publications

Article

Enhancing Institutional Publication Data Using Emergent Open Science Services

David Walters [1] **and Christopher Daley** [2,*]

[1] Open Access Officer, Information Services, Brunel University London, London UB8 3PH, UK; david.walters@brunel.ac.uk
[2] Research Development Officer, Research Support and Development Office, Brunel University London, London UB8 3PH, UK
* Correspondence: christopher.daley@brunel.ac.uk; Tel.: +1-895-265-314

Received: 5 March 2018; Accepted: 14 May 2018; Published: 16 May 2018

Abstract: The UK open access (OA) policy landscape simultaneously preferences Gold publishing models (Finch Report, RCUK, COAF) and Green OA through repository usage (HEFCE), creating the possibility of confusion and duplication of effort for academics and support staff. Alongside these policy developments, there has been an increase in open science services that aim to provide global data on OA. These services often exist separately to locally managed institutional systems for recording OA engagement and policy compliance. The aim of this study is to enhance Brunel University London's local publication data using software which retrieves and processes information from the global open science services of Sherpa REF, CORE, and Unpaywall. The study draws on two classification schemes; a 'best location' hierarchy, which enables us to measure publishing trends and whether open access dissemination has taken place, and a relational 'all locations' dataset to examine whether individual publications appear across multiple OA dissemination models. Sherpa REF data is also used to indicate possible OA locations from serial policies. Our results find that there is an average of 4.767 permissible open access options available to the authors in our sample each time they publish and that Gold OA publications are replicated, on average, in 3 separate locations. A total of 40% of OA works in the sample are available in both Gold and Green locations. The study considers whether this tendency for duplication is a result of localised manual workflows which are necessarily focused on institutional compliance to meet the Research Excellence Framework 2021 requirements, and suggests that greater interoperability between OA systems and services would facilitate a more efficient transformation to open scholarship.

Keywords: open access; scholarly communication; repositories; compliance; REF 2021; Research Excellence Framework; research information systems; UK funder policies

1. Introduction

Described within Suber's seminal monograph on open access [1], the history of open scholarly communication can be traced back almost two decades. The most succinct articulation of the open access movement was found through the 2002 Budapest Open Access Initiative [1,2]. The clear message delivered by this declaration led to the adoption of OA policies at an institutional and funder level, with OA becoming, in recent years, a notable area of focus for librarians, research offices, and funding bodies. Such enthusiasm has resulted in an increasingly complex policy terrain whereby institutional policies exist alongside those of various funding bodies. At the time of writing, the Registry of Open Access Repository Mandates and Policies (ROARMAP) lists 928 distinct policies by research organisations and funding bodies [3]. Most recently, UK institutions have been required to implement the Higher Education Funding Council England's (HEFCE) OA policy for the 2021 Research Excellence

Framework (REF) [4] (from 1 April 2018, HEFCE has been replaced by a new body named Research England). This policy requires all journal articles and published conference papers to be deposited into an institutional or disciplinary repository within three months of acceptance for publication. The policy contains a series of exceptions, which include those papers which have been made OA via the Gold route—a preference for researchers in receipt of the Research Councils UK (RCUK, from 1 April 2018 now named UK Research and Innovation) grant funding [5]. Failure to comply may see research papers ineligible for submission to REF 2021 or researchers denied future funding opportunities.

HEFCE's policy has positively embedded OA firmly within the minds of researchers and ensured that practical steps have been taken to make publications openly available. Some library professionals, notably Kingsley [6], have nonetheless commented on the difficulties in providing local services to researchers within a complex UK policy terrain and the international, collaborative research environment. However, so far, there has been little analysis of institutional repository (IR) workflow interactions with existing channels for open dissemination.

Alongside recent policy developments, there has also been a growth in online services that we refer to as emergent open science services. These are typically global services or platforms, used by researchers and university staff around the world to support research processes, collaboration and dissemination. The '101 Innovations in Scholarly Communication' project [7] documented many of these and demonstrated the proliferation and application of tools and services within various research lifecycles and workflows. Publishers and other private entities have developed some of these innovations while others are publicly maintained and led by research communities and libraries. A number of services enable programmatic access to data, particularly where this concerns scholarly outputs. Typically, REST interfaces enable access to resources, maximising the potential re-use and application of data. Data from these services can be used to provide insights into OA publishing and scholarship.

Commercial products that provide data on OA works include 1findr (1science) and the freemium Lantern service (Cottage Labs) [8]. Non-commercial products include CORE (Jisc/The Open University), the Bielefeld Academic Search Engine (Bielefeld University Library), Dissemin (CAPSH), and the Open Access Button (SPARC). These organisations are based in the US, UK, Germany, and France. OA information is of intense and growing interest to many stakeholders in the global research community.

Three such services were selected to provide data for this study: Sherpa REF, Unpaywall, and CORE. Sherpa REF resources indicate possible OA locations from serial policies. Unpaywall and CORE services aggregate information on OA content and location, for example, of outputs hosted by institutional repository systems. These services were selected because of the size and type of data holdings (for example, OA/publications/serials), confidence in data quality (for example, use in mainstream research and information system tools), and the reputation of the service or organisation provider. The data they hold also aligns with the aims and objectives of the study, described towards the end of this section.

One complexity of the modern scholarly communication landscape is that individual works may be available to be accessed from several 'OA locations' on the internet. Unpaywall describes an OA location as 'the particular location where a given OA article is found. The same article is often available from multiple locations. There may be differences in format, version, and licence from one location to another, even if it is the same article in all cases' [9]. An instance of an OA location may be where it is hosted on a publisher platform or in an institutional or disciplinary repository.

Sherpa REF [10] holds data on specific REF compliance information taking account of REF Panel requirements. The service also holds data about the number of possible OA locations that authors have available when publishing in a scholarly journal or conference proceeding, exclusive of REF policy requirements. Therefore, Sherpa REF data indicates the potential availability of multiple OA locations for a single publication.

Sherpa REF builds on the functionality of its sister SHERPA RoMEO service (Jisc). SHERPA RoMEO is a reputable and heavily used service by the scholarly community, holding data on self-archiving policies for individual journal titles [11]. Sherpa REF is still in Beta, but its data model quantifies machine-readable OA permissions data. This contrasts to the SHERPA RoMEO service, which provides qualitative insights through free-text fields that must be interpreted by humans during a typical permissions assessment process.

CORE is an aggregation service that harvests full-text repository content and makes metadata and outputs available via their API [12]. CORE's repository connector, developed for the OpenMinTeD project [13], means they currently hold the most extensive datasets for text mining Open Access content. CORE data include geolocation fields such as country codes and physical coordinates. In 2016, the authors of this study conducted a 10 years analysis of global OA activity relating to Brunel University, London, using CORE [14]. At the time, this resource indicated growth in Gold and Green OA beyond the visibility of local institutional systems.

Unpaywall was released in 2017 and has been quickly adopted, in part due to the popularity of the related browser extension that provides users with seamless access to OA works, but also because the reuse of the data is not restricted. The Unpaywall service independently monitors content host locations, including Gold OA journals, Hybrid journals, institutional and disciplinary repositories. A study by Bosman [15] applied the availability of Unpaywall data in the Web of Science database to provide a detailed breakdown of OA by type for universities in the Netherlands. Piwowar and Priem (et al.) recently conducted research using Unpaywall to reveal a global 'state of OA' whereby 45% of recent scholarly works were found to be legally available through Gold or Green publishing models [16]. Additionally, a similar study conducted by Universities UK found 37% of UK publications were open access in 2016 [17]. Piwowar and Priem (et al.) comment on the fluidity of OA definitions and different subtypes identified by researchers, and the challenges this presents in delivering acceptable data to the sector [16]. At the highest level, Piwowar and Priem (et al.) define OA as 'free to read online, either on the publisher website or in an OA repository' and 'all articles not meeting this definition were defined as Closed' [16]. Nonetheless, Piwowar and Priem (et al.) also define several OA subtypes to classify data from Unpaywall. These classifications correspond with the terminology used by institutions and funding bodies. These are:

- Gold: Published in an open-access journal that is indexed by the DOAJ (Directory of Open Access Journals—a directory indexing open access peer-reviewed journals).
- Green: Toll-access on the publisher page, but there is a free copy in an OA repository.
- Hybrid: Free under an open licence in a toll-access journal.
- Bronze: Free to read on the publisher page, but without a clearly identifiable licence.
- Closed: All other articles, including those shared only on an Academic Social Network or in Sci-Hub [16].

Bronze is a new classification singled out by Piwowar and Priem (et al.) to highlight the emerging trend for publishers to release 'free-to-read' content while retaining copyrights, which corresponds with Suber's well-known 'Gratis' definition [1]. Piwowar and Priem (et al.) describe Bronze as 'articles made free-to-read on the publisher website, without an explicit Open licence' [16]. This type increases access to outputs by researchers, but paradoxically contravenes reusability benefits advocated by RCUK, HEFCE, and the Charity Open Access Fund (COAF) which drive the UK agenda. The Bronze classification holds properties of the Gold and Hybrid models, in the sense that it is a comparative form of publisher hosted OA. For example, at the time of writing 403 journal titles listed in the DOAJ publish under a 'free-to-read' licence. These serials specify the 'journal licence' as 'Publisher's own licence.' Piwowar and Priem (et al.) discovered that half of 'Bronze' serials were 'hosted on journals that published 100% of content as free-to-read but were not listed on the DOAJ and did not formally licence content', which they refer to as 'Dark Gold' [16].

Piwowar and Priem's (et al.) study was acknowledged by Himmelstein (et al.) [18], where they imply that the universal adoption of a Bronze model by publishers for 'closed' works may remedy the dangers to the ecosystem posed by piracy sites. This landscape presents challenges for automated tools, which can check the access 'state' of papers at the article level, particularly regarding Hybrid publications. For example, works might become freely accessible when subscription journals release content on their platform after an embargo period ('Delayed OA' according to Piwowar and Priem), or promotional access, where subscription publishers enable access to content for a limited period to increase subscriber levels—a current example of this practice is the American Dental Association, who are presently offering time-limited 'free access' to the *Journal of Prosthodontists* [19]. Piwowar and Priem's (et al.) results may suggest that the monitoring of OA data should be continuous to track publisher activities and ensure perpetual access to research where reuse rights are restricted. We applaud the authors for bringing great attention to this issue and agree that longitudinal research should be undertaken on a global scale to assess these trends. The availability of the Unpaywall dataset through a REST API and other mechanisms [20] enables this level of scrutiny without manual effort, at different levels of scale.

A recent study by Martín-Martín (et al.) [21] utilised Google Scholar (GS) as a source of data to analyse OA levels across all countries and fields of research. Data from GS is extracted with difficulty because GS does not maintain a public API. Martín-Martín (et al.) discovered OA levels comparable with other recent studies in the field, but they also provide new insights on the open availability of research through other routes—mainly due to the academic social network ResearchGate, but also personal websites and harvesters. They discovered 23.1% available as Gold, Hybrid, Delayed, or Bronze OA, 17.6% available as Green OA and 40.6% available from other sources. The study is a further indication of the multiple host locations available to researchers when disseminating their work. This study aims to enhance Brunel University London's publication data by using software to retrieve and process data from open science services. The data produced develops from previous studies by exploring the complexities in OA choice for authors wishing to submit papers and also investigates OA volume and multiplicity. This encompasses the following objectives:

- Create OA classification schemes to interpret the data; a 'best location' hierarchy to measure publishing trends and whether open access dissemination has taken place, and a relational 'all locations' dataset to examine whether individual publications appear across multiple OA dissemination models.
- Investigate possible OA locations from serial policies described in Sherpa REF resources.
- Investigate OA locations from CORE resources and plot the geolocation of works.
- Investigate OA locations and volume from related Unpaywall and CORE resources.

From the data and classifications, this paper explores if local institutional systems capture the full range of OA publishing activity made visible by these new services and considers whether a more vibrant picture of OA activity may be recorded. HEFCE's policy emphasises the importance of on-acceptance deposit for manuscripts and metadata for discoverability. We, therefore, apply the data capture methodology and classifications to published and accepted research outputs in the sample to provide a comparative view of OA dissemination by Brunel researchers at the point of acceptance in addition to publication.

This research is important because UK higher education institutions (HEI) are currently seeking to accurately report on both Green and Gold OA activities by their researchers. Additionally, a comprehensive understanding of OA behaviours amongst researchers is crucial for institutions as they seek to provide tailored scholarly communication training and services to their academic staff. Localised and manual processes, which are detailed and resource intensive, are currently necessary to provide an institutional picture of OA compliance. This study explores how the adoption of emerging open science services that hold OA metadata and manifestations may streamline processes and reduce the administrative burden for academics and support staff.

The data generated highlights wide-ranging engagement with OA across a variety of models. However, it also indicates potential duplication of effort in OA workflows with single publications appearing in multiple locations. From the data, we consider whether current workflows—and particularly those surrounding HEFCE's OA Policy—may assume such heavy focus on local IRs that they do not entirely register other methods for compliance such as deposits within disciplinary repositories, the IRs of co-authors, or Gold routes. From an institutional perspective, the intense focus on IRs is understandable as this currently provides the most robust and controllable mechanism for reporting on compliance levels at a given institution. This paper does not, therefore, criticise UK institutions for their responses to the current policy environment, but instead investigates how new open science services can be applied to increase the accuracy of institutional reporting on OA compliance, especially in helping to validate existing manual work undertaken by library staff.

2. Materials and Methods

Brunel University London manages a Symplectic Elements Current Research Information System (CRIS) to record all publication and research activity by the organisation. The system links data from authority sources including ORCID, CrossREF, Scopus, and Web of Science. Records may also be input manually, thereby increasing the local visibility of research activity for disciplines where coverage in authority sources is incomplete. For the study, the institution provided a data sample from their Elements system; the complete set of journals articles and conference proceedings from the period January 2014 to December 2017. This comprised outputs of 5570 published and 403 accepted works. There is more data available for the on-publication view as the organisation began collecting on-acceptance metadata in 2016 to coincide with the activation of HEFCE's OA policy. Works are defined as published when they have an early-online or publication issue date in the metadata. Works are defined as accepted, when they have no publication date, but have included a date of acceptance. Data were retrieved from the open science services using attributes from the Brunel sample by a small software program customised by the authors.

The Elements dataset contains some OA data attributes used in the analysis. These are:

- DOAJ; matched on ISSN
- SHERPA RoMEO; identifies publisher policy 'banding', matched on ISSN
- Records and file locations from some large repositories; Europe PMC, ArXiv
- Deposits in local IR (Brunel University Research Archive)

Sherpa REF service data provides possible OA locations from journals in the sample. We investigate only possible locations that enable OA and ignore embargo, cost, or funder policy restrictions. We omit 'submitted' article version types from the results due to the community focus on accepted/published versions. CORE service data identifies OA locations and provides additional location information such as the names of host resources where Brunel content is held and geolocation data.

Unpaywall service data identify a 'best' OA location hierarchy. Unpaywall qualifies a 'best' OA location for an OA work by prioritising publisher-hosted content first (for example, Hybrid or Gold), and then versions closest to the version of record [9]. A 'best location' can help libraries answer the typical compliance question of 'is it Gold or Green?' All location objects that relate to works are retrieved from the service to ascertain the total OA volume. 'Closed' access publications are inferred where there is no discoverable OA location from any data source.

Using a small Java SE software program [22], attributes from the data sample (DOIs and serial ISSNs) provided the inputs to retrieve resources from the Unpaywall, CORE, and Sherpa REF REST APIs and return OA data at the article or serial level. The software manages data in the publication and open access domain. The software replicates data from adjacent systems (that is, business systems like Symplectic Elements) and uses data attributes to make HTTP requests and access the resources of

OA RESTful web services. This data is persisted in a MySQL relational database using the Hibernate framework [23]. Entities and relationships are described in the data model in Figure 1.

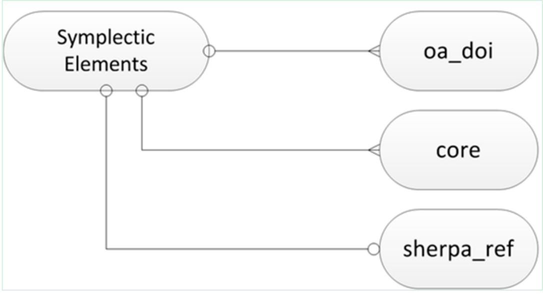


Figure 1. The conceptual data model, showing all entities and relationships used in the data analysis. Unpaywall stopped using the OaDOI brand in January 2018, after the data for this study was collected.

All software [24] and related data outputs [25] produced during the study are published online and described in the Supplementary Data Section. Data analyses are the product of SQL relational queries on the MySQL database. For example, one SQL query produced an enhanced OA publication dataset by joining the Elements, CORE, and Unpaywall entities together in a relational query [26]. Logical conditions test attribute values and this identifies the type of OA according to the classifications defined in the methodology outlined below. As mentioned in the Introduction, the study uses two classification schemes; 'best location', which highlights publishing trends and whether OA dissemination has taken place, and 'all locations', indicating the multiplicity of works across OA dissemination models.

The 'best location' classification extends Unpaywall's definitions but with some fundamental differences between the Gold, Green, and Closed access subtypes (see Table 1). The types enable us to measure publishing trends and open access dissemination.

Table 1. The open access classification (1) by 'best' available location.

Hierarchy	Classification (1): "Best Location"	Description
1	Gold	A publisher hosted 'Gold' OA location
2	Hybrid	A publisher hosted 'Hybrid' OA location
3	Green (External)	A repository hosted OA location (external system)
4	Green (Brunel IR)	A repository hosted OA location (Brunel IR)
5	Closed (Green option)	The paper is not OA, but a Green OA option is available
6	Closed (publisher)	The paper is not OA, and no Green OA option is available

Like Unpaywall, we monitored the OA publisher content first (Gold OA and Hybrid), followed by external disciplinary and institutional repository locations and, finally, Brunel's institutional repository.

Unpaywall data does not explicitly distinguish Bronze OA from Gold and Hybrid classifications in its API. Rather, it attributes it in its data model (including host location descriptors, publication licences, and Gold journal indicators) and enables these types to be derived. For example, the service matches Gold locations by ISSN entries in the DOAJ. The API documentation says that this indicator for Gold OA journals is 'useful for most definitions of Gold OA. Currently this is based entirely on the inclusion in the DOAJ but eventually may use additional ways of identifying all-OA journals' [9]. In our method, we subsume Piwowar and Priem's (et al.) Bronze classification within the Gold and Hybrid types because Bronze characteristics are prevalent within both publishing models, that is, documents that are free-to-read from the publisher, but do not have an explicit OA licence. As discussed in the introduction, we do not wish to conflate Bronze with Gold or Hybrid OA. Therefore, in this study, we examine the licence data retrieved by the Unpaywall service to inform our discussions on any Bronze OA characteristics displayed by publications in the sample.

SHERPA RoMEO data, sourced within Elements, use a Yellow and Green banding scheme for serial titles, which is used to indicate whether an archival option for peer-reviewed content is available. Otherwise, we consider no compliant archival option as available. In this way, RoMEO data distinguishes closed access works between 'publisher' and 'author,' that is, where a journal policy prohibits archiving and where an author has apparently failed to take up a possible Green OA location. This distinction can help libraries target researchers to achieve OA where this is possible and highlight publisher stakeholders who are not engaged in the movement. The 'best location' classification is derived from the dataset by comparing the values of OA attributes from the various data sources. Supplementary Data [27] describes the tests performed against the attributes to derive the classification.

Unpaywall data describes unique publisher host locations and the multiplicity of repository hosts as location objects. The total number of location objects enable an OA 'count' of published works. To distinguish disciplinary repositories and IR locations, we excluded outputs found by Unpaywall that have an oai:bura identifier (of Brunel's local IR—the Brunel University Research Archive). We instead use internal records via Elements to measure engagement with Brunel's IR. This is due to embargoed works not being returned by Unpaywall and because the aim of this study is to maximise the identification of OA locations.

By establishing a comprehensive, relational dataset on all OA publishing activity discovered by the services, we can describe an 'all location' classification to explore the possible effects of UK's Gold and Green policies (see Table 2):

Table 2. The open access classification (2) by all available locations.

Classification (2): "All Locations"	Description
Gold	A publisher hosted OA location ('Gold'/'hybrid') and no repository OA locations
Gold and Green	A publisher hosted OA location ('Gold'/'hybrid') and 1 or more repository OA locations ('External'/'Brunel IR')
Green	No publisher hosted OA location and 1 or more repository OA locations ('External'/'Brunel IR')

From this data and the classification schema, we may infer whether duplication is evident in OA dissemination for Brunel University London. UK funder policies may encourage this outcome because in combination they preference both publisher (RCUK) and IR (HEFCE) OA locations and this effect may be exacerbated by research collaborations across institutions and territories. The variance in publisher archival policies may also increase the likelihood of multiple OA locations for a single work.

3. Results

This section is partitioned by the data and analyses of the specific open science services used in the study.

3.1. Sherpa REF Data

The Sherpa REF data highlights the scale of possible OA locations available to authors when making dissemination decisions for their research. The Brunel data sample comprised of 5570 published journal articles and conference proceedings; 4475 records had ISSN or ESSN attributes. Of these, there were 2154 distinct serial ISSN or ESSNs. Sherpa REF returned 1808 records. The service covered 84% of serials in the sample.

The possible locations derive from the 'route', 'version', and 'type' attributes within the data. Combining these values reveals the count and description of possible locations that may be selected by authors and that are described by publisher policies:

- Open access 'route': 'Publish', 'Hybrid', and 'Archive', which refer to 'Gold', 'Hybrid', and 'Green' OA publishing models respectively
- Article 'version': Accepted or Published

- Repository 'type': Disciplinary, Institutional, Other.

From the serials in the sample, the most common OA location is the archiving of an accepted manuscript into an institutional repository. The results demonstrate that many publisher policies align with the requirements of UK funding bodies. However, the data shows 34 total OA scenarios for authors [28]. There is an average of 4.767 permissible open access options in the publications selected by our authors. A total of 6% of titles had 8 or more possible OA locations available to authors. Additionally, 99% of journals in the sample had at least two possible OA locations. As shown in Table 3, a small number of publications had as many as 13 possible OA locations, such as the *Journal of Applied Physics* (0021-8979).

Table 3. The Sherpa REF results: open access dissemination options for the *Journal of Applied Physics* (AIP Publishing).

Count	Sherpa REF Results
1	Archive, Accepted Version, Author Website
2	Archive, Accepted Version, Department website
3	Archive, Accepted Version, Institutional Repository
4	Archive, Accepted Version, Other
5	Archive, Accepted Version, Disciplinary repository
6	Archive, Published Version, Author Website
7	Archive, Published Version, Department website
8	Archive, Published Version, Institutional Repository
9	Archive, Published Version, Other
10	Archive, Published Version, Disciplinary repository
11	Hybrid, Published Version, Author Website
12	Hybrid, Published Version, Institutional Repository
13	Hybrid, Published Version, Disciplinary repository

The results may indicate that authors are engaged in many OA workflows, given the availability of options per publishing policy. The number of institutions and authors involved in collaborations may increase this possibility. We conclude that the complexities and range of publisher policies, combined with funder and institutional policies, may create opportunities for duplication in OA locations.

3.2. CORE Data

CORE data finds multiple OA locations hosting content affiliated to Brunel University London. The data shows 131 hosts with Brunel research holdings [29]. The service found 1894 OA locations for 1073 distinct publications. CORE records contain additional geolocation data, enabling physical host locations to be plotted on a map.

Figure 2a plots data from repository hosts, which tend to include coordinates specifying the physical location of the institution or repository. Readers of this study may wish to access the interactive online publication of this map, which plots the coordinates along with the names of OA works and host locations—see Supplementary Data 'Collection B: Artefacts'.

Figure 2b,c plots data from publisher hosts, which tend to include country code data, assumed to reflect their general base of operations. This data is available here and published online [29–31]. The visualisations provide some insight into the global distribution of Brunel's research.

The highest concentration of OA locations is in European territories. The results show collaboration in OA dissemination, likely reflecting the international research process and author networks. The results discovered 370 distinct works in multiple OA locations. CORE found one publication, 'Improving the experience of dementia and enhancing active life-living well with dementia: Study protocol for the IDEAL study' (2014), available in 9 OA locations. This data may be incomplete because the service harvests only full-text PDF content from hosts and no other document formats. For example, a Google Scholar search for this paper finds 24 OA locations.

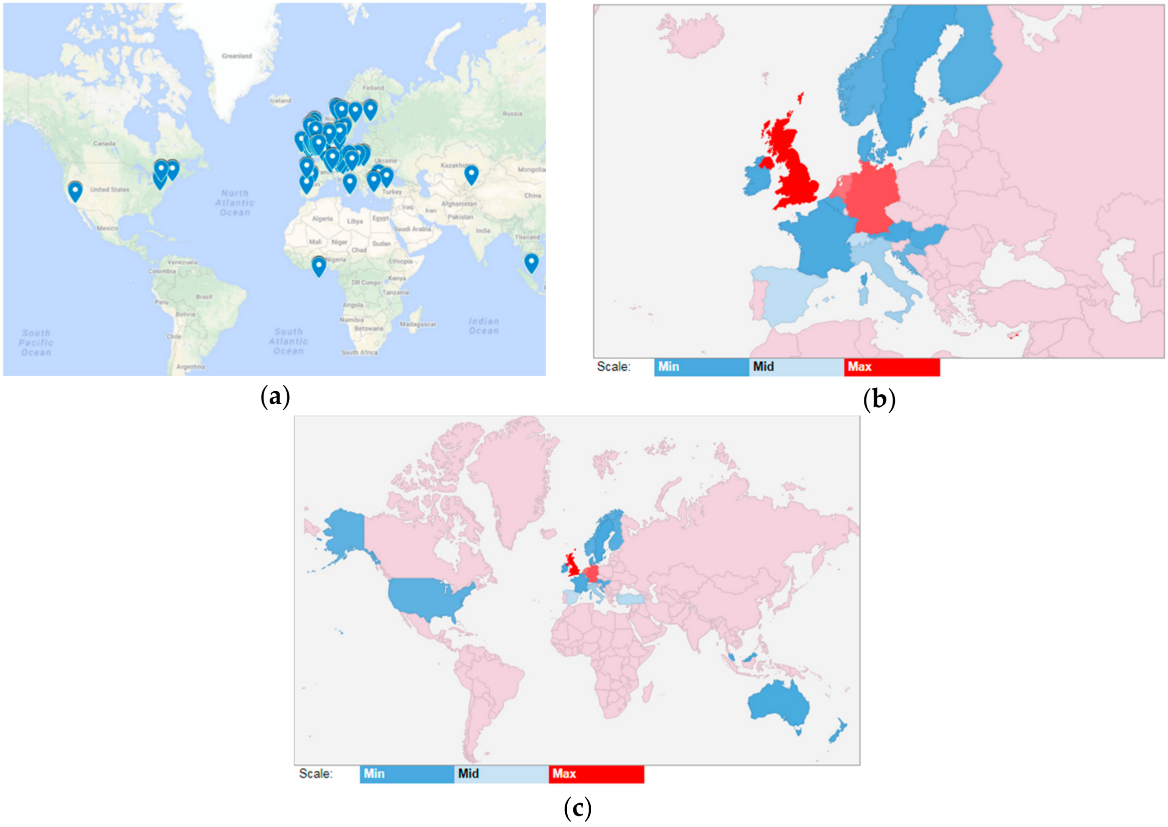


Figure 2. The geolocation data for Brunel publications found within CORE (**a**) Brunel publications by repository host locations (latitude and longitude coordinates); (**b**) Brunel publications by repository host locations in Europe (country code); (**c**) Brunel publications by repository host locations globally (country code).

3.3. Enhanced Publications Dataset and Unpaywall Data

The results in this section describe Brunel's institutional sample enhanced by open science services [32]. This combined Unpaywall, CORE and Elements data to investigate complete OA locations for the institution's research.

Unpaywall found 1589 publisher hosts with OA locations and 1563 repository hosts [33]. The results also show collaboration in OA dissemination. A total of 821 publications were discoverable in multiple OA locations. In this data, the most highly replicated paper was from the CMS collaboration entitled 'Search for pair production of excited top quarks in the lepton+ jets final state', which was available in 9 OA locations. Like CORE, we found that this data may also be incomplete. A comparison with search results from Google Scholar for this paper finds 57 OA locations.

3.3.1. 'Best Location' Classification Data

Using the 'best available' OA classification scheme, we produced a data set of 5973 institutional publication records enhanced using Unpaywall data [32]. An 'on-publication' view shows increasing trends of Gold, Hybrid, and Green. Figure 3a shows an increase mainly in the IR deposit and Hybrid publications in 2017. Gold levels appear stable across time.

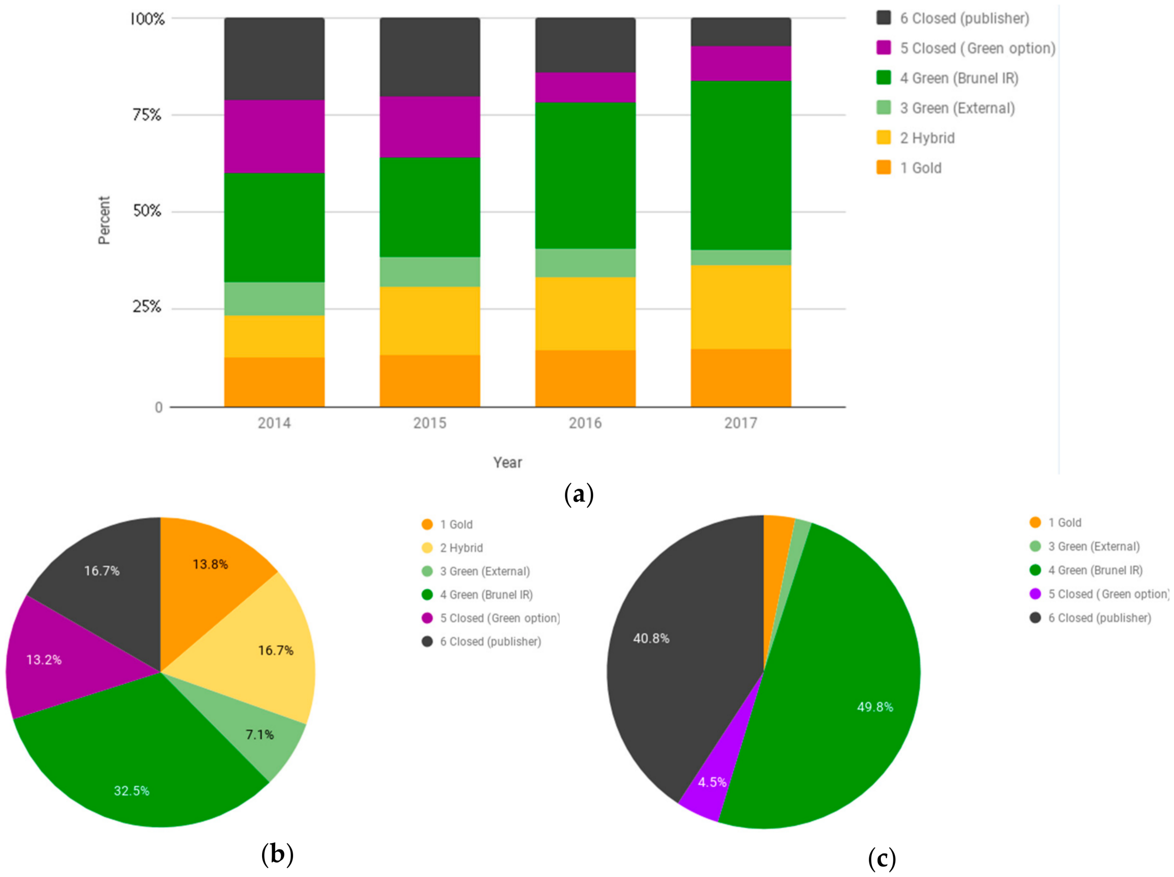
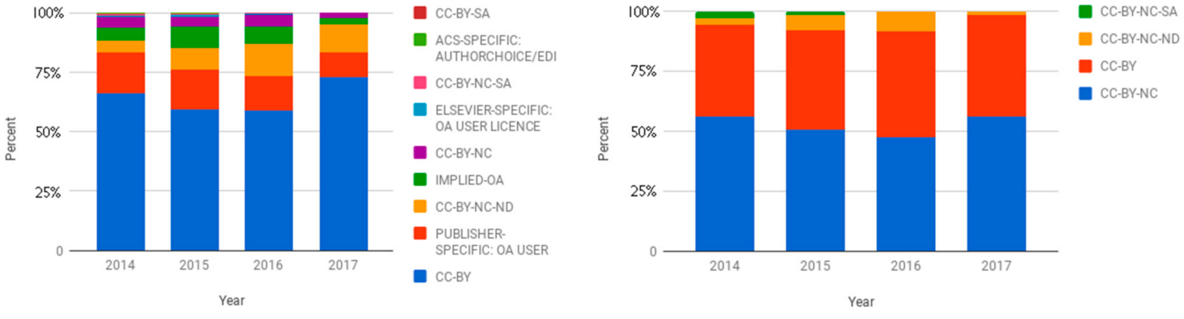


Figure 3. The enhanced publication dataset ("Best available" classification) 2014–2017, (**a**) Brunel OA dissemination by year (percentages on publication); (**b**) Brunel OA dissemination 2014–2017 (percentages on publication); (**c**) Brunel OA dissemination 2014–2017 (percentages on acceptance).

Figure 3a,b show that the institution has levels of OA consistently higher than Piwowar and Priem's (et al.) global baseline of 45% OA for recent works [16]. A small subset of publications require action to be made Green open access—that is where Sherpa RoMEO indicates a compliant archival option and where an OA location has not been established.

Figure 3b,c show the comparative view of OA for published and accepted works, respectively, from the sample of enhanced records. The 'on publication' view shows 13% with a possible Green OA location apparently not taken by the authors—in 2016 (when HEFCE's policy became active), this was just 7.7%. Comparatively, the 'on acceptance' view indicates a much-reduced volume and diversity of OA dissemination. This may be because manual entries made by authors, to meet institutional compliance workflows, lack the critical attributes that enable the automated, hierarchical checks that are possible on publication, such as DOIs and ISSNs.

Unpaywall data include OA publishing licences across publisher and repository hosts. Figure 4a shows the 'Creative Commons with Attribution' (CC-BY) as the dominant licence selected for Gold and Hybrid works, which may indicate the impact of UK and European funder policies. Nonetheless, many demonstrate Bronze OA properties with some publishers clearly retaining rights [32] (see 'Elsevier specific: OA user licence'), and others where licences cannot be established (see 'Implied-OA'). We found that 28% of Gold or Hybrid works in the sample are available with Bronze OA characteristics, as described by Piwowar and Priem (et al.) [16].



(**a**) (**b**)

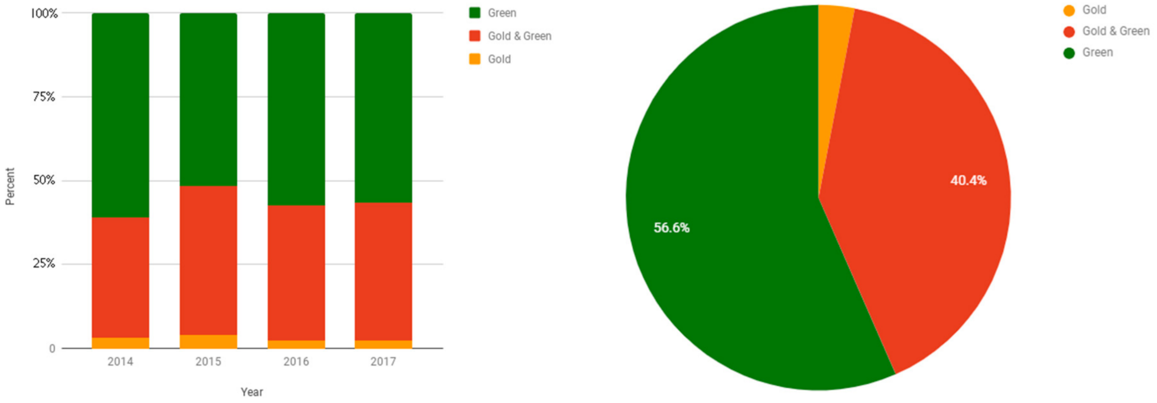
Figure 4. Unpaywall: Licence adoption by publisher and repository host locations: (**a**) publisher licences; (**b**) repository licences.

Of the 1563 repository host locations returned by the Unpaywall service, 1122 did not return any licence data [33]. This accounts for 71% of repository hosted OA locations found by Unpaywall. Figure 4b shows that where this data was identifiable by the service, non-commercial licences are the most widely adopted. The lack of Green repository licences suggests there may be difficulties in Unpaywall identifying these or it may indicate that repositories adopt stricter licences on reuse than the RCUK and COAF standards. Where copyright allows, the Brunel University Research Archive applies a CC-BY licence to deposits.

3.3.2. 'All Location' Classification Data

The 'all location' OA classification scheme [32] demonstrates the effect of Gold and Green policies on access workflows. Total OA locations for individual published works are also extracted from the data.

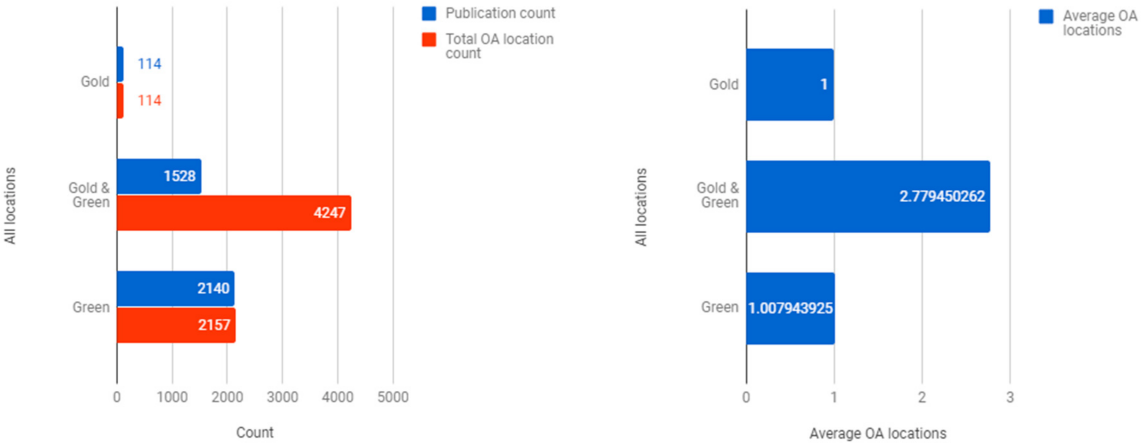
Figure 5 shows that 40% of OA outputs are available in Gold and Green host locations. Most Gold published outputs are available in other Green locations.



(**a**) (**b**)

Figure 5. The enhanced publication dataset ('All location' classification) 2014–2017, (**a**) Brunel OA dissemination by year (percentages on publication); (**b**) Brunel OA dissemination 2014–2017 (percentages on publication).

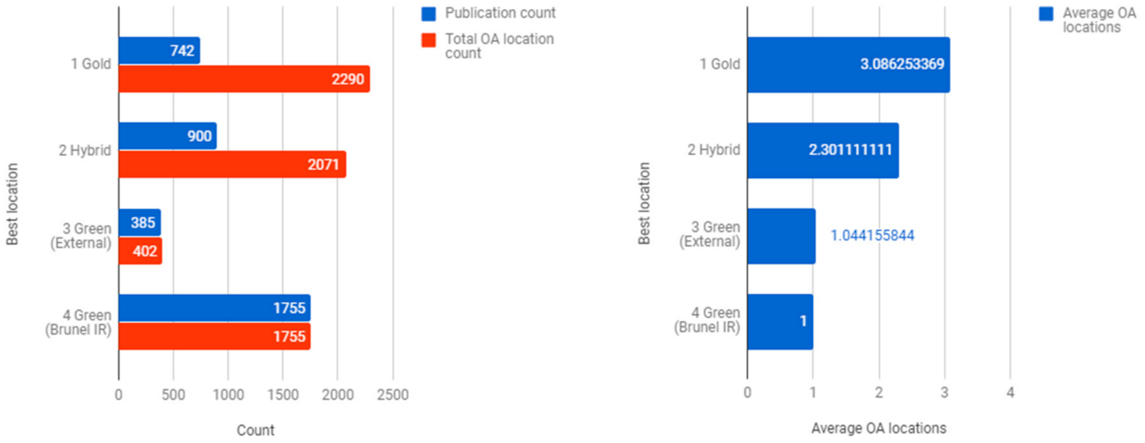
Figure 6a shows a comparison of the total OA locations against the count of distinct publications. There appears to be a significant volume of OA locations where Gold and Green hosts are available. In this scenario there is an average replication of OA works in 2.78 locations.



(**a**) (**b**)

Figure 6. The total OA locations ('All location' classification) 2014–2017, (**a**) total OA locations; (**b**) average OA locations.

Figure 7 repeats this approach using the 'best location' classification. This shows Gold publications with the highest levels of replication, with an average availability of 3.08 locations. Hybrid has an average availability of 2.3 locations. The data on exclusive Green publications does not indicate any duplication.



(**a**) (**b**)

Figure 7. The total OA locations ('Best available' classification) 2014–2017, (**a**) total OA locations; (**b**) average OA locations.

In combination, the figures show a significant proportion of outputs singularly available via Green locations—which underlines the importance of the HEFCE OA policy and the repository networks—and a significant replication where a Gold publication has taken place.

4. Discussion

This study aimed to enhance Brunel University London's publication records using data from open science services. The results show high levels of OA for the institution. The results also reveal OA activity external to institutional systems. This study resultantly contributes to two broad discussions for UK university staff involved in OA compliance reporting and service delivery.

Firstly, dialogue may focus on localised systems and whether they can be wholly relied upon to provide an accurate picture of OA behaviours amongst academic staff. Secondly, this study's ability to use emergent open science services to enhance institutional data suggests that there may be the opportunity for greater interoperability between emerging services such as Unpaywall, CORE and Sherpa REF and established providers of IRs and research information systems.

An objective of the study was to investigate possible OA locations from serial policies described in related Sherpa REF resources. This provided insight into the complexities of OA choice prior to OA dissemination. The data in Section 3.1 demonstrates that authors have multiple OA options available to them, which can be attributed to the success of the OA movement in producing a range of platforms for the open dissemination of research. The variety of dissemination channels available may increase the likelihood of duplication, with multiple versions of any given paper potentially existing on a mixture of disciplinary repositories, IRs, social platforms and publisher websites. Some publications in the sample offer as many as 13 possible OA locations for peer-reviewed works.

This study also generated a 'best location' hierarchy to measure publishing trends and to determine whether open access dissemination had taken place, and a relational 'all locations' dataset to examine whether individual publications appear across multiple OA dissemination models. As set out in the introduction, there are challenges when classifying OA works, particularly in light of Piwowar and Priem's (et al.) Bronze OA definition. In Section 3.3.1 we used licence data to investigate reuse trends across Gold and Hybrid types. The results may indicate a positive, cultural impact of RCUK policy as CC-BY is the dominant licence selected for Gold and Hybrid works.

However, as with Piwowar and Priem's (et al.) discoveries, we found that 28% of Gold or Hybrid works in the sample display Bronze OA characteristics. Access by authors to other funding sources and the collaborative research environment may leave little recourse for intervention by institutional services. It may also be the case that in some circumstances researchers may not be granted any form of open licence by their journal. Such reuse limitations may prevent replication of the final published version in repository hosts. However, the evidence provided in this study suggests this may not be the case as we found that almost all Gold and Hybrid works in our sample are available in other OA locations, irrespective of Bronze publication indicators. It may be that, in practice, reuse is not prevented by rightsholders. The reasoning behind Gold OA publishers retaining copyrights to works is not fully understood by the sector and requires further study.

Although providing the greatest levels of access, the reusability of repository content is much less clear from the data. Section 3.3.1 shows that 71% of repository hosts discovered by Unpaywall do not return any licence information and that non-commercial licences were the most common of the remaining set. HEFCE's policy advises a CC-BY-NC-ND license to satisfy minimum reuse requirements [4]. Concerning repository hosts, RCUK's policy does not require a specific licence for green OA, only that there are no publisher restrictions on non-commercial re-use. This can be met by a Creative Commons Attribution Non-commercial (CC-BY-NC) licence. Further research may be useful to explore the effects of stricter restrictions on the use of repository contents. Such investigations may inform future policy direction.

Described in the methodology, the 'best location' classification is inclusive of closed OA works, distinguishing author and publisher. This is an important consideration often overlooked by OA research. In Section 3.3.1, the 'Closed (Green option)' category indicates the 'gap' to 100% legal Green OA, accounting for just 7.7% of Brunel publications in 2016. By targeting support towards the authors of these works, institutional services may have the greatest opportunity to improve compliance figures and boost the OA agenda.

In contemplating how we may drive greater levels of access for the 'Closed (publisher)' category (that is, where no legal Green archival option was found to exist), we must also consider the lack of parity in access to Gold and Hybrid funding by researchers that is exacerbated by rising APC and subscription costs [17]. Additionally, there is a specific challenge presented by small society publications, who may argue that they require subscription models to sustain their activities. Initiatives like SCOAP3, the Open Library of the Humanities, and offsetting deals like the 'Springer Consortium agreement' may begin to redress this imbalance by shifting costs from authors to institutions. However, financial difficulties may continue to prevent the transformation of the remaining literature to 'open'. Advocacy activities by libraries may help authors to consider these issues when selecting their publisher.

The 'all location' classification examined OA location multiplicity and possibly highlights the impact of mandatory policies at an institutional level. Section 3.3.2 found that 40% of OA publications have taken Gold and Green routes. Whilst this duplication is not problematic *per se*, it does raise questions about the extent to which local institutional workflows fully capture the diversity of OA activity. The data in Sections 3.2 and 3.3 show that aggregation services may help to increase institutional compliance figures by increasing the visibility of OA dissemination. Indeed, emerging open science services may run concurrent to localised workflows necessarily imposed by UK institutions to ensure compliance. CORE data shows 131 hosts that have Brunel research holdings [29].

Section 3.3.2 found Gold works to have the highest number of OA locations, on average, followed by Hybrid. This may be due to publisher-driven archival policies. Gold models increase the likelihood that the journal will share content directly with disciplinary repositories. For example, Biomed Central mirrors all content into the PubMed platform [34]. However, significant replication of Gold and Hybrid publications may also indicate that the policy terrain is driving a duplication of effort in access workflows. Further study may explore whether the most affected are funded authors in the UK, who have specific mandatory requirements. We found no evidence of duplication within exclusively Green works. This indicates that HEFCE's policy may be successfully driving OA access to works that might otherwise be closed.

Presently, certain aspects of policy and the effectiveness of internal procedures are only measurable by engagement with local IRs/research information systems. This is partly because certain administrative data are not defined within global bibliographic and metadata standards. For example, the 'date of deposit' is not part of the RIOXX standard, and has only become important recently for UK institutions due to HEFCE's policy and audit requirements. At the time of writing, there are no indicators that this data will be adopted within global metadata standards, as it regards system administration and has yet to become a mainstream descriptor of a scholarly work. New, existing, and growing OA cultures in disciplinary repositories and other researcher-led communities may now be disregarded by UK institutions because only internal workflows that engage with local IRs/research information systems can capture the necessary administrative data to mitigate the perceived risk to the institution.

Section 3.3.1 highlights the lack of OA information available for accepted works. This lack of information may drive a sense of risk within institutions and possibly inform manual and localised workflows and strategies. Tools like Unpaywall currently only accept DOI identifiers that are maintained by DOI providers and these correspond to published works. This particular finding may, therefore, strengthen calls for publisher engagement with the Jisc Publications Router, a service to transfer on-acceptance metadata and manuscripts between publishers and research systems. However, at the time of writing, just five publishers are listed on the website as content providers and three of these are Gold OA publishers [35].

A further challenge facing institutions is that authors frequently change employers and import publications and records into local systems as they change jobs. Further studies could look at the impact of author mobility on OA compliance, and will probably add further weight to a need for global OA solutions. CORE data [29,30] show the majority of external host OA locations are spread across external IR systems in the UK/EU such as 'Spiral' (Imperial College London) and King's College London's Research Portal.

Sections 3.2 and 3.3 describe a spot check comparison with Google Scholar. This indicated many more OA locations available than shown within the CORE and Unpaywall data. Martín-Martín's (et al.) research [21] has highlighted the extent to which these OA locations are funder assured, such as disciplinary and institutional repositories, and not social networking repositories, such as ResearchGate, or uploads to project websites. Further study may explore the extent to which 'other' OA locations benefit the discovery of research. Large-scale and reproducible data in this area may also inform UK and other funder policies, such as ongoing reviews by Research England and the Wellcome Trust [36,37].

At the broadest level, this paper has highlighted a tension between the reality of research as a highly collaborative enterprise, with partnerships that cross institutional and territorial boundaries, and the requirements of nationally focused policies. This makes monitoring OA activity for compliance difficult as institutions tend to rely on local IRs and research information systems to report back to national funders. Approaches by institutions—in the UK at least—are therefore understandably driven by reactions to RCUK (now named UKRI) and HEFCE (now named Research England), with the former displaying a preference for Gold OA and the latter favouring repository usage. This study acknowledges the significance of these policies, which have been crucial in embedding OA within the minds of authors. However, the OA landscape comprises of a vast network of repositories, social systems, and open initiatives by new and longstanding publishers, which is not easily captured within local systems. The results of this paper, therefore, point to the need for interoperability between local systems and emergent open science services which are attempting to aggregate OA activity. Such two-way communication has the potential to enhance data for compliance reporting, limit the possibility of manual duplication, and free up library staff to target effort on those who have not yet engaged with OA.

Supplementary Materials: The data listed below is available online within the following collections:

Collection A

- Walters, D.; Daley, C. Exploring researcher engagement with open access using emergent open science services: Software Artefacts, 2018, *figshare*. doi:10.6084/m9.figshare.c.3966030

Collection B

- Walters, D.; Daley, C. Exploring researcher engagement with open access using emergent open science services: Data Artefacts, 2018, *figshare*. doi:10.6084/m9.figshare.c.3966027

Collection A: Artefacts

- Walters, D. Open access publishing data: File conversion and retrieval software, 2018, *figshare*. doi:10.6084/m9.figshare.5774244
- Walters, D. 'Enhanced' OA publications data SQL query, 2018, *figshare*. DOI: 10.6084/m9.figshare.5808375
- Walters, D. 'Best location' OA classification tests, 2018, *figshare*. doi:10.6084/m9.figshare.5892748
- Walters, D. Sherpa REF: SQL query and data sample, 2018, *figshare*. doi:10.6084/m9.figshare.5799339
- Walters, D. CORE: SQL query and data sample, 2018, *figshare*. doi:10.6084/m9.figshare.5799336
- Walters, D. Elements: SQL query and data sample, 2018, *figshare*. doi:10.6084/m9.figshare.5799222
- Walters, D. Unpaywall: location SQL query and data sample, 2018, *figshare*. doi:10.6084/m9.figshare.5799303
- Walters, D. OA database MySQL dump-table structure, 2018, *figshare*. doi:10.6084/m9.figshare.5765499
- Walters, D. Report: A Java application to create and persist objects from XML data, and interact with 'open access' RESTful web services, 2018, *figshare*. doi:10.6084/m9.figshare.4887011

Collection B: Artefacts

- Walters, D. Live google map plotting co-ordinates of repository outputs for Brunel University (2014–2017), 2018. figshare. doi:10.6084/m9.figshare.5947855
- Walters, D. Google sheets: CORE location data and figures, 2018, *figshare*. doi:10.6084/m9.figshare.5820753
- Walters, D. Google sheets: 'Enhanced' OA publications data and figures, 2018, *figshare*. doi:10.6084/m9.figshare.5799342

Author Contributions: Conceptualisation: D.W. and C.D. Methodology: D.W. Software: D.W. Formal Analysis: D.W. and C.D. Data Curation: D.W. Visualisation: D.W. Writing-Original Draft Preparation: D.W. and C.D. Writing-Review and Editing: D.W. and C.D. Project Administration: C.D.

Acknowledgments: The study received no sources of funding. The authors wish to thank Monique Ritchie and Margaret Weaver for their diligent feedback and copyediting of the draft.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suber, P. *Open Access*; MIT Press: Cambridge, MA, USA, 2012.
2. Budapest Open Access Initiative, 2002. Available online: http://www.budapestopenaccessinitiative.org/read (accessed on 15 May 2018).
3. ROARMAP. Available online: http://roarmap.eprints.org/ (accessed on 1 February 2018).
4. Higher Education Funding Council England. Policy for Open Access in Research Excellence Framework 2021, November 2016. Available online: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2016/201635/HEFCE2016_35.pdf (accessed on 19 February 2018).
5. Research Councils UK. RCUK Policy on Open Access and Supporting Guidance, 12 September 2017. Available online: https://www.ukri.org/files/legacy/documents/rcukopenaccesspolicy-pdf/ (accessed on 9 May 2018).
6. Kingsley, D. Could the HEFCE Policy Be a Trojan Horse for Gold OA? *Unlocking Research Blog*, 2016. Available online: https://unlockingresearch.blog.lib.cam.ac.uk/?p=488 (accessed on 19 February 2018).
7. Kramer, B.; Bosman, J. 101 Innovations in Scholarly Communication—The Changing Research Workflow. *Figshare* **2015**. [CrossRef]
8. Walters, D.; Daley, C. Brunel's 10 Year Journey towards Open Scholarship: Measuring 'Openness' over Managing Mandates. *Hindawi Opinion Blog*, 2016. Available online: https://about.hindawi.com/opinion/brunels-10-year-journey-towards-open-scholarship-measuring-openness-over-managing-mandates/ (accessed on 19 February 2018).
9. Unpaywall API Documentation Version 2. Available online: http://unpaywall.org/api/v2 (accessed on 19 February 2018).
10. JISC Sherpa REF. Available online: https://www.jisc.ac.uk/rd/projects/sherpa-ref (accessed on 19 February 2018).
11. Tennant, J.; Mounce, R. Open Research Glossary. *Figshare* **2015**. [CrossRef]
12. Knoth, P.; Zdrahal, Z. CORE: Connecting Repositories in the Open Access Domain. CERN workshop on Innovations in Scholarly Communication (OAI7): Geneva, Switzerland, 2017. Available online: http://oro.open.ac.uk/32560 (accessed on 19 February 2018).
13. Knoth, P.; Anastasiou, L.; Basile, G.; Pearce, S.; Pontika, N. Machine Accessibility of Open Access Scientific Publications from Publisher Systems via ResourceSync. *OAI10*, 2017. Available online: http://oro.open.ac.uk/50181 (accessed on 19 February 2018).
14. Walters, D.; Daley, C. 'Measuring' and Managing Mandates. *CORE Blog*, 2016. Available online: https://blog.core.ac.uk/2016/07/07/measuring-and-managing-mandates/ (accessed on 19 February 2018).
15. Bosman, J.; Kramer, B. Open Access Levels: A Quantitative Exploration Using Web of Science and oaDOI data. *PeerJ Preprints* **2018**, *6*, e3520v1. [CrossRef]
16. Piwowar, H.; Priem, J.; Larivière, V.; Alperin, J.P.; Matthias, L.; Norlander, B.; Farley, A.; West, J.; Haustein, S. The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* **2018**, *6*, e4375. [CrossRef] [PubMed]
17. Jubb, M.; Plume, A.; Oeben, S.; Brammer, L.; Johnson, R.; Butun, C.; Pinfield, S. Universities UK. *Monitoring the Transition to Open Access*, 2017. Available online: http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2017/monitoring-transition-open-access-2017.pdf (accessed on 9 April 2018).
18. Himmelstein, D.; Romero, A.R.; Levernier, J.G.; Munro, T.A.; McLaughlin, S.R.; Tzovaras, B.G.; Greene, C.S. Sci-Hub provides access to nearly all scholarly literature. *eLife* **2018**, *7*, e32822. [CrossRef] [PubMed]
19. ADA News. Open Access to Select Articles in Prosthodontic Journal until May 31, 2018. Available online: https://web.archive.org/web/20180416110028/https://www.ada.org/en/publications/ada-news/2018-archive/april/open-access-to-select-articles-in-prosthodontic-journal-until-may-31 (accessed on 12 April 2018).
20. Priem, J.; Piwowar, H. The Unpaywall Dataset. *Figshare* **2018**. [CrossRef]
21. Martín-Martín, A.; Rodrigo Costas, T.; Emilio, D. Evidence of Open Access of Scientific Publications in Google Scholar: A Large-scale Analysis. *ArXiv* **2018**. [CrossRef]
22. Walters, D. Open access publishing data: File conversion and retrieval software. *Figshare* **2018**. [CrossRef]
23. Hibernate. Hibernate ORM. Available online: http://hibernate.org/ (accessed on 8 May 2018).
24. Walters, D. Analysing the 'State of Open Access' at Brunel University London (2018): Software Artefacts. *Figshare* **2018**. [CrossRef]

25. Walters, D. Analysing the 'State of Open Access' at Brunel University London (2018): Data Artefacts. *Figshare* **2018**. [CrossRef]
26. Walters, D. 'Enhanced' OA publications data SQL query. *Figshare* **2018**. [CrossRef]
27. Walters, D. 'Best location' OA classification tests. *Figshare* **2018**. [CrossRef]
28. Walters, D. Sherpa REF: SQL query and data sample. *Figshare* **2018**. [CrossRef]
29. Walters, D. CORE: SQL query and data sample. *Figshare* **2018**. [CrossRef]
30. Walters, D. Live google map plotting co-ordinates of repository outputs for Brunel University (2014–2017). *Figshare* **2018**. [CrossRef]
31. Walters, D. Google sheets: CORE location data and figures. *Figshare* **2018**. [CrossRef]
32. Walters, D. Google sheets: 'Enhanced' OA publications data and figures. *Figshare* **2018**. [CrossRef]
33. Walters, D. Unpaywall: Location SQL query and data sample. *Figshare* **2018**. [CrossRef]
34. Biomed Central. What Is the Relationship between BioMed Central, PubMed Central and PubMed? Available online: https://web.archive.org/web/20160708212906/https://old.biomedcentral.com/about/faq/pubmed (accessed on 19 February 2018).
35. Jisc Router. Current Content Providers. Available online: https://pubrouter.jisc.ac.uk/about/providerlist/ (accessed on 19 February 2018).
36. Research England. Research England Launches Real-Time REF Review, 11 April 2018. Available online: https://re.ukri.org/news-events-publications/news/research-england-launches-real-time-ref-review/ (accessed on 13 April 2018).
37. Wellcome Trust. Wellcome Is Going to Review Its Open Access Policy, 5 March 2018. Available online: https://wellcome.ac.uk/news/wellcome-going-review-its-open-access-policy (accessed on 13 April 2018).

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).