Chapter 1: Digital Repositories and Discoverability: Definitions and Typology

George Macgregor, University of Strathclyde

http://orcid.org/0000-0002-8482-3973

Abstract
Digital repositories typically improve the efficacy with which GLAMR organizations can manage their digital content and ensure its discoverability. Repositories are consequently a key enabling technology and have come to dominate the digital service offerings of many organizations. This chapter will provide a background to digital repositories, their emergence, and their typical characteristics. We will first seek to define digital repositories and will go on to explore several prominent repository types, resulting in the presentation of a 'digital repository typology'. The concept of discoverability is central to our understanding of repositories and is a core concern of this chapter too. We will therefore explore the principal attributes of repository discoverability, including consideration of the most significant digital repository discoverability technologies and users' typical routes to repository content discovery.

Introduction

Digital repositories have established themselves as a key enabling technology of GLAMR institutions (Galleries, Libraries, Archives, Museums & Records) over the past two decades, with a particular acceleration to be noted in their adoption and development in more recent times (Gupta & Sharma, 2018). Within some communities of practice, such as research libraries, repository technology has come to dominate digital library strategy and has become a key enabler of emerging visions of so-called 'Library 3.0' (Kwanya et al., 2013), including an increased emphasis on externally facing digital services focused on open scholarly publishing (Adema et al., 2017). By typically supporting established and open technical protocols, as well as rich interoperable metadata models, digital repositories have become the principal focus of digital object collection, storage, and re-use in many organizations. They present convenient platforms for the parallel management, curation, and preservation of digital objects and – within the context of this volume – there are opportunities for better exposing digital content to search agents, thereby facilitating superior levels of discoverability.

As repositories have evolved technically, so their scope and functions have also diversified, such that numerous distinct types have emerged. This chapter will introduce and explore the concept of digital repositories, including their purpose and characteristics, before presenting a typology of these varying repository systems. But before this we will then seek to define digital repositories and explore the notion of discoverability within repository contexts.

Defining digital repositories

The origins of the 'digital repository' concept could be said to have emerged from computer software development. Digital software repositories are used to provide a location for the deposit of software packages or code libraries whereupon software is stored, maintained, and re-used, with metadata attached to support discovery. Such repositories became popular in software development methodologies in the 1980s (Boisvert et al., 1996) and are central to programming approaches today. Notable present-day software repositories include Ruby Application Archive or Comprehensive Perl Archive Network (CPAN). The digital repositories to be discussed in this chapter, and the wider

This is an accepted manuscript of a book chapter scheduled for publication in April 2023: Macgregor, G. (Accepted/In press – 22 October 2021). Digital repositories and discoverability: definitions and typology. In Discoverability in Digital Repositories : Systems, Perspectives, and User Studies Routledge.

monograph, are an extension of this software repository concept, and merge the concept with critical digital library innovations and functions, such as digital archiving and discovery.

Precise definitions of digital repositories can vary in the literature (Bicknese, 2003; Björk, 2014; C. Lynch, 2006). This is partly because there has been rapid and significant diversification of the concept within GLAMR organizations since the early 2000s. This diversification will be addressed in a later section but, in their most general form, we can assert that digital repositories are information systems or platforms that support the ingest, storage, management and exposure of digital content (C. Lynch, 2006). Their capacity for content storage and long-term content management means that they often support aspects of digital preservation (Xie & Matusiak, 2016), making them suitable systems for digital archiving.

Typical features of a digital repository will include:

- A digital location, with front and backend functionality supporting the deposit and ingest of digital content (Clobridge, 2010). The nature and heterogeneity of this content will depend on the type of digital repository but could include open scholarly articles, research datasets, digitized collections, multimedia assets, learning objects, complex digital objects, and so forth. Rich and extensible metadata schema will typically support the description and management of digital content (Mering & Wintermute, 2020), some of which may be exposed through the repository discovery protocols below or through complementary semantically aware approaches, such as Linked Open Data (LOD) (Candela et al., 2019).

- Support for the exposure, visibility and discovery of digital content – often open – thereby generating (re)use and impact, especially of open scholarly research content (Arlitsch et al., 2014). Repository systems therefore typically support several established technical standards and protocols, all designed to ensure interoperability with discovery agents and enable participation in the wider, distributed global repository network. Though many repositories will vary in their observation of 'discovery' standards, almost all will at least support the keystone protocols of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Simple Web-service Offering Repository Deposit (SWORD). Increasingly repositories will support emerging protocols associated with the Confederation of Open Access Repositories (COAR) Next Generation Repository initiative, such as ResourceSync (a successor to OAI-PMH), Signposting (Rodrigues et al., 2017), and more recently, Notify, based on Linked Data Notifications (Shearer et al., 2021).

- Support for the management of digital content (or 'assets') over time, normally using open-source technologies thereby obviating software obsolesce. This is undertaken to ensure access, identification, and persistence of digital content. In some repositories this also extends to the digital preservation and curation of digital content. Such management is especially important to the maintenance of unique digital collections held in trusted locations (Bak, 2016; Corrado, 2019); but it is also an increasingly important instrument in maintaining the 'digital scholarly record', an issue being confronted in scholarship where less stable academic publishing technologies have been employed (Klein et al., 2014).

The number of digital repositories has grown considerably since the mid-2000s as GLAMR organizations have sought convenient information systems with which to manage growing digital content collections. According to the global Directory of Open Access Repositories (OpenDOAR), there are close to 6,000 active digital repositories (Jisc, 2021). This figure is likely far higher since OpenDOAR is not exhaustive and does not necessarily record all repository types (Ali et al., 2018).

Discoverability: resource discovery within digital repository contexts

Our ability to communicate knowledge is what gives information, and ergo information resources, their value and their ability to satisfy users' information needs (Smucker, 2011). Definitions of *resource discovery* have been numerous over the years but most are unified in their recognition that the concept entails the systematization of information resources in order to provide users with an intuitive, organized view of resources (Bowman et al., 1993). Seminal informatician, Clifford Lynch, has noted that the identification of potentially relevant information resources is the principal task in discovery, with resources intuitively organized and ranked, and their subsequent browsing made possible via results expansion or filtering tools. Though predating the emergence of digital repositories by many years, Lynch cites the 'searching of various types of directories, catalogs or other descriptive databases' as typical examples of resource discovery, all of which have obvious parallels to our current era of digital repositories (C. A. Lynch, 1995). The concept of *resource discovery* is therefore central to our understanding of how users explore, navigate, locate and retrieve information resources; but is also central to the way in which our information technologies facilitate that discovery, i.e. *discoverability*, the ease with which these resources can be discovered by users or machines (Beyene, 2016). We can be even more specific by suggesting that *discoverability* is a measure of the extent to which those information systems or technologies purporting to be discoverable – in our case digital repositories – are technically optimized to ensure it. For example, eliminating all possible discoverability barriers or optimizing technologies to interoperate with specific discovery tools or search agents, thereby ensuring maximum ease in discovery.

The mechanisms that facilitate discoverability are the focus of numerous chapters throughout this book. It is nevertheless worth considering discoverability within the context of digital repositories by providing a conceptual overview. Figure 1 provides a conceptual diagram of users' routes to digital repository content discovery. These routes are multiple, and all are contained within the ellipse and ultimately link to the digital repository, the content origin. Users are situated outside the ellipse, and are presented with a multiplicity of discovery routes, some of which they may use knowingly while others they may use unknowingly. The multiplicity of discovery routes goes some way to demonstrating the high levels of discoverability that digital repositories generally display – much of which is made possible through repository observance of discovery standards and protocols. But it also emphasizes the importance of identifying and eliminating discoverability barriers, since users' routes to repository discovery can be difficult to anticipate.

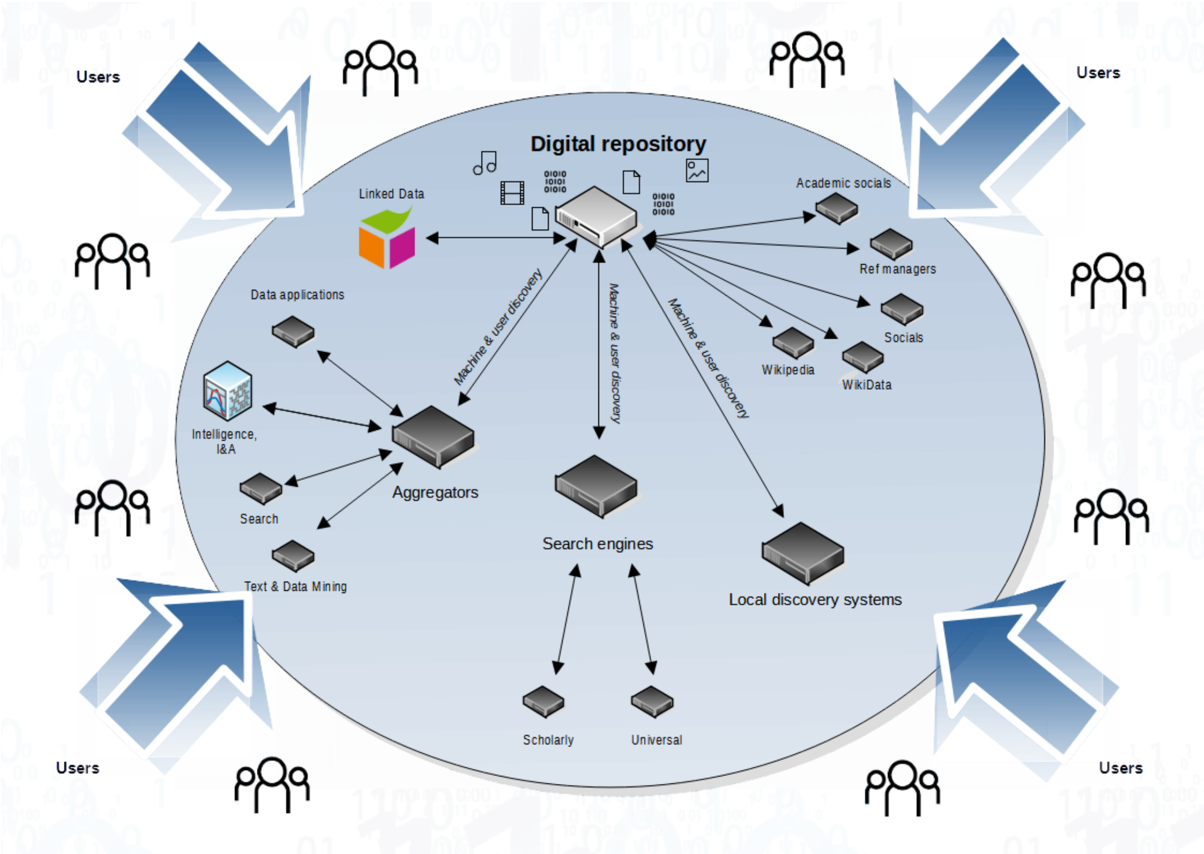

Figure 1: Conceptual diagram of digital repository discoverability and users' routes to content discovery.

Using Figure 1 as a reference, we can observe that discoverability of digital repository content will tend to be achieved via the following routes:

- **Native**: Users may elect to use the digital repository directly. This route will typically be employed in known-item queries, particularly in those repositories with a large corpus of rich, digital content or content less well exposed via the alternative routes below. These alternatives are generally more conducive to unknown-item queries as they enable de facto federated searching of multiple digital repositories simultaneously.
- **Local discovery systems**: Digital repositories will often feature within local discovery tools as possible search targets. For example, the search layer of a university library management system (e.g. ExLibris Primo), or an archival description platform (e.g. AtoM), may enable the discovery of content held within the university's other digital repository solutions, normally through repository support for OAI-PMH.
- **Search engines**: It is increasingly necessary for digital repositories to optimize their discoverability in third-party search tools, especially given their ubiquity in users' search behavior (Pulikowski & Matysek, 2021). 'Universal' search tools (e.g. Google, Bing, DuckDuckGo, etc.) contribute significant user traffic in repository content discovery but, depending on the type of the repository, an increasing proportion may also arrive via 'scholarly' search engines, such as Google Scholar (Macgregor, 2020), powered by repository support for structured data, etc.
- **Aggregators**: Aggregators will be described in more detail in a later section of this chapter. In the meantime, we can note that such tools will routinely harvest content from multiple digital repositories, aggregating content and data in a single location. This single aggregated collection can be searched by users. However, such large aggregations make possible other discovery routes, such as intelligence and indexing/abstracting tools, Text and Data Mining (TDM)

applications, or discovery applications such as CORE Discovery or Unpaywall, providing single-click access to open access versions of research papers whenever users encounter paywall restrictions (Hoy, 2019).

- **Linked Data**: Repositories contributing Linked Data to the 'web of data' promote alternative forms of aggregation (Freire et al., 2019). Re-use of linked repository data in this way promotes its integration within new or novel services, often outside the traditional domains of GLAMR organizations, and can consequently drive the discovery and usage of repository content (Pennington & Cagnazzo, 2019). Linked Data is a growth area in repository discovery and will be given fuller treatment in Chapter 4.
- **Social**: The importance of social networking services ('socials') as a route to discovering digital repository content is now well established (Mohammadi et al., 2018). However, these routes have diversified to include academic social networks (e.g. ResearchGate, Academia.edu) but also reference management software applications (e.g. Mendeley, ReadCube, etc.), many of which provide repository discovery functionality and the ability to share or search reference collections curated by other users, thereby presenting a more informal means of resource discovery. We may also include Wikimedia community projects, such as Wikipedia and WikiData (to be discussed in Chapter 8), both of which increasingly depend on open repository content to populate services but from which repositories often maximize the value and discovery impact of their services (Lubbock, 2018).

The above is merely a conceptual overview of digital repository discoverability and will be interrogated in subsequent chapters. However, we should note that the nature of repository discoverability can also be associated with repository type. As digital repositories have evolved in their functionality, so there has also been a diversification in the varieties of repository. Specific digital repositories have therefore emerged, each corresponding with specific use cases, organizational affiliations, or community of practice. The following sections will summarize some of these, before going on to propose a typology of repositories to aid understanding of the domain.

Digital repository types

Institutional repositories: scholarly communications & government

Institutional repositories (IRs) are digital repositories that have been established by an institution or organization, typically a higher education institution or research institute. Their purpose can be multifarious but there is typically a focus on delivering a repository service supporting the management and dissemination of digital content created by the institution and the community it serves (C. A. Lynch, 2003). IRs therefore represent an institutional commitment to the stewardship of the digital content gathered by the repository, as well as a commitment to its dissemination. For this reason, IRs are principally concerned with supporting open scholarly communication, serving textual content such as Open Access academic research articles, conference papers, research theses, and so forth. IRs are uniquely designed to promote the discovery and ergo impact of this scholarly content; the software they use (e.g., DSpace, EPrints, Islandora, Invenio) demonstrates this, and will be discussed in Chapter 2. But they also function as a counterbalance to the costs associated with the 'legacy' scholarly publishing industry by providing access to scholarly content that might otherwise reside behind a paywall (Björk et al., 2014; Tennant et al., 2016). Owing to their focus on literature, and the limited filetype scope that this infers, IRs tend to display only low-to-moderate levels of content heterogeneity, with generally low levels of object complexity.

A critical mass of this open scholarly content has emerged in recent years as IRs have grown to accommodate the requirements of national and funder-specific Open Access policies. IRs are therefore ubiquitous within the global digital repository landscape, accounting for 90% of all known digital repositories (Jisc, 2021). The increase in IR numbers is not only driven by policy compliance and the corresponding number of institutions which stand to benefit from their creation, but also because some institutions have evolved to support several IRs in parallel, thereby better accommodating massive growth in collections and/or simplifying management of digital content, e.g. separate IRs for academic articles, research theses, grey literature, etc. This is particularly notable in the UK where a mature policy framework promotes research openness, and where the deposit of peer-reviewed manuscripts is mandated by both research funders and the UK's national research assessment exercise. The impact of these policy instruments is visible in Open Access data reported by the Centre for Science and Technology Studies (CTWS) Leiden Rankings, in which UK institutions report Open Access levels more than 80%, and some in excess of 90% (de Castro, 2021).

Research libraries have been exploring alternative scholarly publishing models, especially those enabling organizational ownership over publishing infrastructure. In an extension to the scholarly communication aim, IRs – alongside subject-based repositories, to be discussed below – are increasingly used in the publication of so-called 'overlay journals' (Marušić et al., 2019; Whitehead et al., 2019). Though not entirely a new concept, overlay journals are nevertheless an innovation in the institutional publishing space because they harness repository functionality and discovery potential to provide an alternative model of scholarly communication. In such a scenario, an editorial board of an overlay journal may accept submissions to a journal through the IR. Peer-review is then undertaken within or outside the repository, and revised manuscripts are approved for deposit and inclusion in issues of the journal, to which a journal website might link. Notable examples include the Journal of Data Mining and Digital Humanities (JDMDH) and ST-OPEN, both of which are overlays of IRs (Thornton & Kroeker, 2021). However, the growth of overlay titles using subject-based repositories, especially those that enjoy a global user base such as arXiv, has been far greater in recent years, particularly within physical science domains (Marra, 2017).

While an IR is typically a declaration of a commitment to the stewardship of institutionally created digital content, IRs have historically been less concerned with the digital curation and preservation of that content (Y. Li & Banach, 2011; Xie & Matusiak, 2016). As we shall see in following sections of this chapter, other repository types emerged that specialize in these aspects of digital content management, especially at larger GLAMR institutions. Nevertheless, as time has progressed and the fragility of digital content has become more apparent, most repository types have conceded that some digital curation of content is necessary. In IRs the need to prioritise persistence in the scholarly record has motivated more recent attempts to improve digital curation (Macgregor & Neugebauer, 2020; Neugebauer et al., 2018; Xie & Matusiak, 2016), in parallel with the increased deployment of persistent identifiers (PiDs) to describe and link to that content (Ananthakrishnan et al., 2020; Bunakov & Madden, 2020; Klump & Huber, 2017). Unfortunately, recent systematic reviews of the current state-of-the-art continue to suggest that insufficient digital curatorial activity is taking place within IRs and that repository managers are avoiding their responsibilities to 'assure the long-term access to the assets they store' (Barrueco & Termens, 2021).

It has been convention to make a distinction between IRs and so-called 'governmental' digital repositories (Xie & Matusiak, 2016); but it is apposite to note that many governmental digital repositories are themselves IRs, established under the auspices of government organizations, i.e. a repository service supporting the management and dissemination of digital content created by the institution and the community it serves. The Fiskeridirektoratets Digitalarkiv (Norwegian Directorate of Fisheries - https://fdir.brage.unit.no/) or the Irish Health Publications Archive (https://hselibrary.ie/), both of which discharge the digital archiving and dissemination responsibilities of their respective Norwegian and Irish government departments, are prototypical examples. Fewer than 3% of digital repositories listed on OpenDOAR are described as governmental, and many of these are government or nationally sponsored GLAMR repositories. It can be argued that such government focused IRs are more receptive to the needs of digital curatorial action, especially in instances in which IRs are disseminating government records or supporting aspects of e-government (Aas et al., 2014; Kulovits et al., 2012). Nevertheless, while we maintain the repository distinction for the typology presented in the following section, it is increasingly becoming a redundant distinction to be made.

Subject-based and preprint repositories

Subject-based repositories – also known as 'disciplinary repositories' – are those repositories which coalesce around a specific subject or disciplinary area instead of a specific institution or organization (Björk, 2014). Their principal function is almost identical to that of IRs insofar as they provide a mechanism for the dissemination of open research content thereby supporting the goals of Open Access, including providing infrastructure for the creation of overlay journal titles. Notable examples include arXiv (https://arxiv.org/), AgEcon (https://ageconsearch.umn.edu/), RePEc (https://econpapers.repec.org/), and E-LIS (http://eprints.rclis.org/). Subject-based repositories enable content creators to engage with their knowledge community (i.e. peers) when sharing their work and reach users who seek discover new content within that knowledge community, e.g. (Kuperberg, 2020); but they also provide access to repository infrastructure for those creators either unable or unwilling to use an institutional repository alternative (Emery, 2018). As well as being well indexed by a wide variety of discovery tools, some subject-based repositories are at such a scale that they are the principal source of literature for scholars (Clement et al., 2019).

The governance of subject-based repositories is an additional point of difference with IRs. Owing to their community focus, most subject-based repositories operate transparently through advisory boards, elected steering groups and other models of community participation (Adamick & Reznik-Zellen, 2010). These governance mechanisms are used to determine questions surrounding repository collection policies, funding, development paths, and so forth; but also helps to guard against private interests in the management of open repository content, as has historically been the case with some privately operated platforms, e.g. (T. Li, 2019).

There is increasing cross-over between subject and preprint repositories. The essential concept of preprints is not new, with progenitors existing for many decades within some disciplines (Brown, 2001; Cobb, 2017). A preprint can be described as a 'precursor' to a research article that may ultimately find publication in a peer-reviewed journal title (Brown, 2001). Preprints are therefore open research papers that have not undergone peer-review, but which instead enjoy rapid results dissemination and the possibility of community feedback. Preprints have increasing acceptance within scholarly communities despite the criticism that they lack certification via peer-review (Johansson et al., 2018). For example, it is now widely acknowledged that those preprints documenting science surrounding COVID-19, the 'Severe Acute Respiratory Syndrome Coronavirus 2' (SARS-CoV-2), were critical to the global public health response to the COVID-19 pandemic. By circumventing conventional research publication routes, it was possible for research results to be disseminated rapidly thereby accelerating clinicians' understanding of virus and its human impact (Fraser et al., 2021). Notable repositories in this instance of preprint dissemination included medRxiv (https://www.medrxiv.org/) and bioRxiv (https://www.biorxiv.org/), both of which took inspiration in their naming from perhaps the oldest and most well-known subject-based repository, arXiv.

Data repositories

Opportunities for the digital archiving of data and datasets is almost as old as the web itself (Hahnel & Valen, 2020). In more recent times mature repository solutions – data repositories – have emerged to support the management of often complex, multi-object research datasets. Since a dataset can be as simple as a single 30kb `.csv` data file but as sophisticated as, say, a complex 100TB dataset comprising millions of related data components, these repositories are optimized for data, harnessing rich metadata schema to capture sufficient descriptive, administrative, structural, and technical metadata about the data which are being managed. Not only is such metadata necessary to facilitate discovery and reuse of datasets but it is essential to ensure datasets remain intelligible to users and machines, and that they can be subject to digital preservation actions. Typically, data repositories will attract deposits of structured quantitative or qualitative data; although we should acknowledge that definitions of 'data' can vary across communities of practice such that, depending on the academic context, objects such as software, media, or research instruments may constitute data or a dataset. Furthermore, the heterogeneity of content within data repositories and the sophistication of the associated digital objects is far more expansive than in literature repositories. Open file standards may be managed alongside lesser known, proprietary data formats associated with, say, microscope software, like `.opju`.

Research data management, as well as open data more generally, has become a key focus at research intensive organizations but also at GLAMR institutions with a research support purview. The culture surrounding research integrity and public accountability has demanded improved data openness and transparency to support the goals of research verification and reproducibility, all largely delivered via data repositories. As in the case of institutional repositories, the need to ensure persistence in the scholarly record means that research datasets are increasingly subject to digital curation requirements. Researchers may also have the opportunity of reusing an existing dataset, resulting in research and knowledge efficiencies as duplication is avoided. This, in turn, has resulted in greater linkage between research literature and its underlying data, with research bodies and academic peers insisting on explicit linking between the two, or the inclusions of data availability statements (Colavizza et al., 2020). Linkage functions as an important dataset discovery path for users, since dataset search tools remain experimental (Mannheimer et al., 2021; Sansone et al., 2017) and awareness of the existence of data may only arise through reading an associated research article (Singhal & Srivastava, 2017).

Like literature repositories (institutional, subject, preprint), data repositories have diversified to encompass both institutional and subject-based varieties, with latter examples often better fulfilling disciplinary metadata expectations. Popular open-source platforms used within the literature repository space (e.g. DSpace, EPrints, etc.) are available in datacentric releases but there is an increasing preference for optimized solutions, such as Dataverse, CKAN, and Samvera. Commercial solutions, such as Figshare and Mendeley Data, demonstrate some popularity too but are often deemed inconsistent with the requirements of open infrastructure (Bilder et al., 2020). Variability in the implementation of data repositories has resulted in notional global agreement that such repositories should be FAIR, demonstrating defined levels of 'findability, accessibility, interoperability, and reusability' (Wilkinson et al., 2016). The so-called 'FAIRness' of data repositories, and the data they store, has become an expectation of data manager practitioners and research funders (Bahim et al., 2020).

Mega-repositories

Mega-repositories are large-scale repositories which have few discernible restrictions on the type of material deposited, its format or size, its peer-review status, or its discipline. Such repositories tend to be generalist in their scope, allowing participation from a wide variety of academic users. For our purposes we can define a mega-repository as one containing more than 2 million heterogeneous objects, submitted by creators originating from multiple, geographically distributed organizations and disciplines. As this definition might suggest, there are few repositories falling into this category or have the resources to sustain their operation; but those that do are among the most significant repositories of any type in the world. A good example of a mega-repository would be The European Organization for Nuclear Research (CERN) repository, Zenodo (https://zenodo.org/), which accommodates heterogenous objects of varying complexity, from simple preprints to more complex objects like software and learning objects (Peters et al., 2017). Built on CERN's own open-source repository platform, Invenio, Zenodo can also function as a publication platform for overlay journal titles and conference proceedings. Repositories like Zenodo are of such significance that they assist in the governance and development of core repository infrastructure, such as DataCite, ORCID, etc. However, the generalist approach of mega-repositories can mean that suboptimal metadata modeling occurs where specificity might be sought for object types. For example, the earth and planetary science data repository, PANGAEA (https://www.pangaea.de/), is likely to better satisfy the metadata requirements of an environmental science dataset than Zenodo, which necessarily takes a one-size-fits-all approach.

Trusted Digital Repositories

The digital curatorial focus at GLAMR institutions has typically been on items of artistic or cultural value rather than the traditional outcome of scholarly research (e.g. academic manuscripts, etc.). This reflects their longer tradition in collecting, conserving and ensuring long-term access to materials (Bak, 2016). The emergence of the trusted digital repository (TDR) concept (sometimes termed 'trustworthy digital repository') emerged in recognition of these GLAMR traditions and was an acknowledgement that such institutions would increasingly accumulate potentially fragile digital collections, many of which would require dedicated curatorial attention, e.g. (Research Library Group / OCLC Working Group on Digital Archive Attributes et al., 2002). Most repository types set out to be reliable locations for the management of digital content, however a TDR – often managed under the auspices of a much larger GLAMR or national memory institution – extends the commitment of reliability to one of trust. A TDR is therefore one 'whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future' (RLG/OCLC Working Group on Digital Archive Attributes et al., 2002).

Early work by the RLG and OCLC (Ibid.) set out the expected attributes of a TDR. Resource discovery may be a component of the TDR context; but the priority of TDR management processes is instead to steward digital content over time and in accordance with recognized standards. This entails compliance with the Reference Model for an Open Archival Information System (OAIS) thereby guaranteeing a basis for long-term maintenance and access (Consultative Committee for Space Data Systems (CCSDS), 2012), as well as demonstrable evidence of 'organization viability', 'financial stability', 'administrative responsibility', 'technological suitability and procedural accountability', and 'system security', the latter emphasizing the importance of disaster preparedness, digital content recovery procedures, data integrity actions, and so forth. Despite TDR adherence to recognized 'trustworthy frameworks' there is recent concern that such approaches are technocratic or do not set sufficiently objective measures of trust (Bak, 2016). Certification of the trustworthiness of TDRs is therefore increasingly sought through mechanisms such as the 'CoreTrustSeal' or 'DIN 31644', awarded to repositories satisfying rigorous trustworthy requirements (Corrado, 2019). Owing to the large digital corpora that TDR-hosting institutions tend to curate, it is not uncommon for TDRs to serve highly heterogeneous collections and for objects to be highly complex, particularly as many will be multipart and require preservation metadata.

Learning Object Repositories

Potential complexity in the digital objects managed by repositories is demonstrated particularly by open learning object repositories (LORs). Sometimes known as 'open educational resource repositories', LORs are typically used to share learning objects with others, normally teaching communities (Ochoa & Duval, 2009). A 'learning object' is a digital object, or a collection of digital objects, which can be reused to support learning. Such learning objects may be as simple as a PDF document detailing a lesson plan for a teacher to reuse in an online course, or as complex as a multipart digital learning object containing multiple related objects of various formats (e.g., textual documents, multimedia, assessments, interactive tests, etc.), all designed for delivering an entire postgraduate module. By openly sharing learning objects via a LOR it enables learning materials to be reused and improved, thereby avoiding duplication of effort. These objects may then be re-shared for digitally supported teaching and learning purposes. As digital learning pioneer Stephen Downes noted in 2001: "…the world does not need thousands of similar descriptions of sine wave functions available online. Rather, what the world needs is one, or maybe a dozen at most" (Downes, 2001).

LORs provide a convenient technology for describing, organizing, presenting, and sharing learning objects. Prominent LORs are often established by higher education institutions (HEIs) where – notwithstanding a desire to share objects openly – there exists a concern about capturing the valuable knowledge assets staff create in the delivery of curricula. Capturing them enables potential local teaching efficiencies to accrue as content is reused and for HEIs to be more responsive to students' real time learning requirements (Sampson & Zervas, 2013). Although large global LORs exist such as OER Commons, it is not uncommon for LORs to use optimized versions of institutional repository software (Cervone, 2012). For example, EdShare at the University of Glasgow (https://edshare.gla.ac.uk/) and OpenEd@UCL at University College London (https://open-education-repository.ucl.ac.uk/) both use the EPrints 'EdShare' repository 'flavour'.

It is important to note that management of learning objects within an LOR typically requires considerable extension to LOR metadata models to facilitate the specialized organization and discovery requirements of LORs. In addition to typical descriptive, administrative, and technical metadata elements, LORs must capture adequate descriptive metadata concerning the teaching or learning context of the object as these form an important discovery avenue for potential reuse (Palavitsinis et al., 2014). This may include metadata pertaining to the intended learning outcomes of the learning object, intended audience and their education level, pre-requisite study requirements, client software requirements, and so forth. In this respect there is some commonality between LORs and data repositories insofar as both have breached traditional digital description boundaries; although, the complexity of LOR metadata – often expressed using Learning Object Metadata schema (Learning Technology Standards Committee (LTSC), 2020) – is frequently cited as one reason why metadata quality issues arise within LORs (Palavitsinis et al., 2014). A typology of repository types will be presented in the next section of this chapter; but it is noteworthy that LORs are now of such diversity that separate LOR typologies have been posited in the literature, e.g. (McGreal, 2008).

Aggregating repositories

Aggregating repositories are the final repository type to be considered as part of this chapter. We can identify two principal subtypes: *machine-aggregated*, and *user-aggregated*. Machine-aggregated repositories will typically harness repository technical protocols, such as OAI-PMH and/or ResourceSync, to harvest metadata and digital content from distributed repositories to aggregate it centrally. CORE (https://core.ac.uk/) and BASE (https://www.base-search.net/) are two prominent examples of this model; both routinely harvest and aggregate as much content as possible from thousands of repositories, irrespective of their subject scope, affiliation, content heterogeneity or object complexity (Knoth & Zdrahal, 2013). It is therefore apposite to note that smaller, more selective aggregations can and do occur, e.g., based on a subject or topic. Public Health Scotland, for example, created a Covid-19 aggregating repository to unify Scottish Covid-19 related research made available through IRs or subject repositories (https://www.publichealthscotland.scot/). User-aggregated approaches perform a similar function to machine-aggregated but instead rely on user action to build the aggregation, normally by providing content to deposit. DataverseNL (https://dataverse.nl/) in the Netherlands provides a user-aggregated data repository, for example – with DataverseNL aggregating all data content from Dutch consortia member universities.

Whichever the subtype, aggregating repositories seek to 'aggregate' content in a single digital location – with the location being a digital repository. The aggregating repository approach is cognate to the more established union catalog approaches which have been prevalent in digital libraries for decades (Dunsire, 2008), and even those examples emerging prior to the advent of the Web, such as WorldCat (Salmon, 1982). The benefits of providing these digital aggregations are manifest to users in their potential size, diversity and the way in which they can support users' resource discovery needs (Hudson-Vitale, 2017). In addition, such aggregations of digital content and metadata present unique opportunities for large-scale computational text and data mining (TDM) too, as well as providing the foundation of countless novel applications, including recommendation engines (CORE, 2019; Knoth et al., 2017).

Proposing a typology of digital repository systems

In this section we formalize the aforementioned 'repository types' by placing each of them within a typology of digital repository systems. We also use several of the previously highlighted repository characteristics as typology facets, described below. Such a typology provides a useful conceptual aid to understand the repository landscape, how specific repository types differ (or not) and will help to contextualize subsequent chapters. The typology is presented in Table 1.

Type	Volume	Object complexity	Content heterogeneity	Metadata curation	Typical governance
Preprint	Large	Low	Low	User	Community
Institutional	Variable	Low	Moderate	Mixed	Organizational
Subject	Large	Low	Moderate	User	Community
Data	Moderate	High	Moderate	Mixed	Community / Organizational
Mega	Large	High	High	User	Community
Trusted	Large	High	High	Organizational	Organizational
Learning Object	Moderate	High	Moderate	Organizational	Organizational
Governmental	Moderate	Low	Low	Organizational	Organizational
Aggregating	Large	Moderate	Moderate	Mixed	Community

Table 1: Typology of digital repositories.

- **Type** refers to the repository type, whether this be an IR or TDR or some other repository type. A total of 9 repository types were outlined in previous sections.
- **Volume** refers to the typical size of the digital corpus served by the repository type. This is not a reference to digital storage size and instead denotes the volume of digital objects typically held by a specific repository type. For example, in our discussion of mega- repositories we noted that they are characterized by serving a *large* digital corpus, normally more than 2 million digital objects. By contrast – and owing to the nature of the digital content they manage – data repositories are *moderate* and typically enjoy a slower rate of repository growth.
- **Object complexity** is an indicator of the complexity of the digital objects typically deposited in the repository type. Repositories that manage predominantly textual content, such as IRs and subject repositories, can be said to have *low* object complexity. The objects are similar in nature and their intellectual content is typically self-contained, e.g., within a PDF file or similar. This is unlike, for example, data repositories in which a single dataset may comprise millions of data files, organized according to a specific structure. Such datasets are highly complex, multipart objects and may also have similarly complex metadata describing the complete data object to facilitate data interpretation or reproducibility (Soiland-Reyes & Goble, 2021). Their object complexity could therefore be considered *high*.
- **Content heterogeneity** attempts to characterize the extent of variety in the digital objects managed by an archetypal repository type and defines whether a repository typically manages a digital collection exemplifying low or high levels of heterogeneity. If we take the example of preprint repositories, we can observe that such repositories serve relatively homogenous textual

content, suggesting a *low* level of content heterogeneity. Mega-repositories are almost the inverse in this respect, serving digital content of all varieties, file formats and size – this can include everything from preprints to datasets to learning objects. Mega-repositories therefore typify *high* levels of content heterogeneity. As might be expected, repositories can demonstrate close alignment on content heterogeneity and object complexity such that a direct variation relationship can be said to occur. This is because as digital repositories accommodate content of ever-increasing heterogeneity, the complexity of the digital objects managed will also tend to increase (Fig. 2). As examples of generalist repositories, mega-repositories demonstrate this principle clearly. Similarly, as content becomes more homogeneous, object complexity declines correspondingly.

- **Metadata curation** helps to define the typical approaches adopted by repository types in their description, organization and management of metadata associated with digital objects. Great variation exists between repositories in how metadata curation is approached, with some relying almost exclusively on *user* generated metadata (e.g. preprint, subject, mega) and accepting the limitations that this poses. Those repositories with higher object complexity and content heterogeneity (e.g. Trusted) demand *organizational* approaches in which metadata are managed by skilled professionals. Other repository types adopt a mixed approach.
- **Typical governance** denotes the way the repository type is usually governed. As we noted in previous sections, some repository types demonstrate strict adherence to forms of *community* governance to ensure transparency and guarantee openness. Others are founded, managed, and maintained by *organizational* entities, whether this is a library, university, or some form of national memory institution.

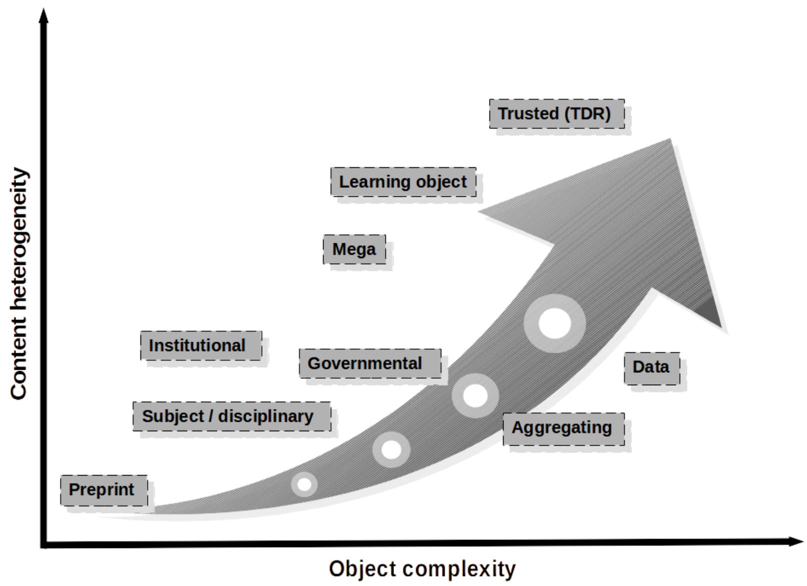

Figure 2: Illustration of the relationship between content heterogeneity and object complexity in digital repositories.

Conclusion

In this chapter we have attempted to define and characterize the nature of digital repositories, and we have explored the multifarious ways in which repositories can support digital content discovery. The extent to which digital repositories are discoverable is of core concern to GLAMR organizations and although repository software tends to demonstrate satisfactory levels of discoverability, it is always necessary for repositories to optimize discoverability. The typology presented formalizes our understanding of the digital repository types to be discussed in subsequent chapters. All typologies are subject to change and it is apposite to highlight that the COAR Next Generation Repository initiative (Rodrigues et al., 2017) may, over time, modify repository behaviors in unanticipated ways, thereby necessitating future revisions to the typology. In the meantime, we can report that the typology presents a useful conceptual tool to aid understanding of the current and near-future repository landscape. Many of the themes and topics raised in previous sections will be explored in more detail in subsequent chapters. Chapters 2, 3 and 4 will continue discussion of repository functionality, metadata schema, and discoverability optimization approaches via Linked Data.

References

Aas, K., Delve, J., Vieira, R., & King, R. (2014, December). Integrating e-government systems with digital archives: Paper - iPES 2014 - Melbourne. *Proceedings of the 11th International Conference on Digital Preservation*. https://hdl.handle.net/11353/10.378127

Adamick, J., & Reznik-Zellen, R. (2010). Trends in Large-Scale Subject Repositories. *D-Lib Magazine*, *16*(11/12). https://doi.org/10.1045/november2010-adamick

Adema, J., Stone, G., & Keene, C. (2017). *Changing Publishing Ecologies: A Landscape Study of New University Presses and Academic-Led Publishing* (p. 103). Jisc. https://digitalcommons.unl.edu/scholcom/80

Ali, M., Loan, F. A., & Mushatq, R. (2018). Open Access Scientific Digital Repositories: An Analytical Study of the Open DOAR. *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, 213–216. https://doi.org/10.1109/ETTLIS.2018.8485265

Ananthakrishnan, R., Chard, K., D'Arcy, M., Foster, I., Kesselman, C., McCollam, B., Pruyne, J., Rocca-Serra, P., Schuler, R., & Wagner, R. (2020). An Open Ecosystem for Pervasive Use of Persistent Identifiers. *Practice and Experience in Advanced Research Computing*, 99–105. https://doi.org/10.1145/3311790.3396660

Arlitsch, K., OBrien, P., Kyrillidou, M., Clark, J. A., Young, S. W. H., Mixter, J., Chao, Z., Freels-Stendel, B., & Stewart, C. (2014). *Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories* (pp. 1–10). Montana State University. https://scholarworks.montana.edu/xmlui/handle/1/8924

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., & Stall, S. (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal*, *19*(1), 41. https://doi.org/10.5334/dsj-2020-041

Bak, G. (2016). Trusted by whom? TDRs, standards culture and the nature of trust. *Archival Science*, *16*(4), 373–402. https://doi.org/10.1007/s10502-015-9257-1

Barrueco, J. M., & Termens, M. (2021). Digital preservation in institutional repositories: A systematic literature review. *Digital Library Perspectives*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/DLP-02-2021-0011

Beyene, W. M. (2016). Resource Discovery and Universal Access: Understanding Enablers and Barriers from the User Perspective. *Studies in Health Technology and Informatics*, *229*, 556–566.

Bicknese, D. (2003). Institutional repositories and the institution's repository: What is the role of university archives with an institution's on-line digital repository? *Archival Issues*, *28*(2), 81–93.

Bilder, G., Lin, J., & Neylon, C. (2020). *The Principles of Open Scholarly Infrastructure*. [GitHub]. https://doi.org/10.24343/C34W2H

Björk, B.-C. (2014). Open access subject repositories: An overview. *Journal of the Association for Information Science and Technology*, *65*(4), 698–706. https://doi.org/10.1002/asi.23021

Björk, B.-C., Laakso, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the Association for Information Science and Technology*, *65*(2), 237–250. https://doi.org/10.1002/asi.22963

Boisvert, R., Browne, S., Dongarra, J., & Grosse, E. (1996). Digital software and data repositories for support of scientific computing. In N. R. Adam, B. K. Bhargava, M. Halem, & Y. Yesha (Eds.), *Digital Libraries Research and Technology Advances* (pp. 103–114). Springer. https://doi.org/10.1007/BFb0024606

Bowman, M., Danzig, P. B., Manber, U., & Schwartz, M. F. (1993). *Scalable Internet Resource Discovery: Research Problems and Approaches ; CU-CS-679-93* [Technical Report]. University of Colorado Boulder. https://scholar.colorado.edu/concern/reports/rx913q821

Brown, C. (2001). The E-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science and Technology*, *52*(3), 187–200. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1586>3.0.CO;2-D

Bunakov, V., & Madden, F. (2020). Integration of a National E-Theses Online Service with Institutional Repositories. *Publications*, *8*(2), 20. https://doi.org/10.3390/publications8020020

Candela, G., Escobar, P., Carrasco, R. C., & Marco-Such, M. (2019). A linked open data framework to enhance the discoverability and impact of culture heritage. *Journal of Information Science*, *45*(6), 756–766. https://doi.org/10.1177/0165551518812658

Cervone, H. F. (2012). Digital learning object repositories. *OCLC Systems & Services: International Digital Library Perspectives*, *28*(1), 14–16. https://doi.org/10.1108/10650751211197031

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). *On the Use of ArXiv as a Dataset*. arXiv.org. http://arxiv.org/abs/1905.00075

Clobridge, A. (2010). 1—Introduction. In *Building a Digital Repository Program with Limited Resources* (pp. 3–11). Chandos Publishing. https://doi.org/10.1016/B978-1-84334-596-1.50001-8

Cobb, M. (2017). The prehistory of biology preprints: A forgotten experiment from the 1960s. *PLOS Biology*, *15*(11), e2003995. https://doi.org/10.1371/journal.pbio.2003995

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Consultative Committee for Space Data Systems (CCSDS). (2012). *Reference Model for an Open Archival Information System (OAIS) [ISO 14721:2012]*. CCSDS Secretariat / ISO. https://public.ccsds.org/pubs/650x0m2.pdf

CORE. (2019, May 31). *CORE content and discovery*. https://www.youtube.com/watch?v=kp4cvk7aJjs

Corrado, E. M. (2019). Repositories, Trust, and the CoreTrustSeal. *Technical Services Quarterly*, *36*(1), 61–72. https://doi.org/10.1080/07317131.2018.1532055 de Castro, P. (2021, January 25). *Recent developments in Open Access: An Overview* [Other]. International Federation of Catholic Universities (IFCU/FIUC). https://strathprints.strath.ac.uk/75158/

Downes, S. (2001). Learning Objects: Resources For Distance Education Worldwide. *The International Review of Research in Open and Distributed Learning*, *2*(1). https://doi.org/10.19173/irrodl.v2i1.32

Dunsire, G. (2008). Collecting metadata from institutional repositories. *OCLC Systems and Services*, *24*, 51–58.

Emery, J. (2018). How green is our valley?: Five-year study of selected LIS journals from Taylor & Francis for green deposit of articles. *Insights*, *31*(0), 23. https://doi.org/10.1629/uksg.406

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, *19*(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Freire, N., Voorburg, R., Cornelissen, R., de Valk, S., Meijers, E., & Isaac, A. (2019). Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network. *Information*, *10*(8), 252. https://doi.org/10.3390/info10080252

Gupta, D. K., & Sharma, V. (2018). Analytical study of crowdsourced GLAM digital repositories. *Library Hi Tech News*, *35*(1), 11–17. https://doi.org/10.1108/LHTN-07-2017-0055

Hahnel, M., & Valen, D. (2020). How to (Easily) Extend the FAIRness of Existing Repositories. *Data Intelligence*, *2*(1–2), 192–198. https://doi.org/10.1162/dint_a_00041

Hoy, M. B. (2019). New Tools for Finding Full-Text Articles Faster: Kopernio, Nomad, Unpaywall, and More. *Medical Reference Services Quarterly*, *38*(3), 287–292. https://doi.org/10.1080/02763869.2019.1629215

Hudson-Vitale, C. (2017). The Current State of Meta-Repositories for Data. In L. R. Johnston (Ed.), *Curating Research Data. Volume One: Practical Strategies for Your Digital Repository* (pp. 251–261). Association of Research Libraries. https://openscholarship.wustl.edu/lib_papers/19

Jisc. (2021). *OpenDOAR*. https://v2.sherpa.ac.uk/opendoar/

Johansson, M. A., Reich, N. G., Meyers, L. A., & Lipsitch, M. (2018). Preprints: An underutilized mechanism to accelerate outbreak science. *PLOS Medicine*, *15*(4), e1002549. https://doi.org/10.1371/journal.pmed.1002549

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE*, *9*(12), e115253. https://doi.org/10.1371/journal.pone.0115253

Klump, J., & Huber, R. (2017). 20 Years of Persistent Identifiers – Which Systems are Here to Stay? *Data Science Journal*, *16*(0), 9. https://doi.org/10.5334/dsj-2017-009

Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N., & Bayer, V. (2017, June). *Towards effective research recommender systems for repositories*. Open Repositories 2017, Brisbane, Australia. http://oro.open.ac.uk/49366/

Knoth, P., & Zdrahal, Z. (2013). CORE: Aggregation Use Cases for Open Access. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 441–442. https://doi.org/10.1145/2467696.2467787

Kulovits, H., Rauber, A., Gamito, R., Barateiro, J., Borbinha, J. L., Mazive, M., & João, D. (2012, May). Archives and digital repositories in an eGovernment context: When the subsequent bird catches the worm. *IST-Africa 2012*. http://repositorio.lnec.pt:8080/jspui/handle/123456789/1003522

Kuperberg, G. (2020). Using the arXiv. *Notices of the American Mathematical Society*, *67*(02), 1. https://doi.org/10.1090/noti2022

Kwanya, T., Stilwell, C., & Underwood, P. G. (2013). Intelligent libraries and apomediators: Distinguishing between Library 3.0 and Library 2.0. *Journal of Librarianship and Information Science*, *45*(3), 187–197. https://doi.org/10.1177/0961000611435256

Learning Technology Standards Committee (LTSC). (2020). *IEEE 1484.12.1-2002—IEEE Standard for Learning Object Metadata*. https://standards.ieee.org/standard/1484_12_1-2002.html

Li, T. (2019). *SSRN and open access for non-institutional scholars* [Preprint]. LawArXiv. https://doi.org/10.31228/osf.io/dq57y

Li, Y., & Banach, M. (2011). Institutional Repositories and Digital Preservation: Assessing Current Practices at Research Libraries. *D-Lib Magazine*, *17*(5/6). https://doi.org/10.1045/may2011-yuanli

Lubbock, J. (2018). Wikipedia and libraries. *Alexandria*, *28*(1), 55–68. https://doi.org/10.1177/0955749018794968

Lynch, C. (2006). *Digital Repositories*. Digital Curation Centre (DCC). https://www.dcc.ac.uk/sites/default/files/documents/resource/briefing-papers/digital-repositories.pdf

Lynch, C. A. (1995). Networked information resource discovery: An overview of current issues. *IEEE Journal on Selected Areas in Communications*, *13*(8), 1505–1522. https://doi.org/10.1109/49.464719

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libraries and the Academy*, *3*(2), 327–336. https://doi.org/10.1353/pla.2003.0039

Macgregor, G. (2020). Enhancing content discovery of open repositories: An analytics-based evaluation of repository optimizations. *Publications*, *8*(1), 8. https://doi.org/10.3390/publications8010008

Macgregor, G., & Neugebauer, T. (2020). Preserving digital content through improved EPrints repository integration with Archivematica. *UK Archivematica User Group*, 1–10. https://strathprints.strath.ac.uk/73978/

Mannheimer, S., Clark, J., Hagerman, K., Schultz, J., & Espeland, J. (2021). Dataset Search: A lightweight, community-built tool to support research data discovery. *Journal of EScience Librarianship*, *10*(1), 1189. https://doi.org/10.7191/jeslib.2021.1189

Marra, M. (2017). Astrophysicists and physicists as creators of ArXiv-based commenting resources for their research communities. An initial survey. *Information Services & Use*, *37*(4), 371–387. https://doi.org/10.3233/ISU-170856

Marušić, M., Tomić, V., Gudelj, D., Wager, E., & Marušić, A. (2019). University repository overlay journal – increasing the quality and visibility of student research at the University of Split, Croatia. *European Science Editing*, *45*(2). https://doi.org/10.20316/ESE.2019.45.19007

McGreal, R. (2008). A Typology of Learning Object Repositories. In H. H. Adelsberger, Kinshuk, J. M. Pawlowski, & D. G. Sampson (Eds.), *Handbook on Information Technologies for Education and Training* (pp. 5–28). Springer. https://doi.org/10.1007/978-3-540-74155-8_1

Mering, M., & Wintermute, H. E. (2020). MAPping metadata. *Journal of Digital Media Management*, *9*(1), 71–85.

Mohammadi, E., Thelwall, M., Kwasny, M., & Holmes, K. L. (2018). Academic information on Twitter: A user survey. *PLOS ONE*, *13*(5), e0197265. https://doi.org/10.1371/journal.pone.0197265

Neugebauer, T., Simpson, J., & Bradley, J. (2018, June 5). *Digital Preservation through EPrints-Archivematica Integration*. International Conference on Open Repositories, Bozeman, Montana, USA. https://spectrum.library.concordia.ca/983933/

Ochoa, X., & Duval, E. (2009). Quantitative Analysis of Learning Object Repositories. *IEEE Transactions on Learning Technologies*, *2*(3), 226–238. https://doi.org/10.1109/TLT.2009.28

Palavitsinis, N., Manouselis, N., & Sanchez-Alonso, S. (2014). Metadata quality in learning object repositories: A case study. *The Electronic Library*, *32*(1), 62–82. https://doi.org/10.1108/EL-12-2011-0175

Pennington, D., & Cagnazzo, L. (2019). Connecting the silos: Implementations and perceptions of linked data across European libraries. *Journal of Documentation*, *75*(3), 643–666. https://doi.org/10.1108/JD-07-2018-0117

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. I. (2017). Zenodo in the Spotlight of Traditional and New Metrics. *Frontiers in Research Metrics and Analytics*, *2*, 13. https://doi.org/10.3389/frma.2017.00013

Pulikowski, A., & Matysek, A. (2021). Searching for LIS scholarly publications: A comparison of search results from Google, Google Scholar, EDS, and LISA. *The Journal of Academic Librarianship*, *47*(5), 102417. https://doi.org/10.1016/j.acalib.2021.102417

RLG/OCLC Working Group on Digital Archive Attributes, Research Libraries Group, & OCLC. (2002). *Trusted digital repositories: Attributes and responsibilities : an RLG-OCLC report.* RLG. https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

Rodrigues, E., Bollini, A., Cabezas, A., Castelli, D., Carr, L., Chan, L., Humphrey, C., Johnson, R., Knoth, P., Manghi, P., Matizirofa, L., Perakakis, P., Schirrwagen, J., Selematsela, D., Shearer, K., Walk, P., Wilcox, D., & Yamaji, K. (2017). *Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group*. Confederation of Open Access Repositories (COAR). https://doi.org/10.5281/zenodo.1215014

Salmon, S. R. (1982). The Union Catalogue: Functions, Objectives and Techniques. *Cataloging & Classification Quarterly*, *2*(1/2). https://doi.org/10.1300/J104v02n01_03

Sampson, D. G., & Zervas, P. (2013). Learning object repositories as knowledge management systems. *Knowledge Management & E-Learning: An International Journal*, *5*(2), 117–136.

Sansone, S.-A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J. S., Xu, H., Fore, I. M., Lyle, J., Gururaj, A. E., Chen, X., Kim, H., Zong, N., Li, Y., Liu, R., Ozyurt, I. B., & Ohno-Machado, L. (2017). DATS, the data tag suite to enable discoverability of datasets. *Scientific Data*, *4*(1), 170059. https://doi.org/10.1038/sdata.2017.59

Shearer, K., Klein, M., & Walk, P. (2021, March 22). *Notify: The Repository and Services Interoperability Project*. https://vimeo.com/527538892

Singhal, A., & Srivastava, J. (2017). Research dataset discovery from research publications using web context. *Web Intelligence*, *15*(2), 81–99. https://doi.org/10.3233/WEB-170354

Smucker, M. D. (2011). Information representation. In I. Ruthven & D. Kelly (Eds.), *Interactive Information Seeking, Behaviour and Retrieval* (pp. 77–93). Facet.

Soiland-Reyes, S., & Goble, C. (2021, March 18). *RO-Crate: Describing and packaging FAIR Research Objects*. Scottish Covid-19 Response Consortium (SCRC). https://doi.org/10.5281/zenodo.4633655

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, Chris. H. J. (2016). The academic, economic and societal impacts of Open Access: An evidence-based review. *F1000Research*, *5*, 632. https://doi.org/10.12688/f1000research.8460.3

Thornton, G., & Kroeker, E. (2021). Overlay Journals: Overlooked or Emergent? *Proceedings of the Annual Conference of CAIS / Actes Du Congrès Annuel de l'ACSI*. https://doi.org/10.29173/cais1199

Whitehead, M., Shearer, K., Matthews, C., & Rieger, O. (2019, November 26). *Adding value to repositories through overlay journals*. The 14th International Conference on Open Repositories (or2019), Hamburg, Germany. https://doi.org/10.5281/zenodo.3554328

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Xie, I., & Matusiak, K. K. (2016). Chapter 9—Digital preservation. In I. Xie & K. K. Matusiak (Eds.), *Discover Digital Libraries* (pp. 255–279). Elsevier. https://doi.org/10.1016/B978-0-12-417112-1.00009-0