>
<tr>
<td>Fragility</td>
<td>The distribution system would be relatively robust, assuming that common technology standards can be agreed. There would be no central point of failure in the distribution of data – although if the supply of data from individual publishers were unreliable, this would lead to a systemic problem.</td>
</tr>
<tr>
<td>Technology</td>
<td>The technologies required to distribute data in this way exist, but allowing practical use of the data, would require harmonisation between participating publishers. The key elements to agree are the data format, structure, interchange standard and the data harvesting protocol. The challenge in this de-centralised model is where would discussions to harmonise the technologies occur and what incentive is there for publishers to co-operate? CrossRef provides a potential model for a body that could provide a forum for such discussions but may be unwilling to take on the role itself.</td>
</tr>
<tr>
<td>Trust</td>
<td>The non-participation of some small publishers would not prevent the remainder of the system from functioning. However, if a significant number of publications were not available either through the non-participation of one or more major publishers, or many smaller ones, there would be significant problems of trust in the data.</td>
</tr>
</tbody>
</table>

Table B-6: Distribute: open metadata and citation data option assessment

Self-archiving (Green OA)

An alternative to Gold OA (pay-to-publish) is so-called Green OA – where an author can lodge papers in an appropriate repository including subject and institutional repositories, but potentially simply on his or her own website. There is overlap with non-traditional publications, in that many of those publications are in practice operated as a repository.
Traditional publishers typically embargo articles that they publish eg for a 6-month period after publication. In some cases these embargoes may be different for institutional as opposed to subject repositories.

Table B.7 sets out the assessment for this option.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Self-archiving is <em>ad hoc</em> and inconsistent. Articles lose the context of their publication, and may be deposited in several archives, in several forms, by different individuals or organisations. Anyone interested in the content of these articles can obtain it, and could parse the contents to extract citations. The content held in these publications is typically in a presentation format such as PDF or LaTeX, and does not have references marked up (in XML, for example).</td>
</tr>
<tr>
<td>Benefits</td>
<td>The infrastructure to support Green OA is already in place. It is possible that this system could operate in parallel with other distribution models, but it is unclear what benefits would be delivered.</td>
</tr>
<tr>
<td>Complexity</td>
<td>Any system relying on repositories as a supply of citation data would have to identify, parse, aggregate, and de-duplicate content from many thousands of repositories. This would be an extremely complex system, highly dependent on the quality and quantity of metadata within repositories, which is variable.</td>
</tr>
<tr>
<td>Data</td>
<td>Full text articles are available, but the current level of self-archiving in repositories is low. Changes to funding body and organisational policies may drive uptake, but will not resolve the significant challenges in distributing citation data with this process. Back files are unlikely to be made available on OA repositories, meaning that this distribution model would only support future publications.</td>
</tr>
<tr>
<td>Disruption</td>
<td>Transition to using this distribution model exclusively would be extremely disruptive – it would not be possible to deliver reliable services using data harvested only from OA repositories.</td>
</tr>
<tr>
<td>Financial</td>
<td>Although access to the content is open, the repository itself must be funded. Some repositories are operated by or on behalf of funding bodies, others are operated by institutions, and others are forming new business models.</td>
</tr>
<tr>
<td>Fragility</td>
<td>The distribution system would be relatively robust. There would be no central point of failure in the distribution of data – although differences in approaches to metadata between different repositories lead to a risk of developing technical incompatibilities.</td>
</tr>
<tr>
<td>Technology</td>
<td>No new technologies are required for distribution, but infrastructures to harvest and process the data collected would require development.</td>
</tr>
<tr>
<td>Trust</td>
<td>Due to the difficulties in assuring completeness of the data or even understanding how complete the data is, this model would probably have a significant negative impact on trust, at least until essentially all academic publications are available in OA repositories. Many repositories include mixtures of post-print, pre-print, and potentially non-published content, which leads to significant problems in understanding provenance and the value of citations within these articles.</td>
</tr>
</tbody>
</table>

Table B.7: Distribute: self-archiving (Green OA) option assessment

**B.5 Collate & index**

The collate and index elements are addressed together here, as in practice it is not sensible to collate data without indexing it.

This includes the work done to bring together the citation data from the various publications within the desired coverage, and the work done to process the raw citation data from publications, to produce useful information. This may include identifying citations within publications, extracting, interpreting and normalising their formats, and creating a data store that can then be used to process this data.
Existing systems

B.5.3 A range of systems already exist for collation and indexing of citation data; these are discussed in detail at Section 4. The key points to note are:

- All of the existing systems have established relationships with publishers, although some also take data from ‘open’ websites
- All of the existing systems rely strongly on bibliographic data. None treat citation data as separate from the bibliographic data that describes the articles
- All bar one of the existing systems (the Jisc Citations Data Open Data Corpus) are designed to index and enable discovery of academic literature, and treat citation data as one piece of information about the literature
- The services which are provided on a charged basis (Scopus and Web of Knowledge) are selective regarding their coverage, and both take data from a range of distribution channels including around 35% in paper format
- Some publishers raise revenue by charging A&I providers for access to their data. Most publishers see benefit in getting increased traffic and thus sales for their products
- This system would work with ‘open’ data as well – it does not depend on the data being closed

B.5.4 The bulk of the effort undertaken by the existing services is in parsing and indexing citations that are provided in many thousands of different formats, and with varying quality. Unique identifiers for researchers (ORCID) should ease the challenges of accurately assigning works to authors – but it will not help in assigning citations to works. Further, ORCID is user-centric, and is not designed to be retrospective: only active researchers are likely to register with ORCID.

Open centralised database

B.5.5 Publishers currently make their bibliographic and citation data available through several routes. The current ‘closed’ systems are primarily historical artefacts – it is only recently that there has been significant discussion of the benefits that open data can bring.

B.5.6 The infrastructure already exists for publishers to make metadata and citation data available without charge – CrossRef already holds essentially all publication metadata, and a large proportion of current and recent citation data. The key challenges to making data open through CrossRef are:

- The publishers must decide that opening their citation data is worthwhile, given that it may reduce their income from selling this data to A&I providers. This may be particularly challenging for publishers that run A&I services (clearly including Elsevier, but also others)
- A major revenue stream for CrossRef is charging for access to the article metadata. If publishers want to release metadata openly, this will affect CrossRef’s business model. The existing CrossRef Cited By service explicitly excludes general use of citation data – it is strictly for the use of participating publishers to retrieve citations to their own publications.

B.5.7 Note that although the study team has engaged with CrossRef, this model was developed independently of CrossRef. There would probably be significant challenges for CrossRef in adopting this model, and at this stage, it is not clear whether CrossRef would be interested in supporting such a move.
Table B-8 sets out the assessment for this option.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Access to citation data would be improved, at least for current and recent publications.</td>
</tr>
<tr>
<td>Benefits</td>
<td>Opening access to collated and indexed citation data such as that held by CrossRef could reduce the cost of accessing citation data. This would enable some usage scenarios – but the limitations on the coverage of CrossRef data would probably make this only a partial solution.</td>
</tr>
<tr>
<td>Complexity</td>
<td>The system would be relatively straightforward – the organisational infrastructure already exists and is relatively robust.</td>
</tr>
<tr>
<td>Data</td>
<td>Data would most likely only be available for current and recent publications, with citation data available to CrossRef in XML form. On the other hand, publications could be included that are not within the coverage of the Scopus or Web Of Knowledge services.</td>
</tr>
<tr>
<td>Disruption</td>
<td>This would likely exist in addition to the current system. Metadata and citation data by themselves do not enable the broader A&amp;I activities that other providers conduct, and the limitations to coverage would likewise limit use.</td>
</tr>
<tr>
<td>Financial</td>
<td>There would be significant challenges in establishing a business model. CrossRef derives income in part from charging for access to metadata, and balancing this requirement with a mission to open citation data would be challenging. Revenue streams for other existing providers may be challenged.</td>
</tr>
<tr>
<td>Fragility</td>
<td>CrossRef would represent a single point of failure in this system. More serious is the requirement to persuade all or most publishers to participate, as many usage scenarios depend on broad coverage.</td>
</tr>
<tr>
<td>Technology</td>
<td>No new technologies would be required. Some publishers may need to alter workflow systems to provide citation data to CrossRef.</td>
</tr>
<tr>
<td>Trust</td>
<td>While this would provide a broader base of data, there might be resultant issues regarding assessments of the quality of publications because of the potential removal of the selective element.</td>
</tr>
</tbody>
</table>

---

**Federated search**

It is possible to envisage a system that does not depend on a central coordinating body to collate citation data. Rather, ‘collation’ would be conducted in real time by searching APIs provided by data holders – traditional and non-traditional publishers. This is known as a federated search. This option would make most sense if distribution is through an ‘Open metadata and citation data’ distribution channel (paragraph B.4.11), but in principle it would be possible to include open and closed sources within a federated search.

Table B-9 sets out the assessment for this option.
Factor | Notes
---|---
Access | Federated search would be available to any organisation that could build an infrastructure capable of querying the many thousands of data sources that are likely to be of interest.
Benefits | No central database or organisation to support.
Complexity | The complexity is extremely high. More than 4000 publishers are included within the existing major citation databases, so a federated search would likely need to cover this number of endpoints (or more).
Data | In practice, it will be necessary for publishers to agree on API syntax and data formats to enable this type of environment. Data would be available from all publishers who provide an API for their data. However, the nature of citation data is that all data sources must be queried to return the citations to a single paper. Querying so many large datasets over internet APIs is simply not practical. The likely and somewhat counter-intuitive result is that relying on federated search would actually reduce the availability of usable data.
Disruption | Disruption would be substantial, as this option makes impossible a range of usage scenarios which depend on complex analyses of data. It is likely that major players would still create centralised databases in order to fulfil these usage scenarios – thus operating as they do at present, or as per the CrossRef Cited By option above.
Financial | Revenue streams for existing providers may be challenged, but it is extremely unlikely that federated search would replace the monolithic systems in any significant manner.
Fragility | Fragility would be extremely high, as searches are conducted in real time, requiring all data sources to be available and to respond in good time.
Technology | Existing technologies could in theory support this kind of infrastructure, but in practice reliability and responsiveness would make such systems unusable. Whereas linked data approaches such as RDF are often designed to work with heterogeneous and distributed data sources, the scale of this problem would make existing technologies unworkable. The volume of data transferred from each data provider would probably become unsustainable, because of the increasing numbers of systems collecting data. In particular, more-complex queries also become unrealistic in this kind of distributed system without transferring and caching large data volumes locally – queries on large datasets are notoriously slow. They are also typically conducted on high performance systems (Husain, et al., 2010). An asynchronous search, essentially harvesting data from each publisher regularly, rather than in real time would enable the creation of a centralised dataset that could process the volume of data required, but this would essentially operate as at present, or as per the CrossRef Cited By option above.
Trust | It would be difficult to generate trust in this model, as there is essentially no quality control in the data – it must be taken as it is provided by the publisher. It would be impossible to conduct any data-cleaning activities during a federated search. Furthermore, as data is collected from so many independent sources in real time, there is no way to guarantee reproducibility of results, and likewise no record of the history of the dataset. This would significantly hinder usage scenarios that depend on reproducibility and reliability.

Table B-9: Collate & index: federated search option assessment

B.6 Analysis of future business models

B.6.1 This sub-section sets out several 'end-to-end' models of how citation data can be created, processed, and made available to users. These options have been chosen to represent three realistic and interesting cases:

- The current model
- An open model, where article metadata, citation data and abstracts are freely available from each publisher
- A central, open distribute model, where article bibliographic metadata and citation data are freely available from an open centralised database; one possibility is that CrossRef might undertake this role, building on the data already held. This option is intermediate between the current and open models.
These end-to-end models are described at Table B-10 in terms of the process elements considered at Table B-1. Not all process element options have been included. For example create: structured citation capture during authorship is considered to provide only a variation of these three options and the federated search option is technically infeasible.

<table>
<thead>
<tr>
<th>Process element</th>
<th>Current model</th>
<th>Open model</th>
<th>Central, open distribute model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
<td>Existing processes</td>
<td>Existing processes</td>
<td>Existing processes</td>
</tr>
<tr>
<td>Publish</td>
<td>Traditional publication</td>
<td>Non-traditional publications</td>
<td>Traditional publication</td>
</tr>
<tr>
<td>Distribute</td>
<td>Traditional distribution (inc Gold OA)</td>
<td>Open metadata and citations</td>
<td>Traditional distribution (inc Gold OA)</td>
</tr>
<tr>
<td>Collate &amp; Index</td>
<td>Existing systems</td>
<td>A range of providers may offer centralised or distributed options</td>
<td>Open centralised database</td>
</tr>
<tr>
<td>Analyse &amp; Exploit [24]</td>
<td>See Usage scenarios (Section 3)</td>
<td>See Usage scenarios (Section 3)</td>
<td>See Usage scenarios (Section 3)</td>
</tr>
</tbody>
</table>

Table B-10: Comparison of end-to-end options considered

These end-to-end options are discussed in the remainder of this sub-section with Table B-14 providing an overall comparison.

**Current model**

The main flow of the current system is a combination of the existing processes described above:

1) The article is drafted in a word processor, and citations are included in text form  
2) The article is peer reviewed and edited/formatted by a publisher  
3) The article data and metadata are distributed to a range of organisations, including the major A&I/citation data providers Scopus and Web of Knowledge, as well as Google Scholar and Microsoft Academic Search. This is in a range of formats including in paper form, PDF and XML  
4) The A&I providers bring together data from many thousands of publishers, and extract citation data from the publications they receive  
5) The A&I providers match the text format citations to papers, authors, and institutions  
6) Customers access data within the A&I systems; this may be free or charged depending on the service

Table B-11 sets out the overall assessment for this option.

---

24 Alternatives are out of scope. The analysis and exploitation of citation data depend strongly on the requirements of the specific user. The usage scenarios considered are discussed in Section 3.
Access to Citation Data: Cost-benefit and Risk Review and Forward Look

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Access to citation data is through a range of services, and in a range of forms depending on what the customers require.</td>
</tr>
<tr>
<td>Benefits</td>
<td>A very wide range of usage scenarios are fulfilled by the current system – data is available in many forms – as long as the user is able to pay for it. A reasonable range of data is available at no cost (for example through Microsoft Academic Search and Google Scholar), but this data has undergone far less rigorous selection and QA processes, and is less trusted by users. Unless publishers make access to their metadata and citation data open, this is the only realistic process for generating usable citation data.</td>
</tr>
<tr>
<td>Complexity</td>
<td>The system is large, but relatively simple – a small number of large organisations collate &amp; index data from a large number of publishers. Complexities come in understanding the different coverage of the different systems, and the different approaches to QA.</td>
</tr>
<tr>
<td>Data</td>
<td>The existing processes are very good at enabling access to citations within mainstream, traditional academic publications. Citations of and within books are increasingly being included by the major services. Different services have different coverage - Non-traditional publications (eg direct to arXiv) are not well covered by Scopus or Web of Knowledge but are included within Microsoft Academic Search and Google Scholar, for example.</td>
</tr>
<tr>
<td>Disruption</td>
<td>N/A</td>
</tr>
</tbody>
</table>
| Financial  | The key revenue streams in the current system are:  
  - A&I service providers charging users for access to the data  
  - Publishers charging some A&I services for access to data  
  The key costs are incurred by:  
  - Collating citation data from many thousands of publishers  
  - Inputting any paper-form data  
  - Processing the collated data to identify and parse citations  
  - Allocating citations to works, authors, and institutions, with associated quality assurance |
| Fragility  | The current system is reasonably robust, with four major providers arranged into two groups of two: high quality/high price (Scopus and Web of Knowledge) and automated/low cost (Google Scholar and Microsoft Academic Search). Relationships between the publishers and the service providers are strong, and are typically based on contractual agreements. The service providers in the high quality/high cost space have contractual relationships with their users, which can include service level agreements. |
| Technology | N/A                                                                                                                                                                                                  |
| Trust      | Stakeholders trust the current system. In our interviews, all participants confirmed they thought the current system was robust and largely satisfied their needs. There are failings – for example, coverage of non-traditional publications, and publications in the arts and humanities – but these primarily come down to choices of coverage. |

Open model

B.6.6 If the ‘open’ option is chosen at each stage, the flow would be somewhat different:

1) The article is drafted in a word processor, and citations are included in text form
2) The article is peer reviewed and edited/ formatted by a publisher
3) The publisher makes the article and metadata available openly
4) A wide range of organisations and individuals access the data from publishers, through federated search or by establishing their own centralised databases
5) Those organisations process the data and make it available to their users according to their own strategies, but at least some are likely to be free to end users
Table B-12 sets out the assessment for this option.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Access to citation data would be improved - at least for current and recent publications - as a range of alternative providers could consume article citation and metadata to create new services.</td>
</tr>
<tr>
<td>Benefits</td>
<td>The cost of access to raw citation data – at least for current publications – would be reduced. This could enable new entrants to create services based on this data. There would be broader benefits to the discovery of academic publications. Increased discoverability is a benefit for the publishers as well as users of the data.</td>
</tr>
<tr>
<td>Complexity</td>
<td>The complexity of the system would be high. Publishers would make their own metadata available. The aggregation requires knowledge of which publishers to harvest data from. Each data user would need to draw data from all the publishers of interest to them. For citation data, this usually means all publishers. This comprises many thousands of organisations, and tens of thousands of publications. Agreeing and specifying the technologies required for this open data will take time, and will lead to significant inertia in changing the structures in future.</td>
</tr>
<tr>
<td>Data</td>
<td>Major publishers would be able to produce data in a usable format with minimal extra effort – their production workflows already manage multiple data formats and export paths. Smaller publishers may not be able to provide data in a structured format. Issues may arise regarding data for historical articles, where different publishers have digitised different quantities of their back files, and will no doubt have different metadata available for these historical publications.</td>
</tr>
<tr>
<td>Disruption</td>
<td>The existing A&amp;I services use full-text data to enable effective processing. Despite the moves in the UK to promote open access, on a global scale only a small minority of content is available in this way, and little historical content. Thus, these services will provide value that cannot be met by open metadata and citation data. The current A&amp;I providers would probably see challenges to their business models, and may find that they come under pressure from free or lower-cost alternatives. Both Scopus and Web of Knowledge focus on the quality of their data, and both ingest and maintain data that is not currently available in electronic form from the publisher – so it is unlikely that open publisher data would disrupt these systems immediately.</td>
</tr>
<tr>
<td>Financial</td>
<td>Revenue streams from A&amp;I services to publishers would by definition be eliminated. This model considers the ‘fully open’ model, but it is important to recognise that the existing A&amp;I services would probably continue to operate alongside any potential open alternatives – they compete primarily on quality rather than price. Enabling lower-cost alternatives to the A&amp;I providers may cause some pricing pressure, but would not directly challenge their services, which exploit full text for discovery and analytics. It seems likely that both Scopus and Web of Knowledge would shift their focus toward analytics and business intelligence services. Conducting analytics on datasets of the size of either of these services is compute - and storage-intensive, and the barriers to entry – even with free, well-structured metadata – are significant. A ‘lightweight’ federated search would not meet the range of usage scenarios relevant to citation data – such services are more likely to meet resource discovery rather than citation data usage scenarios, and then only with very serious performance and quality issues.</td>
</tr>
<tr>
<td>Fragility</td>
<td>The distribution system would be relatively robust, assuming that common technology standards can be agreed. There would be no central point of failure in the distribution of data – although if the supply of data from individual publishers were unreliable, this would have systemic problems.</td>
</tr>
<tr>
<td>Technology</td>
<td>The technologies exist, but enabling practical use of the data would require harmonisation between participating publishers. The key elements to agree are the data format and structure, and the data harvesting protocol. The challenge is that in this de-centralised model, where would discussions to harmonise the technologies occur and what incentive is there for stakeholders to participate.</td>
</tr>
<tr>
<td>Trust</td>
<td>The non-participation of some publishers would not prevent the remainder of the system from functioning. In practice, existing A&amp;I services would probably still obtain data from them using the current system. However, if a significant number of publications are not available either through the non-participation of one or more major publishers, or many smaller ones, then there could be real problems of trust. Trust overall will depend on the approach taken by the service providers. Centralised, collated and indexed data sources could enable trust, whereas federated search tools are much less likely to do so.</td>
</tr>
</tbody>
</table>

Table B-12: End-to-end open model analysis
Open distribute

B.6.8 This model describes a situation where citation data is distributed openly, through a common centralised service.

1) The article is drafted in a word processor, and citations are included in text form
2) The article is peer reviewed and edited/formatted by a publisher
3) The article data and metadata is distributed to a range of organisations, including an open central database of citation data
4) A wide range of organisations and individuals access the data from the central database
5) Those organisations process the data and make it available to their users according to their own strategies, but at least some are likely to be free to end users

B.6.9 It would be expensive to establish a new organisation to deliver this service – but CrossRef represents an opportunity, as it already has an established system and workflow for collating and indexing citations.

B.6.10 Note that although the study team has engaged with CrossRef, this model was developed independently of them. There would probably be significant challenges for CrossRef in adopting this model, and at this stage it is not clear whether CrossRef would be interested in supporting such a move.

B.6.11 Table B-13 sets out the assessment for this option.
Access to citation data would be improved – at least for current and recent publications, as a range of alternative providers could consume article citation and metadata to create new services. This would enable some usage scenarios – but the limitations on the coverage of CrossRef data would probably make this only a partial solution. There would still be a role for more comprehensive A&I services operating through other channels.

It is possible that A&I services would make use of the CrossRef data as an input for their services, perhaps to cover publishers that they do not include, or to provide an additional QA stage.

Opening access to collated and indexed citation data such as that held by CrossRef could reduce the cost of accessing citation data.

The system would be relatively straightforward – the organisational infrastructure already exists and is relatively robust. If CrossRef were unable or unwilling to adopt this role, it would be extremely challenging to establish a new organisation to do so.

Data would probably only be available for current and recent publications, with citation data available to CrossRef (or an alternative provider) in XML form. On the other hand, publications could be included that are not within the coverage of Scopus or Web of Knowledge services.

This would probably exist in addition to the current system. Metadata alone does not enable the broader A&I activities that other providers conduct, and the limitations to coverage would likewise limit use.

This system would be totally dependent on CrossRef, which therefore represent a single point of potential failure in this system. More serious is the requirement to persuade all or most publishers to participate, as usage scenarios depend on broad coverage.

No new technologies would be required. Some publishers may need to alter workflow systems to provide citation data to CrossRef.

Using CrossRef data would remove the selective element that both Scopus and Web of Knowledge apply to the publications that are indexed. This would provide a broader base of data, but with resultant issues regarding assessments of the quality of publications that are included.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Access to citation data would be improved – at least for current and recent publications, as a range of alternative providers could consume article citation and metadata to create new services. This would enable some usage scenarios – but the limitations on the coverage of CrossRef data would probably make this only a partial solution. There would still be a role for more comprehensive A&amp;I services operating through other channels. It is possible that A&amp;I services would make use of the CrossRef data as an input for their services, perhaps to cover publishers that they do not include, or to provide an additional QA stage.</td>
</tr>
<tr>
<td>Benefits</td>
<td>Opening access to collated and indexed citation data such as that held by CrossRef could reduce the cost of accessing citation data.</td>
</tr>
<tr>
<td>Complexity</td>
<td>The system would be relatively straightforward – the organisational infrastructure already exists and is relatively robust. If CrossRef were unable or unwilling to adopt this role, it would be extremely challenging to establish a new organisation to do so.</td>
</tr>
<tr>
<td>Data</td>
<td>Data would probably only be available for current and recent publications, with citation data available to CrossRef (or an alternative provider) in XML form. On the other hand, publications could be included that are not within the coverage of Scopus or Web of Knowledge services.</td>
</tr>
<tr>
<td>Disruption</td>
<td>This would probably exist in addition to the current system. Metadata alone does not enable the broader A&amp;I activities that other providers conduct, and the limitations to coverage would likewise limit use. Nonetheless, the open availability of well-indexed citation data may support experimentation, and potentially some secondary services based on this data, most likely services designed to monitor current publications and citations, rather than to generate comprehensive and reliable data. This could potentially draw some marginal users away from the comprehensive A&amp;I services – but it is more likely that these users are currently satisfied with the data from Google Scholar or MS Academic Search and do not subscribe to one of the charged services.</td>
</tr>
<tr>
<td>Financial</td>
<td>There would be significant challenges in establishing a viable business model. CrossRef derives income in part from charging for access to metadata, and balancing this requirement with a mission to open citation data would be challenging. Establishing this model using a host other than CrossRef would incur substantial costs in building and running a database and the relationships with publishers that will be necessary to enable data sharing.</td>
</tr>
<tr>
<td>Fragility</td>
<td>This system would be totally dependent on CrossRef, which therefore represent a single point of potential failure in this system. More serious is the requirement to persuade all or most publishers to participate, as usage scenarios depend on broad coverage.</td>
</tr>
<tr>
<td>Technology</td>
<td>No new technologies would be required. Some publishers may need to alter workflow systems to provide citation data to CrossRef.</td>
</tr>
<tr>
<td>Trust</td>
<td>Using CrossRef data would remove the selective element that both Scopus and Web of Knowledge apply to the publications that are indexed. This would provide a broader base of data, but with resultant issues regarding assessments of the quality of publications that are included.</td>
</tr>
</tbody>
</table>

Table B-13: End-to-end open distribute model analysis

Overall assessment

Table B-14 summarises the analysis of the end-to-end options considered. The overall assessment is that these options are all viable. In practice, the situation is likely to be complex with these options existing in parallel and interacting. The options described are not deterministic. The study team does not believe that any of these options is a likely or preferable future environment.
## Access to Citation Data: Cost-benefit and Risk Review and Forward Look

<table>
<thead>
<tr>
<th>Factor</th>
<th>Current situation</th>
<th>End-to-end Open</th>
<th>Open distribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Good</td>
<td>Improved for current and recent publications</td>
<td>Improved for current and recent publications</td>
</tr>
<tr>
<td>Benefits</td>
<td>Usage scenarios fulfilled Pragmatic with closed data Range of providers</td>
<td>Cost of accessing current and recent data reduced. Broader discoverability benefits for publishers</td>
<td>Cost of accessing current and recent data reduced</td>
</tr>
<tr>
<td>Complexity</td>
<td>Large but not overly complex; generally well understood by all stakeholders</td>
<td>High</td>
<td>Simple, if CrossRef leads</td>
</tr>
<tr>
<td>Data</td>
<td>Good, limitations in coverage</td>
<td>Challenges for smaller publishers and particularly for historical content. Coverage could be broader</td>
<td>Challenges for smaller publishers and particularly for historical content. Coverage could be broader</td>
</tr>
<tr>
<td>Disruption</td>
<td>No particular issues</td>
<td>May challenging existing value propositions, but only in the long term, as cannot fulfil all usage scenarios</td>
<td>Low. May challenging existing value propositions, but only in the long term, as cannot fulfil all usage scenarios</td>
</tr>
</tbody>
</table>
| Financial    | Revenue streams:  
- A&I service providers  
- Publishers  
Key costs:  
- Collating citation data  
- Inputting any paper-form data  
- Processing the collated data  
- Allocating citations to works, authors, and institutions, with associated QA | Revenue to publishers from A&I services eliminated  
- Costs for aggregating and processing data remain high  
- Costs to conduct analytics on data remain high  
- Price of trustworthy data aggregations likely to remain high | Significant challenge in establishing business model  
- If CrossRef not involved, very high costs in establishing system |
| Fragility    | Robust                                                                           | Robust due to distributed nature, but risks to trust if individual data sources are unreliable | Central provider single point of failure. Necessity to persuade all or most publishers to participate |
| Technology   | No particular issues                                                             | Data formats and interchange standards need to be agreed                          | No new technologies, but perhaps some workflow changes for publishers           |
| Trust        | High, although some concerns over coverage                                        | Non-participation of publishers would significantly reduce trust. Overall trust depends on approaches taken by service providers in the system | Central organisation provides focus for trust. Coverage could be broadened, increasing utility and trust in some subject domains |

_Table B-14: Analysis of end-to-end options considered_
This report was commissioned by Jisc and produced by Dr Geoff Curtis, Dr Max Hammond and Professor Charles Oppenheim